Powering the Academic Web

NATTRASS, CARL

Powering the Academic Web

By

Carl Nattrass

School of Engineering and

Computing Sciences

Durham University

A thesis submitted for the degree of

Doctor of Philosophy

March 2015

Abstract

Context: Locating resources on the Web has become increasingly difficult for users and poses a number of issues. The sheer size of the Web means that despite what appears to be an increase in the amount of quality material available, the effort involved in locating that material is also increasing; in effect, the higher quality material is being diluted by the lesser quality. One such group affected by this problem is post-graduate students. Having only a finite amount of time to devote to research, this reduces their overall quality study time.

Aim: This research investigates how post-graduate students use the Web as a learning resource and identifies a number of areas of concern with its use. It considers the potential for improvement in this matter by using a number of concepts such as: collaboration; peer reviewing and document classification and comparison techniques.

This research also investigates whether by combining several of the identified technologies and concepts, student research on the Web can be improved.

Method: Using some of the identified concepts as components, this research proposes a model to address the highlighted areas of concern. The proposed model, named the Durham Browsing Assistant (DurBA) is defined, and a number of key concepts which show potential within it are uncovered.

One of the key concepts is chosen, that of document comparison. Given a source document, can a computer system reliably identify other documents which most closely match it from other on the Web?

A software tool was created which allowed the testing of document comparison techniques, this was called the Durham Textual Comparison system (DurTeC) and it had two key concepts. The first was that it would allow various algorithms to be applied to the comparison process. The second concept was that it could simulate collaboration by allowing data to be altered, added and removed as if by multiple users.

A set of experiments were created to test these algorithms and identify those which gave the best results.

Results: The results from the experiments identified a number of the most promising relationships between comparison and collaboration processes. It also highlighted those which had a negative effect on the process, and those which produced variable results. Amongst the results, it was found that:

1. By providing DurTeC with additional source documents to the original, as if through a recommendation process, it was able to increase its accuracy substantially.

2. By allowing DurTeC to use synonym lists to expand its vocabulary, in many cases, it was found to have reduced its accuracy.

3. By restricting those words which DurTeC considered in its comparison process, based upon their value in the source document, accuracy could be increased. This could be considered as a form of collaborative keyword selection.

Conclusion: This research shows that improvements can be made in the accuracy of identifying similar resources by using a combination of comparison and collaboration processes. The proposed model, DurBA would be an ideal host for such a system.

Table of Contents

Table of Figures

Abbreviations

CA	Client Agent
CNS	Carl Nattrass Synonyms
CSCL	Computer Supported Collaborative Learning
DOS	Degrees of Separation
DSL	Defined Synonym List
DurBA	Durham Browsing Assistant System
DurTeC	Durham Textual Comparison System
DWL	Defined Word List
IP	Internet Protocol
IT	Information Technology
LMS	Learning Management System
LSI	Learning Style Index
OAI	Open Archive Initiative
PLE	Personal Learning Environment
PR	Percentage Rule
RRF	Resource Representation File
RDF	Research Description Framework
SA	Server Agent
SAs	Structured Abstracts
SCORM	Shareable Content Object Reference Model
SS	Secondary Sources
URL	Uniform Resource Locator
VLE	Virtual Learning Environment
WBES	Web-based Educational Systems
XML	Extensible Mark-up Language

Chapter 1 Introduction

1.1 Research Overview

It is difficult to accurately quantify the size of the Web, it is currently estimated to be around 4.7 billion pages (www.worldwidewebsize.com/), and is expected to continue to grow for the foreseeable future. It offers users an unparalleled source of materials, and within education it is considered to be extremely important. One such group embracing this resource is post-graduate students who are increasingly making use of the Web as their primary source of study material [ERDE02].

Teaching styles are changing and many students are encouraged to be more in charge of their own learning; instead of simply being handed books with highlighted chapters, they are increasingly being set tasks whereby they must locate their own material, acting on initiative, and managing their resources. This change in researching style, from more traditional sources such as academic papers, journals and books, to Web based resources, has been shown to impact on students in a number of ways; Mioduser and Nachmias [MIOD02] argued that students needed new skills, including improved information organisation, representation, and handling capabilities, in order to undertake effective Web research.

These changes leave the student with a number of issues regarding Web researching:

1. Locating the academic materials on the Web. As the size of the Web increases, the haystack in which students must sift through is growing. This makes it increasingly difficult for students to find the materials they are after; searching time is increasing and thus impacting on researching

time. Only limited information may be available about such resources 'at first glance', and much of these materials are badly described online.

2. Synchronous Research. It may be that students undertaking the same research are synchronously expending time locating similar resources which may have already been found by one of their peers, unbeknown to them.

3. Identifying the quality of materials found on the Web. Even when students locate materials on the Web, it may be difficult to identify the academic value of the given material. Historically, where an article was sourced from an academic journal, the journal could be used as a guide to the quality of that article. But where an article is sourced from a Web page, the student may have trouble in establishing the academic value of such an article based on the Domain it was drawn from. Much of the material found on the Web is of small pedagogical value to the students, is too specific, sometimes sensationalised or purely false [LOZA06].

4. Gaining physical access to materials hosted on the Web. A large proportion of the academic material on the Web is actually hidden behind 'paid access' portals. Where once, articles may have been reproduced from a journal, paid for by a subscription through a University, there are now too many of these 'paid access' portals for Universities to subscribe to them all, if not more than a few.

This research aims to investigate these issues, and answer the question of whether it is possible through using technology that student research can be made more productive.

1.2 Research Contributions

The research contributions in this thesis are as follows:

1. The definition of a Model and a Process (DurBA) to aid postgraduate students to collaboratively locate relevant research material thus reducing their time spent on searching for material and increasing their quality study time.
2. Investigation of textual comparison methods (DurTec) that form an integral aspect to the DurBA model. The research highlighted a number of important aspects of the DurBA model and textual comparison was chosen for further investigation.

The research concludes that by utilising a model such as DurBA, and combining a Textual Comparison System with collaboration, such as with DurTeC, the viability of postgraduate research can be significantly improved.

This research is aimed at post-graduate research students. Where it refers to materials found on the Web, they should be considered as academic quality materials such as published articles, conference papers and books.

1.3 Thesis Outline

Chapter 2 firstly provides an overview of how the Web is currently used by students as a learning resource. It discusses current trends in student Web researching, and compares these with more traditional methods. Then, a discussion of the potentiality and problems associated with this 'rapidly available resource', such as how the formally designed environs of a Web page may affect the presentation standards of students.

As part of this research, it will highlight those problem areas which show the most potential for improvement. Then, the chapter discusses Web-based researching and learning software and highlights potentially useful technologies.

Chapter 3 discusses standard practices for experimental design, and Chapter 4 continues on from this by proposing the research method employed. Chapter 4 undertakes some pilot studies examining the way students understand and compare text.

Chapter 5 proposes a model called the Durham Browsing Assistant (DurBA) which utilises a number of the identified technologies, which will help solve the problems highlighted in Chapter 1. It then goes on to examine a sub area of research within the DurBA model chosen for its potential for improvement; Textual Comparison Systems. A system called Durham Textual Comparison system (DurTeC) is then proposed and implemented based on those improvements.

Chapter 6 discusses the design and implementation of the DurTeC system.

In Chapter 6, the DurTeC system is used in two experiments. Those experiments are:

1. How, by adjusting the system through the use of different comparison algorithms, can the DurTeC system be optimised?
2. Comparing the human results from one of the previous pilot studies with results from the DurTeC system to consider how accurate it is.

Chapter 7 leads a discussion into the results of the two experiments and considers the impact of the results.

Chapter 8 concludes and summarises the work contained with the thesis and suggests future work which could develop the state of the art.

1.4 Criteria for Success

The method of judgement made upon this work will be in its success of identifying and documenting the effect of applying a software system to a student researching paradigm. Through this process, the student researching 'experience' will be significantly improved, but this cannot be guaranteed. However, the level of success is not the only factor, and the research will also focus on establishing that which is learned through the effect of the utilisation of this software system by the students.

The criteria for success for this research are:

1. To propose a model which can provide a viable solution for the identified issues.
2. Highlight any pertinent areas of research identified in the proposed model, and undertake further research into it.

Figure 1.1 Criteria for Success

The level of success this research has in achieving these criteria will be discussed in the final Chapter of the Thesis.

1.5 Thesis Scenario

This research shall suggest that there are a number of problems that students are encountering using the Web as an academic resource. To help visualise and illustrate how these problems could be addressed, a Scenario shall now be presented, and this Scenario shall be referred to throughout this research. This Scenario concentrates on students studying at degree level or higher.

Consider two students researching very similar topics but working independently of one another. They may be in the same classroom, or on the other side of the world from one another. Both students are studying for a degree in similar subjects.

As part of their research they both spend quite a large amount of time browsing the Web looking for relevant research documents. At some point, Student A locates a research paper which she finds pertinent to her research. At this point, she undertakes some action to highlight her interest in it, such as bookmarking it on her browser, printing or emailing it.

Sometime later, Student B, whilst browsing the Web for materials for his coursework is presented with a pop-up window displaying, amongst the many other suggestions, a link to the research paper which Student A found of interest. He also finds a number of students who have also found this particular research useful who have added their own comments about it. These comments may include, for example, particular sections which they have specifically drawn upon.

Many of these suggestions have been added by students studying similar coursework. As well as these, there are a number of suggestions which have been included by the system itself; they have been identified autonomously as they most closely match the topic of research.

This is of course a purely hypothetical scenario, for without additional software running on all of the student's machines, a standard browser would fail to operate in such a way. Having said that, technology far exceeds what is needed for this to be a viable, real-life situation.

The key concepts here are that:

1. If Student A finds material of interest whilst researching for her subject, then Student B should also find that material being recommended to him by some mechanism.
2. Students can rate and make comments about located materials.
3. The system will autonomously seek out further relevant materials based on the student's browsing activities

Chapter 2 Literature Review

2.1 The Web as a Learning Resource

2.1.1 Introduction

The Web is being relied upon increasingly in an academic supporting role; it has moved from one of a peripheral tool, to being highly integrated into universities. The Web has driven education to experience a radical change [ZAFA04], and its impact has been most profound; it has altered the way tutors teach, how students learn and think, and what their expectations are.

It impacts upon:

1. Curricula.
2. Libraries and resources.
3. Teaching and learning skills.
4. Academic priorities.
5. Department structure and funding.
6. The quality of academic standards, quantity of materials, and plagiarism.
7. Social and cultural aspects of students.

The main paradigm shift lies in the Web's ability to merge both 'currency of information' and 'supply-on-demand'. This has caused education to go through more changes, and change more rapidly than at any time during the last decade [COLL09]. The explosion of Web content, its ease of access, games consoles, data connection speeds, global sharing and collaboration, social networking, electronic communication and mobile devices have all had an impact upon students' attitudes to learning.

Learning is moving from a modular, insular, structured and progressive approach to a more fragmented, 'on-demand' and collaborative style. The Partnership for 21st Century Learning Skills (www.p21.org) argue that academia must "bridge the gap between how students live and how they learn". This is because students are learning to react quickly to the latest tools and concepts, and are moving to a more 'self-driven' and ‘life-long’ attitude for learning [COLL09]. Lifelong learning has become widespread in professional organisations, spurred on by the need for employees to update their, mainly, computing skills to be successful in work. Wessner [WESS02] wrote that "lifelong learning has become one of the major challenges in today’s industrial societies".

Longworth and Davies [LONG96] defined lifelong learning as:

The development of human potential through a continuously supportive process which stimulates and empowers individuals to acquire all the knowledge, values, skills and understanding they will require through their lifetimes and to apply them with confidence, creativity and enjoyment in all roles, circumstances, and environments.

The concept of traditional education does not fit well with this new world and needs to change [ZAFA04]. However, some argue that many universities still remain in the 20th century [PREN05] and are lagging behind technology, and students’ expectations and skills [IVAN10]. Veen & Vrakking [VEEN06] wrote that users such as students are 'digital' whilst Schools and universities are still 'analogue'.

With a broader view, Chickering [CHIC96] identified seven principles for good practice in education, they are as follows:

1. Encourage contact between students and faculty
2. Develop reciprocity and cooperation among students.
3. Use active learning techniques.
4. Give prompt feedback.

5. Emphasise time on task.
6. Communicate high expectations.
7. Respect diverse talents and ways of learning.

Many of these concepts lend themselves to being supported through the use of technology. Encouraging team working, active learning through involvement, prompt feedback, communication and acknowledging differing learning styles are already aspects of the Web, so it is understandable how it could be used to enhance and support Chickering’s principles.

This Chapter shall examine and establish the following:

1. What is driving the changes in the way students learn?
2. How is the Web affecting their research and learning?
3. To what degree will education benefit in the longer term?

In short, it will discuss the current uses of the Web for academic purposes, its potential, and its problems.

2.1.2 Motivating Factors

There are numerous motivating factors behind the way the Web is being used by students as an academic resource. One approach to consider this is to examine is the way students currently use it.

Many students rely on communication tools such as e-mail, wikis, blogs, bulletin boards, chat applications, messaging and video conferencing on a daily basis. Exposure to these 'quick fix' technologies has led students to feel they need information to be fast, available, and exciting whilst allowing selection to be ad-hoc and fractured. Dawson [DAWS07] stated that students have "hypertext minds".

One of the key differences between students and their tutors is that students have more time to expend in experimenting with these technologies, and as a result, and in many cases, the tutors are learning these trends from the students in a game of catch-up.

However, students are not passively engaged with the Web, they are proactive in its uses; they are rearranging the borders of education at an alarming rate; Smith and Salaway's [SMIT09] study noted "a little over a third of the respondents said they like to learn by contributing to websites, blogs, wiki's etc." This is a unique way of learning, with little comparison historically other than informal discussion groups.

Spires [SPIR07] argued that students today “are creating understanding and knowledge in new and different ways". Students already are widely adopting systems for communicating (instant messaging), sharing (blogs), buying and selling (eBay), exchanging (peer-to-peer technology), creating (Flash, Dreamweaver, Joomla), meeting (3D worlds, Google Groups), collecting information (downloads, torrents, peer-to-peer), contributing (wikis), evaluating (reputation systems), searching (Google), analysing (SETI), reporting (camera phones), programming (modding), socialising (chat rooms), and even learning (Web surfing) [PREN05]. If the tutors’ responsibilities include providing for, and preparing students for, a knowledge society [REST07], they need to be equipped with the skills themselves in order to do this.

Might it be that students are driving the changes in education?

Another motivating factor behind the academic Web is Money. The Web has indeed the potential to save universities money [MOXL01] [BHAT10]. Conversely, to position a university in such a way as to be advantaged to other universities through the Web is a costly business; technology needs to be updated constantly.

By replacing a university's physical library with a virtual library potentially shared by multiple universities, there exists an attractive prospect from a financial point of view. But to many academics, the concept that students find using a library a thing of the past is seen as a distinct cause for concern. However, if it is quicker to locate the contents of a book or article on the Web, than to walk to a library and physically find it, it can be seen as a significant benefit. Petroski [PETR05] found that "unless they are specifically constrained against doing so, students assigned to write a research paper will begin and end their research on the Internet"; many academics would consider this not to be good practice and might have trouble coming to terms with this new way of researching.

Another approach universities use to attain more students, as well as to potentially make more money is by opening up to external coursework and distance learning. This allows the university to attain a wider potential student base, including evening courses for full-time workers, and those students too far away to attend regularly.

Others argue that universities are driving the need for change in education; increasing standards, flexibility, access and resources. In a world where universities are increasingly competing for students, results, and recognition, embracing the Web and being ahead, and being seen to be ahead are all important aspects to a universities profile. Web based courses, distance learning, multi-national affiliations, pioneering research, and the use of the latest trends are all great selling-points with the potential to attract more students. Virtual Learning Environments (VLE), also known as Learning Management Systems (LMS) and Personal Learning Environments (PLE) are commonplace in universities and all come under the umbrella of e-learning. E-learning includes applications and processes such as Web based learning, computer-based learning, virtual classrooms, and collaboration. VLEs are becoming "ubiquitous technology" in higher learning [MACH07], and their development has been led firstly, through the private business sector and secondly, through students widespread usage and acceptance of similar systems in social networking.

VLEs would make the transition from more traditional approaches of learning to one of electronic-based learning easier to support. There has long been a perceived novelty effect associated with IT, and with the Web in particular [WELL10], and students by their very nature are affected by such novelties. The use of the word *novelties* belies their tangible values, and many of these novelties mature into accepted, productive viable methods and tools. It can be envisaged how the given Scenario could be supported by a VLE being installed on each user's computer.

Working practices are sometimes seen as a motivating factor. Lifelong learning is widely accepted as a convention in many professional fields, and the learning is mainly technologically based. Wessner [WESS02] wrote "the need for a modern workforce to acquire new knowledge is increasing". Students going into the professional world need to be equipped with the skills to be able to continue to teach themselves once out of university, but it is not just the 'skills to learn' that are important; "we are dealing with [the] required skills and competences of teachers and students" also [IVAN10]. Modern working patterns involve more collaboration between individuals, departments, and even organisations; indeed, these are encouraged.

Higher level research is also following this trend; increasingly, consortia are being used to solve research questions, Su et al. [SU2007] wrote "modern research needs to span multiple organisations to be effective". One such consortium is the Teaching and Learning Research Program's 'Technology Enhanced Learning Project' (www.tlrp.org/) which is investigating the potential of technology to improve the quality of formal and informal learning. The project has been split throughout a number of universities into smaller areas of expertise for better focus.

Research shows that students support collaborative learning on the Web [PUMI07], and rate it between 'good' and 'excellent' for developing group processing, accountability and cooperative social skills. Others argue that there are limited advantages where costs and benefits are unevenly distributed amongst the participants [SHAH14]. It may be that collaboration is itself being driven by technology itself and in particular by Web 2.0 [WANG07]. Web 2.0 can be considered as the 2^{nd} generation of Web technologies which provides a framework to collaborate and share information through. This is very relevant to this research as it highlights the fact that the technologies are present and easily accessible to support collaborative learning; progress should not therefore be hindered by technology.

2.1.3 The Web's Role in Teaching and Learning

Introduction

As the Web is being used increasingly as a teaching and learning tool in education, it is necessary to consider its impact upon the quality and availability of education. It is unquestionable that there is a shift from more traditional teaching and learning styles towards delivering knowledge electronically through electronic text, multimedia, organisational systems, and teaching systems.

Although the Web is not always accurate and current, it has the potential to be both these things. There are great rewards for organisations capable of implementing on-line educational systems which can be seen by the increase in the number of people undertaking education. The quantity and quality of research being undertaken in this field is increasing; Brusilovsky and Peylo [BRUS03] stated that “an interest to provide distance education over the Web has been a strong driving force behind these research efforts”. These driving forces offer new techniques for learning providing the material is pedagogically sound and structured, and Gorder [GORD01] highlighted these benefits.

In order to consider educational systems, it would be appropriate to define the term *education*. The Oxford Dictionary defines *education* as follows:

'The theory and practice of teaching'

The term 'theory and practice' implies that there is a *process* involved in the learning of knowledge; indeed it is a learning process. Whilst discussing education software and systems, special attention should be paid to the processes involved.

In this thesis, the process of learning and the quality of material is paramount, whilst actual topics of study are irrelevant. A pedagogically successful piece of software can be adapted to deliver a number of topics [SONW05] [POPE10], in fact there is a whole research area dedicated to Adaptive Learning Systems.

The term e-learning refers to any learning which may take place on a computer, and more recently this applies to learning undertaken over the Web or by email. It has been adapted from computer based training in the early 1970's where material was provided on disks for distribution, and it mainly relies on the student's self-motivation to complete learning tasks. This means that some students are better suited to e-learning than others. In some scenarios, e.g. distance learning, the tutor is almost completely replaced by the computer, whilst in others where e-learning is purely an addition to traditional learning methods, there is still tutor to student interaction.

The Web offers vast potential for academia, but much material is simply not visible from search-engines; this has become known as *The Hidden Web* [RAGH01]. Chakravarty and Randhawa [CHAK06] stated "the web can be divided into two parts – the Visible web and the Invisible web". The *visible* Web or the *surface* Web is what we see in the result pages from general web search-engines. It's also what we see in almost all subject directories. The d*eep* or *invisible Web* is what we cannot retrieve or see in the search results and other links contained in these types of tools. Much of the deep Web's materials are held behind pay-per-view portals.

This makes it difficult in identifying, extracting and linking sources of learning materials. As will be seen, there also lies difficulty within the direction of the utilisation of the Web; during the late 1990's and early 2000's there was huge enthusiasm for the possibilities of the Web, much of which has now been curbed by the practicalities encountered.

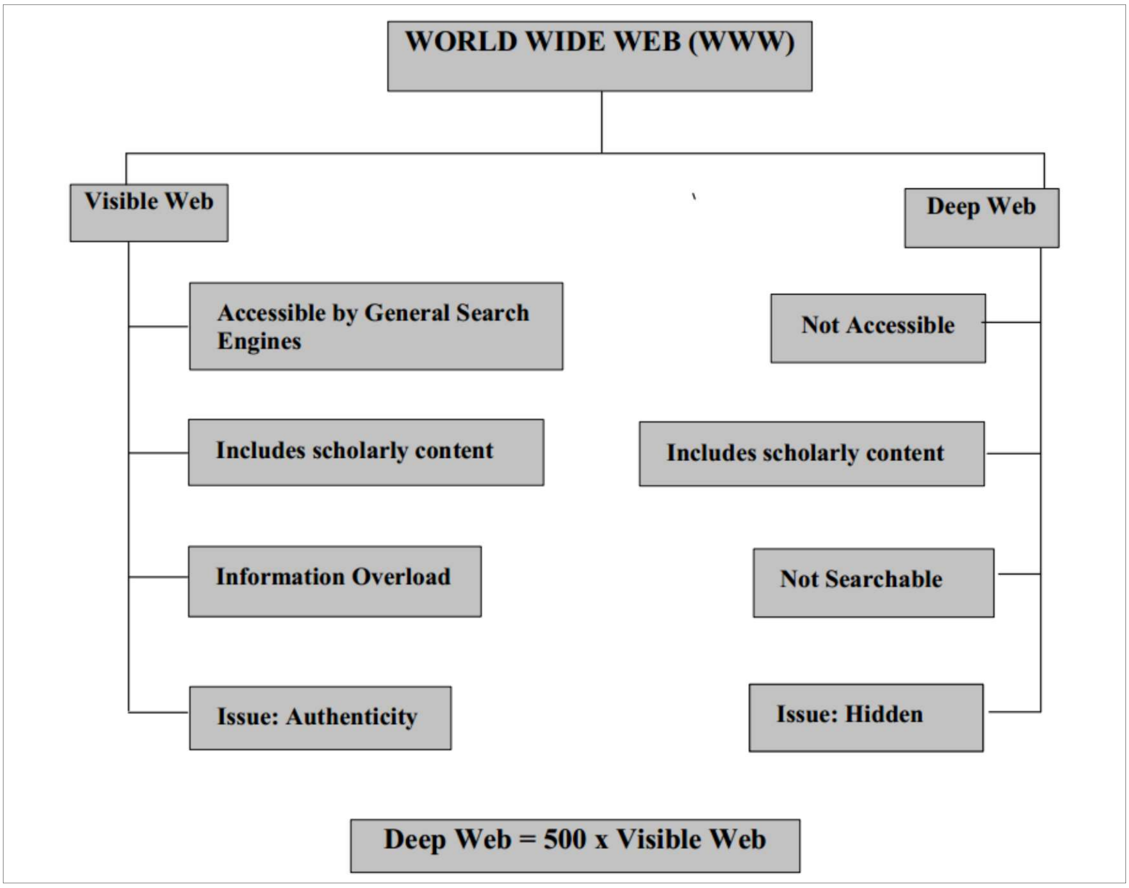

Figure 1.2 Chakravarty's Deep Web

It can be seen in the image above that Chakravarty argues that the deep Web is hugely larger than the visible Web; although much of this will be private and corporate data. In the Scenario, located materials would be proposed to targeted users thus making the Hidden Web slightly more accessible.

Classifying and Orientating Web-based Educational Systems

Identifying and classifying educational systems is a difficult task; this is due to the expanse and variety of those available. The differing technologies, target audience, subjects, objectives and geographics within education all collide to create complex entities. Many attempts have been made to define and create definitions and classification methods; one such, an early attempt by Harasim et al. [HARA93] described seven modes which are either expert based (e-lecture, ask-an-expert, mentor-ship, tutor-support) or student based (access to information, peer interaction and structured group activity).

Collis [COLL99] classified educational software systems into purposes:

1. Publications, information and dissemination.
2. Communication.
3. Collaboration.
4. Information and resource handling.
5. Teaching and learning purposes.

Ausserhofer [AUSS97] used a different approach as shown following.

Software	Description

Online books	These are quite a distance from the concept of educational software; pure on-line books are an electronically displayed version of text. They may allow hyper-links, but require minimum programming to utilise. These are not considered in this discussion.
Edutainment	Educational software which used entertainment has been recognised as a productive method for teaching children. It is a difficult field to successfully create software for. This shall not be considered in this discussion.
Educational software	Users find learning in an interactive environment more interesting than reading a book. Software can be designed to structure student's learning to. There is a large market in educational software and they are generally off-the-shelf packages which are self-contained with regard to information. These are not considered in this discussion.
On-line education There are two main types of on-line educational systems;	1. Hypermedia systems. These have been around for a number of years and allow the student to make selections in what they wish to learn. They may make use of media such as Podcasts.
	2. Asynchronous presence systems. These allow tutors to set tasks, monitor students' work, and provide material whilst the student benefits from 24 hour access, feedback and an environment in

	which they can develop their work. These form the main area of this research.

Figure 1.3 Ausserhofer's Classification of Educational Systems

Ausserhofer's classification splits software designed for educational purposes into four main groups; on-line books, edutainment, educational software and online education.

Ausserhofer's classification has advantages over Collis' for a number of reasons:

1. It is aimed at currently available Web resources
2. It abstracts away, and accepts the technologies inherent in the Web such as communication methods
3. It does not consider the manner in which information is held.

One study based purely on pedagogical goals, undertaken by Ausserhofer was aimed at addressing those concerns listed below. The resulting system called *Surf the Web and Learn* (S.E.A.L.) offered a teaching environment which could adapt to different teaching paradigms and proved successful within its domain of teaching programming languages.

As part of his study, Ausserhofer listed educational software functionality guidelines as:

1. Has overall educational goals. The system should be created with the aim that it should enhance the pedagogical experience and learning capability of the user.
2. Material management. Allows tutors to store, manipulate and display materials.
3. Allows collaboration. Both students and tutors should be able to share and present information. Stahl [STAH95] described such a tool called

Agentsheet, and discussed how to make better use of social interaction, student encouragement, and encouraging less disciplinary and more open learning.

4. Keep information overload to a minimum. Students are using the software to learn a subject, they do not want to be slowed by having to learn an over-complicated system first. This caveat was discussed by Monthienvichienchai [MONT05] in his presentation on the differences between what can be perceived as expected from adaptive hypermedia software, and what is currently available.
5. Offers a user friendly interface and offers individuality. The interface must be easy to understand and learn, flexible and navigable.
6. Adapts to a variety of subjects. The system should be able to adapt to a variety of subjects and its importance was supported by Kareal and Klema [KARE06] and Monthienvichienchai et al. [MONT05], the latter stated that "current and future Adaptive Hypermedia systems need to be able to support the current teaching and learning practice in today's classroom, in order for them to be adopted successfully on a wide scale".
7. Extensibility. The system should be extensible, expandable and adaptable and thus considers the complexities of learning. Learning style affects learning effectiveness. One of the key benefits of adaptive hypermedia is that it can be dynamically suited to students [CONL05].

Finally, from another perspective it is worth considering Mioduser and Nachmias [MIOD02] who suggested that classification should be approached by process; content delivery, instruction delivery, communication support, and creation support. Their approach was an aggregation of several previous perspectives and concentrated on process and flow rather than from a student and teacher perspective.

The History of Educational Systems

Educational systems began their development in earnest in the 1960's, when due to limited technology, material presentation was generally fixed and repetitive and output was restricted through monitor quality. Pankratius et al. [PANK04] stated that with regard to e-learning systems, “the [current] architectures suffer from drawbacks like poor scalability or low availability”.

During the 1970's, role playing games and simulations led to more advanced tutoring systems, some of which were available commercially. However, many of the educational systems available were not adaptive and were no more than electronic books.

There have been a number of attempts at establishing theoretical approaches to educational support systems such as that suggested by Modritscher et al. [MODR04] who identified 4 historical main approaches:

1. Aptitude Treatment Interaction. As early as the 1950's, e-learning was being considered, Cronbach [CRON57] envisioned an adaptable environment to allow differing student learning styles. His Aptitude Treatment Interaction (ATI) approach proposed differing instructions and media types for different students. Although this research was fairly unique at the time, and was not solely aimed at e-learning, a number of learner characteristics where identified which are still in use today. Another aspect of ATI which is still important today in e-learning was concept of the user having control over his/her style of course instruction.
2. Macro Adaptive. This was first introduced in the 1970's [GLAS77] and concentrated on the adaptation of macro level instructions. Users were able to make selections from a few main components such as learning objectives, detail level and system delivery.
3. Micro Adaptive. By monitoring the users responses to tasks, future

instructions can be optimised to the user's abilities and learning style through a diagnosis process. This concept was first introduced by Pressey in 1926 [PRES26].

4. Constructivist-Collaborative. This approach is the most recent, and unlike the previous three, is considered from the point of e-learning. Initially, it began with the concept of adaptive instructions focussing mainly on knowledge and procedure, and more recently it has concentrated more on construction, collaboration and motivation; here, the learner plays a part in defining his/her own style of tuition.

Currently, e-learning may be categorised into a number of main areas:

1. Adaptive Instructional Systems. As listed in Item 4 above.
2. Computer-Managed Instructional Systems. This is similar to Item 1, but autonomously controlled by the computer.
3. Intelligent Tutoring Systems. A more advanced artificial intelligence allows this type of system to be more adaptive than those stated in Item 2. A major positive aspect of this type of software is its ability to act as a tutor in a one-to-one environment. There are some reservations as to the programmed learning principles and it remains an active research topic.
4. Adaptive Hypermedia. This combines content-level adaptation and navigation adaptation within a user-model based interface. Eklund [EKLU97] proposed InterBook; a hypermedia based textual system which allowed any type of subject to be catered for. Brusilovski [BRUS01] has also undertaken a large amount of research in this area, and was heavily involved in the Web-based, ELM-ART project.
5. Other Technologies. There are a number of other technologies which specialise in more defined areas of e-learning such as Collaborative Learning, Virtual Research Environments, and Recommender Systems. It can be envisaged that a model created to encompass the Scenario would

fall into this category.

The level of accomplishments within this pedagogical sphere is not as high as is widely thought and expected; Mioduser and Nachmias [MIOD02] stated that “the impressive pace of growth of Web educational implementations is accompanied by high expectations regarding the potential of the technology for teaching and learning, as well as by a certain level of deception in view of actual level of accomplishments”.

The technology of the Web has been fairly static for a number of years; underlying technology i.e. TCP/IP, HTTP, networks and Web browsers have not changed dramatically. Most of the growth in this sphere is spurred on by network speeds and capacity.

Lennon and Maurer [LENN94] examined the historical developments of educational technology and used various criteria to rank their effectiveness. They concluded that multimedia delivery “enables students to work actively' and represents a “fundamental shift from traditional lecture theatre passivity”.

Research into using the Web for educational purposes has also been less than dynamic [KOPE04] and it is probable that these work hand-in-hand. The technological aspects of Web-based educational systems have been well documented; McCormack and Lester [MCCO97] investigated the benefits, pitfalls, methods, and installation and operational aspects of it. They found that building a Web Based Educational System (WBES) is an evolutionary process, and although they argued that “the medium in which learning takes place does not significantly affect the outcome”, they found that approaches in learning helped improve outcomes. Technological advancements had been expected to put e-learning onto a new level, and the result of this is evident in the drive towards the creation of the Semantic Web [BERN01], a parallel Web which holds additional information defining web page contents.

The Benefits of Web-based Educational Systems

Many see a vision of a Web which works in conjunction with education as one still of futuristic expectation and enthusiasm [AROY04]. Mioduser and Nachmias [MIOD02] stated that “the impressive pace of growth of Web educational implementations is accompanied by high expectations regarding the potential of the technology for teaching and learning, as well as by a certain level of deception in view of actual level of accomplishments”. Technology is seen as the key to Web based education [FORE03] as well as the key to future learning. In this section, the benefits of Web based education are discussed.

The expectations for technology enhanced learning are a result of the expansion of the Web and the shift in expectations by, and of students; communications, collaboration, and problem solving. One of the key aspects to the Web's success lies in its ability to interest users to returning to it day after day, and to rely on its contents. There are a number of reasons behind this; interaction, dynamic content and currency. Many academics agree that interaction is a very effective way of helping people learn. Marshall and Mitchell [MARS03] argued that there are many learning benefits by using resources that “encourage students to engage in meaningful interaction”. Interaction may be seen as the ability of a student to annotate academic materials, or being able to submit work and have it returned with observations attached.

A difficult aspect to the selection of e-learning systems by an organisation is in the evaluation and the quantifying of the benefits of individual systems; the paradigm is still relatively new and evaluation systems may have differing aims and may be designed for differing target groups. The role of evaluation may be best addressed by partitioning a system into functions, such as the user interface, expandability, the variety of domain subjects and content, and adaptability to target groups. One such approach to e-learning system evaluation was proposed by Granic et al. [GRAN07] through their eXtended Tutor Expert System, a method for testing Web-based educational systems. It supported scenario-based, end-user testing, reviewing of the system using a set of tests and guidelines, and examining the user's subjective satisfaction with the interaction. The process runs alongside the system undergoing the test and is an iterative in nature.

Because studies of educational systems have been inconsistent [MACG99], Jen-Hwa Hu et al. [JENH07] undertook an experiment to evaluate the learning effectiveness, perceived course learn-ability, community support, and learning satisfaction associated with such systems. Jen-Hwa Hu found significant improvement in learning effectiveness and that all learning style groups benefited at least slightly from using the educational system but may have found the course learn-ability lower. This is on the periphery of this study as the Scenario does not concern itself with the pedagogical aspect of materials, only in the locating and delivery of them.

It is interesting to note the relevance of Kolb et al. [KOLB90] who argued the benefits to the conceptualising of material by revisiting it, and Jen-Hwa Hu [JENH07] who found that those students who tended to learn through conceptualisation, perceived educational systems easier to use as a learning medium.

Research has been produced which shows that e-learning is highly successful, but there has also been some which argue overwhelmingly against it. Spurlock-Johnson et al. [SPUR04] undertook a study which found that “the majority of students preferred the traditional environment that they were used to learning in”. In her questionnaire to the students, she found that up to 70% preferred traditional teaching methods. Her choice of student was quite distinct in its cross section and some would argue that they are exactly the type of student who would benefit from better learning systems.

There is a belief that by using visual support, interaction in the learning process knowledge retention is greatly increased; Ambron [AMBR90] stated that “people find it easier to learn and remember knowledge visually” rather than being fed passively. A study confirming this point from the student's point of view was undertaken by Fidel et al. [FIDE99]. They found that “[students] whilst searching multimedia resources spent more time and effort browsing and exploring than their counterparts using printed papers, their perception of the value added by their resource was higher”. Carlson [CARL05] reported that in his research he found some students arguing that they can learn better using multimedia materials, such as video presentations from Web archives than from sitting through a lecture.

The majority of older Web-based Education Systems rely on hypertext to provide this ability to alter the course content to suit individual learning styles and to suit the learner's profile. The profile will define the course material and level to which it is taught. A Web-browser is normally used to display the material and is linked to course material held on a database through a scripting language such as PHP, Java or Perl. An advancement to this is Adaptive and Intelligent Web-based Education Systems which offer levels of intelligent tutoring and collaborative learning.

There is also the question of costs associated with Web based educational systems; these are often a driving factor in the decision of introduction of such systems, with pedagogical improvements secondary. Kemp et al. [KEMP02] undertook a study amongst students and found that some students see the introduction of computer-based educational systems as unhelpful particularly where interaction is concerned. The introduction of course-ware was considered by some students as attempts by universities to absolve themselves of their teaching duties and were thought to be part of cost-cutting measures. Despite this, in his study Kemp found no significant change in student learning as the result of replacing traditional labs with virtual labs. In contrast, there have been a number of studies to suggest that e-learning offered poorer quality learning such as Spurlock-Johnson et al. [SPUR04] who found that students felt less satisfied, found it harder to learn, expected worse grades, felt they learned less and would not recommend this method of study.

There are three main areas of costs involved in the implementation of Web based systems. These are similar to most computer systems; hardware/software, development, and maintenance and support. It is difficult to quantify individual estimates where costs lie in general terms due to the amount of factors involved. One organisation may easily be capable of adding a new course to their curriculum with little effect in the way of costs, where another may need to consider new hardware, software or employees. What are certain are the benefits made by an organisation in using the Web as a method by which they can promote their product, in this case the courses on offer; because the courses are made available to a wider audience? It remains to be seen if this will lead to an increase in enrolment to courses or advances in educational standards of future generations, but this does seem likely. Koper [KOPE04] highlighted the direct correlation between increases in course flexibility and the use of e-learning, to the increases in staff workload. However, Koper stated that there were many benefits associated with these potential changes.

2.1.4 The Problems facing Web-based Research, and Educational Systems

As was highlighted previously, the Web offers a huge potential with teaching and learning. This however, does not come without problems which are intensified where exposure to it is more intensive such as with student research. Generic problems have emerged with using the Web in academic environments, and increasingly, as demands upon it are increased, they will become more and more apparent.

Let us consider the incompatibilities between current academic systems and currently available technology. Collins and Halverson [COLL09] undertook a study into these incompatibilities and found the following:

1. Uniform learning vs. customisation. Curriculum and teaching methods offer a one size fits all schooling, whereas technology enables customisation by respond to the particular interests and difficulties of learners
2. Tutor as expert vs. diverse knowledge sources. Schooling is built on the concept that the tutor is an expert, whose job is to pass on their expertise to students. Tutors do not like to see their authority challenged by students who find contradictory information or may challenge them. In contrast, video and computers provide many different sources of expertise and views.
3. Standardised assessment vs. specialisation. The assessment technology employed in evaluating students uses multiple-choice and short answer items, in order to provide objective scoring. Technology encourages students to go in their own direction and to work collaboratively.
4. Knowledge in the head vs. reliance on outside resources. There is a deep academic belief that students should not rely on outside resources, especially in tests. This is the converse of real working life, where employees are often judged their ability to be flexible and intuitive.
5. Coverage vs. the knowledge explosion. Universities tend to pursue the goal of covering all the important knowledge people might upon a subject. As knowledge has grown exponentially, textbooks have also grown. It has become difficult to cover all the important material, and so curricula may increasingly become wider and shallower. Given the explosion of knowledge, people cannot learn in universities all they will need to know in later life. And so they need to learn how to learn and how to find the information and resources they need.

6. Learning by acquisition vs. learning by doing. Deeply embedded in the culture of schooling is the notion that students should learn a large body of facts, concepts, procedures, theories, and works of art and science that have accumulated over time. In contrast, technology fosters a more hands-on, activity-based education and computers are highly interactive and provide a variety of tools to accomplish meaningful tasks.

This highlights a disparity between how universities are operating, and the potential there is through embracing technology. Presnky [PREN05] wrote that "schools are stuck in the 20th century. Students have rushed into the 21st. How can universities catch up and provide students with a relevant education?" Morgan and Bullen [MORG13] argued that educators are scrambling to understand and adjust to the rapidly changing technological landscape.

There are also wider challenges. Groff and Mouza [GROF07] cited a list of what they said would be a diverse set of obstacles to being able to successfully utilise technology-based projects in the classroom:

1. Lack of concrete research and consensus among experts on the objectives and outcomes of technology integration into the academic curriculum.
2. Assorted hardware and software available with unclear support on which ones meet the university’s needs.
3. Lack of tutor input on the development of innovations for instructional use.
4. Pressure and insufficient support in the form of resources, time, professional development, and human and technological infrastructure from the administration, community, and policy-makers to use the technology.
5. Inadequate academic culture necessary to cultivate technology-based

project success.

6. Tutor beliefs, attitudes, and concerns about classroom technology use, i.e. experience with technology, the shift of pedagogical practices, management issues, and the possibility of new roles and teaching styles.

7. Socio/economic challenges associated with technology-based projects such as university culture/goals, compatibility with existing resources, and alignment with prior tutor experiences.

7. Student attitudes, concerns, and experience with technology in general and as an instructional tool, as well as background in the new roles associated with student-centred projects.

9. Problems inherent in technology and computers themselves, such as unreliability and network speeds.

Groff and Mouza argued that while many tutors use IT to increase their own efficiency and productivity, many do not strive to find effective applications for instructional use.

There are a number of problems facing Web based educational systems and e-learning systems in general. Romiszowski [ROMI04] described the areas in which failure in e-learning is most likely to occur:

1. The Product Level. Poor course design, poor classroom design, inadequate technology infrastructure and slow instructor response.
2. The Learner Level. This includes lack of time, interest and motivation, poor study and time management skills and lack of computing skills.
3. The Organisational Level. Poor equipment, feedback, lack of training, low managerial support and a lack of reward structure.

Where researching is the responsibility of the student, other concerns emerge. Because of the Web's structure and size, it is difficult for users to locate quality material on the web; there is no systematic guide showing where a student or tutor should look on the Web for information. Once material has been located, it may not be readily available; much research material is accessible through pay-per-view or toll-gates which restrict their circulation. Although there is no suggestion that the Scenario would be able to overcome all of these issues, it would certainly go some way in addressing the problem in locating quality academic materials through the recommendation system.

Subscription to these journal portals is increasingly a burden on universities [BERG02] where libraries may incur high charges for access to commercially owned academic journals. This view was highlighted through the SHERPA project [MACC02] which sought to place these repositories at the institutional level by creating a national repository. One solution to this problem is in authors self-archiving their works, the best way of doing this previously was in placing the articles on their own web-sites however this is not a particularly effective method. With a system of Open Access Initiative compliant archiving, there is a way to improve scholarly communication and to increase the likelihood of the sought after notion of peer-reviewed material as has been demonstrated by the SHERPA project. This raises the problem of material corroboration; the Web does not inherently provide any way in which a user may be able to corroborate the integrity of material. Though, through peer-recommendation, this would be addressed in the Scenario.

Grimes and Boening [GRIM01] addressed this issue by developing a set of criteria for evaluating the quality of Web resources; authorship, currency, recommendations, perspective, audience, style and tone, quality of content, and organisation of information. However, the manner in which to implement these guidelines is still open for investigation. Their study also considered the gap between the quality of Web resources expected by academics and the reality; they found that Web resources when located on academic Web-sites generally fulfilled the evaluation criteria, whereas those taken from non-academic sources fell below that expected.

Another concern is the method of delivering learning material to students. Gorder [GORD01] argued that students using the Web learned equally as efficiently as those learning from hard copy material, providing that the material was displayed in an ordered method. However, there are those who disagree to this, Dunwoody [EVEL04] for example, stated that “learning on the Web is difficult, mostly because surfers can't devote their full attention to reading. Instead, they must constantly make decisions; which text to read, which hyper-links to follow, or whether to scroll down a page”.

Adapting materials to enable learning is a most difficult task in systems engineering, but is regarded as a key aspect to the success of a system. Clark [CLAR94] even argued that any improvement in learning will come from instructional design and not the type of media delivery. This point was supported by McCormack and Lester [MCCO97] in their study (see section 2.3.3). MacGregor [MACG99] undertook a study into the effects of hypermedia navigational design on cognitive processing which showed that the majority of previous research into educational systems and their effects had been ad-hoc. It is clear that further investigation is needed into the effect of learning styles on their levels of pedagogical success as current research is conflicting.

Research [JENH07] showed that many students still consider technology-assisted learning tools to be just that; tools. They perceived that these tools should support a course as opposed to being *the course*, and that educational systems were sometimes difficult to operate and offered weaker learning community support. Piccoli et al. [PICC01] argued that such students also exhibited less satisfaction from the learning process and as currently, the majority of students engaged in e-learning use it in isolation, i.e. from a library or from home; if true, it could be problematic.

The majority of Web based educational systems development is done by organisations who feel that they may benefit an economical advantage by not making their progress public; this also leads to the creation of many proprietary *standards*. Varlamis and Apostolakis [VARL06] highlighted this fact and argued that such developments generated confusion and decelerated the growth of e-learning communities. They suggested that sound e-learning standards would help the market to achieve some key goals.

Stahl [STAH04] summarised a study into collaborative learning by saying “we need a vision of how networked computers can facilitate the discussion of all with all that does not require the coordination of a manager or tutor and can support the collaborative building of knowledge that is not restricted to the skills, memories and efforts of individuals”.

The Problems of Acceptance

If Web-based education was purely a business concept, this section would possibly be named 'Market Penetration', as it considers the extent to which Web-based education is accepted and utilised. From an academic point of view, it is about how widespread the understanding and belief of the notion is; generally speaking, the more a concept is accepted and expected to be a success, the more research and usage it will get.

Muhlhauser [MUHL04] identified a large gap between e-learning research and its usage, and his findings suggested that success was hindered by:

1. Projects were undertaken with large financial and temporal support; effort and monetary comparisons were difficult.
2. The projects depended heavily on the personal interest of the specialist researchers involved who could not be classed as 'average users'.
3. Continuous adaptation and improvements were not assured.
4. The projects demanded substantial dedication by those involved for success, future customers who already have a full workload were not considered.

Historically, research undertaken by universities is classed as their own intellectual property and they may attempt to maximise the value of this ownership through restricting its use. This means that some of this knowledge essentially does not get used or tested in real situations.

Organisations have also been discouraged in developing and distributing knowledge, technologies and new systems by the associated costs of such projects, and the difficulty in estimating their actual returns or profit.

Kareal and Klema [KARE06] listed 5 main barriers to the adoption of e-learning:

1. Personal barriers. Personal barriers such as attitudes towards learning, learning style and preferences.
2. Organisation barriers. These include points such as lack of time for study and registration problems.
3. Technological barriers. Barriers such as system management quality, limitations of technical support, and loss of data.
4. Content suitability barriers. These include barriers such as content not being audience specific, poor construction and assessments.
5. Instructional barriers. These include items such as lack of progress and feedback, limited learner engagement, poor instructional design, unclear instructions, and information overload.

To embed the characteristics into a system which can overcome these barriers is difficult. Adaptive hypertext is a good starting point and is classed by some as essential in learning systems [KARE06]. Another aspect which hinders the expansion of e-learning is its perception by users and developers. These can be broken down into a number of key approaches:

1. Technological. This includes the effectiveness of tools in being able to store and retrieve information, and the effectiveness of the software in being able to assist humans successfully, in learning and in the retrieval of information.

2. Conceptual. Sometimes, concepts are more important than technology; the concept of the semantic web is relatively simple to understand, but realising its full potential may prove difficult, and its impact may far outweigh any technical advancement which may emerge in the short term.

Both of these may be difficult to comprehend; analysing, designing and implementing a system for learning is complex and difficult. This is probably why tutors are still employed and have not been replaced by automatons. Ausserhofer [AUSS97] stated that “the actual learning process is, of itself, necessarily complex”. The majority of current educational systems are not yet capable of taking into account the inconsistency of human learning.

2.1.5 The Web's Academic Impact

Considering the impact that the Web has had on academia, there has been relatively little research done into this impact. Having said that, *some* research has been done, some of it from a pioneering perspective considering the short period of time it has been available. Windschitl [WIND97] was one of the first to acknowledge and promote the Web from an academic perspective, and acknowledged its great potential, whilst also recognising its shortfalls.

Windschitl suggested that research should focus on a number of key areas:

1. Using the Web for student inquiry.
2. Student communication.
3. Invoking qualitative research methods to enhance Web-based learning.

The following section attempts to break the paradigm down in to manageable and digestible chunks. It considers the Web's impact on teaching, learning, researching, resources, social and cultural, and finally the impact it has had on universities.

The Impact on Teaching

As universities are increasingly expecting tutors to utilise and implement Web based tools and contents, it is important to consider its impact upon them and their teaching styles. Studies such as Connolly [CONN05] show that whilst some tutors, with a positive attitude towards computers in the classroom, eagerly integrate these technologies into teaching strategies and curriculum, other tutors have concerns that through its use, they shall see a potential increase in their own training, preparation, and utilisation time. There are a number of reasons why tutors have such varied opinions and acceptance levels; many of these relate to the subjects which they teach, and the level of their own IT literacy.

The impact that the Web has on teaching may prove to be the largest change in teaching style since formal teaching was introduced. Historically, tutors would make use of texts in the classroom over a period of years in order to spread the cost of their purchase. Because the currency of texts lasted longer in the past, there was no need to replace them normally, until they were beyond repair. The Web is increasingly introduced a concept of immediacy and currency; materials over 5 years old are now considered to be out of date in some subjects, and new information is being added at a staggering rate. This impacts on the volume of knowledge that teachers and students have at their disposal. The contents of the information, not necessarily within a curriculum, but relating to it changes rapidly. Prensky [PREN05] argued that students expect curricula to be flexible and that they should be able to have some influence on them.

VLEs allow tutors to create and organise courses in an online environment. This has a number of benefits to the tutor:

1. Materials and resources can be stored and organised in such a way that

they can be released to students on a timely basis. They also allow the content of annually used courses to be iteratively updated more easily.

2. Courses and their materials can be shared amongst a number of departments and universities.
3. Lectures can be created in advance and archived for another time. This can reduce the pressure of being tied to providing weekly presentations to classes.
4. They offer a number of communication channels to and from the student.
5. They offer environments where tests and marking systems can be centrally located.
6. They allow students to work collaboratively on courses.
7. A better variety of content and delivery media and its timing is possible

A few decades ago, blackboards and more recently whiteboards were the main method of displaying information in front of students. Static overhead projectors were widespread but were too inflexible for day to day teaching. Projectors linked to computers are now common place and are used in universities as the standard method for formal and informal presentations. VLEs reflect social networking Web sites and students who have become acclimatised to their visually appealing, multimedia presentations and flexibility, expect no less. Technology has been proven to enhance learning; Zwyno [ZWYN03] stated that "educational technology can accommodate a wider variety of learning style preferences, and thus increase learning". These tools allow tutors to incorporate a degree of autonomy into the curriculum; giving the students wider ranging projects in which they are able to self-learn at their own pace.

Petzold et al. [PETZ10] wrote that "integration of information literacy principles with traditional teaching methods has proven to be an effective way to improve the teaching and learning experience of instructors, students, and librarians". The Web offers the opportunity for tutors to alter the delivery of a course to suit students' pedagogical styles and to be able to augment their learning materials from external learning media sources such as blogs and wikis.

Grimes and Boening [GRIM01] found that librarians still have a role to play in universities; advising students of web sites, databases, and encouraging the use of electronic resources. They noted that the role of a tutor will change with the Web; tutors will need not only to be qualified as tutors, but as the providers of on-line coursework, they will need to be highly computer literate. Owston [OWST97] agreed by suggesting that tutors' roles may shift from being the deliverers of instruction, to that of being the creators of learning experiences.

The Web provides an unparalleled variety and quantity of resource materials for both tutors and students; not just texts, but also video and audio. Information given to the students is more up to date, and more defined than ever before. The variety and quantity of materials is exponential and comes with both benefits and challenges. Teaching has been affected in a number of ways: content, relevancy, delivery and expectations. As the Scenario was centred around the quality and accessibility of materials on the Web, the teaching aspect described in this section is external to it.

Challenges

Many tutors feel that increasingly, they are playing a game of 'catch-up' with technology, as unlike students, they do not have enough time to spend exploring and learning new technologies [SPIR07]. Some academics such as DeGagne and Wolk [DEGA07] argue that technology can be intrusive and distractive in a classroom atmosphere, and some universities pressure tutors into using technology in the classroom because they believe that is what today's students want. Carlson [CARL05] stated that "Faculty members feel pressured to shorten lectures, increase group-discussion time, and ignore the 'multitasking' student who is e-mailing his friends in the back of the room".

Other tutors consider the ingress of IT to be undercutting their own responsibilities and authority [BLAC05]. Ivanisin et al. [IVAN10] stated that “the level of rejection [of technology within teaching] (or acceptance) in some cases, relates to the course being presented; tutors of more technical courses are more likely to accept e-learning than those of less technical courses such as humanities and human resources".

There remain several main areas of concern for tutors:

1. Their own education of new technologies.
2. How e-learning technologies will be implemented.
3. The long-term effects upon students.

E-learning may mean that students need increased comprehension evaluation by tutors because of the reduced peer-tutor time resulting through its use. Teaching needs to be expanded to include into the regular curriculum, guidance on locating, identifying and managing Web resources. One of the key issues here is the technical and social gap between the students and the tutors; Lozano-Nieto et al. [LOZA06] stated that "Some teachers and students have differing thinking patterns because of the way they have experienced the Internet". Students also gain differing experiences through learning which affects their levels of uptake of new technologies and teaching mechanisms. For example, it is more difficult for students to adjust to collaborative learning where they have not been exposed to it during earlier schooling [WANG07]. Some, if not motivated may be marginalised during collaborated learning, resulting in minimal input and reduced learning.

Britain and Lieber [BRIT00] argued that "training staff should not only consist of improving their technical skills and ability to create content, but also in using appropriate pedagogical styles in the environment". This concern does not only relate to teaching staff; support staff are also involved. In Prensky's [PREN05] research, it was argued even at that time, that many universities were struggling to keep up with technological advancements.

Web-based teaching is not ubiquitous; students will always need a tutor to ask questions of, to propose ideas to, and to have enforce boundaries. Degagne and

Wolk [DEGA07] found that 74% of students felt that labs with a teaching staff were a vital asset to their course.

The Web offers huge potential to tutors, but these positive aspects may carry overheads that the tutor will have to bear until they get 'up to speed'. It is safe to say whilst technology can offer many benefits to tutors, current systems and methods are not at their zenith in this aspect. Delivery systems of the future offer great potential, and increasingly as tutors come into the profession and bring with them the latest skills and understandings of the Web, the academic Web will develop hugely.

The Impact on Learning

Learning has been impacted by the Web in a similar way to teaching and it is closely linked; changes have had a profound effect on users. It is interesting to note how tutors and students have been disproportionately affected by the Web's advancement; tutors will have noticed massive change in course content, expectations, and students' attitudes, whilst students who have grown alongside the Web, will barely have noticed. Where tutors consider the Web from the perspective of a tool to make their job easier and more effective, students have different priorities such as the social aspect, and a more 'just-in-time' approach.

The benefits to learning as a result of the Web are felt by many as the most important area of concern. As the main aim of going to a university is to learn, it is seen as the end result to a successful process. The Scenario identified how this learning process could be improved through collaboration and automation. Although the process of learning would not be directly affected, the fact that if a model which could replicate the Scenario could be created, then ultimately it would provide the scaffolding for more productive research.

Using the Web as a learning tool has numerous benefits to students:

1. Convenience. One of the primary benefits to students is that it is possible to learn at their own pace. All materials associated with the course can be hosted on the Web. This means that students can learn at their own pace and that lectures are not so ‘all-important’, as they may have the luxury of being able to watch them online at a later date. Smith and Salaway [SMIT09] stated that "convenience is the most powerful benefit of Information Technology in courses". This shift from more formal to more informal learning is widely acknowledged and has led to a number of areas of research. This is a new concept which could not have been considered twenty years ago; on demand learning, allowing catch-up and curriculum flexibility. Jones [JONE02] noted in a study that 79% of students felt that the Web has a positive effect on their education. This convenience allows students to be more ‘self-directing’ in their studies; "lifelong learners are acting much more self-directed and they are responsible for their own learning pace and path" [DRAC10].

2. Variety. Because of the Web's ability to break down physical barriers, students have more opportunity and scope for studying. The Open University (www.open.ac.uk) is a good example of this, offering more than 570 self-study courses.

3. Adaptability. Informal learning is an old-fashioned term in the pedagogical domain; it has traditionally been used to describe an ad-hoc learning approach outside of universities. More recently, there has been a discussion about informal learning

   (e. g. Naismith et al. [NAIS04], Conole [CONO06]), and universities have started to introduce more informal learning into their curricula. It is seen as a powerful method for students to understand and reinforce concepts, and student’s use of the Web is seen as one of the most effective mediums for informal learning [CROS07]. The Web also offers the opportunity to allow students to alter the delivery of a course to suit their own pedagogical style and to be able to augment their learning materials from external learning media sources such as blogs and wikis etc.

4. Collaboration and cooperation. The Web also offers support for collaborative learning which has been proven to be more successful than lecture driven learning [SLAV95]. Wang et al. [WANG07] argued that Web-based collaboration "is beneficial to all of the students", and Parrenas and Parrenas [PARR93] argued that cooperative learning facilitates “higher student achievement". But what is collaborative learning? In its simplest form, collaboration is ‘the action of working with someone to produce or create something’, for example, teams of students may work together to solve a set problem or project. Collaboration is not a new concept, Vygotsky [VYGO77] stated "students also learn through the problem solving under adult guidance or in collaboration with more capable peers". Attle and Baker [ATTL07] stated that the strategy of collaboration has a positive aspect of "motivational competition through inter-group competition between collaborative teams". Liu and Fang [LIU006] argued that "collaborative learning is much more efficient for learners to improve their recognition than personalised learning". Wiezel [WIEZ97] wrote "the most simple and effective way of improving the quality of a class is simply organising it so it will take advantage of the techniques of cooperative learning".

5. Tools. The Web offers learners a number of tools not only for academic assistance but also from a social aspect. These may range from message-sending and communicative tools, through to fully-featured learning and content delivery systems.

The majority of the concepts listed above require some form of tool to utilise the concept to its fullest. To many, the Web is actually the *tool*. One of the most important range of Web-based tools are VLEs as previously discussed; their use can lead to changes in the pedagogical paradigm such as:

1. The role of the learner as an active, self-directed creator of content, not purely as consumer.
2. Personalisation with the support and data of community members.
3. Learning content as an infinite ‘bazaar’.
4. The big role of social involvement.
5. The ownership of learner's data.
6. The meaning of self-organised learning for the culture of educational institutions and organisations.
7. Technological aspects of using social software tools and the aggregation of multiple sources.

Source: Schaffert and Hilzensauer [SCHA07].

The result of these tools is undoubtedly positive to students and their ability to learn. Not only is the Web increasing the quality of learning [DEGA07], it is increasing the range of skills that students have.

Not only do the academic authorities believe that the Web is important, they accept that the skills needed to negotiate the Web are also essential [USUN03]. However, the majority of students consider themselves to be well situated academically to make the most of the Web; "7 out of 10 students considered themselves very skilled or expert in their ability to search the Internet effectively and efficiently" [SMIT09]. They also are in agreement about the perceived benefits mastering technology; "74.5% of the total sample agreed or strongly agreed that owning a computer has had a positive influence on their academic performance" [DEGA07]. Attle and Baker [ATTL07] stated that "it is not uncommon for students’ academic performance in cooperative-competitive team activities to exceed the initial expectations of the instructor".

However, there appears to be a disparity between how students estimate their own skills, and the opinions of tutors. Heidi and Barker [HEID09] stated that there is a "recognised gap between the importance placed upon information literacy skills, generally within the literature on information literacy and particularly with respect to students' expected skill levels, and the actual skills that students are able to demonstrate". On the same subject, Hashim et al. [HASH09] wrote that "most [students] do not know how to use and search the information according to what they need", and Rockland [ROCK00] wrote "students need assistance" in locating academic resources on the Web.

The use of the Web has impacted not only on how students learn, but also what skills they need in order to learn. Information is now so readily available, and in such quantities that students constantly need to 'self-filter' what they read and learn. This is a skill which was not necessarily needed in previous generations of students, and may need to be formally taught to students. The outstanding majority of students embrace technology and feel that it is beneficial to their studies. However, an important aspect of modern academic life is learning to develop and maintain personal and professional relationships. Modern teaching methods and academic structure mean that students are expected to interact with fellow students, staff and academic professionals all at different levels, which some students may find intimidating. Students may also need to be introduced to new teaching platforms and environments, social patterns, alternative learning resources and communication methods. These area all need further research to establish whether they affect the quality and validity of submitted coursework.

The methods by which the students learn vary depending upon their cognitive styles. People learn differently; in excess of fifty different cognitive learning style theories and models have been proposed by scholars [ARMS00]. One of the most successfully adapted theories for electronic learning environments is the Felder-Silverman Learning Style Index (LSI) [FELD05]. The LSI model focuses on the different methods in which information is acquired, processed, and understood. Felder’s LSI model consists of four dimensions. They are as follows:

1. *Active* learners learn by trying things out and working with others, whereas *reflective* learners learn by thinking things through and working alone.

2. *Sensing* learners like to learn concrete material and tend to be practical, whereas *intuitive* learners prefer to learn abstract material such as theories and their meanings and tend to be more innovative than sensing learners.
3. *Visual* learners remember best what they have seen whereas *verbal* learners get more out of words, regardless whether they are spoken or written.
4. *Sequential* learners learn in linear steps and prefer to follow linear stepwise paths whereas *global* learners learn in large leaps, and are characterised as holistic.

The Scenario It is argued by some that the Web is more appropriately suited to some learning styles than others [KVAV04]; active-sensing learners may not be as well suited to learning from the Web as, for example, sequential-reflective learners might be. Might this mean that some students are being disadvantaged by universities embracing the Web?

Almost certainly, those students with better basic information technology skills will feel more comfortable and productive with electronic media. Hong et al. [HONG03] stated that students with better basic skills in Web-use perceived a more positive attitude toward using the Web in their studies. Kuh and Hu [KUH001] wrote that “some evidence suggests that the effects of computing and information technology use may not be uniform for different types of institutions or students. Institutional affluence, student ability, socio-economic status, and accessibility and use of computing and information technology appear to be highly correlated”.

Smith and Salaway's [SMIT09] study into undergraduate usage of the Web found that in many cases, how students perceived the effectiveness of technology in learning, related to how they perceived their own level of technology adoption. The effects of the designs of e-learning systems are stronger for low-knowledge students and those with lower spatial awareness.

The Web is also useful because of its ability to coordinate collaborative learning. Whilst many academics agree that this mechanism for learning is beneficial to students [WANG07], [PARR93], [ATTL07], others argue that the academic performance of a group relates to individual member's previous performances [WIEZ97]. Students also agree on its benefits [PUMI07]. Whilst higher achieving students may raise the standard of collaborative group work, conversely, lower achieving students may lower those standards. This may result in higher achieving students achieving lower standards that had the individual been working alone. Other research such as that undertaken by Morgan [MORG03] suggested that students will only make a commitment to collaborative working if they are rewarded for their participation.

Despite the outstanding acceptance of Web technologies as beneficial to education, there exists still a few reservations; Smith and Salaway's [SMIT09] reported that one of the students involved in their study stated that "I feel that when PowerPoint is used, the professor is more likely to go faster, write access information on a slide, and allow less flexibility for the students to ask questions and guide the class". This implies that the student felt that technology made presentations less adaptable, this is contrary to the normal understanding of its application.

There is also the concept that where the total content of a course is available online, then students may deem lectures as less important and only optional. This has not been an issue for many universities so far as lecture attendance is considered as compulsory, and in Smith and Salaway's study, the majority of students disagreed with the line "I skip classes when materials from course of lectures are available on line".

Other problems associated with the Web's impact on learning include:

1. The 'Bright Lights' effect. Because of the Web's visually rich appearance, student's research output might mirror the academic and visual standards found on the Web; where aesthetics of materials may be judged to be more

important than the content. Erdelez noted that “presentation and organisation of information on the Web may have influenced the overall structure of student's information seeking – it was iterative, ill structured, time-intensive, and abundant with accidental discovery" [ERDE02].

2. Personal opinions in Web content can affect student's opinions and can blur the facts. These opinions can be more easily affected through presentation standards in Web pages.
3. The availability of material in an electronic form may make plagiarism through 'cut and paste' more widespread.
4. The speed of information retrieval on the Web can make traditional classroom learning seem dull and protracted.
5. The quantity and availability of ‘distracting’ materials available on the Web; Dawson [DAWS07] noted that 29.1% of students used the university's computers to play computer games.
6. Students are becoming increasingly impatient [CARL05].

Relying on the Web as a source for academic material makes for a challenge as it is such a dynamic field; students increasingly need to learn new skills to work with it. They need to be able to learn new tools quickly, and even work with different versions of the same tool, potentially on different platforms. Slauson at al. [SLAU09] stated that with education, "value is added individually and collectively when a person improves his/her ability to adapt quickly to new requirements".

Universities need to provide as part of their courses, instruction to demonstrate the skills needed to use the Web to an academic standard. Bahiraey [BAHI10] found that some students were uncomfortable with a student-centred approach; he stated that this was a “philosophical issue that needed to be tackled at a different level". Without training in locating, identifying, qualifying and managing resources, there is the potential for two tiered learning; those who know and understand IT, and those who don't. A study found that people with a prior qualification were more likely to use IT for learning, regardless of the course content [ATTE05].

Many universities embrace the Web but are addressing this shortfall in instruction; a 2006 study showed that "students are still not proficient in oral and written communication and lack discrimination when dealing with the amount of information available on the Internet" [LOZA06]. Students may feel more comfortable than ever with using the Web, but may be unaware of the skills they are lacking; "students are now more comfortable with an online keyword search than they are with a print index" [MANU02]. Groff and Mouza [GROF07] argued that there is a large disparity between classroom, professional, and instructional use of computers.

This lack of skills can manifest themselves most obviously through the difficulty students have in making judgements on the academic quality of materials. In a study, students "were not really sure how they could tell the difference between good and bad information on the Web" [LORE02]. If students are to use the Web as an academic resource, both they and tutors must be confident that it is being used correctly. The Scenario describes how, through a recommender system, students may locate suitable material for their studies and recommend it to similarly places students. It would be possible to extend this feature to quite a depth so that students can almost have a discussion about an academic resource.

Cheating and Plagiarism

Where material is placed into a domain which is as readily accessible as the Web, plagiarism has the potential to exist. Pirating software and data, copying academic materials, and the copying of music files are all on the rise. The 'cut and paste' decision is difficult to resist for some, and plagiarism from academic resources as well as from peers is on the increase [RUMB01]. Jones [JONE02] stated that plagiarism from online sources has become a major issue for many academic establishments. Rumbough also found that 17% of students admitted to academic cheating using the Web, and that was over 10 years ago. Lozano-Nieto et al. [LOZA06] confirmed this trend, having written that they had found in their study 30 – 40% of submitted work contained 'cut and paste' materials, as it was simply "too easy".

However, 'cut and paste' is not the only problem which academia is faced with, the following are a suggested list of other activities students can potentially be used to 'cheat' using the web:

1. There is a worrying trend in the availability of pre-created model answers for much of the student coursework.
2. Virtual universities which will provide individuals with qualifications via the minimum of academic work, for a charge.
3. Universities face the problem of being able to identify the author of the submitted work; i.e. has the submitted work truly been completed by the enrolled student.
4. The use of smart phones within examinations and tests to access external data.
5. The practice of sharing results and research between students for submitted work.

Paradoxically, the origins of these problems exist for the same reasons that the Web has become so successful; availability and access to material. Currently, there exists no UK or European regulatory policy concerning Web-based information. This is leading to increased electronic sabotage and content alteration and seems that this trend will continue. There are, however, steps being taken to combat these issues. There are now a range of online services that universities can use to quickly check work submitted by students such as turnitin.com and iThenticate.com

The Impact on Researching

It is in the area of research that the Scenario will be most effective in improving. Sharing located resources, rating and commenting on them, and having relevant resources proposed to the users will all make positive contributions to the researching paradigm. This section aims to examine how the Web is currently being used both by tutors and students and examines how and what students actually use it for, and what impacts it has. This is a most relevant

Head [HEAD07] states that there are 3 stages essential to quality research:

1. Plotting the course for research.
2. Crafting the quality research paper.
3. Preparing the paper, and adhering to grading criteria and citation standards.

As has been identified previously, the over-riding benefit of the Web is in its capacity to provide a storage area for resources. There is no doubt that the Web makes researching easier; resources are essentially in one (virtual) location. Researchers no longer need to physically seek for resources amongst collections and libraries. Universities no longer need to purchase hard copies of materials, be concerned with their physical storage, their condition, or location (i.e. library loans).

Increasingly, academic content is being made available on the Web, and potentially, all influential research will eventually be made available.

To many students, using a library has become a thing of the past. It is quicker to locate the contents of a book or article on one of the computers at the entrance to the majority of libraries much faster than to walk and physically find it; it is even quicker from their own bedrooms. This observation warrants the question that if students are no longer going to rely on libraries as their main research resource, then how do they intend to make use of the Web as their main resource and how are universities best suited at guiding their usage?

So, the main benefits to researching using the Web include:

1. There is a larger variety and quantity of available material; international material, non-mainstream and obscure material.
2. Faster access to materials; less physical searching and quicker downloading to a local computer.
3. Ease of access to materials. Although much material is behind pay-per view Web sites, much of it is also in the public domain.
4. Ease of access from the student location. The student can work from many locations through the use of a laptop.
5. Currency. Materials can be replaced quicker and easier on the Web.
6. Storage. Material is easier to archive in virtual libraries and for students to keep a copy.

All of these points add up to a more productive researching paradigm, and although the Web is less than perfect, students are in an enviable position now compared to a generation ago.

It is without doubt that the use of the Web makes access to academic material quicker and easier. There is however a number of concerns with this situation from a researching point of view. Firstly, the ease by which materials are accessible is seen by some as a threat to researching standards; Lozano-Nieto et al. [LOZA06] stated that "fewer than 5% [of students] used specialised journals and textbooks", and went on to state that "almost all of the students employed the Web as their only source of information". This is clearly a disturbing trend as resources are still yet, not all available on the Web and this could lead to incomplete research and lost skills of researching. It can be seen how in the case of the Scenario, this would be a concern to many academics as it might actually exacerbate the procedure. Therefore, need the academic community have reservations about students not using texts at all for research? Sapp and Van Epps [SAPP06] stated that "engineering students have frequently been known to go four years without stepping a foot in the door of a library". However, this is only problematic if their research is incomplete or badly directed. Heidi and Barker [HEID09] argued that “students are generally unsophisticated information seekers in academic contexts” and that “sophisticated information literacy skills are beneficial to academic success”.

Head [HEAD07] stated in his research that students felt their chances of succeeding in a research project would be improved by:

1. The opportunity of turning in draft papers which are reviewed and returned.
2. Individual sessions with librarians for help with narrowing down topics.
3. One-to-one coaching with professors focussing on how to overcome obstacles.

In her study, she found that the majority of students were not as reliant upon search-engines as prior studies had suggested. Only 1 in 10 students reported using search-engines for conducting research [note: I would think this is now out of date]. This could be a result of the overwhelming amount of information returned for search. This is supported by her statement ‘at first, a majority of students in our discussion groups reported using Yahoo or Google as their first step in their research process. However, further discussion with the participants revealed their search-engine searches often proved useless’.

This is not to say that students do not have methods for researching, even students with no formal Web research training have devised what they believe to be effective methods. In a study undertaken by Erdelez [ERDE02] on graduate students, she found that all students have some form of strategy for information discovery.

These strategies generally involve:

1. A broad and exploratory initial search.
2. Following leads to other sources.
3. Completing the research by focussing on specialised resources such as websites and databases.

However, to formalise a strategy is not as simple as it first appears. Firstly, educational researchers often have different goals and perspectives [REST07], and students need to feel that there is a need to research in a particular modus rather than just because they are directed to do so. Heidi and Barker also argued that students gave less emphasis on the method of finding information, and concentrated on the end product of the search. Secondly, combining the difficulty of formalising a research strategy with the concept of developing collaborative studying using the Web will make for a difficult research question. Resta and Laferrière argued that collaborative learning is a ‘complex concept’ and it has not yet been clearly defined.

The study of Information Retrieval has been on-going for over 4000 years. Tables, indexes, repositories, libraries and classification systems are the resulting tools which were originally manually created. Since the invention of computers, it has been possible to automate these processes. Unfortunately, due to the size of the Web compared to the processing capacity of computers, repositories of indexed data have had the choice of either limited more up to date information, or more expansive but less up to date information.

There are methods which help this process such as the use of meta-data and signature files to identify the contents of documents; these are more easily searched than using the whole document.

Information retrieval on the web can be traced back to Wide Area Information Servers which was one of the first tools to allow indexing and searching. The Gopher browser system was then introduced, after which the Web as we know it came into being. The Web was accessible by browsers such as Mosaic which allowed the user to link with search-engines.

There is a three way trade off with Information Retrieval between precision, recall and speed:

1. Precision (the ratio of relevant documents to retrieved documents) = the number of relevant documents / the number of retrieved documents.
2. Recall (the proportion of documents retrieved) = the number of relevant retrieved documents / the number of relevant documents.
3. Speed = the time taken to return the requested information to the user.

Search-engines need to balance the three aspects above to provide an accurate search result quickly. A slower result may be more accurate but may make the search-engine unpopular with its users. The technologies behind more traditional databases do not scale well for Web usage; the quantity of simultaneous users, and the amount of the data. There is also the difficulty in the added dimension of the numerous types of data needing to be accessed, and the problem of Spamming which is the use of inserting keywords numerous times into documents to promote retrieval from search-engines.

In summarising this sub-section, Dror [DROR07] astutely argued that "the difficult and tricky challenge is how to translate theoretical and academic research into practical ways to utilise technology so as to enhance learning". The Web's ability to deliver research material is most certainly a positive aspect to it. However, there are concerns about the ease of availability, the loss of traditional researching skills and the different methods students need to become successful Web-based researchers. One might argue that research is changing from being paper driven to electronic based, and that although Web research is different, it is still real research.

The Impact on Resources

The Web has the potential to offer a platform in which all resources are available to everyone. Unfortunately however, in some cases it is currently being used as a 'shop-front' for some of these resources. The Web offers wider resource choice, quicker access, access on demand, easier and more opportunities to extract and copy content, and social tools. It provides an extraordinarily rich, expansive variety of resources for all aspects of society; indeed its primary purpose is the storage and distribution of information. This has both positive and negative effects; the positive is the richness and variety, the negative is from the information overload.

The range of resources on the Web is vast and they are considered to be an extremely important aspect to higher education; many students employed the Web as their only source of information [LOZA06], [MOYO05], [GARC04]. This is one of the points identified here in this research, together with the question that should research be made easier (or more productive) through a tool such as could be employed to facilitate the Scenario.

Students are also becoming more actively involved in course content; Smith and Salaway's [SMIT09] study noted that "just over a third of the respondents said they like to learn by contributing to websites, blogs, wiki's etc." Using the Web not only as a resource but also as a palette to contribute to is seen as a positive aspect by many students; they feel that their contributions may earn them renown amongst their peers. This aspect of communal contributing has not yet been widely accepted by tutors [SPIR07]. Greenhow et al. stated that "[The Web] is considered to be a platform on which innovative technologies have been built and a space where users are as important as the content they upload". This approach is argued by some as simply an augmentation to the students learning experience, Spires [SPIR07] noted that "many studies have conceptualised Web use in classrooms as an information repository, and students as recipients rather than producers of knowledge". With the expansion of Web and mobile technologies, sharing, and social media, the use of peer assessment is increasingly being used. Peer ratings are a widespread occurrence on the Web, used in all areas where human reviews are possible, such as restaurant, hotel and travel websites. They are also used in social media services such as delicious (www.*delicious.com*)and Flickr (*www.flickr.com).*

The concepts of peer assessment include the following:

1. Source material. How much does peer assessment affect an article's presence on the Web?
2. Peer skills. The impact on the reviewing peers' skills such as team building, social perceptions, analysis and judgement.
3. Peer understanding. The impact on the reviewing peer’s understanding of the source material (in addition to what that peer would have learned purely through reading it).
4. Peer’s own standards. The impact on the standards found in a peer's own work; exposing peers to good or bad solutions.
5. Author’s standards. The impact a peer review might have on the researcher’s future work.

The Scenario relies quite heavily on peer assessment of materials on the Web. Indeed, any materials proposed to the student will need to have been peer-reviewed prior to being presented to them. The perceived success of the Scenario is partially down to the inclusion of this process. Reiber [REIB06] stated that there were at least four reasons that peer assessment in education is effective, namely; students complete assignments ahead of their due date, students review assignment directions a second time, students submit better work when they know peers will be reviewing, and students react better to peer comments than to tutor's comments. Other academics list different, but similar benefits; Reid and Mason [REID06] stated that peer review can have benefits such as:

1. It will broaden the audience to whom the students are responsible.
2. Students may be required to produce a draft earlier than they might otherwise have done.
3. Students gain an appreciation that they should value other students' input.
4. As students review their peers' work, it can create a greater understanding of the standards required of them.
5. It will reinforce the concept that the results of their work are of their own choice, which can be controlled and modified, rather than one of fixed vision.
6. It can help students check each other's essays and, indirectly their own, against the assignment criteria.

However, some research such as Williams et al. [WILL07] acknowledged that the benefits associated with peer assessment are mixed; during their research, in answer to a question when asked whether "knowing you would be graded by your team-mates as well as your professor changed your behaviour or attitude toward working on a team?", not a single student answered "yes". Despite this, in the same research, they found that students generally welcomed structured peer feedback during the project. Scott [SCOT97] stated that "Peer review improves quality, but its use to screen papers has met with limited success"; peer review may not assure quality.

One of the primary obstacles faced by Web users lies in the difficulty in locating the desired information. Search-engines are the main tools used for this purpose, but a substantial number of search-engines are acknowledged as being incomplete in their coverage [METR99]. There are very few centralised information filters relative to the amount of information available, and fewer still are specialised. Erdelez [ERDE02] argued that at the time, the more specialised a subject, the less effective Web research might be.

Another problem associated with Web resources lies in their authenticity and accuracy; by the nature of the Web, any user can add uncorroborated information to it. The main problem lies when the information added is inaccurate, biased or misleading, either unintentionally or otherwise. It is difficult for users to make an informed decision upon content. Sapp and Epps [SAPP06] suggested that dynamic 'hints' in the form of banners could be placed strategically throughout the websites near to resources, assisting students in how to make judgements on those resources. This is one example of how the Web can be manipulated to solve a problem, however temporarily.

Following, is a list of problems associated with using the Web as an academic resource.

1. The difficulty in locating resource on the Web. Generally, a large amount of material must be read through before suitable materials may be found.

2. Judging the academic quality of resources. Unless the resource comes from an academic domain, or has proven academic value (such as being published in a journal), it is difficult to establish its academic quality. Lozano-Nieto et al. [LOZA06] argued that "much of the materials is of small pedagogical value to the students, is too specific, sometimes sensationalised or purely false". Making such a judgement is not easy for students, especially if not all the information is available such as: author's name, qualification, institution present, the publisher's details, affiliations, citation quality, and currency. Few Web sites have established reputations that can aid users in assessing the site’s veracity.

3. Many resources are biased without transparently being so. This is similar scenario as found in Item 2; unless the resource has been reliably peer reviewed, it is difficult to establish its bias. What is needed is more explicit editorial review policies for analysing content and to verify factual information; this point was argued by Su et al. [SU2007] stating that more peer reviewing of materials is necessary to maintain quality. Indeed, there are views that the Web has a positive effect on the quality of writing by

students. Where a student's work is published, for example on the Web, and is visible to peers, it is argued that the student will be more conscious of the quality of their work and will therefore further develop their writing skills [GRAV73].

4. Resources are less permanent that in printed text form: texts, especially journals and conference papers are invariably permanent, whereas Web content can be altered or removed without leaving any paper trail.

5. Students may go for a 'quick-fix' to their research. Where texts are linear and domain specific, Web pages allow users to jump from one item of content to another potentially unrelated item; Lozano-Nieto et al [LOZA06] wrote that "most of the materials obtained from the Web is not read thoroughly before inclusion". 'Cut and Paste' is made easier. Novelty affects the student's attention. This all can lead to an under researched result. "Left to their own devices, …today's students will compile and print bibliographies that consist of nothing but 'www' addresses" [PETR05].

6. Students may consider that the newest material, or even those most attractively displayed are more relevant than older material. Lorenzen [LORE02] stated that "many students judge the validity of a Web site based on how elaborate it looks".

The points highlighted above identify three main areas of concern relating to the academic Web; quality, content and access. Points 5 and 6 could be classed as student discipline problems and need to be addressed by universities setting their own standards. However points 1 through to 4 are problems associated with how the resources are located, described and/or classified, and how they are made accessible on the Web.

A potential solution to these issues lies in the utilisation of the Semantic Web [BERN01]. With the Semantic Web, Web pages have semantic data associated with them which, whilst not being presented to viewers of the page, is accessible to automated processes. This additional data can be used to identify the contents of the page and more accurately describe and index it. Unfortunately, this has yet to become widespread, and the reasons for this are outside the scope of this document.

Thompson [THOM03] accepted that the Web might be adversely affecting students researching styles and expectations and argued that "if college students are being seduced by the instant access to and seemingly endless supply of information on the Web, faculty need to get involved to counteract the power of the Internet."

Research has shown that by better organising the Web, human users benefit, but Erdelez's [ERDE02] study found, against this general understanding, that unstructured free Web resources proved more productive for student research than commercial online databases in this matter. However, he still pointed out that the former were found to be still lacking. One concept under development is the Open Archive Initiative (OAI); this promotes more access to scholarly communications allowing authors to self-archive their own work in a repository. The notion of an open-access database containing research papers is not new, but historically has had little success. As the momentum for support grows, it looks increasingly likely to happen; Hitchcock et al. [HITC02] stated that "authors are well aware of the potential benefits of open access, but how can they be persuaded to act in pursuit of these benefits?"

The Web is an unprecedented resource for learning materials. However, there are concerns about the authenticity, origin, presentation, and legality of the materials on the Web, and what impact these may have on learning standards and presentation skills over time.

Social and Cultural Impacts

Over the past 15 years, students' academic and social lives have been increasingly transformed by the Web and its technologies. The wide acceptance of smart-phones, laptops, social networking, collaboration, immediacy of information, ease of access and team working have all had their effects. Students are now more in-touch whilst potentially being more distant, more informed whilst being potentially more ill-informed, more skilled in some areas whilst potentially being under-skilled in others, more focussed yet being more distracted; and the main reason is because of the Web.

It is arguable whether the Web is driving social change or if it is the other way around. Users certainly create the drive behind a market such as a new social media experience, but are they just responding to the novelty of the market once they are aware of it? The Web brings with it a diversity of social aspects, many of which have not been encountered historically. One of the most significant uses of the Web for students is for social and communication purposes. Social media can be defined as Web-based applications that allow the creation and exchange of user generated content. 75% of Web users are members of one or more social media Web sites. Kaplan and Haenlein [HAEN10] stated that "as of January 2009, the online social networking application Facebook registered more than 175 million active users. Every minute, 10 hours of content were uploaded to the video sharing platform YouTube, and, the image hosting site Flickr provided access to over 3 billion photographs".

It is valuable to consider the motivational effect which the web has on its users. Ivanišin [IVAN10] stated that for students to be motivated they need to be partly sure their engagement in a course will be rewarded. It would be interesting to establish categorically if students felt ‘better rewarded’ through using the Web compared to taking part purely classroom taught courses.

Despite these highlighted problems, the Web is still an outstanding tool, and its use may only become a problem through misuse and lack of understanding. With proactive guidance and training from universities, the Web should become increasingly an asset. From this short study, it appears that although the Web is an opportunity as a research resource, it also brings with it some problems. Locating resources and in making academic judgements on those resources is still a concern, and the degree to which students make use of the Web is also varied; "the division between the digital ‘haves’ and ‘have nots’, has not been entirely overcome" [JISC09]. This study proposes that this change in culture will persist and continue to increase, and like the Scenario, Academies must scaffold this.

Changes in Communication and Knowledge Sharing

With the technology of the Web, methods of communication and knowledge sharing in education will alter, but it is believed that human to human interaction will remain important alongside the use of tools such as e-learning systems and Agents. Although such contact will be minimised in long-distance learning, in universities, e-learning is still considered by some as an additional method to *top-up* learning.

These observations pointed towards the student's opinion of the importance of human interaction rather than the benefits of technological expansion. It may be that the students took for granted the available technology such as the Web, and were concerned about the level of human interaction for the future. In the Scenario, physical human interaction was low, whereas virtual interaction was quite an important aspect. However, Nahl and Harada [NAHL96] found that despite the increased levels of information being available through the Web; many students still lacked language and information seeking skills. This reiterates the need for standard teaching practices as well as knowledge gained from the Web.

Email is one such technology and has become an important aspect of educational establishments, from primary level up to post-graduate. The use of Computer Supported Collaborative Working encourages the organising and recording of interactions and enables the creation of work-groups who can interact with one another in real time. The addition of semantic information to these stored interactions can only help in the indexing, storage and retrieval of such sources.

E-learning Standards and Specifications

It is unfortunate that from the outset, researchers and developers within this field had not cooperated in establishing and adhering to defined standards, protocols and architectures [VARL06]. Although it has been noted that there is still an independent undertaking of research and development, these were the minority and were considerably less widespread than others. By creating and adhering to standards, the community can benefit from:

1. Scalability. The functions within learning technologies can be built upon, much like building blocks.
2. Manageability: This is the ability for teaching to be able to define, assemble and manage a course.
3. Interoperability. Course content from one provider will be usable by other providers' software. Users are not restricted to their initial provider.
4. Accessibility: The learner can access the appropriate content when suitable.

One of the most influential standards in e-learning, SCORM (www.adlnet.gov/scorm/), the Shareable Content Object Reference Model which is composed of a set of standards for storing, distributing and displaying learning materials though computer systems. SCORM compliant learning systems can create, export and import learning objects and can contain a variety of media and it is typically designed to be used as self-paced instruction.

The Future

The future of education will be an exciting place to be; it will be filled with rich and multi-faceted media, dynamic and varied syllabi combined with varied learning approaches. The Web has the power to change how students engage in learning

[GREE09]; they will have more options on what, when and how they learn.

Advances in technology have led to a positive shift towards the delivery of knowledge via electronic means; e-text, multimedia, organisational systems, and teaching systems are becoming more widespread today. The students of today have grown up with the Web and all it entails; blogs, emails, texts, multimedia, social networking, collaborating, sharing and immediacy. Despite some determined concerted efforts, the use of these mediums has not yet fully utilised its potential.

Collaboration, sharing and transparency will replace independent, insular and competitive learning. Just-in-time and on-demand learning will become more commonplace because of the quantity of information available and students will be encouraged to take part in "collaborative, inquiry-based learning, centred on 'real-world' problem based scenarios” [MOLE00].

Because of the Web, universities will move from a position of providing tutors, supplying information and standards to students as it does presently, to providing a scaffolding system to support learning. Although there has been a substantial and rapid growth of research studies focussing on most aspects of technology-supported learning [REST07], some academics argue that there is still much to be done [IVAN10].

The Effects of embracing the Web may be more far-reaching than can be currently anticipated. There may be fluctuations in research and presentation quality, speed of research, social impact, learning skills, and curriculum and funding impact. Can we place a limit upon how literate we want our students to be?

Despite not having a clear understanding of where the Web currently lies within education, and difficulties in understanding its future role, it may potentially offer education a golden future as a provider of information, and as a host of learning environments. Lorenzen [LORE02] argued that the future looks bright providing we acknowledge the Web is “not perfect”.

To paraphrase Collins and Halverson [COLL09]:

"The revolution in education will alter not just the lives of students, but the entirety of modern society"

2.1.6 Summary

This section has highlighted the benefits and problems associated with academia increasingly being reliant upon the Web as a resource.

It has considered both how students and tutors use it, regard it, and foresee its future usage. It has shown that many tutors are concerned about the pace that the Web and reliance upon it is developing and that the majority of students are embracing the Web at a faster pace than them.

It has also suggested that the skills needed to use the Web more effectively need to be part of standard curriculum, and that relying on it as a learning tool may better suit some students than others. It has shown that offers great potential as an academic tool, but can also be a distraction, and it makes plagiarism in many forms easier.

There is a strong relationship between how the Scenario operates and the concepts within this section, and were the Scenario to be modelled and implemented, the salient points would have to be considered.

The research has also considered the Web's effect on source quality and access, and on how pedagogy currently is, and how it might change.

It has highlighted a number of observations and concerns with using the Web as an academic resource:

1. The time it takes to locate academic materials on the Web. As the size of the Web increases, the proverbial haystack in which students must sift through is growing. This makes it increasingly difficult for students to find the materials they are after; searching time is thus increasing, impacting on research time. In many cases, there is limited meta-information available about such resources 'at first glance', and many of these materials are poorly described online.
2. Synchronous or repeated research. In some cases, students are undertaking synchronous research on the Web in their attempts to locate similar resources.
3. Identifying the quality of materials found on the Web. Even when students locate materials on the Web, it may be difficult to identify the academic value of the given material. Historically, where material was sourced from an academic journal, the journal could be used as a guide to the quality of that material. But where material is sourced from a Web page, the student may have trouble in establishing the academic value of such material based on the Domain it was drawn from. Much of the material available is of small pedagogical value to the students, is too specific, sometimes sensationalised or purely false.
4. Gaining physical access to materials hosted on the Web. Much academic material on the Web is actually hidden behind 'paid access' portals. Where once, articles may have been reproduced from a journal, paid for by a subscription through the universities, there are now too many of these 'paid access' portals for such universities to subscribe to them all, if not more than a few.

Learning systems which harness the Web and are used in universities need to be improved; learning environments are still not over-popular with students [WELL06], have had limited success and are lacking in both their design and direction [WILS06]. Some academics have trouble coming to terms with this new way of teaching and learning, and even though its future cannot be predicted, it is with fair certainty that no matter how far the Web misses its target as a learning resource, it will play a major part in the future. As research is still on-going, it seems difficult to draw that research to a conclusion as the technological 'goal-posts' are constantly changing, and studies are changing how we approach and use, and the outcomes of use, of the Web.

2.2 Technologies for Researching and Learning on the Web

Technology reduces the amount of time it takes to do any one task but also leads to the expansion of tasks that people are expected to do.

Juliet Schor – professor of Sociology at Boston College

One of the most useful features of the Web is in its ability to organise and interact; collaboration between users, synchronisation of documents, access of materials, the bringing together of communities, and blogs. These features open up a whole new world of possibilities which can create a truly ‘Global School’ where students can study any curriculum from any location at any time. Behind these features of the Web lie a host of technologies, many of which are becoming commonly used by students. This following list is by no means exhaustive, but aims to highlight the diversity in approaches available.

2.2.1 Academic Learning and Researching Environments

The term Academic Learning Environments can be seen as an umbrella title covering all forms of computer systems that provide a means of allowing student access to learning materials. They are primarily content delivery systems, but may also provide additional services such as communications, storage, and sharing.

Virtual Learning Environments

The Joint Information Systems Committee (*www.jisc.ac.uk*) classes a VLE (VLE) as a computer system who's users, learners and tutors participate in “on-line interactions of various kinds, including on-line learning”. VLEs can also be called learning management systems, and e-learning systems.

They can be generally broken down into a number of more defined sub-system types:

1. Learning Managements Systems. Used by both Students and tutors to manage and present courses, tracking, communing, and activities
2. Content Authoring Systems. Allow tutors to create and distribute learning materials.
3. Adaption Systems. This is a core system allowing the user to tailor the contents to suit their individual style, goal, level, and pedagogical method.
4 Assessment Systems. This supports the creation and maintenance of a marking system and may incorporate facilities for the delivery of tests. They may also include feedback capabilities to highlight where students knowledge is lacking.
5. Content Delivery Systems. These offer the ability to distribute learning materials to students.

A VLE may offer a number of capabilities to its users, both tutors and learner. These may include: course content and management tools, communication tools, peer assessment ability, collaboration, and tracking facilities to name but a few. The first systems that fitted the criteria of a VLE as would be recognised today emerged between 1995 and 1997 [STIL07] with developments such as Cecil (www.cecil.auckland.ac.nz) and WebWork (www.webwork.maa.org).

Often VLEs were, and still are, used for distance learning and are often considered purely as material delivery tools, rather than management systems. They seemed an ideal method of delivering course materials to students. They would allow all of the course material to be made available to students on a timely basis, even on-demand. Tests could be done on-line, as could their marking. They could be seen by administrator as the future of education.

However, VLEs were not, and still are not an over-popular tool amongst students [WELL06]; they are seen by many as highly content-focussed, tutor oriented, technically insular, and generally only adequate. They are also thought to require a steep learning curve in order to make initial use of them, and some have had browser compatibility problems.

Despite this, almost all higher education establishments within the UK operate a VLE in one form or another. Wilson et al [WILS06] considered VLEs to be a “dominant design not optimally used”, and are seen by some as successful mediators between students and teachers [KENN09]. Chou and Liu [CHOU05] stated that "VLEs offer a better chance for students to articulate their thoughts and understandings". Others [STIL07] argue that they are being used by universities merely as a way of keeping 'hold' on education; to stop the Web from its complete domination.

Professional VLEs such as those employed in the medical field have had more success than their educational counterparts due to their ability to match the current business model. However, for whatever reason, budgets being potentially the most influential, academic learning environments have not been able to manage such exactitude.

Attempts at creating VLEs have had limited success [KENN09]. In 2004, the Open University's initiated a study into the viability of creating their own VLE. After a thorough review, it was decided that an amalgamation with another currently available VLE called Moodle (hwww.moodle.org/) was the best option due to cost and time constraints. One major problem for development is that the technology is fluid; new concepts and approaches are being heralded on a regular basis as 'the next big thing'. This creates uncertainty in the departments whose responsibility is their implementation; the value of continual change seen within this area of technology is difficult to quantify. A VLE should be pervasive and support nomadic learning [JACO00] and its implementation should be a continual process [BECT03].

The main benefits of VLEs include:

1. Assessment. Because of the nature of VLEs, they are mainly used for formative assessment. However, if managed correctly, summative assessment is potentially possible. Feedback can be given either automatically or manually by the tutor.
2. Communications. The VLE can allow students and tutors to communicate more easily, and less formally. Whilst some academics argue that VLEs increase the level of communication [CHOU05], others argue that it can be frustrating [SCHU97]. Communication may come in the form of emails, bulletin boards and instant messaging.
3. Delivery of Materials. All types of media can be used in course content; text, audio, video, embedded presentations etc. Content can have timed releases and time scales. Out of hours access to materials is an important factor to the students.
4. Collaboration. VLEs allow students to engage in collaborative work and content sharing. Whether formally, or informally, multiple students can access the same content and edit it in real time. Many VLEs have virtual 'work areas' where grouped students can work together, yet remotely.
5. Management tools. The tutors benefit from VLEs in a number of ways especially with regard to managing a course. All the content can be located in one place and released to students when needed. An entire study course can be installed into place before it actually begins so that only minor changes need to be done. The tutor also benefits from tools such as online marking of work, immediate feedback to students, as well as tracking and monitoring tools.
6. Student tools. The student may also get a set of tools to work with such as a diary, and data storage.

VLE uses and roles vary from one institution to another to support distance learning, virtual learning, or are blended with traditional learning. It has been acknowledged that theories of learning have shifted from ‘behaviourist’ towards ‘social constructivist’ models [CMAL03]. There is now an increased emphasis on social learning which can be suitably supported by VLEs. This theory shift has also affected how students perceive these systems; LMSs are most frequently used by educational institutions and tutors, but they are considered different from social networks tools by students who use these on a daily basis.

The majority of VLEs support the Shareable Content Object Reference Model (www.scorm.com) [ZAFA04] which is a set of standards which at their simplest guide a developer as to how their code should cope with interoperability.

There are a number of VLEs currently available and they are mainly categorised in open and closed source versions; as a rule, open-source versions are generally freely available. Such VLE's include Moodle (www.moodle.org), SAKAI (*www.sakaiproject.org), Blackboard (www.blackboard.com), and SharePoint (*www.*sharepoint.microsoft.com). With regard to how widespread VLEs are,* Pishva et al. [PISH10] stated that VLS's “have become an integral component of the education systems in most universities”.

Collaborative Environments

Collaboration can take place in many ways; from simply giving a ‘thumbs-up’ on a website to show that something was liked, to using specially designed Collaborative Environments (CE). Web sites that make use of a user rating system can cover all manner of content; movie reviews, recipes, news, and music.

CEs offer a more formal approach to users giving their opinions, and they allow users to be actively involved in a given project. An environment in this context can be seen as a piece of software, or collection of Web pages which give the users a virtual area in which to focus their efforts. The boundaries between collaborative environments and some VLEs can often be blurred. There is also a blurring of boundaries between those specifically designed for sharing, and those where sharing is an incidental by-product. The majority of collaborative environments are Web-based because it offers such an excellent network distribution hierarchy.

Collaborative environments offer a number of facilities such as:

1. Creation of groups.
2. Storage of materials.
3. Provision of materials.
4. Allocation of responsibilities.

CEs provide unique, immediate, pervasive, and semi-public access, based on their preferences to materials, and are seen as a key tool in learning, increasing knowledge availability which was traditionally embodied in publications, teaching, and presentations. They may be considered to be a dynamic knowledge base.

Capra et al. [CAPR10] stated that in their research, participants not using specifically designed collaborative environments reported using a variety of other methods to communicate, collaborate, and share information found from searches with each other such as:

1. Email (70%).
2. Face-to-face (67%).
3. Telephone/conference calls (37%).
4. Printouts (37%).
5. Blogs/webpages (17%).
6. Shared network hard drives (13%).
7. Instant messaging (10%).

This list gives some idea of what is needed in an ideal collaborative environment.

The level of collaboration can vary from tool to tool. One approach to collaboration is to allow users to join specific content/interest groups and then enable them to make reviews or leave feedback upon those groups. The term computer-supported collaborative learning was used as early as 1979 by O'Malley and Scanlon [OMAL79] and was recognised by Koschmann [KOSC96] as an important area of research focus in 1996. The primary aim of CSCL is to provide an environment that supports collaboration between students to enhance their learning processes.

There are a number of collaborative tools currently available such as:

1. CiteULike (www.citeulike.org), an academically orientated social bibliography site.
2. Sakai (www.sakaiproject.org), an 'Open Academic Environment'.
3. Delicious (www.delicious.com/), a social bookmarking service.
4. Zotero (www.zotero.org), a freeware academic bibliography program that allows compiling of academic citations from library databases.
5. Endnote (www.endnote.com), which is a bibliographic tool aimed at publishing researchers and writers.
6. CiteLine (www.citeline.mit.edu/) organises and shares collections of bibliographies and citations.

Collaborative learning has been seen for many years by the majority of academics as wholly beneficial to the student learning process; Astin [ASTI93] wrote "research shows that students who participate in successful group projects, collaborative learning and cooperative learning exhibit a higher level of academic achievement, better critical thinking skills, a deeper understanding of learned material, and more positive and supportive relationships with peers". An important aspect to collaboration is the kudos that users can achieve for themselves. This can be attractive to both students as well as tutors.

Wiezel [WIEZ97] stated that his research demonstrated "the superiority of [the collaborative] teaching system" when compared to traditional teaching methods. Bahiraey [BAHI10] argued that hypotheses of these types would be difficult to support because of the lack of any formal methods in measuring quality of learning.

That aside, because of technological advances in the recent past, "new possibilities for the design of our next generation collaborative educational environment to harness collective intelligence" [SU2007] have been created. These tools provide immediate, ubiquitous, and semi-public access to an academic knowledge that previously had been largely contained in tutors' heads, in filing cabinets, or in self-contained computer software. The key to these emerging, open, student-oriented technologies is that the expertise usually embodied in one's publications, teaching, and presentations now becomes part of a much larger, collaborative, and dynamically updating online knowledge base.

Collaborative learning and in particular collaborative learning environments are not without their critics. Stiles [STIL07] argued that "there are indications that rather than resulting in innovation, the use of VLEs has become fixed in an orthodoxy based on traditional educational approaches". It has also been argued that collaboration does not suit all academic styles. Kolb et al. [KOLB90] demonstrated a quadrant approach to learning styles, and one of those compartmentalised styles was called 'assimilating'. Students who approach learning using this method may not find collaboration an easy approach. More obvious still is that other students such as those who are shy, socially under-experienced, or non-native speakers may find it difficult to take part in collaborative learning. A concern is that many of these tools do not attempt to qualify academic value or source of the materials.

An interesting point to note is that some academics (Blackhall 2006 [BLAC05], Wilson 2005 [WILS05]) question why universities need formal systems such as VLEs to replicate what is already taking place on the Web quite successfully. This is a valid point, but can be countered by arguing that those informal systems are distributed and fragmented. One notable proposal which the author encountered was the DiCore environment as described by Su et al. [SU2007]. This proposal for a Web-based environment to incorporate multiple modules for collaborative research, despite only being a proposal, was well thought out. It featured:

1. A Web-based collaborative citation management tool.
2. A Web-based sharable research reading note management tool.
3. A virtual collaborative research group management tool.
4. A collaborative discussion board management tool.
5. A collaborative research project management and tracking tool.
6. A collaborative researcher paper writer.
7. A collaborative recommended research reading list generator.

Su stated that its objectives were:

1. Lowering the costs of large research collaborations.
2. Encouraging collaborative learning, knowledge sharing, and discovery.
3. Relieving researchers from daily house-keeping tasks.
4. Enabling a new way of providing qualified educational materials.

It can be seen how the implementation of such a tool draw together a set of tools which students would find highly useful. It also benefited from being Web-based.

Peer assessment systems (PAS) can be considered to be a set of tools for grading, assessing, and learning [WESS02]. Peer reviews are generally considered to be fair method of assessment, reduce errors in submissions, and increase submission quality.

Whilst peer review and assessment have a history of research associated with it, research into online peer assessment and its affects has been scarce [KNIG11]. However, there have been a number of key pieces of research undertaken into the viability of employing peer assessment in classroom scenarios, and from several viewpoints. These consider the effects on areas such as the author, the assessing peer, material quality, and material availability.

One of the major hurdles to online peer assessment is finding an environment which suits the working style of an organisation. There are currently a number of online systems designed specifically for managing peer assessment such as the Open Journal System (www.pkp.sfu.ca/ojs/) which is an open source web-based system designed for hosting journals. However, the majority of these tools are 'one-size-fits-all' and potentially not flexible enough for most organisations. These types of tools are simple enough in concept, but due to the decentralised nature of peer to peer networks, the paradigms of responsiveness, authenticity and resilience are not as easily satisfied. There are major differences between an assessment system distributed on the open Web, and one hosted in a central location such as with on-line marketplaces like eBay and Amazon.

Another hurdle is security. Not only does this type of system have to handle the atypical type of Web based attacks such as *Denial Of Service* attacks, they also have to be able to handle content based malicious behaviour such as whitewashing attacks, false reviews, trolling of reviewers, discrediting of reviewers; all of these degrade the performance and accuracy of such a system.

Taraborelli [TARA07] argued that peer assessment has been criticised on various grounds and its major threat is scalability. Tian et al. [TIAN07] noted that Structured Supervised Rating systems are more accurate and more responsive than self-managing ones. There have been a number of alternative systems and concepts proposed; Taraborelli [TARA07] proposed a 'soft peer review' system based on the social bookmarking of academic materials. However, he pointed out that it could not offer the same guarantees as standard selection processes.

Where peer assessment systems have been used, there have been both opponents and supporters. Some research ([ARMS91], [SCOT97]) argue that peer reviewing procedures seem to discourage academic papers which scientifically advance a subject; papers with controversial findings or conflict with current beliefs are sometimes harshly reviewed, or judged to have defects. Peer reviewing appears to work more successfully where the subject is better known, although reviewers may have difficulty in reviewing a little known or innovative subject.

Some studies noted that students may feel that the quality of students' comments on each other's essays are, at best vague and unhelpful, and often misleading or incorrect [REID06]. This appears to conflict with Reiber's research [REIB06].

There have been a number of suggestions to improve online peer assessment systems such as incentive mechanisms [TIAN07], blinded assessment [FLET97], and limiting the quantity of peers for example, friends [SINH01].

Recommendation Systems

Recommendation Systems (RS) have been researched in depth in the past by notable names such as Balabonovich [BALA97] and Bollacker [BOLL99]. They have explored how the similarities between users, or between articles can be exploited. One important aspect to the use of computer aided learning environments is in the use of metrics in measuring how effective they are. This is a difficult evaluation as it is a new area, and many of the developments are still on-going. It is also difficult because the systems may have differing aims and be designed for differing target groups.

The role of evaluation may be best addressed by partitioning a system into functions, such as:

1. User interface.
2. Expandability.
3. Variety of domain subjects and content.
4. Adaptability to target groups.

One of the fundamental problems with recommender systems is where, as a result of the systems processing, it recommends inappropriate items. This has a negative effect on the user's opinion of the system.

Virtual Research Environments

Virtual Research Environments (VRE), comprise of digital infrastructure and services which enable research to take place. They are normally associated with an institutional supporting structure and at minimum offer administration and management facilities for the sharing and reuse of tools, data and results. Such sharing assumes the implementation of standards for data representation, indexing and access. In the majority of cases, the VREs are owned and managed by the research communities and isolation would be very unrealistic. One such example of a VRE under development is *The Building a VRE for the Humanities Project* which is being undertaken within the Humanities Division at the University of Oxford. The main aim of this project is not in creating a system in its entirety, but through surveying, establishing a set of priorities for VRE developments. Fraser [FRAS05] states that with regard to the BVREH Project "initial results suggest that the overall priorities …are central hosting and the curation of digital components of research projects, and the potential for a VRE to facilitate communications".

Shaw [SHAW03] stated that there are 6 steps to online research, and they confirm Head's recommendations, breaking them down slightly more:

1. Questioning. The student must understand the assignment.
2. Planning. The duration of the work, where to look, number of sources, who they will be working with.
3. Gathering. Making use of as many sources as possible. Make aware of the issue of reliability of information on the Web.
4. Sorting and sifting. This is ordering information into categories.
5. Synthesising. Mapping all the information into one report, possibly through the use of a concept mapping method.
6. Evaluating. Does a completed paper meet requirements? They should be prepared to re-write parts.

Some students see the introduction of VREs as unhelpful particularly where interaction is concerned. Kemp et al. [KEMP02] stated that the introduction of course-ware was considered by some students as "attempts by universities to absolve themselves of their teaching duties". Despite this, in his research, he found that "there was no significant change in student learning as the result of replacing traditional labs with virtual labs", although he acknowledged that "the focus of this work was concerned with reducing costs in terms of laboratory equipment, laboratory space and demonstrator time" which is not necessarily a pedagogical view.

Although the underlying system behind the described Scenario could not be classed specifically as similar to any of the above, it can be seen how an amalgamation of many of their features could be used to facilitate it.

2.2.2 Resource Sharing within Learning

There has always been a need for academics to remain current in their field of research. Historically, periodicals were used, but as the quantity and diversity of both periodicals and the information printed in these has increased, it has become increasingly difficult to keep them up to date. This prompted the introduction of digital libraries, but still it is necessary for researchers to spend time looking for relevant publications. In the Scenario, materials are offered to students from both other students and from autonomous searches. In effect, this is a simplified version of sharing materials. However, there is a fundamental problem with Web-based materials and that is that the majority of research materials held on the Web are shielded behind toll-gates which restrict their circulation. Subscription to these journal portals is increasingly a burden on universities. One solution to this is in authors self-archiving their works. The best way of doing this previously was in placing the articles on their own web-sites; this is not a particularly effective method.

With a system of OAI-compliant archiving, there is a way to improve scholarly communication and to increase the likelihood of the sought after notion of peer-reviewed material. The SHERPA (*www.sherpa.ac.uk*) which stands for 'Securing a Hybrid Environment for Research Preservation and Access' project seeks to place these repositories at the institutional level by creating a national repository. Search facilities will be created by other projects. The main challenge is a cultural one; convincing academics to be involved in the initiative. There is a widely held view that freely available research material on the Web is of low quality, for this reason, SHERPA aims to place a priority on refereed material.

The easiest solution is that all existing documents and all future created documents are placed onto a central repository. There are a number of problems with this:

1. Needing incentives so that authors readily contact the repository to notify

it of new papers.

2. Copyright infringement issues.
3. Knowledge hiding. Some papers may contain knowledge which the author may not want in the public domain.
4. Storage and scalability issues.

But as Cameron [CAME07] pointed out, such a repository would both ensure that publications in any form are visible, but not necessarily accessible, and, allow fairer competition between publication venues.

There are a number of viable repositories currently in existence:

1. The Berkeley Personal Libraries (*www.berkeleypubliclibrary.org*) provide cataloguing and full search capabilities and provides privacy mechanisms.
2. The Greenstone (*www.greenstone.org*) system is a public, extensible, open source project intended to grow in functionality as people contribute to it.
3. Haystack (www.simile.mit.edu/hayloft) is a repository which can be searched by personalised mechanism so that the resulting set reflects the user's interests.
4. CiteSeer (www.citeseer.*ist.psu.edu*). This website does not require the author to do anything to list a paper. This is done automatically by the CiteSeer Web spiders.
5. Directory of Open Access Journals (www.doaj.org).
6. INFOMINE (www.infomine.ucr.edu).

There are also a number of *in-progress* repositories in existence such as the Edu-

Sharing project based in Fern University in Hagen, Germany [KLEB10], and the Universal Citation Database as proposed by Robert Cameron [CAME07]. The latter would link every piece of academic work ever written, no matter how published, to every work that it cites and every work that cites it. Sources for indexing literature face several main costs:

1. Entering the data about the documents onto a database, costs associated with the purchase of the documents, preparing and entering key text from the document.
2. The addition time and cost of creating and adding the meta-data associated with the indexing of the document.
3. Marketing and distribution.

It is also worth mentioning that there are currently a number of Web-based educational applications which support the exchange of materials such as the European Universal Project (www.ist-universal.org). The Universal Project aims to demonstrate the feasibility of an open exchange system for course units between institutions of higher education.

2.2.3 Agents

De Roure [DERO00] stated that “the definition of Agents is still an unfinished task”.

Agents are software applications created to undertake specific tasks on the Web. The majority of Agents are designed to seek out and keep a log of what they have explored; each Web-page they visit can contain hyper-links leading onto further pages to explore. Considering how a small number of Web pages can soon lead to a far greater number, Agents must be self-restraining in their processing.

A common use for Agents is in Web exploration for use by search-engines. As an Agent explores a domain, it will report back to the search-engine's database, the contents it encounters. In the majority of cases, a second process, possibly on a second server does the actual processing of the found pages; the Agents job is solely to explore and report back.

Up until present, Agents are generally designed as solutions to specific problem domains e.g. medical, or shopping. One of the first such Agents was Jango, created by Netbot, but now defunct, it was a shopping Agent much like those provided by current search-engines. Other services include Autonomy (www.autonomy.com) and AgentSoft (www.Agentsoft.de). It remains to be seen if they can make a profit over a longer period.

Agents are notoriously independent of one another in respect of creation and compatibility. However, the introduction of KQML (www.csee.umbc.edu/csee/research/kqml) and CORBA (*www.corba.org*) is starting to address this issue. Standardisation is the key to the success of Agents as unless they communicate with one another, they are less useful and their take-up will be limited.

Agents may be classified in a number of ways:

1. By target user:

    a. Service Agents: provides a service to the web community.

    b. User Agent: provides a service to the individual. Known as Personal and Information Agents. The big barrier with this is in the lack of research into the Agent's cognitive ability. Maes [MAES94] suggested the Agent learning approach some time ago, but limited progress has been made since then.

2. By architecture:

    a. Multi Agent systems.

    b. Autonomous interface Agents/information Agents.

3. By scope:

    a. Limited scope Agent. Limited to local PC or network.

    b. Global scope. Can traverse networks.

4. By intelligence:

    a. Observer. Where Agent watches and tries to 'learn' the user's needs. This may take a long time to be of use due to the necessary learning curve which the Agent must undertake.

    b. Intelligent systems. Search systems which exhibit an above average degree of learning capability.

    c. Level of supervision. This refers to the level of autonomy the Agent is able to operate with.

Like many paradigms, there is an issue of balance and reality with Agents; if the task is too complex, the Agent may struggle and may take too long to produce viable results, if the task is mundane or too ambiguous, it may be that an Agent is not the best choice of tool.

Agents must have the capacity to:

1. Sense their own environment.
2. Act upon their environment with a view to altering it; such as self-starting.
3. Learn. Be able to 'learn' in some way, possibly through collaboration and feedback.
4. Flexible. They should be dynamic, communicative and adaptive.
5. Have an agenda, be goal oriented.

There are a number of questions about Agents:

1. Will research in time be able to instil sufficient intelligence into the Agent?

2. Does the cost of deployment outweigh the benefit?

3. Will users find it acceptable that an Agent, not necessarily their own, request information from them such as dates of travel for booking holidays?

4. What is the difference between an Agent and a program?

5. Their extra workload traffic on the Web. Menczer [MENC02] stated that “More active robots would improve accuracy at enormous costs in network load”.

Agents need considerable development before they can be employed reliably in academia. However, they show potential, and may come into their own with the expansion of the Semantic Web, especially in the field of educational information retrieval.

One of the most influential Agents in academic history was that created by Lieberman [LIEB95] called Letizia. It was the first time that anyone had considered employing an Agent to observe and try to anticipate the user’s area of interest.

As discussed in Chapter 5, there would be a number of methods of implementing a system such as that described in the Scenario. One such method would be through the use of an Agent.

2.2.4 Document Classification and Text Comparison

If autonomous systems are to assist in Web-based research, document classification techniques are essential. Without accurate classification of documents, it would be impossible to organise them into any order of usefulness. One of the problems encountered by researchers in this field such as Balabonovich [BALA97], is the quantity of documents needed to be classified before any automated classification system can be reliably utilised. An alternative to this was offered by Balabonovich, where the system started weakly with zero documents and learned to provide better and better recommendations as time went on.

This is an important aspect to the Scenario. As part of its ability, the system must be able to form an understanding of what the student is searching for, and fundamentally, it must be able to group students into their research subjects in order to autonomously propose academic resources to them. In order for this to happen, and system addressing the Scenario must have the capability to classify and compare documents.

In other words, given a student and a set of materials that that student has shown interest in, the system must be able to make a judgement about how similar those materials are when compared with others it finds on its searches.

Document classification is used most repository systems. The Citebase (www.citebase.org) Web interface is an experiment into organising research paper references into a usable format. By indexing the references of all articles held in the OAI catalogue, Citebase has created a searchable archive of references, allowing the user to see both backward and forward references.

The increasing output of academic materials combined by the increasing use of the Web for electronic storage has made document classification and comparison an increasingly important, yet time-consuming task. Material classification has traditionally been an important activity set for the librarians of collections of books and articles [COYL07]. However, Lewis and Sebastiani [LEWI01] stated that "there are very few operational document classification systems which are based on machine learning algorithms", and Dumais et al. [DUMA02] suggested that success in its application has had limited success.

Sebastini [SEBA02] wrote that "automatic text classification (categorisation) is the task of building software tools to automatically assign some labels (from a set of pre-defined class labels) to a document based on some selected features of that document". This process is important in the use of queries in Web search-engines; for queries to be effective, the materials need to have been correctly identified initially for the query words to return them.

Identifying Documents through Text Extraction

Text extraction may be used in conjunction with document classification, and indeed with user profiling. It is the process of extracting from a body of text certain contents, more often than not in an attempt to classify or identify the contents of that document. The starting point for any classification system is in the extraction of words from a document. In the majority of cases, the classification groups of categories will be decided dynamically by the algorithm based upon the extracted words. In most cases, the text is normalised, i.e. less important words are removed. Words are ranked according to their occurrence, and are usually normalised against the size of the document to restrict scaling bias and to give each word a metric.

This list of words together with their metrics form the basis of a classification, and are used in conjunction with one or more algorithms to place the document into one or more groups. Often, if a classification process is aimed at a particular topic, documents with lower scores may be omitted to speed the process up and increase accuracy.

One of the simpler and commonly used methods of text extraction is the Term Frequency/Indirect Document Frequency (TF/IDF) method [SALT77] which is described in Section 2.3.

Classification of Documents

The main aim of classifying a document is that through its characteristics, it may be possible to place that document into one or more groups or categories relative to those characteristics. In document classification, a document may be a member of a number of categories, although its membership might be more relative in some than others. The quantity of groups it is a member of is decided by how defined its relativity is i.e. some classification algorithms will only assign categories where a strong relativity is recorded, whilst others may allow a more fuzzy relativity creating a broader coverage.

Classification Algorithms

The development of algorithms designed to allow computers to evolve behaviour in order to learn from given data sets are called Machine Learning Algorithms. Source data is given to the algorithm in order that it may learn from it, and apply that knowledge to the target data. The result in document classification may range from grouping documents, keywords or key phrases into sets, to the ranking of documents against a Comparator.

One of the difficulties assessing the effectiveness of document classification algorithms is by their very nature, open to humans' subjectivity and bias. Where experiments have taken place, their effectiveness is difficult to quantify.

Dumais [DUMA02] wrote "where machine learning techniques are applied to document classification in the real-world, the effectiveness of these operational systems is far below that demonstrated in the experimental settings". In effect where algorithms may provide reasonable results under test conditions, live setting results are not so good.

Salton and Buckley [SALT77] confirmed the benefits of the judicious use of single term identifiers in their research. Nanas et al. [NANA03] stated that "the best weighting methods appear to be those which balance exploiting user input and data from the collection". These views seem to confirm the understanding that in a generalised scenario, the concept of term weighing is a balance between accuracy and speed; the higher the recall and precision, the more processing power and time is needed.

Text Summarisation

Although research on automatic text summarisation dates back around 50 years ago, there is still a lot of research to be done in this field [DAS07]. Radev et al. [RADE02] defined a summary as "text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that".

Automatic summarisation should:

1. Be produced from a single document or multiple documents.
2. Preserve important information.
3. Be concise.

Automatic text summarisation can be done on sentence or document level, and will sometimes use natural language understanding, abstraction and generation which are all challenging processes [YEW00]. TextTiling [HEAR97] is a method for partitioning full-length text into coherent multi-paragraph units which represents subtopics and consists of 3 main parts:

1. Tokenisation. Dividing the text into individual words.
2. Lexical Score Determination. Similarity tests between blocks of text using shared terms.
3. Boundary Identification. The assignation of weightings to sequences of words based on their surrounding words.

A number of key concepts of automatic text summarisation are:

1. Extraction. This is the procedure of identifying important sections of the text and producing them verbatim.
2. Abstraction. This aims to produce important material in a new way.
3. Fusion. This combines extracted parts coherently.

4. Compression/Rationalisation. This aims to remove the unimportant sections of the text.

Analysis of automated text summarisation is open to personal choice, and there are numerous protocols in which the results may be presented. One such standard document classification scheme is the Library of Congress Classification (http://www.loc.gov/catdir/cpso/lcc.html).

Resource Description Framework (RDF)

RDF is a language for representing information about resources on the Web. It was intended for representing metadata of Web resources such as author, title and modification date of a Web page. RDF is intended to be used in conjunction with application which can extract the additional information for processing.

RDF uses Extensible Mark-up Language (XML) to describe objects such as academic articles. The benefit of having an XML file associated with an academic resource is that it can be classified and described in a manner which is machine readable. Ultimately, this would allow materials of similar standard and content to be linked to one another and more easily located by software tools and automated processes. However, at present there are a plethora of technologies and standards relating to this field of study, most competing against each other; such as the Semantic Web [BERN01]. A thorough identification of most of these can be found in Dieter et al. [DIET13].

Structured Abstracts

Structured abstracts (SAs) were introduced into medical journals in the mid-1970s [ADHO79] and have become the predominant mode of abstract found in the major clinical journals [BAYL03]. They are considered by some [HARL04, BUDG07] as a better representation of the material than a basic abstract.

An abstract can be considered as a link between the publication it represents and the reader which allows a timely appraisal of its content suitability. It allows the author to provide a brief summary of the publication's research area, and may include mention of any experiments undertaken, and their results. Harris noted that the abstract should be able to "alert potential readers to the existence of an article of interest". In the majority of cases, researchers rely on the abstract to make an informed decision on whether the publication is worth reading through or should be discarded.

Jedlitschka and Ciolkowski [JEDL07] noted the abstract's importance as it is often "the only part of a publication that is freely accessible", and can therefore be seen as an 'advert' for the publication content. Some editors [EDIT04] agreed with this by stating "the abstract is, arguably, the most important part of an article…because it is the only part of an article that many people read". It may even be seen by some as a 'selling point' on pay-to-view web sites.

Structured abstracts take the concept of an abstract one step further by breaking its contents into defined elements used to describe the study, rather than using paragraph format. It can be seen how these defined elements of the SA could be merged with the concepts found in RDF tokenisation and used in autonomous retrieval.

Budgen et al. [BUDG07] found that "abstracts for software engineering papers are frequently of such poor quality they cannot be used to determine the relevance of papers". It could be argued that the introduction of SAs would make authors focus harder on the abstract, which ultimately could leave a more defined and accurate one.

Shaw [SHAW03] stated that at minimum, these elements should include a background, its objectives, any methods used, any results, and a conclusion. These elements are not fixed, and may depend upon the field of research covered, and they may also be decided upon by the publishing organisation. Each of these elements will contain a section of descriptive text identifying/summarising the publication's content. The need for a self-contained abstract is "beyond any question" according to Jedlitschka and Ciolkowski [JEDL07], and the use of SAs is recommended, and their utilisation is growing [KOST02].

The benefits of SAs can be categorised into a number of main areas:

1. Publication creation. Where publications have been written with the inclusion of a SA, it is argued by some that the publication itself might benefit from that inclusion. Bayley and Eldredge [BAYL03] argued that "preparing SAs also can help you from the very outset of considering research.
2. Reviewing. Some academics argue that SAs make the reviewing process of articles easier. In their research of articles submitted for inclusion into a publication [BAYL03a].
3. Reading. Publications using SAs are considered clearer to understand according to some researchers. Budgen et al. [BUDG07] argued that the structure of the abstract make it more readable and considerably clearer to understand.
4. Access. One benefit of having more information available within an abstract is that more of the publications' content can legally be made available in the public domain.
5. Indexing and Classification. One of the most important benefits in using SAs lies in their ability to better describe and identify the publication. Harley [HARL04] argued that their use made searching through publications easier because the enforced structure allows more accurate electronic assessment.

The drawbacks of using SAs can be categorised into a number of main areas:

1. Reviewing. One observation noted is that SAs are generally larger in size; Budgen et al. [BUDG07] noted a 97% increase on average in their research. This could be seen by some as a drawback; for example, editorial reviewers will have more to read therefore increasing their workload.
2. Reading. Critics of SAs [HARL04] argue that although they are generally welcomed by readers, they may have confusing layouts, and are still prone to the same omissions as are traditional ones.
3. Indexing and Classification. Studies in live scenarios [STEV09] have shown that the use of SAs does not necessarily improve the sensitivity, precision or yield of retrieval systems. In Stevenson and Harrison's research of the MEDLINE archive, this was also found, however, they noted that the use of SAs "have been shown to improve the information provided in abstracts and therefore make it easier for readers to evaluate whether an article is methodologically sound and applicable to their clinical situation".

Although much evidence suggesting that SAs are beneficial to readers, publicists and text-mining software, more research is needed to examine their value, and the format of the structure itself. From an academic standpoint, the better an article is described, the more chance that human review, and automated search mechanisms, will have in being able to identify and make a more informed judgement on suitability.

Looking to the future with the explosion in quantity of academic publications, researchers must increasingly look to automated processes to locate relevant materials. "Advances in text-mining, research-profiling and computer-based document retrieval can only profit from the use of these more informed abstracts" [HARL04], and any processes likely to improve retrieval quality is worthwhile

[STEV09].

However, the use of SAs should not be purely technology led. Many researchers do not ultimately rely on networks to seek out resources; “both electronic and hand-searching should therefore be used in order to maximise retrieval of all the available evidence when undertaking systematic reviews” [STEV09].

As this research involves examining the ability of automated systems in identifying publications, it is worth noting Stevenson and Harrison [STEV09] who stated that SAs could be improved for citation retrieval from electronic data-bases by the addition of more details giving the opportunity for ‘text-word’ searching retrieval. If SAs encourage authors to include details of their study, they may otherwise not have included, indexers may be assisted in selecting more appropriate indexing terms. It is arguable that wherever data is structured more formally; it is more accessible by automated processes.

Term and Document Weighting

Document Weighting was first introduced by Luhn [LUHN57] who suggested the use of terms within source text, combined with those found in the user's queries in order to facilitate automatic textual retrieval. Typically, keywords would be extracted from source text and these would be used in conjunction with the user's queries. In effect, a document would be represented by a series of terms or keywords, this is known as indexing or linearisation.

The process of linearisation can consist of up to 5 steps [GARC05]:

1. Mark-up and format removal. This increases relevancy of terms.
2. Tokenisation. All remaining text is made into lower-case, all punctuation removed.
3. Filtration. This is the process of deciding which terms to use to represent the document. At this point, stop-words are excluded.

4. Stemming. This refers to the process of reducing terms to their root variant; i.e. the words 'computing' and 'computer' would be represented as 'comput'.
5. Weighting. The final stage where a weighting model is applied to the terms.

The more terms that the user has included in their search criteria which are present in this list of extracted terms, the better the match. The aim for a good document weighting model is to provide accuracy, but with speed.

One of the most popular, earlier methods was the Vector Space Model [SALT75], which is used to transform a given text or document into a multi-dimensional vector. The vector values can be used to compare documents for closeness; if the vector values are close, the documents can be considered as semantically related.

Despite the variety of models, they tend to be 'specific to task'; i.e. designed with a particular target task in mind, and do not transfer well. Singhal [MITR97] stated that “it has been observed that the term weighting strategies developed for the small collections do a poor job of ranking articles from the new, large, more varied, full-text collections”.

The parsing of a document may occur at several levels:

1. Corpus level. The collection of documents within the area of research.
2. Document level. The individual document being parsed.
3. Term level. The individual terms found in a document.

In the majority of cases, a weighting process will involve at least two of the levels listed above. Term weighting is the question of how important a word or phrase is in a given text in relation to a given area of study. The main factors which affect the importance of a term are:

1. Term Frequency. Articles which frequently use a term appropriate to the area of study potentially offer a closer match to those which do not.
2. Inverse Document Frequency. Where a term is commonly used in general terms such as 'the', it should contribute less to the weighting of a document.
3. Document Length. The longer a document, the higher the frequency of words, and the larger the dictionary of words used. This effect should be taken into account when quantifying terms.

Document weighting can be used in two differing scenarios; information retrieval, and information filtering. Information retrieval can be seen as a one-time need, for example a search-engine query, whereas Information filtering can be seen as a repeated pattern used on a corpus of documents such as used by a filtering Agent. In being different, they have different approaches. Information retrieval relies on query parameters and the representation of those words within a corpus; i.e. the words used are expected to describe a document. The terms are not generally measured in relation to themselves being specific to a corpus. It is assumed that the user has selected the correct keywords. Information filtering is seen as a more persistent operation [NANA03]; sometimes augmented with a profile and document examination and user feedback.

Salton and Buckley [SALT77] defined the two major concepts in term weighting as:

1. Recall. The proportion of relevant items retrieved measured by the ratio of the number of relevant retrieved items to the total number of relevant items in the collection.
2. Precision. The proportion of retrieved items which are relevant. This is measured by the ratio of the number of relevant retrieved items to the total

number of retrieved items.

To achieve high levels in both precision and recall need differing approaches; the high recall of terms is achieved by using broad, high frequency terms. High precision is achieved by using highly refined terms which can more readily be associated with a document.

2.2.5 Context Awareness

The field of research which is known as Context Awareness is not something which can be covered fully in the limited space available here, however, the following is a brief overview.

Context-aware systems acquire and utilise information on a context in order to provide services that are appropriate to particular people, place, time, event, etc. [ARAN14]. The awareness can be either of a physical context or a notional context. Physical context is when the user's location is known, and notional being that something is known about the user's requirements. Learning can be augmented or led by the given context. For example, when a user moves from one exhibit to another in a museum, their device can display information pertaining to each individual exhibit. In this case, the context is based on the physical location of the device, probably using the museum's Wi-Fi signals as an indicator. The same principle can be employed by devices presenting the user with facilities within their environs based on GPS data, such as listing nearby restaurants. Wu et al. [WU2012] stated that through the use of context aware, ubiquitous learning, “the students’ learning achievements were significantly improved in terms of several cognitive processes in Bloom’s taxonomy of educational objectives”. A context-aware system is one that is able to make use of sensed context information and to adapt through it [CHEV04].

This research is more interested in the notional context of awareness. An example of this is when vendors of books such as Amazon (amazon.co.uk), and Alibris (alibris.co.uk) provide the user with alternatives to the selections when they are browsing for books. Google Ads (adwords.google.com) is also effective at doing this. The mechanism for this type of facility is quite simple to execute within a Web domain area, the difficulty encountered is when the user moves over several domains, such as moving from Amazon to Alibris, and taking the context awareness with them.

There has been a lot more research undertaken on the former context than the latter.

2.2.6 Summary

This section has highlighted many of the tools and facilities made available to both teachers and learner by the Web. It is widely acknowledged that the Web offers huge potential for research; however there are still those who argue that there are no convincing links between the use of technology and learning [OLIV13]. Although the research was not exhaustive, it aimed to identify those technologies which presently and in the future offer the greatest potential to assist in the learning process.

The section identified that any technology used for learning tools should:

1. Have overall educational goals.
2. Allow material management.
3. Keep information overload to a minimum.
4. Offer user friendly interface and individuality.
5. Allow flexibility for a variety of subjects.
6. Encourage the big role of social involvement.

The section has considered where encountered technologies could be used within a mechanism to facilitate the Scenario. It seems after this review, that the pedagogical concepts within researching and learning systems are more important than the technologies utilised to attain them.

2.3 Measuring Similarity between Corpora

2.3.1 Introduction

A corpus is a collection of texts, and corpus-based comparison techniques have become increasingly important in recent years, mainly due to two factors:

1. The increasing amount of academic materials being created, making research on the Web more difficult because of their quantity.
2. The increasing amount of academic materials held in electronic format.

One important employment for measuring the similarity between corpora is in its use in creating metrics which can be compared against one another. Consider a tutor recommending an academic article as a starting point for a research topic. Given that starting article, would it be possible for similar and relevant articles to be identified from a corpus of texts through an automated process? If this same process was able to rank the compared articles in a list, it would be easy to comprehend how this could assist in identifying materials closely matching the starting article.

For example, through an automated process and using some given algorithm, we might find that when an Article A is compared with an Article B, it returns a value showing that the pair are 64% relative. Using the same algorithm, we could then compare Article A with many other articles and begin to build up a table of relativity metrics, which when ordered, would identify those Articles which most closely match Article A. It can be seen how by incorporating an automated textual comparison process such as this into the Scenario, it could begin identify similar articles based on those identified by users.

This chapter firstly considers the relevance of textual comparison, and the notion of automating such a process. It then goes on to consider important aspects of implementing such algorithms, such as the use of word-frequency lists, and a number of the algorithms which can be employed.

Finally, the chapter illustrates the process by giving a worked-example of a text comparison technique.

2.3.2 Background

The question of similarity between texts can be considered from a number of perspectives. Consider “how similar is the Times newspaper to the Guardian newspaper”, or on a more defined basis, “how similar is the text in article A to article B”. In both cases, a form of content analysis and comparison must be employed. For a human, this is not a particularly difficult task, although it may be a lengthy one depending upon the size and complexity of the two materials being compared. However, for a computer, this is a considerable feat, and needs a lot of processing.

The potential of computers comparing ranking judgements on the similarity between documents autonomously is huge; consider how effective this process could be employed on Web based research. In the Scenario described in Chapter 1, it can be seen how, through the use of a comparison and ranking mechanism, automated suggestions could be made of similar topic materials.

One of the most influential early studies in this field was undertaken by Hofland and Johansson [HOFL72] which made a comparison of one million words of American English and one million words of British English. Since then, other studies have progressed this field, and a number of models and methods have been proposed ([LEWI04], [MCCA97], [PEDE96], [CHUR95]). These models vary from simple word frequency comparisons to more complex ones such as The

Hierarchical Poisson Convolution model proposed by Bischof and Airoldi [BISC12]. This provides estimates of the differential use of words across topics as well as their frequencies and is effective in making comparisons across a range of document contents.

Comparing corpora in an automated manner is not an exact science; as with human results, opinions differ. Some researchers, such as Chang et al. [CHAN09], and Mimno et al. [MIMN11] argue that sometimes the results are incoherent or redundant, and might need human modification for completion.

One default pre-processing operation is in the removal of ‘stop-words’. ‘Stop-words’ are a set of words which can be considered as extremely common, frequently used words that are not essential to the content. There is no definitive ‘stop-word’ list, as they are generally created for a particular topic or domain. If a word is common in a topic, it is also important to understand if that word is common in many topics, or relatively exclusive to the topic at hand. Bischof and Airoldi [BISC12] stated that “both measurements are informative: non-exclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic”.

Rayson [RAYS00] highlighted a number of issues which need to be considered when comparing corpora:

1. Representativeness. For a corpus to be representative, it should contain samples of all major text types, and if possible in some way, proportional to their usage in ‘every-day language’.
2. Homogeneity within the corpora. This is important as it may be found that comparison results reflect sections within one of the corpora which are unlike other sections in either of the corpora under consideration.
3. Comparability of the corpora. It is important that corpora should have been sampled in the same manner; it should be built using the same sampling

method and with, if possible, randomised methods of the sample selection.

4. Reliability of statistical tests in relation to the size of the corpora under consideration.

Cavaglia [CAVA02] defines a homogeneous corpus as one that belongs to the same sub-language or domain. A heterogeneous corpus can be seen as the antithesis of those listed above; a more random selection of content, style, layout etc.

This wider aspect, not only in comparing the similarity between two documents but in how similar the set of corpora are to one another is an important factor. Are the content of the compared documents tightly related? For example, are the two academic articles being compared of a highly defined subject, or might they contain quite different content as in comparing the two newspapers?

Homogeneity or entropy is the level of similarity sets of a corpus have to one another, and cross-entropy between two corpora quantifies their similarity. Kilgarriff and Salkie [KILG97] stated that “a judgement of similarity runs the risk of being meaningless if a homogenous corpus is compared with a heterogeneous one”.

There are a number of different approaches to this task, each employing differing algorithms, but with similar aims. To clarify, an algorithm is *a procedure to accomplish a specific task* [SKIE07]. The purpose of an algorithm can be described through the set of instances it must work upon and its output after being executed.

Another definition of an algorithm is *a sequence of computational steps that transform the input into the output* [CORM01]. In the majority of cases, an algorithm is said to be correct if, for every input instance, it halts with the correct output.

2.3.3 Word Frequency Lists

One of the most commonly used methods in corpora comparison is the use of word frequency lists. Word frequency lists have been used by linguists since the 1920's [HORN26], but it was not until the advent of electronic media that their importance was widely accepted. Prior to this, many linguistic studies concerned themselves with the difference between spoken and written English. More recently, their use has become more important especially when considered in context with Web search-engines. Tribble and Jones [TRIB97] suggested that the most effective starting point for understanding a text is by using word frequency lists. A word frequency list is created when a document is analysed and a word count is made of every individual word occurrence in the document. They are normally listed in descending order.

A word frequency list may appear as follows.

Word	**Frequency**	**Percentage**
development	440	1.52
team	437	1.51
test	362	1.25
software	347	1.20
study	307	1.06
system	306	1.06
agile	275	0.97
time	274	0.97
…		

Figure 2.1 A Word Frequency List

In the above example, the word ‘development’ has 440 occurrences in the chosen text, and occupies 1.52% of that text. Through using a set of these words, such as the highest ranking ones, they can be used in calculating similarities between documents.

However, there are those who note that what is more important about words are their meanings [KILG97]. Linguistics would argue that when phrases are reduced to single words, the heart of their meaning is lost. One could argue that instead, word senses should be counted; however, this is a highly complex and still unreliable procedure for a computer as they are unable as yet to understand the complexities of language. Consider how homonyms such as *date*, *rose*, and *bank* would affect a piece of text.

The benefit of using word frequency lists is ease, speed, the amount of data that can be collected, and its *relative* accuracy in use comparing texts. It is arguable that where the word *bank* is used in documents for two different reasons, i.e. *river bank*, and *money bank*, the words *river* and *money* will affect the comparison process.

Church and Gale [CHUR95] showed how Inverse Document Frequency, a measure of the proportion of documents a word appears in, can be used alongside word frequency lists to estimate the importance of a word.

The issue of speed is most important in Web-based, real-time applications where users may be waiting for the response to be returned to them, making word-lists more viable than other alternatives. As Baron et al. [BARO09] stated that although word frequency lists are very useful as a starting point for the analysis of corpora, there are well known problems with using them:

1. The frequencies must be normalised before the lists can be compared directly.
2. High frequency words at the top of any word frequency list are generally of no further interest to those trying to examine the content of corpora.

3. Comparing the ranking of words may be misleading.

3. Multi-word expressions and inflectional variants of the same lemma are not counted together.

It is worth noting that not only words can be used in frequency lists, duplets, triples, and even phrases can also be used and comparisons of corpora can be between word frequency lists, part-of-speech collections, or semantic tag frequency lists.

The majority of text comparison techniques are similar in their use of word frequency lists; how they differ is the method the lists are employed. Rayson and Garside [RAYS00] proposed a method for establishing the most and least similar relative frequency words in the two corpora. This is useful in identifying those most frequent words found in the two corpora, but does not highlight a level of similarity between them; this research seeks a metric to represent their similarity.

An example of the creation of a Word Frequency List can be found in Section 5.7.

2.3.4 Utilising Word Frequency Lists

There have been a number of approaches developed to use word frequency lists. The following section is not exhaustive, but identifies the most commonly used ones.

Log Likelihood

Dunning [DUNN93] suggested the Log likelihood ratio as a starting point for corpora comparisons. Log Likelihood allows a comparison of the frequencies of word form occurrences in two texts. It gives a statistical measure of the significance of the differences. A log likelihood contingency table looks like the following:

	Corpus A	Corpus B	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

Figure 2.2 The Log Likelihood Method

With regard to that table above, the value of ‘a’ is the number of times the word occurs in the source text. The value of ‘b’ is the number of times the word occurs in the Comparator text. The value of ‘c’ is the total number of words in the source text. The value of ‘d’ is the total number of words in the Comparator text.

This calculation produces a list with the largest values at the top of the list representing the word which has the most significant relative frequency difference between the two corpora. By using this calculation on the top *n* most occurring words in the two corpora, it would be possible to calculate a metric to identify their relative similarity.

Yule's K statistic

Hofland and Johansson [HOFL72] used Yule's K statistic to measure the similarity between corpora and was seen as an important step forward in measuring resemblance. Yule's K [YULE44] returns a metric representing the richness of vocabulary, based on the assumption that the occurrence of a given word is based on chance and can be regarded as a Poisson distribution. Hofland and Johansson employed the Yule’s K characteristic for measurement of large pieces of text. Through its ability to measure the likelihood of two nouns with a text, being the same, it could offer a measure of the complexity of that text, as well as its repetitiveness.

The equation for Yule's K Statistic Coefficient appears like:

$$\frac{Freq_{LOB} - Freq_{Brown}}{Freq_{LOB} + Freq_{Brown}}$$

Hofland and Johansson carried out one of the largest early studies comparing word frequency profiles; a comparison of one million words of American English (the Brown corpus) with one million words of British English (the LOB corpus) using the coefficient found in above.

Chi Squared statistics (X^2)

Kilgarriff [KILG97] stated that the Chi Squared statistics (X^2) using word frequencies were shown to be the best of those tested in their research. The research noted that "the measure of the difference in a word's frequency between two corpora, and, while the measure tends to increase with word frequency, in contrast to the raw frequencies it does not increase by orders of magnitude". The Chi Squared test was used by Oakes & Farrow [OAKS07] to examine differences in vocabulary in corpora. The equation for Chi Squared coefficient appears like [SOTO13]:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Spearman Rank Correlation Co-efficient

The Spearman Rank Correlation Co-efficient [SPEA04] is used to identify and test the strength of a relationship between two sets of data, for example two word frequency lists. It is easy to compute, and is independent of corpus size. A basic calculation would be as follows:

1. Using the two Word Frequency Lists (the source and the Comparator), give each word a rank based on its occurrence count.
2. Calculate the difference between the two ranks of each individual word, and then square the resulting figure.
3. Calculate the r rank value using the 4.10 coefficient Where d = rank 1 - rank 2.

The equation for Spearman's rank correlation coefficient [SOTO13] appears like:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

The Spearman correlation coefficient is often described as being *nonparametric*, i.e. not adhering to a fixed structure or not relying on data belonging to any particular distribution. It is usually used when it is not convenient or possible to define actual values for variables, but relies on a rank order of variables.

Kendall's Tau Rank Correlation Coefficient

Kendall's Tau Rank Correlation Coefficient [KEND37] is a statistic used to measure the association between two measured quantities. It is similar to Spearman's coefficient as it is a measure of rank correlation; the similarity of the orderings of the data when ranked by each of the quantities.

It has a slight advantage over Spearmans' as the occurrence of 'tied' ranks has a more natural effect (Spearman's tends to inflate 'tied' ranks). Scatterplots are mostly used to illustrate results from correlation coefficients [SPSS04].

The equation for Kendall's Tau Rank Coefficient [KEND90] appears like:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

Degrees of Separation/Degrees of Freedom

Degrees of separation can be employed to provide a metric representing the difference in position that a word holds in two frequency lists. Hofland and Johansson [HOFL72] used this process as part of their research comparing British English with American English. There are a number of ways that this method can be implemented and it is sometimes combined with the Chi Squared method [BARO09]. The following example illustrates the two most commonly used methods. For this example, the two Frequency lists following shall be used and the word under consideration is 'software'.

Frequency List 1

Position	Word	% of text
1	**software**	**0.91**
2	engineering	0.74
3	paradigm	0.72
4	method	0.70
5	process	0.66

Frequency List 2

Position	Word	Word
1	review	0.97
2	develop	0.90
3	lifecycle	0.77
4	**software**	**0.61**
5	process	0.57

Figure 2.3 Frequency Lists 1 and 2

Method 1: Direct comparison of the word's position in each list.

By comparing the two tables above, the direct comparison method shall be illustrated. It can be seen that the word 'software' is located in position 1 of the

Frequency List 1, and its position in Frequency List 2 is 4. From this, it can be calculated that its degree of separation between the two frequency lists is 3 (4 - 1). This comparison can be done on the top n words found in the text and added together to give a metric which can represent how similarly matched the two texts are. The lower the total metric, the better the match.

Method 2: Comparison of the word's percentage of the text

A second method of doing this is by comparing the percentage of the document that each word occupies. In this example, the word 'software' occupies 0.91% of Frequency List 1. In Frequency List 2 text, it occupies 0.61% of the text. This means that the degree of separation is 0.3. Again, this comparison can be done on the top n words resulting in a metric representing the two texts similarity. The lower the metric, the better the match.

Method 2 is considered to be slightly more accurate as it considers percentages of text rather than just rankings.

Term Frequency/Indirect Document Frequency

Term Frequency/Indirect Document Frequency (TF/IDF) calculates values which reflect how important a word is in a document given a corpus. Each word in the document has a value calculated through the inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF/IDF values imply a strong relationship with the document they appear in. This can be used by Search-engines; if that word were to appear in a query, the document could be considered as related and suggested to the user.

Given a document collection D, a word w, and an individual document d є D, the TF/IDF coefficient can be found following:

$$wd = fw, d * \log (|D|/fw, D) \quad (2)$$

Referring to the above, fw and d equal the number of times w appears in d, |D| is the size of the corpus, and fw, D equals the number of documents in which w appears in D.

TF/IDF may be expanded to improve its efficiency in a number of ways such as by adding the Chi Squared calculation [LIU012] to reduce the quantity of words (or features) and reportedly improve the accuracy significantly.

Top N Comparison

The top *N* comparison would take the top *N* most frequently occurring words in the source text and see how often they appeared in the Comparator text.

Each Comparator text is compared in turn with the source text. For each of these comparisons, the method iterates through the top *N* words in the Source's word frequency list, and for each word it does the following:

1. It checks the Comparator document's word frequency list to see if the word is present.
2. If the word is present it adds the percentage that the word occupies in the source text to the percentage the word occupies in the Comparator text to give a figure.
3. The figure is added to the 'running total'.

When each word in the top *N* words of the Source's word frequency table have been checked, the 'running total' will represent the similarity value between the two texts.

By repeating this process for each Comparator text, a set of 'running totals' will be collected; the higher the value of a 'running total', the better the match. These metrics will be able to be used when compared against the human opinions.

2.3.5 Summary

This section has established that there are two main concepts in automating the process of identifying similarities in texts, these are document classification and textual comparison. This section concentrated on textual comparison as it is more relevant to this research. Automated textual comparison methods are reliant upon a comparison process such as comparing word frequency lists. The effectiveness of this increases, once the less important words (stop-words) are omitted, or a closely defined domain of topic is used. Comparing two word lists representing two documents leaves the researcher with a set of figures. In the majority of cases, the values of these figures are useless in isolation, but offer potential where the process is run iteratively against multiple corpus documents when compared to a single source document. Using the figures, those documents which most closely match the source can be identified.

It is worth noting that some of the mentioned concepts allow comparisons of entire corpora, while others enable the identification of individual features, and that most of the concepts outlined in this section were not used in the same context as this research, i.e. in comparing document similarities, they were used mainly in the identification of the different uses in vocabulary amongst native English speakers such as British, American and Australian. The study of Natural Language Processing and Linguistics is a massive area, and is only touched upon here.

2.4 Latest Trends

In more recent years, there has been an exponential uptake in the use of social, web-based, academic and research systems. This can partly be accredited to the uptake in social media in general, but also in the growth of both the number of academics and the quantity of academic materials becoming available.

A decade ago, Citeseer, CiteULike and Google Scholar were the main, free directories for academic publications. ACM Digital Library, IEEE and Elsevier were popular in their role as subscription portals.

More recently, there has been the emergence of a number of key resources and tools, almost all being web-based, and many combining the socio-collaborative ethos which current students are seen to be embracing. Many of these websites can be envisaged similarly to the likes of Facebook, Twitter and LinkedIn. Indeed, some have openly been promoted themselves as ‘mashups’ of the three.

Although there is currently litigation occurring about ownership rights amongst a number of these platforms, it has not stopped the spread of them.

Following is a list of the most popular and more influential platforms.

ResearchGate (www.researchgate.net)

Founded in 2007, and through steady growth, ResearchGate has become one of academia most important resource sharing sites. Currently stating that it has 70 million publications and 7 million users, it has grown somewhat since its origins. ResearchGate allows users to follow a research interest and to receive and make recommendations on those interests in a number of ways. There is a facility to make peer reviews, collaborate on projects, and to ask and answer questions, and chat rooms enable users to share and edit documents. The whole system is nucleated around the notion of users gaining scientific reputation through involvement.

Mendeley (www.mendeley.com)

Mendeley is both a web-based and desktop-based system for sharing research papers and is organised around references. Recently, it came into considerable criticism from the academic community for selling out to Elsevier, a publishing company. Mendeley allows users to be members of groups, and allows them to tag and add searchable comments and metadata to PDFs.

Scorpus (www.scopus.com)

Also owned by Elsevier, Scorpus is promoted as the largest abstract and citation database of peer-reviewed literature. It is similar to Mendeley but is strengthened by its peer-reviewed information which users may run complex searches against. It lacks the 'social' aspect of many of the other platforms here.

Web of Science (www.wokinfo.com)

Web of Science is a scientific citation indexing service. It relies on its citation index system to serve as links between similar research items. It offers little in the way of collaboration, but has substantial searching capabilities.

Academia (www.academia.edu)

Academia is a social networking site for over 20 million academics. It is a supporter of open access movements and is considered by many to be the academic's version of Facebook. It allows users to follow specific interests and also other users. It shows which users have been viewing both the users' own profile and their academic work, and allows social connections.

Zotero (www.zotero.org)

Zotero is a free web-based site which helps academics collect, organise, cite, and share research sources. It allows the creation of groups in order to connect and collaborate with others. There are a number of mechanisms such as email and blogs for users to keep up to date with connections.

ArXiv (www.arxiv.org)

ArXiv is a website-based repository for pre-published academic works. It is mainly for research in Physics, Maths, and Computer Science. Documents are uploaded by the author and are free to read by all users. It is seen by many as offering cutting-edge access to the latest materials, but by some as non-reviewed, lower standard materials.

Figshare (www.figshare.com)

Figshare is an online digital repository which allows users to upload and manage their research in various formats. It enables users to make their research public and is considered similar in concept to an online, multiuser version of Microsoft OneNote.

2.5 Conclusion

This Chapter has considered a number of concepts, many of which are pertinent to the technology behind implementing the Scenario described in Chapter 1. They are:

1. The Web as a learning resource.
2. Technologies for researching on the Web.
3. Measuring similarity between corpora.
4. Latest Trends

The chapter began by examining how the Web is currently being used as an academic resource; the motivating factors behind its advancement and its current role for both learners and teachers.

It found that the impact of the Web is far-reaching, from some underlying yet influential trends like the decline of library use, to more ambiguous ones such as the use of virtual avatars as student representations. Benefits can range from being vague and arguable, to being extremely tangible, this research found that:

1. There is a larger variety and quantity of academic material; international material, non-mainstream and obscure material.
2. There is faster access to materials; less physical searching, and quicker downloading to local computer.
3. There is better access to actual materials. Although much material is behind pay-per view Websites, much of it is in the public domain.
4. There is better geographical access for students. The student can work from almost any location through the use of a laptop.
5. There is higher currency. Materials can be replaced quicker and easier on the Web; they are more up to date.
6. There is better storage. Material is easier to archive in virtual libraries and for students to keep a copy.

However, the Web is not without problems. This chapter highlighted a number of areas of concern with using the Web as an academic resource. The purpose of this review was to identify those problems with a view to them being address in the Scenario. These problems included:

1. The time it takes to locate academic materials on the Web. As the size of the Web increases, it is becoming increasingly difficult for students to find the materials they are after.
2. Synchronous or repeated research. In many cases, students are undertaking synchronous research on the Web in their attempts to locate similar resources.

3. Identifying the quality of materials found on the Web. When students have located materials on the Web, it may be difficult to identify its academic value, or even its bias.
4. Gaining physical access to materials hosted on the Web. Much academic material on the Web is hidden, or available on a pay-per-view basis.
5. Resources on the Web can be less permanent than those in print.
6. Students may go for a 'quick-fix' to their research due to their exposure to 'immediate' technologies
7. Students may consider that the newest material, or even those most attractively displayed, are the most relevant.

Figure 2.4 Areas of Concern with Using the Web as an Academic Resource

This chapter also found that whilst using the Web as a learning resource, tutors should encourage:

1. Developing reciprocity and sharing among students.
2. Use active learning techniques, allowing the learner to be the self-directed creator of content, not purely as a consumer.
3. Encourage collaboration and cooperation as part of the learning process.
4. Move from a tutor-centred approach to one of student-centred, and from instruction to construction.
5. Peer assessment, because of the positive learning benefits associated with it.
6. The development of new skills for digital age learners.

The chapter then went on to consider an area much over-looked, but closely coupled to many of the concepts identified in the first section; measuring similarity between corpora.

This is an important paradigm considering that we, as Web users are increasingly relying on computers to find us information, having given them minimal knowledge to work from such as search-words. The research identified two main conceptual areas; document classification and textual comparison, both of which are a key area in creating automating document comparison techniques.

Finally, the chapter considered a number of trending platforms being used by academics such as academia.edu and figshare.com.

Whatever personal feelings individuals have about the Web, it is here to stay and its impact on education will be massive. However, caution must be taken where education relies upon an entity so blindly; studies already show that we may already have e-educational overconsumption [IVAN10]. Should universities rely on the Web to such an extent that a simple power cut may halt learning completely?

The Web

"It's becoming a librarian's nightmare—or paradise, depending on how you look at it"

[GILB96]

Chapter 3 The Pilot Studies

3.1 Introduction

Formal experiments are conducted to allow cause and effect conclusions to be identified. Experiments, like software development itself require a great deal of care and planning if they are to provide meaningful, useful results [PFLE95]. McGowan [MCGO11] stated that “a well-designed experiment is the best method for establishing efficacy of any intervention, be it medical, behavioural, or educational in nature”. More simply, an experiment is a process or study that results in a collection of data, from which something can be learned.

Creating an experiment is not always an accurate science; initial results may highlight shortfalls in the approach, and so the experiment may need to be done a number of times before the researcher is convinced that it is viable. Scientific learning is always an iterative process [SELT12] therefore no design method is ever complete, but the aims of an experiment remain fairly static; to produce meaningful and defensible evidence to support or critique a theory.

So, what are experiments? Consider the following [OEHL00]: Experiments allow us to create a direct comparison between the treatments of interest:

1. We can design experiments to minimize any bias in the comparison.
2. We can design experiments so that the error in the comparison is small.
3. We are in control of experiments, and having that control allows us to make stronger inferences about the nature of differences that we see in the experiment. Specifically, we may make inferences about causation.

According to Fisher [FISH25], experiments are characterised by:

1. The manipulation of one or more independent variables.
2. The use of controls such as randomly assigning participants or

experimental units to one or more independent variables.

3. The careful observation or measurement of one or more dependent variables.

Computer experiments are a relatively new type of empirical investigation and the complexity of many systems has made them prohibitive through physical, time, and financial constraints [LEHM04]. A computer experiment consists of a number of runs of the code with varying inputs. A feature of many computer experiments is that the output is deterministic [SACK79].

The two main difficulties in computer experiments are:

1. They are very time consuming to run
2. The multi-dimensional aspect to programs make experiments difficult to define and encapsulate.

3.2 Definitions and Terminology

There are a number of terms used in the experiments in this research which are defined following.

Control variables. These are also called dependent variables and are metrics associated with the experiment which can be set and maintained by the person undertaking the experiment. These variables can affect the outcome of experiments directly.

The terms qualitative and quantitative refer to the type of data generated from an experimental process [GARB09]. Quantitative research produces data in the form of numbers and originated in the natural sciences such as biology, chemistry, physics and geology. Qualitative research tends to produce data that are stated in prose or textual forms and are generally used in the social sciences such as psychology, sociology and anthropology. In order to produce different types of data, qualitative and quantitative research tend to employ differing methods, and they are rarely combined [FLIC04].

Because in Chapter 4, the aim of the pilot studies is to collect metric data associated with comparisons, they will take quantitative approaches.

Response variables. These generally measure the outcome of a system experiment. In most cases of experiments, they are a set of outputted variable quantities which provide us with an answer to the experiment question. Their values are dependent upon other variables within the experiment such as control variables.

The control group. This is the set which is unaffected by the experiment. In most experiments, control variable data is processed against the control group to provide response variables.

The experimental group. This is the set affected by the experiment, thus giving a dependent variable illustrating the size of effect. These collectively form the response variable set.

Participants. As the experiments involved the participants comparing the contents of English written Computer Science related academic articles. The participants of all of the experiments in this Research are all educated to at least graduate level in Computer Science disciplines. Each individual participant will only take part in an experiment once.

Sex, Age and gender were not noted as part of the experiment.

This section would include the quantity of participants taking part.

3.3 Common Statistical Analysis Methods

Siegel [SIEG57] suggested that "the choice among statistical tests which might be used with a given research design should be based on three criteria". Those criteria are as follows

1. The statistical model of the test should fit the conditions of the research.
2. The measurement equipment of the test should be met by the measures used in the research.
3. From among those tests with appropriate statistical models and appropriate measurement requirements, that test should be chosen which has greatest power- efficiency.

We must remind ourselves that some of these criteria may only be attained through an iterative consideration of the experiments. In some cases, the identification of the correct criteria may be the result of an experiment.

Experiments are normally based around sample sets. There are several main sample sets that can be used in experiments:

1. Two independent samples. This is where two different conditions are applied to two different sets of subjects.
2. Two dependant samples. This is where a single set of subjects takes part in the two conditions.
3. Several independent samples. This is where there are more than two groups of participants and more than two sets of conditions.

Whichever samples are used, they can be applied in two ways:

1. Parametrically. Tests are considered parametric when the defining properties of the population used is considered to be normal, and therefore assumptions are made.

2. Non-parametrically. This second type of test does not make assumptions about the population distribution, and so may be classed as non-normal.

Another aspect to experiments is the nature of measurements used. This can be applied to both inputs to and outputs from the experiment. Siegel [SIEG57] stated that there are four levels of measurement, they are as follows:

1. Nominal of classificatory scale such as gender or ethnic background.
2. Ordinal or ranking scale such as hardness of rocks, beauty.
3. Interval scale such as Celsius.
4. Ratio scale such as speed, height or weight.

3.4 Common Mistakes in Experimentation

Jain [JAIN07] stated that there are a number of common mistakes in experimentation:

1. Ignoring variation due to experimental error.
2. Failure to control important parameters.
3. Failure to isolate effects of different factors.
4. Use of simple one-factor-at-a-time designs
5. Ignoring interactions between factors
6. Conducting too many experiments

3.5 Experimental Template

The aim of experimental design is to create a set of protocols which obtain results from the fewest experiments; this minimises time and costs. Essentially, experimental design is a plan or template for the experimental process to meet specified objectives.

The plan also allows the user to be able to quantify effects from different factors using analysis, i.e. through an iterative process. It may also allow the user to determine whether a factor's effect is significant and/or controllable.

There are a number of recommended procedures for performing experiments ([LIGH90], [MCGO11], [SLAV74] and [WU2004]). However, the following steps form a template which each experiment (and pilot study) in this research will adhere to.

1. Overview

Each experiment will use an overview to summarise its method and protocols. This will take the format as can be seen following.

Aim	What is the aim of the experiment? What will be gained through undertaking it?
Hypothesis	What is the expected result of the experiment. This can be seen as an educated guess.
Variables and Sets	This lists the variables and sets affecting and being affected by the experiment.
Inputs	This is the type of data being used to affect the experiment.
Procedure	A description of the experiment process.
Format of Results	The format of the results from the experiment.

Figure 3.1 Experimental Template Overview

2. Aim

This section will provide an understanding of why the experiment is being undertaken.

3. Hypothesis

An important step to an experiment is to define the hypothesis or research question. The hypothesis should be tailored to suit the level of knowledge already attained by other research; it is pointless re-running experiments already done.

The hypothesis should be as clearly focussed as possible to provide a narrow answer. For example, rather than stating "*Can the Web can be best utilised to improve student learning by using a recommender system?"*, it could be stated that "By combining a recommendation system with a document identification system, can a substantial increase in document comparison accuracy can be attained?"

The hypothesis for each individual experiment will be defined in this step.

4. Variables and Sets

Control Variables:	These are the variables which affect the experiment.
Response Variables:	These variables provide a measure of the outcome of a system experiment.
Control Group:	This is the set unaffected by the experiment.
Experimental Group:	This is the set affected by the experiment.

5. Inputs

The treatment is the cause of the effects of an experiment. An experiment allows the implementation of a treatment, and the level of exposure can be varied to affect the outcome. Examples of such variations could be time of exposure and the level of exposure. It is important to understand how the outcome of an experiment relates to the treatment variables.

In order to produce different types of data, qualitative and quantitative research tend to employ differing methods, and they are rarely combined [FLIC04].

The inputs to the following experiments differ slightly and will be described individually in this section.

6. Procedure

The implementation of the undertaken experiments had a number of influencing factors and constraints.

1. Timing. In the case of educational experiments where participants are required, timing is a major factor. The experiments will most likely need to be run through university term time, during term hours, and when students are available. This might impact on both the students and their courses. Other stakeholders and constraints may include the course tutors who won't wish their carefully planned courses to be affected, network and computer resources, lab building resources and student's workloads.
2. Effect. Experiments should not adversely, or conversely effect sets of students where others have been excluded from it. In other words, students taking part in an experiment should not be given an advantage over others in their course work, through the experiment.
3. Communications. Communications with all the stakeholders is important as they need to be kept informed and 'on-side'.

Pfleeger [PFLE95] wrote that “measures and treatments should be applied consistently” during the execution process.

Each experiment would differ slightly, such as parameters, data, and participants, thus the use of this template to illustrate those differences. However, in each case, the nature of the actual experiment would be similar, and each would take the following form:

Step 1: Explain purpose of experiment to participants.

Step 2: Issue the following paperwork:

1. A source document containing text.
2. A set of Comparator documents containing text.
3. A scoring guideline.
4. A results sheet.

Step 3: Explain the task to participants.

Step 4: Tutorial – any questions.

Step 5: The participants perform the task.

Step 6: Collect paperwork – Thank participants.

Step 7: The software tool performs its task.

Step 7: Comparison and evaluation.

7. Results

In this section, the results from the experiments will be presented. The terms qualitative and quantitative results refer to the type of data generated from an experimental process [GARB09]. Quantitative research produces data in the form of numbers and originated in the natural sciences such as biology, chemistry, physics and geology. Qualitative research tends to produce data that are stated in prose or textual forms and are generally used in the social sciences such as psychology, sociology and anthropology.

Results from these experiments will be quantitative, in the form of metrics presented in spread sheets.

The outcome of a given experiments are in the majority of cases, either null hypothesis or alternative hypothesis. *Null hypothesis* describes an experiment where the outcome demonstrates no improvement and/or effect. *Alternative hypothesis* describes an outcome where differences in the response variables expected.

The selection of a high-quality experiment instruments is important for ensuring that data returned is valid, reliable and of usable. The use of standardised assessments methods should be used where possible. It is also important to produce outcomes which can be worked with and understood not just by the experimenter.

7. Discussion

The Discussion section shall examine the data from the Results section.

The data collected during the experiment provide the input to the analysis process. The first step in the analysis is to understand the data by using descriptive statistics to provide a visualisation of the data. These help us to understand and interpret the data informally.

The next step is to consider whether the data set should be altered or reduced in order to clarify the analysis process; it was stated previously that experiments were in many cases, iterative.

Analysis can vary depending on the experiment type; however, there are generally a number of parts to this [PFLE95]:

1. Any data collected should be reviewed for validity and usefulness. This should be organised into such a way that makes it easily understood.
2. Organise data into sets as part of the testing process
3. Data should be analysed according to established statistical principles.
4. Decide whether null hypothesis is supported or refuted.

The result from an experiment can be used in a number of ways:

1. To support a hypothesis or theory such as an approach to software design.

2. To explore a relationship between sets of data.
3. To ascertain the accuracy of a model or expectation.
4. To validate a measure or input.

This section shall contain a discussion on any points noted from the results.

9. Conclusions

Any conclusions drawn from the Discussion section shall be presented in this step. Whether the Hypothesis has been met shall also be discussed.

Gosling and Noordam [GOSL11] wrote that “you should be able to evaluate the relationship between the predicted result as stated in the hypothesis and the actual results, and be able to conclude whether the explanation on which the prediction was based is supported by the data. Or not”.

There are a number of important aspects to include in a conclusion:

1. Whether the hypothesis is correct or not; if not, are there other possible answers to your question?
2. Summarise any difficulties or problems encountered during the experiment.
3. Are there any changes needed in the procedure, or does the experiment need repeating? How would the approach change?
4. List any other things learned.
5. Attempt to answer other related questions that have arisen.
6. Discuss any experimental errors.

3.6 Pilot Study 1 – Comparing the Titles of Academic Articles

The definition of a pilot study, according to The Oxford Dictionary is:

"A test of the methods and procedures to be used on a larger scale."

The aim of these pilot studies are as follows:

1. To test the Experimental Template as defined in Section 3.5 against smaller sets of data.
2. To identify any improvements which can be made prior to the main experiment.
3. To create a benchmark set of data demonstrating how similarly participants understood and related the meaning behind given texts. This data can be used as comparison data in the main experiment.
4. To identify if the size and content of texts could affect the outcome of participant comparisons.

This research undertook four pilot studies, using two sets of participants and two sets of data. Pilot Studies 1 and 2 used academic materials for their data. Pilot Studies 1.1 and 2.1 used text taken from online blogs for their data.

3.6.1 Overview

The overview for Pilot Study 1 is as follows:

Aim	To ascertain how similarly participants understood and related the meaning behind given texts.
Hypothesis	Participants' opinions would be similar.
Variables and Sets	Control variables: None.

	Response variables: Participants opinions. Control group: A source text. Experimental group: 35 Comparator texts. Participants: 10.
Inputs	The participants' opinions when comparing 35 Comparator texts with a source text.
Procedure	The participants are asked to make judgements over similarity between the titles of academic articles.
Format of Results	Metrics representing the participant's opinions.

Figure 3.2 Pilot Study 1 Overview

3.6.2 Aim

The aim of Pilot Study 1 was to create a benchmark for how similar participants understood and related the meaning behind given texts, in this case academic article titles. Participants would be asked to compare a single source title with a set of Comparator titles, and give a metric to represent their opinion on how closely they thought each Comparator title matched the source. The resulting metrics would be used to identify trends and observations. Would the participants' opinions closely match one another? What observations could be made from the resulting data?

3.6.3 Hypothesis

The following table lists the hypothesis for Pilot Study 1.

Hypothesis ID	Hypothesis
H1	The participants' opinions would be similar.

3.6.4 Variables and Sets

The following variables and sets are used in this experiment:

Control variables: There are no control variables within this experiment.

Response variables: The response variables take the form of the metrics given by the participants to represent their opinions.

Control group: A source text.

Experimental group: 35 Comparator texts.

Participants: 10.

3.6.5 Inputs

For this pilot study, participants would be asked to compare a single source text with a number other pieces of Comparator text giving a rating each time. In each case, the texts were titles of academic articles. Each of the titles would be drawn from articles of similar contexts.

The source text was the title of an academic paper:

Empirical studies of agile software development: A systematic review

The Comparator texts were the following titles of academic papers:

02 Agile software development methods

03 New Directions on Agile Methods: A Comparative Analysis

05 Chrysler goes to extreme

06 Web-Based Agile Software Development

09 Extreme Programming Explained 1st Ed.

10 Extreme Programming Explained 2nd Ed.

11 Get Ready for Agile Methods, with Care

15 Selecting a Project's Methodology

17 An Introduction to Agile Methods

19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines

22 Are Two Heads Better than One? On the Effectiveness of Pair Programming

24 Evidence-Based Software Engineering for Practitioners

26 On the Effectiveness of the Test-First Approach to Programming

29 How to Read a Paper

30 Agile manufacturing: A framework for research and development

33 Checklists for Software Engineering Case Study Research

34 Extreme Programming Considered Harmful for Reliable Software Development

36 Guidelines for performing Systematic Literature Reviews in Software Engineering

37 Preliminary Guidelines for Empirical Research in Software Engineering

40 Iterative and Incremental Development: A Brief History

42 Questioning Extreme Programming. Should we optimize our software development process?

43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases

44 Agile Software Development: Adaptive Systems Principles and Best Practices

46 Theoretical Reflections on Agile Development Methodologies

47 Challenges of Migrating to Agile Methodologies

50 A Practical Guide to Feature-Driven Development

52 Lean Software Development: An Agile Toolkit

53 Changing the Paradigm of Software Engineering

54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical

56 Agile Project Management with Scrum

59 A Survey of Controlled Experiments in Software Engineering

61 Extreme Programming Refactored: The Case Against XP

64 Assumptions Underlying Agile Software Development Processes

66 Agile Software Development: It's about Feedback and Change

68 Flexible and Distributed Software Processes: Old Petunias In New Bowl

Each text was numbered for identification purposes; there would be 35 Comparator texts, although it is worth noting that their numbers are not sequential as they were taken from a larger set. The participants were notified that due to the limited amount of information at their disposal, that there were no *correct* answers to the pilot study. It would be a reflection of their opinions.

3.6.6 Procedure

The study took place over two days, each a two hour session. Students could attend any time during these periods.

The format of the pilot study was quite simple. Participants were given:

1. A sheet of paper with brief instructions.
2. The title of the source article.

3. A sheet with a listing of all of the 35 titles of the Comparator articles.

A copy of these documents can be found in Appendix 3.

After reading the titles, participants were asked to use a three point scoring system to record their opinions. The scoring system was as follows:

1. Relevant. Place a tick next to those Comparator titles which most closely matched the source title.
2. Irrelevant. Place a cross next to those Comparator titles which did not match the source title.
3. Neutral. Leave blank any titles which they felt fitted neither of the above.

3.6.7 Results

Figure 3.3 lists the results. Each Comparator paper is represented by a row. Each row contains:

Column 1: The row number

Column 2: The article number (used for ease of identification purposes) and the article's title.

Columns 3-12: Ten columns representing the results for that title from each of the ten participants.

Column 13: A sum total of the participants' results. This total was calculated by giving:

1. +1 for those articles ticked.
2. -1 for those articles crossed.
3. Zero for anything else.

Column 14: An averages figure based on the above sum total divided by the number of participants.

Row	Article number/ Title	Participant											
		1	2	3	4	5	6	7	8	9	10	Total	Average
1	56 Agile Project Management with Scrum	y		y	y	y	y	y	y	y	y	9	0.9
2	17 An Introduction to Agile Methods	y	y	y	y		y	y		y	y	8	0.8
3	44 Agile Software Development: Adaptive Systems Principles and Best Practices	y	y	y	y		y	y		y	y	8	0.8
4	64 Assumptions Underlying Agile Software Development Processes	y	y	n	y	y	y	y	y	y	y	8	0.8
5	66 Agile Software Development: It's about Feedback and Change	y	y	n	y	y	y	y	y	y	y	8	0.8
6	43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases		y	n	y	y	y	y	y	y	y	7	0.7
7	61 Extreme Programming Refactored: The Case Against XP			y	y	y	y	y	y		y	7	0.7
8	10 Extreme Programming Explained 2nd Ed.			y	y		y	y	y		y	6	0.6
9	11 Get Ready for Agile Methods, with Care	y	y	n	y	y	y	y			y	6	0.6
10	02 Agile software development methods	y	y	y	y		y	y	y	y		8	0.8
11	26 On the Effectiveness of the Test-First Approach to Programming			y	y	y	y		n	y	y	5	0.5
12	34 Extreme Programming Considered Harmful for Reliable Software Development		y	n	y	y	y		y		y	5	0.5
13	37 Preliminary Guidelines for Empirical Research in Software Engineering	n	y	y	n		y	y	y	y	y	5	0.5
14	50 A Practical Guide to Feature-Driven Development		n	y	y	y	y			y	y	5	0.5
15	06 Web-Based Agile Software Development	y	n	n	y	y	y	y	y		y	5	0.5
16	19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	y	n	n	y		y	y	y		y	4	0.4
17	52 Lean Software Development: An Agile Toolkit	y	n	n	y	y	y	y			y	4	0.4
18	54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical Models			y	y		y		n	y	y	4	0.4
19	03 New Directions on Agile Methods: A Comparative Analysis	y	y	n	y	y	y	y	n	n	y	4	0.4
20	36 Guidelines for performing Systematic Literature Reviews in Software Engineering	n	y	y	n	y	y		y	y		4	0.4
21	05 Chrysler goes to extreme		y	y	y	n	n			y	y	3	0.3
22	09 Extreme Programming Explained 1st Ed.		y	y	n		y	y	n		y	3	0.3
23	15 Selecting a Project's Methodology	n	y	n	y	y	y				y	3	0.3
24	30 Agile manufacturing: A framework for research and development			y	n	n	y	y	y		y	3	0.3
25	46 Theoretical Reflections on Agile Development Methodologies	y	n	n	y	y	y	y	n		y	3	0.3
26	53 Changing the Paradigm of Software Engineering		y	n	y	y	y		n		y	3	0.3
27	24 Evidence-Based Software Engineering for Practitioners	n	n	y	y		y	y	y	y	n	3	0.3
28	47 Challenges of Migrating to Agile Methodologies	y	y	n	y	n	n	y		n	y	1	0.1
29	42 Questioning Extreme Programming. Should we optimize our software development process?		y	n	y	n	y	y			n	1	0.1
30	40 Iterative and Incremental Development: A Brief History	n	y	n	y		y		n		n	-1	-0.1
31	59 A Survey of Controlled Experiments in Software Engineering	n		n	n	n	y		n	y	y	-2	-0.2
32	68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	n	n	n	y		y				n	-2	-0.2
33	22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	n	n	n	y	n	y	y	n		n	-3	-0.3
34	33 Checklists for Software Engineering Case Study Research	n	n	y	n	n	n	y		n	n	-5	-0.5
35	29 How To Read A Paper	n	y	n	n	n	n	n	y	n		-5	-0.5

Figure 3.3 Pilot Study 1 Results

Figure 3.3 ordered the results by the Average column, or in other words in ascending order of relativity to the source title; the most relative titles are at the top. Participant's choices are represented by ticks, crosses, and blanks. For example, it can be seen that participant 1 identified Title 37 (the 13th row down), as unrelated as it contains a cross.

The participants' Marking Sheets can be found in Appendix 4.

3.6.7 Discussion

There were a number of observations which could be noted from these results.

The first interesting observation was that the participants' top article, *Agile Project*

Management with Scrum is actually a book examining a sub area of Agile methods called Scrum. Although it had a good introductory chapter, it concentrates on the detail of *Scrum* at the expense of other methods. This made it in the Researcher's opinion, knowing more of the book than the title, less similar to the source article than some of the others.

Another interesting observation was that the participants had ranked *Chrysler Goes to Extreme* in the bottom half of the table) alongside *Extreme Programming Explained 1st Edition.* This does not seem entirely logical as the former paper's title is much more vague than the latter's, which sounded like a highly targeted book. Also, they had ranked Extreme Programming Explained 2nd Edition 14 places above the 1st edition!

It was however, reassuring to see that apart from a number of anomalies, the participants had generally identified the most relevant reading materials given the source article fairly well in the top half of the table. One exception was Participant 3. This can be seen in Figure 3.3 where there are a number of crosses which stand out towards the top of the table, where the other participants have used ticks; in particular, Articles 64, 66 and 43. The same participant however, included the 1st and 2nd edition books, where others such as Participant 4 had not.

Referring back to Figure 3.3, it could also be argued that in some cases where an understanding of the subject was present, synonyms were being used to affect the results. An example of this is where six participants identified the value of *On the Effectiveness of the Test-First Approach to Programming* (article 26, row 11) when none of the key terms in this title were used in the source title. Conversely, Participant 7 failed to identify this, possibly due to a lack of understanding of the subject. The same could be said for Comparator 43 (row 6) titled *Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases*, where participant 3 identified it as a poor match.

One aspect to consider was that of collection size. Was using a collection of 35 articles too many to allow the participants to accurately consider them? In a normal researching period, would a student expect to consider the titles of 35 articles? Was there a question of information overload?

Another point to consider was whether the choice of text in the titles accurately reflected the content of the articles? For example, would someone not learning computer programming be expected to relate the terms *extreme* with the term *agile*? If a participant did not know that this term may be used in this context, it could be misleading. Some titles were plainly unclear such as *Chrysler goes to Extreme which* title tells the reader very little about the article's content.

Degrees of Separation Matrix

In an attempt to clarify the highlighted anomalies, a matrix was produced from the data in Figure 3.3. The aim was to identify how varied (or how similar) the participants opinions were.

To create the matrix, figures were calculated by taking each individual participant's ranking for each title and seeing how far it was from the average for that title. To illustrate this process, 2 examples have been created.

Example number 1. Referring to Figure 3.3, it can be seen that:

1. Participant 1 ranked Title 56 (row 1) as Relevant (a tick), therefore a +1 value.
2. The average for Title 56 was 0.9.
3. The variance between the two figures was 0.1, which equates to a small variance.

This figure would be entered into the Degrees of Separation Matrix at the same intersection between Participant 1 and Title 56.

Example number 2. Referring to Figure 3.3, it can be seen that:

1. Participant 3 ranked Title 64 (row 4) as Irrelevant (a cross) therefore a -1

value.

2. The average for Title 64 was 0.7.

3. The variance between the two figures was 1.7, which equates to a large variance.

This figure would be entered into the Degrees of Separation Matrix at the same intersection between Participant 3 and Title 64.

Figure 3.4 shows the matrix containing all the variance figures entered. When we compare all these figures, we can start to understand a number of things about the participants and the titles. Each row contains:

Column 1:	The row number.
Column 2:	The article number (used for ease of identification purposes) and the article's title.
Columns 3-12:	Ten columns containing the variance figures.
Column 13:	The total variance figure for that article.

Row	Article number/ Title	Participant										
		1	2	3	4	5	6	7	8	9	10	Total
1	56 Agile Project Management with Scrum	0.1	1.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1.9
2	17 An Introduction to Agile Methods	0.2	0.2	0.2	0.2	0.8	0.2	0.2	0.8	0.2	0.2	3.2
3	44 Agile Software Development: Adaptive Systems Principles and Best Practices	0.2	0.2	0.2	0.2	0.8	0.2	0.2	0.8	0.2	0.2	3.2
4	64 Assumptions Underlying Agile Software Development Processes	0.2	0.2	1.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	3.6
5	66 Agile Software Development: It's about Feedback and Change	0.2	0.2	1.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	3.6
6	02 Agile software development methods	0.2	0.2	0.2	0.2	0.8	0.2	0.2	0.2	0.2	0.8	3.2
7	43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	0.7	0.3	1.7	0.3	0.3	0.3	0.3	0.3	0.3	0.3	4.8
8	61 Extreme Programming Refactored: The Case Against XP	0.7	0.7	0.3	0.3	0.3	0.3	0.3	0.3	0.7	0.3	4.2
9	10 Extreme Programming Explained 2nd Ed.	0.6	0.6	0.4	0.4	0.6	0.4	0.4	0.4	0.6	0.4	4.8
10	11 Get Ready for Agile Methods, with Care	0.4	0.4	1.6	0.4	0.4	0.4	0.4	0.6	0.6	0.4	5.6
11	26 On the Effectiveness of the Test-First Approach to Programming	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.5	0.5	0.5	6.0
12	34 Extreme Programming Considered Harmful for Reliable Software Development	0.5	0.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	6.0
13	37 Preliminary Guidelines for Empirical Research in Software Engineering	1.5	0.5	0.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	7.0
14	50 A Practical Guide to Feature-Driven Development	0.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	6.0
15	06 Web-Based Agile Software Development	0.5	1.5	1.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	7.0
16	19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	0.6	1.4	1.4	0.6	0.4	0.6	0.6	0.6	0.4	0.6	7.2
17	52 Lean Software Development: An Agile Toolkit	0.6	1.4	1.4	0.6	0.6	0.6	0.6	0.4	0.4	0.6	7.2
18	54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical	0.4	0.4	0.6	0.6	0.4	0.6	0.4	1.4	0.6	0.6	6.0
19	03 New Directions on Agile Methods: A Comparative Analysis	0.6	0.6	1.4	0.6	0.6	0.6	0.6	1.4	1.4	0.6	8.4
20	36 Guidelines for performing Systematic Literature Reviews in Software Engineering	1.4	0.6	0.6	1.4	0.6	0.6	0.4	0.6	0.6	0.4	7.2
21	05 Chrysler goes to extreme	0.3	0.7	0.7	0.7	1.3	1.3	0.3	0.3	0.7	0.7	7.0
22	09 Extreme Programming Explained 1st Ed.	0.3	0.7	0.7	1.3	0.3	0.7	0.7	1.3	0.3	0.7	7.0
23	15 Selecting a Project's Methodology	1.3	0.7	1.3	0.7	0.7	0.7	0.3	0.3	0.3	0.7	7.0
24	30 Agile manufacturing: A framework for research and development	0.3	0.3	0.7	1.3	1.3	0.7	0.7	0.7	0.3	0.7	7.0
25	46 Theoretical Reflections on Agile Development Methodologies	0.7	1.3	1.3	0.7	0.7	0.7	0.7	1.3	0.3	0.7	8.4
26	53 Changing the Paradigm of Software Engineering	0.3	0.7	1.3	0.7	0.7	0.7	0.3	1.3	0.3	0.7	7.0
27	24 Evidence-Based Software Engineering for Practitioners	1.3	1.3	1.3	0.7	0.3	0.7	0.7	0.7	0.7	1.3	9.0
28	47 Challenges of Migrating to Agile Methodologies	0.9	0.9	1.1	0.9	1.1	1.1	0.9	0.1	1.1	0.9	9.0
29	42 Questioning Extreme Programming. Should we optimize our software development process?	0.1	0.9	1.1	0.9	1.1	0.9	0.9	0.1	0.1	1.1	7.2
30	40 Iterative and Incremental Development: A Brief History	0.9	1.1	0.9	1.1	0.1	1.1	0.1	0.9	0.1	0.9	7.2
31	59 A Survey of Controlled Experiments in Software Engineering	0.8	0.2	0.8	0.8	0.8	1.2	0.2	0.8	1.2	1.2	8.0
32	68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	0.8	0.8	0.8	1.2	0.2	1.2	0.2	0.2	0.2	0.8	6.4
33	22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	0.7	0.7	0.7	1.3	0.7	1.3	1.3	0.7	0.3	0.7	8.4
34	33 Checklists for Software Engineering Case Study Research	0.5	0.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5	0.5	7.0
35	29 How To Read A Paper	0.5	1.5	0.5	0.5	0.5	0.5	0.5	1.5	0.5	0.5	7.0
		20.3	25.2	32.9	23.1	19.9	21.3	16.9	22.5	16.1	20.5	218.7

Figure 3.4 Pilot Study 1 Degrees of Separation Matrix

It can be seen in Figure 3.4 that Participant 9 had the lowest combined Total Variance of 16.1 which meant that the opinion of Participant 9 most closely matched the other participant's opinions, on average. Conversely, Participant 3, who was identified earlier, had the highest variance.

It can also be seen that generally, those titles which appeared in the top of the table, and as such were deemed to be of a closer match to the source Title, had the lower variances, whereas the lower half of the table had the higher variances. This implies that where participants thought titles were a more definite match to the source, their opinions tended to match more closely.

Average Variance Figure

To give the Degrees of Separation Matrix more meaning an average variance across the 10 participants and the 35 texts could be calculated as follows:

average variances = total of all 10 variance figures / 10 participants / 35 titles

Therefore:

Total of all 10 variance figures = 218.7

Divided by 10 participants

218.7 / 10 = 21.87

Divided by 35 titles

21.87 / 35 = 0.62

Therefore, the average variance figure is 0.62. As the three point scoring system went from -1 to +1, a total range of 2, the figure 0.62 equated to a 31% variance.

3.6.9 Pilot Study 1 Conclusion

The results show that in many cases, the participants made generally similar choices. However, the results were skewed slightly by the following anomalies:

1. Those participants who understood the topic and who substituted synonyms. It would be interesting to see if synonym substitution had been disallowed, would the level of variance have altered?
2. Those participants who did not understand the topic. If the corpora of participants had been limited to those more expert in the topic, or those who were completely new to the topic, might the accuracy have changed?
3. Those participants who might not have read the titles accurately. Might the accuracy have increased if a smaller set of comparison titles had been used, for example only 10?

The pilot study highlighted the following:

1. Collectively, the participants generally agreed the on the ranks of titles.
2. The participant's opinions were more similar to one another where they were most positive of a higher relativity between titles.
3. The titles were not all good representations of the articles.
4. Participants opinions do vary even when based on such a small set of text.
5. The average variance was 31%.

The results show that the participants had a general relational understanding between the words found in the source document title, and those found in the Comparator. Where the word *agile* was used prominently in a title, the results reflect that fact. It is the researcher’s opinion that the hypothesis in Figure 3.1 has only been partially met due to the highlighted anomalies.

3.7 Pilot Study 1.1 – Comparing the Titles of Social Blogs

3.7.1 Aim

The aim of Pilot Study 1.1 was to confirm and extend what had been learned in the previous Pilot Study. This was to compare the meaning behind given texts, in this case title taken from Social Blogs. Participants would be asked to compare a single source title with a set of Comparator titles, and give a metric to represent their opinion on how closely they thought each Comparator title matched the source. The resulting metrics would be used to identify trends and observations.

The hypothesis, variables and sets and procedure would remain the same as in Pilot Study 1 except for the fact that there was 20 participants which was double those previously.

3.7.2 Inputs

For this pilot study, participants would be asked to compare a single source text with a number other pieces of comparator text giving a rating each time. In each case, the texts were titles of Social Blogs identified on the Internet. Some of the titles would be drawn from Blogs of similar contexts.

The source text was:

80s music – Boy George – need I say more?

The Comparator texts were:

A Eighties on the airwaves

B The birth of the metrosexual

C Why I miss eighties music

D Have I got Spitting Image for you?

E The path to rock mediocrity

1 A decade of decades

2 Bring back the eighties

3 The ascendency of Hip-Hop

4 Social Media and its Social Effects

5 The science of music

A copy of these documents can be found in Appendix 5.

3.7.3 Results

Figure 3.5 lists the results. Each Comparator paper is represented by a row. Each row contains:

Column 1: The text identifier

Column 2: The text.

Columns 3-22: Twenty columns representing the results for that title from each of the twenty participants.

Column 23: A sum total of the participants' results.

Column 24: An averages figure based on the above sum total divided by the number of participants.

Article	Title	Participant																					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total	Ave
2	Bring back the eighties	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	3	1	2	1	24	1.20
A	Eighties on the airwaves	1	1	1	1	2	1	1	2	2	2	2	1	1	1	1	2	1	2	1	1	27	1.35
C	Why I miss eighties music	2	1	1	1	1	1	1	2	2	3	1	1	1	2	4	1	2	2	1	1	31	1.55
1	A decade of decades	2	3	3	2	1	2	1	2	2	2	3	2	3	2	2	2	2	2	3	3	44	2.20
5	The science of music	3	2	3	2	3	3	3	3	3	2	2	3	3	2	2	2	2	2	1	3	49	2.45
E	The path to rock mediocrity	3	2	3	2	2	2	3	2	2	2	4	2	2	2	3	3	4	4	2	3	52	2.60
3	The ascendency of Hip-Hop	4	2	2	3	4	4	3	4	3	2	2	3	2	2	2	3	4	3	4	3	59	2.95
B	The birth of the metrosexual	4	4	4	3	4	3	4	3	1	4	4	4	3	4	4	4	3	4	3	4	71	3.55
D	Have I got Spitting Image for you?	4	4	4	3	4	3	3	3	4	4	4	4	3	4	4	4	3	3	4	4	73	3.65
4	Social Media and its Social Effects	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	3	4	78	3.90

Figure 3.5 Pilot Study 1.1 Results

Figure 3.5 ordered the results by the Average column, or in other words in ascending order of relativity to the source title; the most relative titles are at the top. Participant's choices are represented by the number 1 to 4. For example, it can be seen that participant 1 identified Article A (the 2nd row down), as a 1; very related.

The participants' Marking Sheets can found in Appendix 6.

3.7.4 Discussion

There were a number of observations which could be noted from these results.

The first observation was that as was expected, those Titles with the word ‘eighties’ in them would rate most highly. This confirms the concept of word association; where a word is recognised from the source text is found in the comparator text, it is likely to take precedence over other words. Texts B, D and 4 were added for particular reasons. Text B was added because it is widely acknowledge that Metrosexuals first emerged in the eighties. Text D was selected as ‘Spitting Image’ and ‘Have I Got News for You’ were TV programs both born in the eighties. Text 4 was chosen for its complete lack of relevance. It was reassuring to see them at the base of the table as they were all non-music related.

Degrees of Separation Matrix

Figure 3.6 shows the matrix containing all the variance figures entered. When we compare all these figures, we can start to understand a number of things about the participants and the titles. Each row contains:

Column 1: The article identifier.

Column 2: The article's title.

Columns 3-21: Twenty columns containing the variance figures.

Column 22: The total variance figure for that article.

Article	Title	Participant																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	Social Media and its Social Effects	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.90	0.10	0.90	3.50
2	Bring back the eighties	0.20	0.20	0.20	0.20	0.80	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	1.80	0.20	0.80	6.60
D	Have I got Spitting Image for you?	0.35	0.35	0.35	0.65	0.35	0.65	0.65	0.65	0.35	0.35	0.35	0.35	0.65	0.35	0.35	0.35	0.65	0.65	0.35	8.75
A	Eighties on the airwaves	0.35	0.35	0.35	0.35	0.65	0.35	0.35	0.65	0.65	0.65	0.65	0.35	0.35	0.35	0.35	0.65	0.35	0.65	0.35	8.75
1	A decade of decades	0.20	0.80	0.80	0.20	1.20	0.20	1.20	0.20	0.20	0.20	0.80	0.20	0.80	0.20	0.20	0.20	0.20	0.20	0.80	8.80
5	The science of music	0.55	0.45	0.55	0.45	0.55	0.55	0.55	0.55	0.55	0.45	0.45	0.55	0.55	0.45	0.45	0.45	0.45	0.45	1.45	10.45
B	The birth of the metrosexual	0.45	0.45	0.45	0.55	0.45	0.55	0.45	0.55	2.55	0.45	0.45	0.45	0.55	0.45	0.45	0.45	0.55	0.45	0.55	11.25
C	Why I miss eighties music	0.45	0.55	0.55	0.55	0.55	0.55	0.55	0.45	0.45	1.45	0.55	0.55	0.55	0.45	2.45	0.55	0.45	0.45	0.55	12.65
E	The path to rock mediocrity	0.40	0.60	0.40	0.60	0.60	0.60	0.40	0.60	0.60	0.60	1.40	0.60	0.60	0.60	0.40	0.40	1.40	1.40	0.60	12.80
3	The ascendency of Hip-Hop	1.05	0.95	0.95	0.05	1.05	1.05	0.05	1.05	0.05	0.95	0.95	0.05	0.95	0.95	0.95	0.05	1.05	0.05	1.05	13.25
		4.10	4.80	4.70	3.70	6.30	4.80	4.50	5.00	5.70	5.40	5.90	3.40	5.30	4.10	5.90	3.40	7.80	4.60	7.40	96.80

Figure 3.6 Pilot Study 1.1 Degrees of Separation Matrix

It can be seen in Figure 3.6 that there was a low level of variance amongst the highest and lowest ranked texts. This means that, as with the previous study, where participants were more sure of the relevancy, they were more cohesive in their opinions. Through the mid-section of Figure 3.5, especially E and 3, the variance increased.

Average Variance Figure

The average variance figure is 0.48. With the four point scoring system, this meant that the figure 0.48 equated to about a 12.5% variance.

3.7.5 Pilot Study 1.1 Conclusion

This pilot study highlighted the following:

1. Collectively, the participants generally agreed the on the ranks of most titles.
2. The participant's opinions were more similar to one another where they were most positive of a higher relativity between titles.
3. Participants opinions do vary even when based on such a small set of text.
4. Those central bands of text in Figure 3.5 were more likely to have a heightened variance.
5. The average variance was 12.5%.

3.8 Pilot Study 2 – Comparing Structured Abstracts of Academic Articles

3.8.1 Overview

The overview for Pilot Study 2 is as follows:

Aim	To ascertain how similarly participants understood and related the meaning behind collections of texts.
Hypothesis	Participants' opinions would be similar. There would be an improvement in results seen against Pilot Study 1.
Variables and Sets	Control variables: Text format and content. Response variables: The participants' opinions. Control group: A source Structured Abstract. Experimental group: 10 Comparator Structured Abstracts. Participants: 12.
Inputs	The participants' opinions when comparing 10 Comparator texts with the source text.
Procedure	The participants are asked to make judgements over similarity between the abstracts.
Format of Results	Metrics representing the participant's opinions.

Figure 3.7 Pilot Study 2 Overview

3.8.2 Aim

The aim of the second pilot study was to expand on the understanding gained from the first. Again, participants would be asked to compare a single source text with a set of Comparator texts, but in this case the texts would be Structured Abstracts. As in Pilot Study 1, they were asked to give a score, in this case a metric, to represent their opinion on how closely they thought each Comparator matched the source. The resulting metrics would be used to identify trends and observations and be compared with Pilot Study 1.

This pilot study differed to the first pilot study in a number of ways:

1. The collection size would be reduced. In Pilot Study 1, the participants were asked to consider 35 Comparator article titles. In this pilot study, only 10 Comparator articles would be used.
2. Instead of using the articles' titles, Structured Abstracts were used in the comparison process.
3. A four point scoring system, as defined in Section 3.2.6, was used.

3.8.3 Hypothesis

The following table lists the hypotheses for this experiment.

Hypothesis ID	Hypothesis
H1	The participants' opinions would be similar.
H2	An improvement in the level of similarity between participants' opinions when compared with Pilot Study 1.

Figure 3.8 Pilot Study 2 Hypothesis

3.8.4 Variables and Sets

Control variables:	The content and layout of the texts given to the participants was different in this study to the previous one.
Response variables:	The participants' opinions based on the new texts.
Control group:	A source Structured Abstract.

Experimental group: 10 Comparator Structured Abstracts.

Participants: 12.

3.8.5 Inputs

For this pilot study, the source Structured Abstract was the same as the previous pilot study, the academic paper:

Empirical studies of agile software development: A systematic review

There was also a group of 9 Comparator Structured Abstracts, all representing articles within the domain of *Agile Development*. Where Structured Abstracts were not readily available for the selected academic articles, one was created using the abstract section from that article. The Structured Abstracts would not contain the title of the article, and so, each would be allocated an identifier number.

A further Structured Abstract was added which was considerably less related to the subject domain and would be used as an Indicator Abstract. This abstract was in the medical domain, although it used a number of terms which might be related to a Computing domain such as *system, systematic, recommendation, and communication*. This Indicator Abstract would be Abstract number 9. There was therefore 10 Structured Abstracts.

3.8.6 Procedure

The pilot study took place over two days, each a two hour session. Students could attend any time during these periods. The procedure for the pilot study was quite simple. Participants were given:

1. A sheet of paper with brief instructions.

2. A Structured Abstract of the source article.
3. A set of 10 Structured Abstracts from the Comparator articles.
4. A scoring sheet.

A copy of these documents can be found in Appendix 7.

Having read each Structured Abstract, the participants were asked to make a judgement on the similarity between each Comparator abstract to the source abstract using a four point scoring system which would use the following form:

1. Very similar.
2. Fairly similar.
3. Vaguely similar.
4. Not similar.

3.8.7 Results

Figure 3.9 lists the results. Each Comparator abstract is represented by a row. Each row contains:

Column 1: The abstract identifier number.

Column 2-13: Twelve columns, one for each participant containing their similarity scores for each of the 10 abstracts.

Column 14: The average figure for each Structured Abstract based upon the twelve scores.

Column 15: The rank based on Column 14.

Participant	1	2	3	4	5	6	7	8	9	10	11	12	Avg.	Rank
Abstract #														
Abstract 04	1	2	2	1	1	2	2	2	1	3	3	2	1.83	1
Abstract 03	1	3	3	1	2	3	2	3	4	1	2	1	2.17	2
Abstract 07	1	1	4	2	2	2	1	1	4	4	2	2	2.17	2
Abstract 02	1	1	3	3	4	1	1	1	2	4	3	3	2.25	4
Abstract 05	1	1	3	2	2	4	2	1	4	2	3	2	2.25	4
Abstract 06	1	1	2	3	4	2	3	3	1	4	1	3	2.33	6
Abstract 08	3	3	3	2	3	3	3	1	2	4	3	3	2.75	7
Abstract 01	2	4	2	1	4	4	3	4	2	3	2	4	2.92	8
Abstract 10	4	4	3	2	3	2	3	3	2	3	2	4	2.92	8
Abstract 09	4	4	3	3	4	4	3	4	4	3	1	4	3.42	10

Figure 3.9 Pilot Study 2 Results

Figure 3.9 displays the Pilot Study 2 results ordered by the Average column, or in other words in ascending order of relativity to the source article; those ranked as most relative are towards the top. Participant's choices are represented by metrics ranging from 1 to 4, 1 being a very similar match. For example, it can be seen that Participant 5 ranked Abstract 4 as the only *very similar* abstract.

3.8.8 Discussion

As can be seen in Figure 3.9, the participants found, on average, that Structured Abstract 4 most closely matched the source article's Structured Abstract. Out of the 12 participants, 11 of them judged it as either *very* or *fairly* similar, giving it a '1' or a '2'. This can be seen by observing the Average figure which was 1.73, lower being better.

One notable observation was those figures from Participant 9 who had three '4' ratings in the top half of the table, and three '2' and one ‘1’ rating in the lower half of the table. These 'bucked' the general trend of the other ratings, and may well have affected the overall accuracy of the study. We can see how this Participant has effected variance below.

Using the same method as described in Section 3.1.7, a Degrees of Separation

Matrix was created from the data in Figure 3.9 to create Figure 3.7. From this, it can be seen in the right most columns, the total variance each abstract gained.

Participant	1	2	3	4	5	6	7	8	9	10	11	12	Total
Abstract #													
Abstract 08	0.25	0.25	0.25	0.75	0.25	0.25	0.25	1.75	0.75	1.25	0.25	0.25	6.50
Abstract 04	0.83	0.17	0.17	0.83	0.83	0.17	0.17	0.17	0.83	1.17	1.17	0.17	6.67
Abstract 10	1.08	1.08	0.08	0.92	0.08	0.92	0.08	0.08	0.92	0.08	0.92	1.08	7.33
Abstract 09	0.58	0.58	0.42	0.42	0.58	0.58	0.42	0.58	0.58	0.42	2.42	0.58	8.17
Abstract 05	1.25	1.25	0.75	0.25	0.25	1.75	0.25	1.25	1.75	0.25	0.75	0.25	10.00
Abstract 03	1.17	0.83	0.83	1.17	0.17	0.83	0.17	0.83	1.83	1.17	0.17	1.17	10.33
Abstract 07	1.17	1.17	1.83	0.17	0.17	0.17	1.17	1.17	1.83	1.83	0.17	0.17	11.00
Abstract 01	0.92	1.08	0.92	1.92	1.08	1.08	0.08	1.08	0.92	0.08	0.92	1.08	11.17
Abstract 06	1.33	1.33	0.33	0.67	1.67	0.33	0.67	0.67	1.33	1.67	1.33	0.67	12.00
Abstract 02	1.25	1.25	0.75	0.75	1.75	1.25	1.25	1.25	0.25	1.75	0.75	0.75	13.00
													96.17
Variance	9.83	9.00	6.33	7.83	6.83	7.33	4.50	8.83	11.00	9.67	8.83	6.17	96.17

Figure 3.10 Pilot Study 2 Degrees of Separation Matrix

Another observation to note was the position of Abstract 9, the unrelated Indicator Abstract. Apart from Participant 11, all other participants ranked Abstract nine as either 'vaguely similar' or 'not similar'. It had been anticipated that all participants would have identified this abstract as 'not similar', but only 7 out of the 12 participants did.

It can be seen that Participant 7 had the lowest combined variance of 4.50 which meant that the opinions of Participant 7 most closely matched the other participant’s opinions, on average. Conversely, Participant 9, identified earlier, had the highest variance from the other participants.

Average Variance Figure

The average variance across the 12 participants and the 10 abstracts can be calculated as follows:

Total of all 10 variance figures / 12 participants / 10 abstracts = average variances

Total Variance Figure = 96.17

Divided by 12 participants

96.17 / 12 = 8.01

Divided by 10 titles

7.01 / 10 = 0.80

Therefore, the average variance figure is 0.80. As the four point scoring system went from 1 to 4, a range of 3, the figure 0.80 equated to a 27% variance.

If we were to remove Participant nine, highlighted as the candidate with the highest variance, the variance figure reduces from 0.80 to 0.71, or less that 24%.

3.9 Pilot Study 2.1 – Comparing the Content of Social Blogs

3.9.1 Aim

The aim of Pilot Study 2.1 was to confirm and extend what had been learned in the previous Pilot Study 1.1. This was to compare how participants conceived the meaning behind given texts. In this pilot study, the participants would be comparing the content from a set of Social Blogs. Participants had been asked in Pilot Study 1.1 to make a judgement based purely on the title associated with the blogs. In Pilot Study 2.1, the actual content of those blogs would be compared but the participants would not be given the titles. The resulting metrics would be used to identify trends and observations.

The hypothesis, variables and sets and procedure would remain the same as in Pilot Study 2 except for the fact that there was 20 participants which was double those in Pilot Study 2.

3.9.2 Inputs

For this pilot study, participants would be asked to compare a single source text with a number other pieces of comparator text giving a rating each time. In each case, the texts were the contents of Social Blogs identified on the Internet. Some of the texts would be drawn from Blogs of similar contexts.

A copy of both the source and comparator texts can be found in Appendix 8.

3.9.3 Results

Figure 3.11 lists the results. Each Comparator paper is represented by a row. Each row contains:

Column 1: The text identifier

Column 2-21: Ten columns representing the results for that title from each of the twenty participants.

Column 22: A sum total of the participants' results.

Column 23: An averages figure based on the above sum total divided by the number of participants.

Article	Participant																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total	Ave
C	2	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	2	4	1	2	28	1.4
A	2	1	1	2	2	1	1	1	3	2	2	2	1	1	2	2	3	2	3	3	37	1.85
1	3	2	2	2	1	3	2	1	2	1	1	2	2	2	2	3	1	3	1	2	38	1.9
B	3	2	2	3	3	2	2	2	1	3	2	1	2	2	2	2	4	2	3	2	45	2.25
5	2	3	2	2	2	2	3	2	3	1	1	2	3	2	3	3	3	3	3	4	49	2.45
E	3	2	3	2	3	2	2	2	2	2	3	3	3	2	3	2	3	1	2	4	49	2.45
3	3	3	3	3	2	3	3	2	2	3	2	3	3	2	3	3	3	1	3	3	53	2.65
2	3	3	4	4	2	3	3	3	3	2	4	3	2	3	3	3	2	2	2	3	57	2.85
D	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	3	78	3.9
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	79	3.95

Figure 3.11 Pilot Study 2.1 Results

Figure 3.11 ordered the results by the average column, or in other words in ascending order of relativity to the source title; the most relative titles are at the top. Participant's choices are represented by the number 1 to 4. For example, it can be seen that participant 1 gave text C a 2; fairly relevant.

The participants' Marking Sheets can found in Appendix 9.

3.9.4 Discussion

There were a number of observations which could be noted from these results.

The first observation was those texts which were intentionally included because they were quite unrelated, where identified overwhelmingly correctly.

It was also worth noting that the top choice of the participants in Pilot Study 1.1 “Bring back the eighties” was rated as only seventh here. This was a better reflect on its content as despite what the title implied, the blog was about TV shows, not music.

Degrees of Separation Matrix

Figure 3.12 shows the matrix containing all the variance figures entered. When we compare all these figures, we can start to understand a number of things about the participants and the titles. Each row contains:

Column 1:	The article identifier.
Column 2:	The article's title.
Columns 3-21:	Twenty columns containing the variance figures.
Column 22:	The total variance figure for that article.

Article	Participant																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
4	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.95	0.05	0.05	1.90
D	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.90	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.90	3.60
3	0.35	0.35	0.35	0.35	0.65	0.35	0.35	0.65	0.65	0.35	0.65	0.35	0.35	0.65	0.35	0.35	0.35	1.65	0.35	0.35	9.80
2	0.15	0.15	1.15	1.15	0.85	0.15	0.15	0.15	0.15	0.85	1.15	0.15	0.85	0.15	0.15	0.15	0.85	0.85	0.85	0.15	10.20
1	1.10	0.10	0.10	0.10	0.90	1.10	0.10	0.90	0.10	0.90	0.90	0.10	0.10	0.10	0.10	1.10	0.90	1.10	0.90	0.10	10.80
B	0.75	0.25	0.25	0.75	0.75	0.25	0.25	0.25	1.25	0.75	0.25	1.25	0.25	0.25	0.25	0.25	1.75	0.25	0.75	0.25	11.00
C	0.60	0.40	0.40	0.40	0.60	0.40	0.40	0.60	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.60	2.60	0.40	0.60	11.20
E	0.55	0.45	0.55	0.45	0.55	0.45	0.45	0.45	0.45	0.45	0.55	0.55	0.55	0.45	0.55	0.45	0.55	1.45	0.45	1.55	11.90
A	0.15	0.85	0.85	0.15	0.15	0.85	0.85	0.85	1.15	0.15	0.15	0.15	0.85	0.85	0.15	0.15	1.15	0.15	1.15	1.15	11.90
5	0.45	0.55	0.45	0.45	0.45	0.45	0.55	0.45	0.55	1.45	1.45	0.45	0.55	0.45	0.55	0.55	0.55	0.55	0.55	1.55	13.00
Variance	4.25	3.25	4.25	3.95	5.05	4.15	3.25	4.45	4.85	5.45	5.65	4.35	4.05	3.45	2.65	3.55	6.85	9.65	5.55	6.65	84.50

Figure 3.12 Pilot Study 2.1 Degrees of Separation Matrix

It can be seen in Figure 3.12 that those four texts with the lowest variance can be seen in Figure 3.11 at the bottom of the table. This means that the four least relevant texts were the four most agreed upon ones by the participants. It is also worth noticing the size of the variance from Participant 18's opinions; on average more than double any other participant.

Average Variance Figure

The average variance figure is 0.42. With the four point scoring system, this meant that the figure 0.42 equated to about a 12% variance.

3.9.5 Pilot Study 2.1 Conclusion

This pilot study highlighted the following:

1. Collectively, the participants generally agreed the on the ranks of most titles.
2. The participant's opinions were more similar to one another where they were most positive of a higher relativity between titles especially in the case of the highly unrelated texts.
3. The average variance was 12%.
4. Individual participants can skew the figures.

3.10 Summary

This Chapter began by introducing the idea that the pilot studies would be undertaken in collect a set of data which could be used in the main experiment, and to investigate how similarly participants conceive the same piece of text.

Using a set format, each pilot study was described in turn, and the corresponding findings, discussion and conclusions drawn.

These pilot studies demonstrated the following:

1. From Pilot Study 1 to Pilot Study 2, there was a small decrease in the variance found in participant's opinions; from 31% to 27%. This meant that by increasing the description of a document, from a document's title to a structured abstract, participants were slightly more likely to make similar judgements upon the content. The figure in the Pilot Study 2 would be improved further by removing Participant nine, the spurious candidate, to a figure of 24%.
2. From Pilot Study 1.1 to Pilot Study 2.1, there was a small decrease in the variance found in participant's opinions; from 12.5% to 12%. This meant that by using the blog itself rather than just the title, participants were slightly more likely to make similar judgements upon the content. This was markedly less that had been expected.
3. It found that individual participants could affect the data in such a way as to skew the findings such as Participant 18 in pilot Study 2.1.
4. By using simpler content, the variance could be decreased considerably. From 31 and 27 percent respectively in Pilot Studies 1 and 2, to 12.5 and 12 percent in Pilot Studies 1.1 and 2.1.

However, it was found that in the Pilot Studies, as with many experiments involving humans, results were variable and could be affected by their results. It was worth recognising that the opinions of many participants were more often than not more reliable than individual's opinions.

One of the main aims of these Pilot Studies was to test the Experimental Template against smaller sets of data and identify any improvements that could be made to it. One concern identified with the pilot studies was that by limiting the opinion of a participant to ranges of either a tick, a cross or a blank, as in Pilot Study 1, or a range (of 1 to 4) as in Pilot Study 2, this may be reducing the accuracy of their selection. It was decided that in the experiments, ranks would be used instead of the above method. This would allow the participants to be able to place their selection in ranked order rather than have to allocate values to them.

The pilot studies also highlighted that where a text was better described through its content, participants would make a more cohesive opinion on its relativity.

Chapter 4 The Durham Browsing Assistant

4.1 Introduction

People accessing the web generally fall into two categories; those searching for something specific, and those who are browsing for something new or interesting. The Web offers huge amounts of informal and semi-structured information; these are among the main reasons for its success.

Historically, a person wishing to learn a subject would get an appropriate book or a set of academic papers. The process of obtaining the book or papers may mean that by the time it is read, it may be out of date by some years. In many subjects this is not so apparent, but in subjects such as medicine and computing, it becomes quite apparent due to their rate of change in research. Using the Web for academic research is almost essential, offering the widest variety, and the most up to date material. Typically though, however convenient, research material on the Web is dispersed and difficult to locate within the plethora of other information.

Chapter 2 discussed the Web as a learning resource, and discussed those technologies available which could be of use to both consumers and producers of the Web. Chapter 3 undertook some pilot studies which highlighted that if articles were better described, students could make a more precise judgement about their suitability to their own research.

This chapter will build upon this by considering specifically which technologies may be utilised to enhance the researching potential of the Web. It begins by reviewing the areas of concern highlighted in Figure 2.1, and by using the Scenario as a guide; it proposes a solution to address those areas of concern.

This solution is called DurBA – the Durham Browsing Assistant. DurBA is a model to support Web-based researching and is described in detail, and then a number of specific areas of research were highlighted. One of these areas, the *textual* *comparison process,* is selected for further research and is considered in isolation for improvement.

4.2 Issues: Identified Areas of Concern

Chapter 2 identified that there are a number of areas of concern with using the Web in an academic environment. Many of these concerns cannot be addressed in isolation, being highly cohesive, whilst others can be considered as more independent of one another.

One such area which can be considered in relative independence is in the use of the Web as an academic resource, which shall now be considered.

Chapter 2 highlighted a number of areas of concern with using the Web as a learning resource. From that list of 7, the following were chosen as they all relate to students using the Web as a learning resource:

1. The time it takes to locate academic materials on the Web. As the size of the Web increases, the proverbial haystack in which students must sift through is growing. This makes it increasingly difficult for students to find the materials they are after; searching time is thus increasing, impacting on research time. In many cases, there is limited meta-information available about such resources 'at first glance', and many of these materials are poorly described online.
2. Synchronous or repeated research. In some cases, students are undertaking synchronous research on the Web in their attempts to locate similar resources.
3. Identifying the quality of materials found on the Web. Even when students locate materials on the Web, it may be difficult to identify the academic value of the given material. Historically, where material was sourced from

an academic journal, the journal could be used as a guide to the quality of that material. But where material is sourced from a Web page, the student may have trouble in establishing the academic value of such material based on the Domain it was drawn from. Much of the material available is of small pedagogical value to the students, is too specific, sometimes sensationalised or purely false.

4. Gaining physical access to materials hosted on the Web. Much academic material on the Web is actually hidden behind 'paid access' portals. Where once, articles may have been reproduced from a Journal, paid for by a subscription through the universities, there are now too many of these 'paid access' portals for such universities to be subscribed to them all, if not more than a few.

5. Resources on the Web can be less permanent than those in print. Information may not as permanent as other information such as that stored in a library. This problem can manifest itself as ‘broken links’ in Web-pages.

Figure 2.1 listed a total of 7 items, but only the above 4 were considered to be viable for encapsulation within software. This research shall now propose a model which will be capable of addressing these issues, highlighting how it aims to do this. Such a model, called the Durham Browsing Assistant (DurBA) will now be described in detail.

4.3 Requirements

To implement a system such as one that would address the problems identified above and create functionality such as those described in The Scenario in Section 1.5, a model would be created.

This model must should:

1. Be able to operate independently of chosen Internet Browser.
2. Be non-invasive and operate transparently to the user.
3. Operate over a standard Internet connection.
4. Gather information from the user's browsing activities.
5. Categorise and group materials into areas of research.
6. Enable users to collaborate through sharing and recommending materials.
7. Autonomously undertake web searches based on areas of research.
8. Display relevant materials to individuals based on their area of research.

Most importantly, the model must be realistic and it should be a viable solution to the problem.

4.4 The Model: DurBA, The Durham Browsing Assistant

The Durham Browsing Assistant (DurBA) is designed to address the highlighted areas of concern as listed in Section 4.2, and at the same time, fulfil the Scenario as described above. DurBA is a model for a collaborative, interactive system; collecting information from, and giving information to, its users. It allows multiple users to share suitable materials they find on the Web pertinent to their Area of Research with one another.

4.4.1 Physical Distribution

The physical distribution of DurBA, as can be seen in the image following, is such that each user (client) has a piece of software, called a Client Agent (CA) running on their computer which communicates with a Server Agent (SA) via the Web. The Server Agent stores, manipulates and distributes data from and to the Clients. The Client Agent 'watches over' the user and also displays any feedback such as materials found by other users.

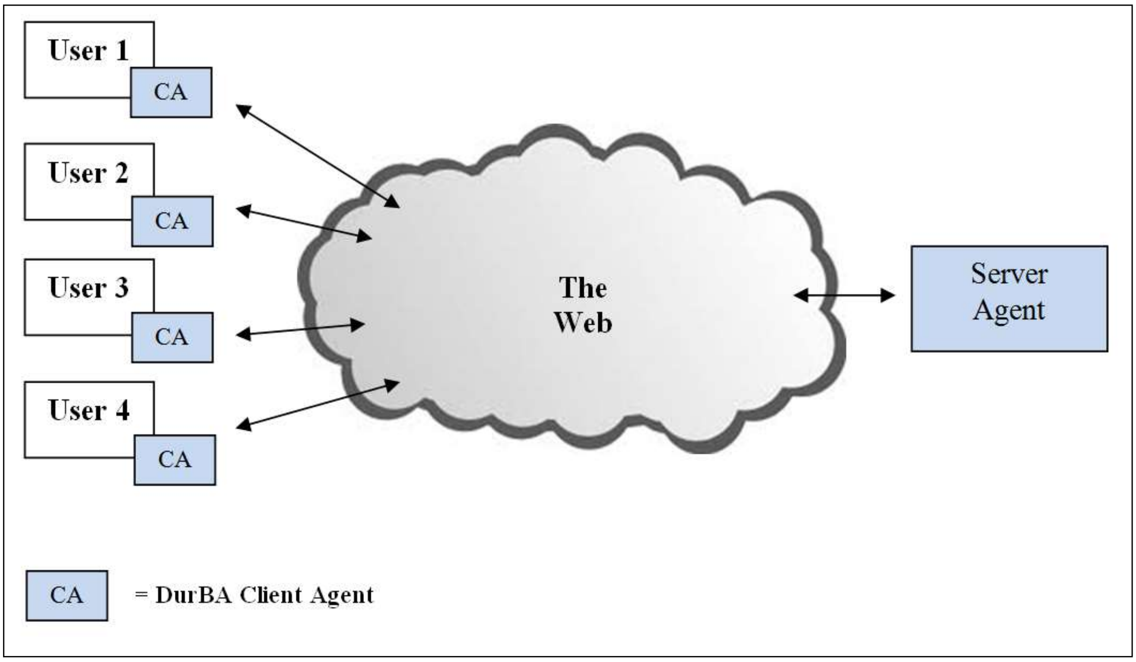

Figure 4.1 DurBA Physical Distribution

There were two network models that could be used to implement DurBA; client based and server based. From the outset, a combined client and server model approach was considered the most appropriate. By considering how the data flows in the model, it can be clearly seen as the best approach. One of DurBAs key features was that it would 'observe' and monitor the users as they went about their browsing and researching tasks. This would have been difficult to do based on a Server only implementation.

The model which was eventually decided upon would have a small Agent sitting on each of the user's PCs, communicating with the server over the Internet.

To implement a server based only version of DurBA, it would have needed all browsing traffic from all of the users to be passed through a proxy server to gather the relevant information even before the processing aspect of the model had begun. This would have caused a scalability problem as the more people who used it, the more the strain would be placed on the server. An alternative was to implement two servers, one as a proxy server gathering information, and a second as a processing server.

It was deemed a better architecture to have Agents in place on each PC to gather and observe, and periodically send the data to the server. By this method the server could process in batches making it more efficient. This method also spread the workload of 'proxying' the web requests and thus the load would be spread over as many PCs as there were users.

4.4.2 High-level View

DurBA in its entirety can be seen from a high-level aspect in the image following. Note how the entire system is portrayed as a single entity with inputs and output as peripherals.

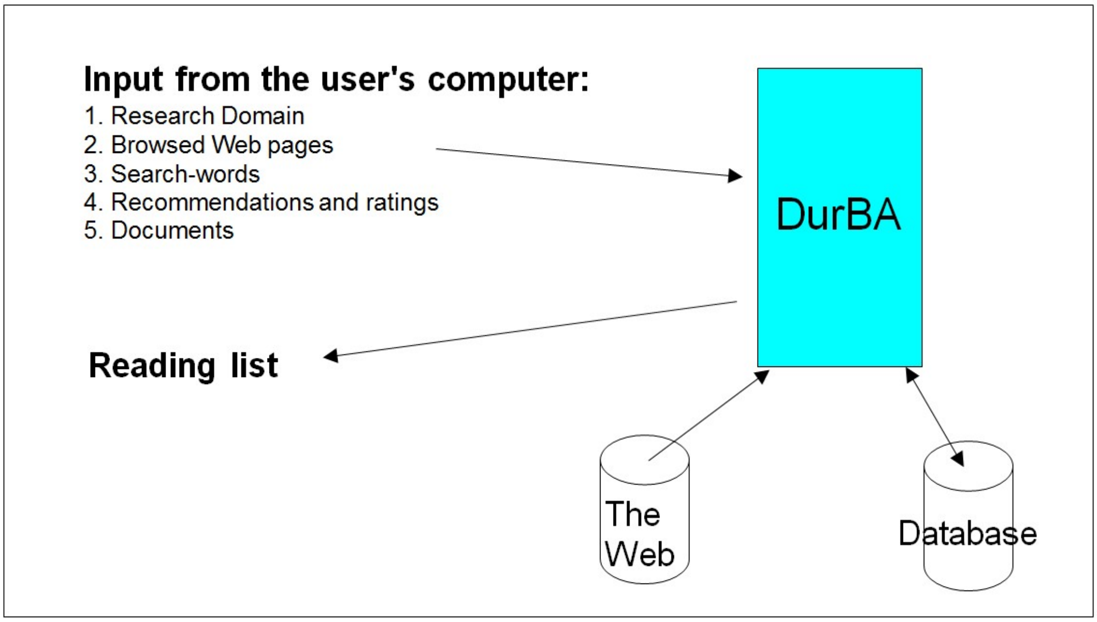

Figure 4.2 DurBA High Level View

Internal inputs from the User's Computer to DurBA are:

1. The Research Domain. This is a ID entered by the user so the Server Agent knows what that individual user's area or research is.
2. Browsed Web pages.
3. Search-words submitted to Search-engines.
4. Recommendations and ratings for suitable materials found on the Web.
5. Documents held on the user's computer.

External inputs from the User's Computer to DurBA are:

1. The Web. In DurBA, the Web is used as a resource.
2. The Database. Any information collected from users will be stored in a database.

The outputs from DurBA are:

1. A Reading List. This is a list of suggested materials most suited to individual areas of research produced by the Server Agent. This is offered to users.
2. The Database. Data is drawn from the database on occasions, such as when a Uniform Resource Locator (URL) of a known resource is requested by a user.

4.3.3 Medium-level Model

As can be seen in the medium-level view in Figure 4.1, DurBA consists of two main components:

1. Server Agent. The Server Agent is located on a remote server accessible from anywhere on the Web.

2. Client Agent. A copy of the Client Agent runs on each user's computer. It consists of a Graphical User Interface (GUI) and communications ability. Each user needs to run a Client Agent.

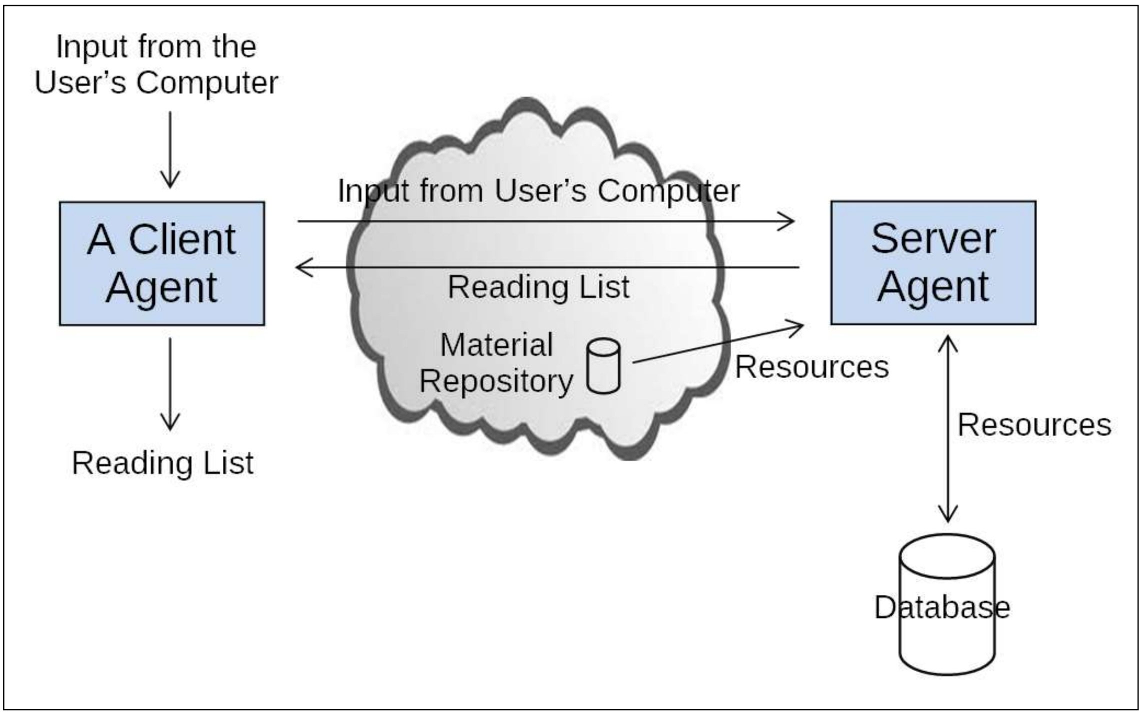

Figure 4.3 Medium-level View of the DurBA

The Client Agent communicates with the Server Agent passing the following data:

1. Input from the User's Computer. This allows the Server Agent to build up an understanding of what the user is researching. This includes any recommendations which the users make.

The Server Agent communicates with the Client Agent passing the following data:

1. Reading List. The relevant Reading List is sent to the Client Agent to allow access by the user.

It can also be seen from this diagram that the Server Agent may make use of the Web as a resource for materials relevant to the Area of Research, and a database to record information in.

Concept: Research Domains

DurBA maps Research Domains to areas of research. Each Area of Research has its own unique Research Domain.

All users researching an Area of Research will log in using a Research Domain. Whenever a Client Agent connects to the Server Agent, it will send its user's Research Domain. This acts as a type of 'log-on' which enables the Server to 'know' what the user is researching.

A Research Domain File is created by the Server Agent based on the information it gathers using input from the users' computers of that Domain. The actual Research Domain a user is currently a member of can be changed manually by the user changing it in the Client Agent GUI. This allows them to be members of multiple Research Domains, although utilising only one at a time.

Research Domains are not to be considered as ephemeral, they are fluid, and may change over time. As new recommendations and Search-words are submitted by users, the Research Domain will adjust and optimise to better represent their Area of Research. However, it would be possible to create some fixed Research Domains to be used in certain situations such as where a tutor wishes to define a research area clearly for a study group or class.

Users will even be able to subscribe to one or more existing Research Domains from a library which they find relevant to their research

4.4.3 The Client Agent

The Client Agent is the interface between the user and the DurBA model. In the low-level view in Figure 4.2, it can be seen that the Client Agent receives a number of inputs, both from the user's computer, and from the Server Agent.

The Client Agent takes the following inputs:

1. Input from the User's Computer.
2. Reading List from the Server Agent.

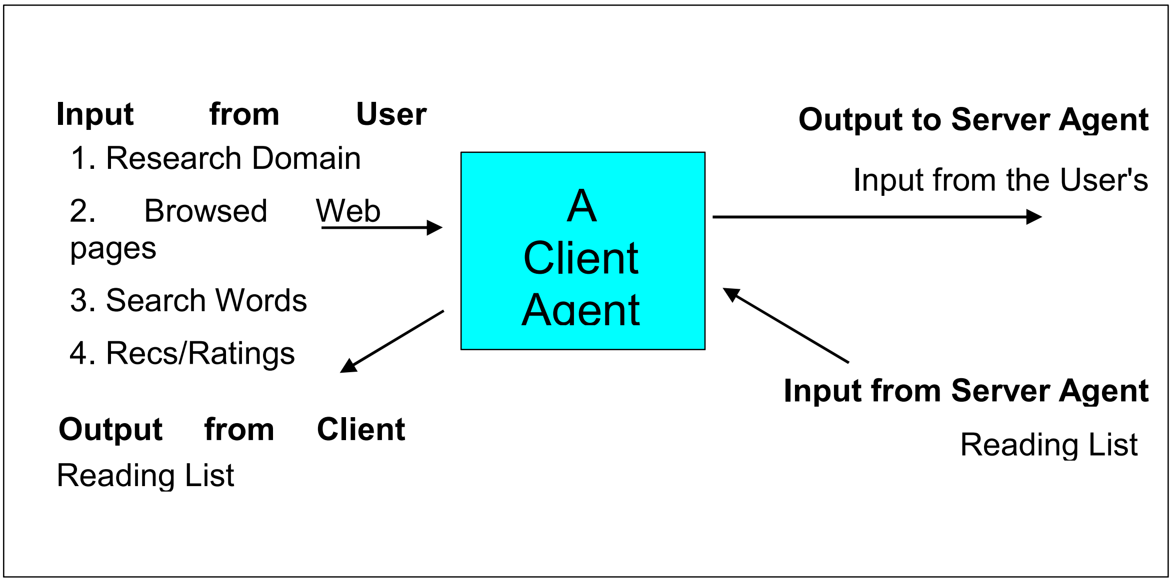

Figure 4.4 Low-level View of the DurBA Client Agent

The Client Agent also has a number of outputs:

1. It passes the input from the User's Computer to the Server Agent.
2. It displays the Reading List to the user.

The Client Agent provides a number of facilities:

1. To provide a graphical user interface to display information to the user, and to provide a means for the user to interact with it.
2. To collect input from the user's computer.
3. To send and receive data to and from a Server Agent.

The Client Agent will be represented through a graphic interface which may appear something like the following.

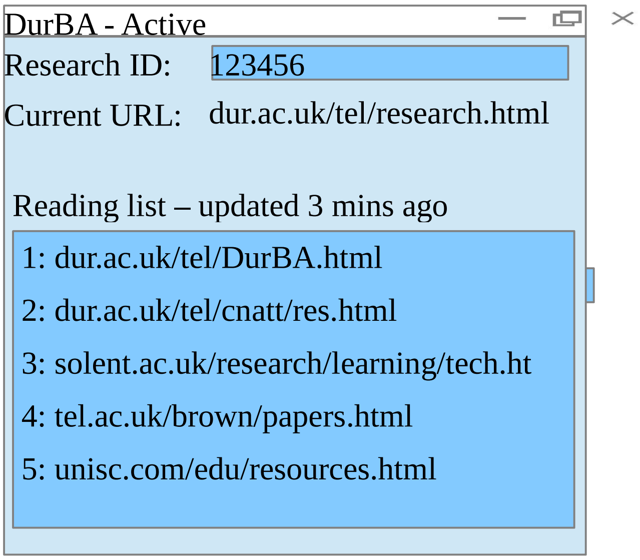

Figure 4.5 A Possible Graphical Interface

The above image illustrates the window which would allow the user to interact with DurBA. In this window, the following can be seen:

1. The Research Domain. This can be changed by the user according to what the user is researching; they may be a member of a number of researching groups.
2. The Current URL. This displays the URL of the currently viewed Web page.
3. The Recommend Button. This would allow the user to recommend the

current Web page; see *Concept: Collaboration* following.

4. The Reading List. This is a 'clickable' list of URLs representing the most appropriate resources currently found by other members of the research group.

As the user browses the Web, the Client Agent 'observes' what the user is viewing, saving, printing, and the time spent viewing documents. This data is all sent to the Server Agent, and it will use this data to make a judgment on how the Research Domain should be defined and redefined to best represent the user's Area of Research.

4.4.4 The Server Agent

Figure 4.3 illustrates the Server Agent. The main aim of the Server Agent is to collect and collate data, and to undertake data processing tasks. This will reduce the impact of running a Client Agent on the user's computer.

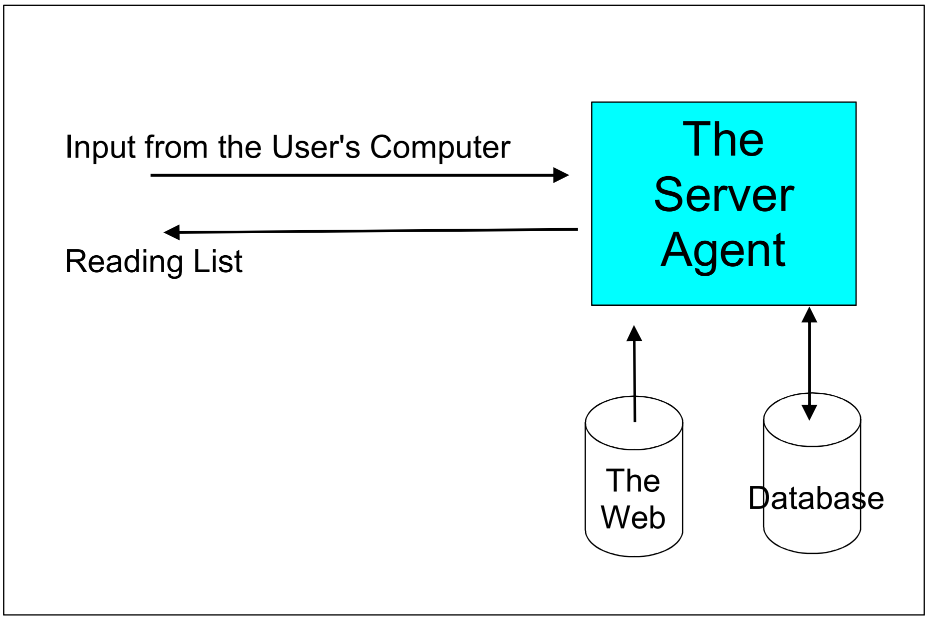

Figure 4.6 Low-level View of the DurBA Server Agent

In the low-level view in Figure 4.6, it can be seen that the Server Agent receives a number of inputs and outputs, these are listed following.

The Server Agent takes a number of inputs:

1. Input from the user's Computer, passes via the Client Agent
2. Input from the Database; information stored previously to the database.
3. Input from the Web. The Server Agent can search the Web for other materials which are suitable given a Research Domain. These may be added to the appropriate Reading List.

The Server Agent has a number of outputs:

1. Those to the Client Agents. A Reading List, which is a list of resources on the Web identified as most suitable to the user's Area of Research. The Server Agent knows what each user's Area of Research is because of their Research Domain.
2. Those to the Database. A variety of information is stored to the Database such as Research Domains, recommendations and Resource URLs.

The Server Agent provides the following functionality:

1. Listen and accept data. The Server Agent will 'listen' for incoming data from Client Agents.
2. Transmit data. It will send out data such as recommendations to Client Agents.
3. Recommendations. It will categorise recommendations from the Client Agents.
4. Store. It will record data to the database such as:
    a. Recommended pages and documents.

   b. Research Domains.

   c. Materials URLs.

5. Research Domains. It will rationalise, adjust and optimise Research Domains depending upon user recommendations and Search-words.

6. Autonomous Web searching. It will undertake occasional Web searching based on user Research Domains in order to locate further relevant materials. Where relevant materials are found, they may be forwarded as recommendations to relevant users. These, in turn may be re-reviewed by users.

Concept: Collaboration

The user will have the facility to write a review for any materials found on the Web. These recommendation will be periodically sent to the Server Agent, and the rating attached to them will affect how highly placed the materials are in the Reading List. The recommendation window may appear similar to that following.

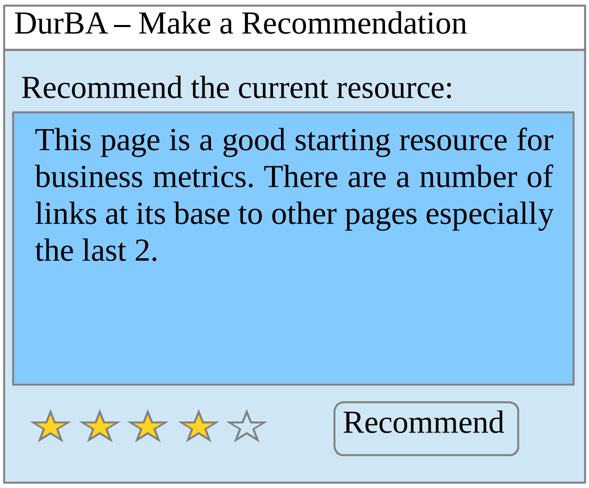

Figure 4.7 A Potential Recommendation Window

As can be seen in the image, using the recommendation window, the user would be able to add a comment and rate the material using rating system. This process will be available not only to Web pages, but also other formats such as PDFs, word processing documents, spread sheets etc. The recommendation may be seen by other users in the same research group, i.e. those using the same Research Domain.

Not only will the user be able to make recommendations, they will also be able to view other user's recommendations. All materials in the Reading List will have at least one review associated with it. Also, when a user browses to a resource which is recognised by DurBA, a notifier would be triggered to identify this fact such as an audio signal. The recommendation window containing the recommendation would then be made available for the user to view. The recommendation window would display all the review comments together with their ratings.

The effect of this 'recognition' process is that once a person has reviewed a piece of material, it may not need reviewing again, and based on the recommendation rating value, it may or may not form part of the Reading List. However, recommended materials can still be re-recommended multiple times which will also affect their inclusion in this list.

Further to this, when a review of a piece of material is sent to the Server Agent, it is accompanied with the URL for that piece of material. This Server Agent downloads and stores a copy of the material together with the rating so that future users would have immediate access to it. This is accessible via the reading list.

The Client Agent broadcasts the Research Domain with any communication it sends to the Server Agent. In this way, the Server knows what Area of Research each individual user is interested in. This is essentially a method of collaboration. When the Server Agent receives a recommendation from a Client Agent, it associates that recommendation with that research group and shares it to other members. All Research Domains and recommendations are stored in a database. The Server Agent may use the Research Domains by autonomously searching the Web (possibly using the domains visited by those users) to undertake assisted searches with a view to finding more relevant materials. Those materials may then be forwarded to the relevant users and displayed to the users.

Concept: The Reading List

Another concept to consider is the Reading List. This is a collection of materials the DurBA suggests to the user which are relevant to their Area of Research.

The resources that make up a user's Reading List will be dynamic, and will change as new materials are located, reviewed and re-reviewed. The Reading List will only be made up from recommendations and ratings made by users of the same Research Domain; i.e. only very relevant materials will be included in the list. The Reading list will be ordered according to ratings, i.e. most highly rated first. When a user selects a resource item in the list, there will be options to view comments and ratings, and to download the resource directly.

4.3.4 DurBA Model Summary

To summarise the DurBA model, it would incorporate the following concepts:

1. It allows users to browse the Web unhindered.
2. A Research Domain is used to identify a user' or group of users' Area of Research.
3. Users may subscribe to one or more existing Research Domain depending upon their current Area of Research, but may only use one at any given time. They may also create their own.

4. By specifying Research Domains, collaboration can occur between users. This approach was identified in Section 2.2.1 of the Literature Review.

5. Users may be part of a researching team, for example a study group, who are all using the same Research Domain. This forms a collaborative aspect to the system.

6. Users can identify relative and non-relative materials, make recommendations, and have recommendations made to them.

7. Where a user browses to some material which has already been reviewed by a peer user previously, a notifier would be triggered to identify this fact, and would display the comments and ratings for that material.

4.5 How DurBA Addresses the Areas of Concern

Chapter 2 identified that there are a number of areas of concern with using the Web as an academic researching environment, highlighted in bold following. The DurBA model would address these areas of concern in the following ways:

1. **The time it takes to locate academic materials on the Web**. Using the Reading List would mean that users have a readily available set of resources 'at hand' which would reduce their searching time.

2. **Synchronous or repeated research**. Where users browse to a Web resource which has already been reviewed by a peer of their Research Domain, they will be presented with a notifier showing its comments and ratings. This will save the user from having to re-read unsuitable materials and thus address this point.

3. **Identifying the quality of materials found on the Web**. Using the recommendation system would go some way in addressing the concern over identifying the quality of academic materials on the Web. Where recommendations have been made, the quality of the material will have

previously been peer-reviewed.

4. **Gaining physical access to materials hosted on the Web**. As the copies of recommended resources are stored by the Server Agent, future researchers would have immediate access to it, via the 'clickable' reading list. This facility would address this point.

5. **Resources on the Web can be less permanent than those in print.** Where possible, DurBA will keep a copy of a found resource in the database. This will solve the problem of ephemeral resources.

6. **Students may go for a 'quick-fix' to their research due to their exposure to 'immediate' technologies.** This is a conditioning and training paradigm. Although it is unlikely that students will move away from the 'quick-fix' mind-set in their social lives, they should be schooled to understand the difference needed in academic work. This is not addressed by DurBA.

7. **Students may consider that the newest material, or even those most attractively displayed, are not always the most relevant.** This is similar to Point 6, and is not addressed by DurBA.

4.6 Areas of Research

The DurBA model has a number of key areas of research:

1. Autonomous Web Searching. The degree and methods which the Server Agent uses to actively seek out similar materials is an important factor. Where the Server Agent knows the Research Domain for a user, it has the opportunity to browse the Web itself, much like a spider, to seek out relevant materials for that user.

2. Collaboration. By combining the research results from users, it will be possible, over a period of time, to create a core of 'working materials' highly

suited to subject areas.

3. Textual Comparison System. One of the most important 'hidden' aspects of the system is the mechanism which the system uses to identify similar materials, namely those related to the Research Domain. Given a set of Web pages, or documents viewed by the user, there must be a method by which the computer can identify similar materials in order to make further suggestions to the user. The most common method is in comparing keywords or text selections. This can be done using a number of algorithms as discussed in Chapter 3.

4. Keyword Selection. The rationale behind human keyword selection is a research topic in its own right; how people associate words with meanings and understandings varies.

5. Recommendation System. This is the mechanism behind how human recommendations are handled by a software system. The success of Recommendation System lies in how they are accepted by users. The graphical user interface is probably what the user relates to most with regard to the Recommendation System as it has the most impact on them. However, how the Recommendation System operates is also important. If using the Recommendation System is an unpleasant experience, it will be unpopular no matter how effective it is behind the scenes. User input will be kept to a minimum, be unobtrusive, and should not use up all of the user's computer's processing power.

6. Recommendation Rewards. How Recommendations are monitored and rewarded are important motivational aspects of a Recommendation System. Just giving the user an attractive GUI will not guarantee success; a recommender System needs to be fluid, reliable, understandable, and rewarding for it to be viable.

4.7 DurTeC: The Textual Comparison System

The DurBA model highlighted a number of areas of research as detailed in Section 5.6. The subject of Textual Comparison Systems was chosen for further research as it was considered closely relevant to the domain of academic research on the Web. A sub-system will be developed in order that it can be studied further. This system will be called DurTeC, standing for Durham Textual Comparison system.

The DurTeC system was designed as a 'plug-in' sub-system to DurBA; any improvements which could be made to it would percolate through from one to the other.

The DurTeC system at its simplest operates as follows; given a Source text or texts, and a set of Comparison texts, it will highlight from the Comparison texts, those which most closely match the Source. In effect, the system will be able to produce a Reading List from a set of Comparison texts, from a corpus of texts. The system would be built in such a way that it would allow different algorithms to be applied to so that their effects could be measured.

4.7.1 Overview

Description

The system will not be concerned with notional areas such as the graphical user interface, communications, recommendation system and architecture, and it will abstract away concepts such as the collection and delivery of information to and from the student. It will concentrate on the comparison process and the concept of being able to apply different algorithms to measure effect. Ergo, given a large set of academic articles, such as could be found on the Web, it will be able to identify those most closely matching the student's Area of Research.

It is possible to visualise the implementation of such a plug-in system into a larger system, such as DurBA which would be of use in the scenario example in Section 4.3.

The system includes a number of assumptions:

1. Area of Research. To simulate a user's Area of Research, the system is supplied with:

    a. Search -words to simulate words as if entered in a Search-engine.

    and/or:

    b. One or more documents containing text to simulate documents and Web pages selected by a user as relevant to their Area of Research

2. Material Repository. To simulate a repository of materials such as can be found on the Web, the system makes use of sets of materials as supplied by the user. In the place of individual Web pages, PDF documents, or word processing documents, the system will have supplied to it a finite library represented in text format.

Firstly, let us take a look at a how the DurTeC system would fit into the DurBA model.

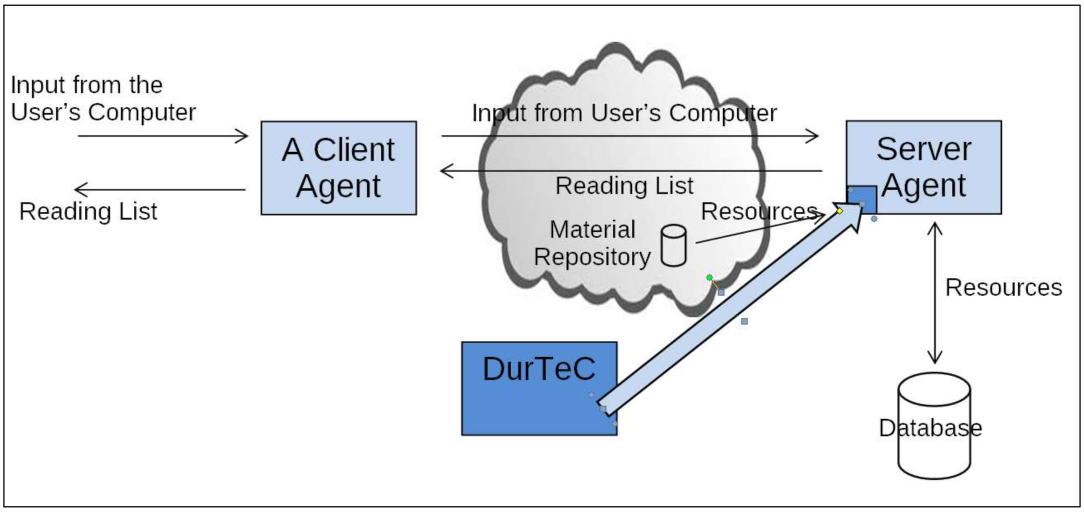

Figure 4.8 How DurTec sits in the DurBA Model

As can be seen, DurTeC would plug-in to the Server Agent. It would need to be activated by the DurTeC system, being passed several inputs, and returning a Reading List to DurTeC. In this implementation, this communication process is represented by supplying DurTeC with a number of inputs, and collecting outputs.

The following is a high-level view of DurTeC:

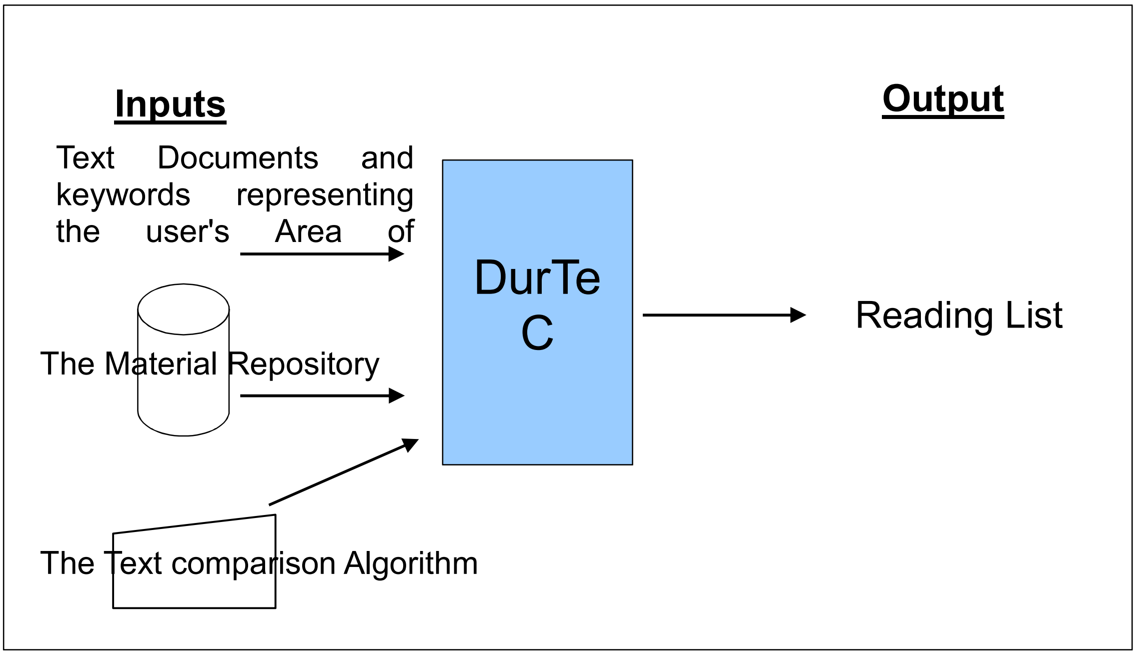

Figure 4.9 A High Level View of DurTeC

DurTeC takes a number of inputs:

1. Text documents and search-words to represent an Area of Research which shall be used with the given algorithm, to create the Research Domain.
2. The Material Repository.
3. The Textual Comparison Algorithm. DurTeC will allow different algorithms to be applied in the textual comparison process in an attempt to improve its accuracy.

It produces only one output:

1. The Reading List. The Reading List is the set of materials which the system identifies as those most suitable from the repository which most closely match the Research Domain.

Phase 1: Select the Source Materials to Represent the Area of Research

For DurTeC to have an understanding of the user's Area of Research, one or more pieces of Source Material and/or Search-words may be submitted to it which are relevant to that Area of Research. Source Materials are submitted in the form of text documents held in a selected folder on the computer, and Search-words are submitted in the form of a string of comma separated words. An example of the contents of one such Source Material text file would be as follows:

<Title> Empirical studies of agile software development: A systematic review, Information and Software Technology

<Author>Dyba, T., Dingsøyr, T.

<Year>2007

<Journal> Information and Software Technology, 50(9-10), Aug.

<Abstract>Agile software development represents a major departure from traditional, plan-based approaches to software engineering. A systematic review of empirical studies of agile software development up to and including 2005 was conducted. The search strategy identified 1996 studies, of which 36 were identified as empirical studies. The studies were grouped into four themes: introduction and adoption, human and social factors, perceptions on agile methods, and comparative studies. The review investigates what is currently known about the benefits and limitations of, and the strength of evidence for, agile methods. Implications for research and practice are presented. The main implication for research is a need for more and better empirical studies of agile software development within a common research agenda. For the industrial readership, the review provides a map of findings, according to topic, that can be compared for relevance to their own settings and situations. However, these skilled and motivated Individuals with high working moral can exhibit high productivity regardless of the methods used, if they are not overly constrained by bureaucracy.

<Keywords>Empirical software engineering; Evidence-based software engineering; Systematic review; Research synthesis; Agile software development; XP;

Extreme programming; Scrum

<Pages>27

<Corpus>...

Figure 4.10 An Example of Source Materials

As can be seen, the text has been formatted into a machine-readable form.

Where multiple Source Materials are submitted to DurTeC, they are combined during the following phase into a single file.

Phase 2: Create the Source Resource Representation File (RRF)

The next stage of the process is for DurTeC to create the Source RRF from the Source Material(s) and/or Search-words submitted to it in the above phase. This is done by submitting the Source Material(s) to a process describe in *Concept: The Resource Representation File Creation Process* following Phase 9 below.

Phase 3: Adjusting the Source RRF to More Accurately Represent the Area of Research

Once the Source RRF has been created, it can be viewed by the user. DurTeC allows the Source RRF to be fine-tuned by adding and removing words. For the text in this example, the following figure represents the Source RRF created from the above text. In the figure, it can be seen that each word has an associated occurrence count, and the percentage it occupies of the rationalised text.

Word	Count	% of Text
study	6	8.11
agile	5	6.76
software	4	5.41
development	3	4.05
review	3	4.05
empirical	3	4.05
research	3	4.05
identified	2	2.70
method	2	2.70

Figure 4.11 An Example of a Source RRF

The DurTeC system allows the user to alter these associated values in a number of ways. For example, the user may only wish to use only the first ten words, as all of the preceding words only have one occurrence each. Alternatively, the user may wish to make some of the words have the same 'percentage of text' value.

Another option available is to alter the percentage of the individual word. For example, the user may wish to change the value associate with the word 'review' from 4.05 to 7.12. This action would place it at the top of the list, thus it becoming the most influential word in the Source RRF.

Phase 4: Select the Material Repository Files

If the Source RRF represents the Area of Research, the Material Repository represents the materials it is to be compared against. In practice, DurBA might use the Web or a database as its repository. In the case of DurTeC, it is pointed towards a set of files to represent those materials.

Phase 5: Create the Comparator Resource Representation Files (RRF)

As with the Source RRF, each item in the Material Repository File has a RRF created to represent it using the same RRF Creation Process as described in *Concept: The Resource Representation File Creation Process* described after Phase 9 below. This allows DurTeC to compare the Source RRF with each Comparator RRF individually.

Phase 6: Algorithm Selection

The system will allow the user to choose which textual comparison algorithm they wish to employ for the comparison process. The default algorithm in DurTeC, and the one used through the experiments in this research is TF/IDF.

As part of this algorithm selection, DurTeC also allows a number of variables to be set such as adding additional source materials and apply percentage rules as described in Phase 3. There are also more details about these, following in the

Concept: Algorithms.

Phase 7: Execution

The next phase is to execute the comparison process. DurTeC uses the selected algorithm to create a metric to represent the similarity between two texts.

As each Comparator RRF is compared against the Source RRF, DurTeC compares those words which are present in both the RRFs. Where words are present in both files, the percentage figures that word occupied in both papers are added together to give a Similarity Metric to represent that word; this metric is in turn, added to a grand total representing the relativity between the two papers.

To illustrate this process, consider the two partial RRFs following.

Source RRF

Word	Count	% of Text
study	6	8.11
agile	5	6.76
software	4	5.14
development	3	4.05
review	3	4.05
…		

Comparator RRF

Word	Count	% of Text
software	4	6.82
development	4	6.82
life-cycle	3	5.16
programming	2	3.41
process	2	3.41
…		

Figure 4.12 An Example of a Source and Comparator RRF

DurTeC found that both the words 'software' and 'development' are present in both of the two partial RRFs.

The word 'Software' would be allocated a Similarity Metric of 12.23 because it occupies 5.41% of Dyba's paper, and 6.72% of the compared paper:

Calculation	Metric
5.41 + 6.72	12.23

The second word 'development', would be allocated a Similarity Metric of 10.77 using the same process:

Calculation	Metric
4.05 + 6.72	10.77

These two 'Metric' figures would be added together to give a total Similarity Metric of 23.10, to represent the similarity between the two RRFs. This figure would be used to represent the Comparator paper against all the others in the Material Repository.

As this was only a partial illustration, there were only two matching words, in reality, there would be many more matching words giving a metric much higher than 23.10.

Phase 8: Reporting

After executing DurTeC, a report in a spread sheet format will be produced listing all the papers from the Material Repository with their associated metrics. These metrics indicate how closely they each match the original Area of Research, the higher the figure, the better the match. The following table is the resulting report from this example.

Paper #	Paper	Metric
15	15 Applying Systematic Reviews to Diverse Study Types: An Experience Report	93.74
22	22 Changing the Paradigm of Software Engineering	79.05
20	20 The Future of Empirical Methods in Software Engineering Research	75.23
…		

Figure 4.13 A DurTeC Report Sample

Phase 9: Results

The resulting reports can be used to consider the results of the DurTeC system processing. The effects of the algorithm selection may be observed.

Concept: The Resource Representation File Creation Process

The creation of a RRF follows an established process called Term Frequency. Term Frequency [SALT77] is a popular method used in information retrieval and text mining to evaluate how important a word is to a piece of text. The importance increases proportionally to the number of times a word appears in the document but is offset by the size of the corpus. The result of using term frequency is the creation of a list of the words present in that text together with its occurrence count and the percentage of the text it occupies.

Before this process can take place, the selected text must be rationalised to adjust it to better represent the chosen Area of Research.

Rationalisation Process

As an example, consider the following text as a resource submitted to represent the Area of Research:

Agile software development represents a major departure from traditional, plan-based approaches to software engineering. A systematic review of empirical studies of agile software development up to and including 2005 was conducted. The search strategy identified 1996 studies, of which 36 were identified as empirical studies. The studies were grouped into four themes: introduction and adoption, human and social factors, perceptions on agile methods, and comparative studies. The review investigates what is currently known about the benefits and limitations of, and the strength of evidence for, agile methods. Implications for research and practice are presented. The main implication for research is a need for more and better empirical studies of agile software development within a common research agenda. For the industrial readership, the review provides a map of findings, according to topic, that can be compared for relevance to their own settings and situations. However, these skilled and motivated Individuals with high working moral can exhibit high productivity regardless of the methods used, if they are not overly constrained by bureaucracy.

The first aspect to notice about this text is the quantity of common words such as ‘the’ and ‘for’. Many would consider these words as superfluous to the content of the text. Indeed if only the four most commonly occurring words in the text were used to represent it, they would not be a great representation of the text.

Word	**Occurrence Count**
The7	
For	6
Studies	6
Agile	5
...	...

Figure 4.14 Occurrence Counts

As can be seen from the top two words in the above table, the text needs to be cleaned up, or *rationalised* before it can be used.

One approach to this is in the removal from the text, a set of words classed as 'stop-words'. Stop-words are words which are classed as unimportant, or unnecessary within a text; words such as 'it', 'and', and 'the' are all common stop-words. Stop-word lists are used in computing where automation is employed to filter out prior to, or after, processing of natural language text. Although there is not one definite list of stop-words, there are a number of them recognised as reliable. In the case of this research, a stop-word list from the University of Glasgow was employed; it can be found at:

http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

The stop-word list can be utilised before or after the creation of the word frequency list. In the case of DurTeC, the process is done before.

Other pre-processing activities which were employed included:

1. Removing punctuation marks.
2. Removing numbers.
3. De-pluralising phrases, e.g. the word ‘themes’ being altered to ‘theme’, ‘books’ to ‘book’ etc.
4. Adjusting tense such as changing ‘reviewing’ or 'reviewed' to ‘review’.

These actions reduced the size of the text from 170 words to just 74; a reduction of approximately 57%. The results of these actions can be seen following.

agile software development represent major departure traditional plan based approach **software** engineering systematic review empirical **study agile software development** including conducted search strategy identified **study** identified empirical **study study** grouped theme introduction adoption human social factor perception **agile** method comparative **study** review investigates benefit limitation strength evidence **agile** method implication research practice presented main implication research empirical **study agile software development** common research agenda industrial readership review map finding topic compare relevance setting situation

Figure 4.15 Text Reduction through Pre-processing

The four most commonly occurring words above are highlighted in bold. Using this text, a resulting word-frequency List is created similar to that found following

Word	Occurrence Count
Study	6
Agile	5
Software	4

Development	3
…	...

It can now be seen how better the text can be represented by the top four words in the list; study, agile, software and development.

This word frequency list now forms the RRF to text it began with. This process is the same for the creation of both Source and Comparator RRFs.

Concept: Collaboration

As the DurBA model aimed to incorporate collaboration into it, so the DurTeC system should also be created with this in mind. To simulate collaboration between users of the system, a number of concepts were included in its design:

1. DurTeC would allow multiple text documents to be submitted to it (see Phase 1 above) to simulate multiple recommendations from users. Each submitted text document would affect the created Research Domain.
2. Search-words could be added to strengthen the Research Domain to simulate when users of DurBA submitted words to Search-engines.
3. The value of words within the Research Domain could be increased or decreased to adjust their effect on the comparison process. This was included to simulate the fluid situation of the Research Domain in DurBA as it may change over time as users submit new materials and it refine the Research Domain accordingly.

Concept: Algorithms

DurTeC allows a choice of algorithm to be applied during its comparison processing for two reasons. Firstly, it was to simulate variables which users could change within DurBA, and secondly, to see what effects they had on results.

Although the basic comparison method, namely TF/IDF remained the same throughout, variables could be applied to create different algorithms. These variables are as listed following:

1. There were two synonym lists available in the experiment. Synonym lists could be used to expand the vocabulary of the DurTeC system. For example, if the word 'study' were to be found in the Source text, the word 'examine' as well as the word 'study' may be used in the comparison process.

    a. The Default Synonym List (**DSL**). This was a Standard English language synonym list of about 140,000 words.

    b. The CNSynonym List (**CNS**). This was a highly defined domain orientated synonym list based on the most popular words found in the subject Software Engineering. This consisted of about 70 words.

2. The Percentage Rule (**PR(X)**). The percentage that a word occupied in the given text was an important factor in the process. DurTeC could be set to only consider those words which occupied above a given figure, e.g. 1%. The words occupying less that that percentage of the given text would be ignored. The X represents the percentage rule applied.

3. A User-Defined Word List (**DWL**). A number of comma-delimited words could be fed into DurTeC to augment the system's understanding of the Research Domain.

4. Secondary Sources (**SS(X)**). DurTeC was originally only to consider one given Source text in order to create a Research Domain to represent the user’s Area of Research. This could be expanded to allow up to 6 documents to be supplied as Source texts. The X represents the number of Source texts used.

4.8 A Small Worked Example

In order to confirm that the results gained from DurTeC were correct, a worked example with limited data was employed. The figures created manually from this worked example would be compared with those from the DurTeC system using the same data.

Employing only a Source RRF Reduced in size to only 5 words, and four Comparator RRFs representing the material repository, also of only 5 words each, it is possible to undertake the calculations manually. These manually created results would then be compared to what the DurTeC system produced for the same data.

The worked example shall use the same phased process as described previously.

Phase 1: Select the Source Materials to Represent the Area of Research

No materials would be used to represent the Area of Research in the worked example. The following phase is used to create the Source RRF manually.

Phase 2: Create the Source Resource Representation File (RRF)

The Source RRF was created manually with the following data.

Word	Count	% of text
study	6	27.57
agile	5	23.71
software	4	19.05
development	3	14.29
review	3	14.29

Figure 4.16 The Source RRF

Phase 3: Adjusting the Source RRF to More Accurately Represent the Area of Research

No adjustment of the Source RRF will take place in this example.

Phase 4: Selecting the Material Repository Files

For this Worked Example, there would be no Material Repository as the Comparator RRFs would be created manually.

Phase 5: Create the Comparator Resource Representation Files (RRF)

For this Worked Example, the Comparator RRFs would be created manually to represent the four papers in the Material Repository. The selected data was chosen in order to simplify the example process.

1st RRF		
Word	**Count**	**% of text**
lifecycle	9	26.47
agile	8	23.53
method	8	23.53
software	6	17.65
paradigm	3	8.82

2nd RRF		
Word	**Count**	**% of text**
software	8	23.53
development	8	23.53
agile	6	17.65
system	6	17.65
formal	6	17.65

3rd RRF		
Word	**Count**	**% of text**
java	7	27.00
basic	6	24.00
development	5	20.00
python	4	16.00
language	3	12.00

4th RRF		
Word	**Count**	**% of text**
apples	7	23.33
oranges	7	23.33
pears	6	20.00
grapes	6	20.00
peaches	4	13.33

Figure 4.17 Comparing the RRFs

Phase 6: Algorithm Selection

In this worked example, a simple term-frequency algorithm is be used. No other variables would be set.

Phase 7: Execution

Each Comparator RRF from the Material Repository would be compared against the Source RRF to test for how closely they matched.

Each comparison shall now be examined in turn.

Comparing the Source RRF with the 1st Comparator RRF

The words that were present in both the Source RRF and the 1st Comparator RRF were ‘software’ and 'agile'.

Source RRF word	% of text		1st Comparator RRF word	% of text		Calculation	Metric
agile	23.71		agile	23.53		23.71+23.53	47.34
software	19.05		software	17.65		19.05+17.65	36.70
						Total	84.04

The total Similarity Metric when comparing the two RRFs is 74.04.

Comparing the Source RRF with the 2nd Comparator RRF

The words that were present in both the Source RRF and the 2nd Comparator

RRF were 'agile', 'software' and 'development'.

Source RRF word	% of text
agile	23.71
software	19.05
development	14.29

2nd Comparator RRF word	% of text
agile	17.65
software	23.53
development	23.53

Calculation	Metric
23.71+17.65	41.46
19.05+23.53	42.57
14.29+23.53	37.72
Total	121.76

The total Similarity Metric when comparing the two RRFs is 121.76.

Comparing the Source RRF with the 3rd Comparator RRF

The only word that was present in both the Source RRF and the 3rd Comparator RRF was 'development'.

Source RRF word	% of text
development	14.29

3rd Comparator RRF word	% of text
development	20.00

Calculation	Metric
20.00+14.29	34.29
Total	34.29

The total Similarity Metric when comparing the two RRFs is 34.29.

Comparing the Source RRF with the 4th Comparator RRF

There were no words present in both the Source RRF and the 4th Comparator RRF.

Source RRF word	% of text

4th Comparator RRF Word	% of text

Calculation	Metric
Total	0.00

The total Similarity Metric when comparing the two RRFs is 0.00.

Totalling up the Similarity Metric

The four manually calculated Similarity Metrics were used to create the following table.

RRF #	Metric
1	84.04
2	121.86
3	34.29
4	0.00

Figure 4.18 A Summary of the Similarity Metrics

Phase 8: Reporting

The results from DurTeC when using the above data was as appeared as in the table below.

File #	Metric
1	84.04
2	121.86
3	34.29
4	0.00

Figure 4.19 DurTeCs Similarity Metrics

Phase 9: Results

The data returned from the DurTeC processing was the same as that in the manually calculated one, confirming the process as correct.

4.8.1 Alternative Algorithms

Further to this confirmation process, the same data as used in Section 4.7 was used with two other comparison algorithms to demonstrate DurTeC's correctness.

A report resulting from using a 'Log-likelihood' algorithm for textual comparison can be found in Appendix 2.

A report resulting from using a 'Degrees of Separation' algorithm for textual comparison can be found in Appendix 1.

4.9 A Worked Example

To confirm that the DurTeC system actually processed correctly under experiment conditions, a dry run was created using the eight phases as highlighted in Section 4.6. To allow a better understanding of this process, a finite selection of materials will be used as the Material Repository for the Worked Example, as this will help clarify the procedure and allow the identity of individual pieces of materials easier. For the same reasons, the materials used to represent the Area of Research shall be limited to one piece.

In selecting suitable materials, a novel approach is used. An academic paper to represent the Area of Research has been selected. The references found within this paper will be used to represent the Materials Repository. By doing so, it will be possible to keep all the materials to a minimum, and within a highly defined Area of Research. To introduce a level of randomness to the Worked Example, five additional non-related papers are being added to this collection.

For the sake of clarity, this Worked Example bases the Source RRF upon the text found in only the Abstract section of these materials. Again, this is only done for the sake of brevity, and could easily be expanded to include all sections of the text. Also, none of the variables as described above in the *Concepts: Variables* section were incorporated. How well can the system identify those materials as unrelated to the Area of Research?

Phase 1: Select the Source Materials to Represent the Area of Research

For this Worked Example, only one item of Source Material shall be submitted to DurTeC which will be the following:

Dyba, T., Dingsøyr, T., 2007. Empirical studies of agile software development: A systematic review, Information and Software Technology, Volume 50 Issue 9-10, Aug. 2007

No Search-words will be submitted.

Phase 2: Create the Source Resource Representation File (RRF)

The next stage of the process is for DurTeC to create the Source RRF from the selection of Source Materials and Search-words given to it. Only one piece of material is being used in this Worked Example and the Source RRF is created from that.

Phase 3: Adjusting the Source RRF to More Accurately Represent the Area of Research

Once the Source RRF has been created, it can be viewed by the user. DurTeC allows the Source RRF to be fine-tuned by adding and removing words, and by altering their associated values. However, for the purpose of this Worked Example, the words shall remain unchanged.

Phase 4: Select the Material Repository Files

In this Worked Example, the Material Repository consists of a set of 51 academic papers in text format, 5 of which were files unrelated to the Area of Research and were chose as indicator files

It is worth noting that some of these papers did not contain an abstract section. This will reduce the size of the examined repository further as they would be skipped by the DurTeC system.

Phase 5: Create the Comparator Resource Representation Files (RRF)

Each individual textual resource in the Material Repository had an RRF created to represent it.

Phase 6: Algorithm Selection

When DurTeC makes a judgement over whether a Comparator RRF from the Material Repository is closely related to the Source RRF, it uses an algorithm in the process. The system allows an algorithm selection to be made, in the case of this Worked Example, a basic TF/IDF is used. It also allows the application of variables to be applied to that algorithm, but none were applied here.

Phase 7: Execution

The next phase is to execute the comparison process. Each Comparator RRF is compared with the Source RRF using the selected algorithm. A Similarity Metric is created to represent how closely they match. Each item in the Material Repository goes through this process, so each gets a representative Similarity Metric attached to it.

Phase 7: Reporting

After executing DurTeC, a report in a spread sheet format will be produced listing all the papers from the Material Repository with their associated metrics. These metrics indicate how closely they each match the original Area of Research, the higher the figure, the better the match. The following table is the resulting report from this example.

Paper	Metric
18 Applying Systematic Reviews to Diverse Study Types: An Experience Report	93.84
38 Changing the Paradigm of Software Engineering	79.05
43 The Future of Empirical Methods in Software Engineering Research	75.23
12 An Introduction to Agile Methods	74.24
2 New Directions on Agile Methods: A Comparative Analysis	72.88
48 Agile Software Development: Its about Feedback and Change	69.18
28 Preliminary Guidelines for Empirical Research in Software Engineering	60.51
25 Checklists for Software Engineering Case Study Research	57.25
27 Guidelines for performing Systematic Literature Reviews in Software Engineer	57.24
1 Agile software development methods	57.03
47 Assumptions Underlying Agile Software Development Processes	56.76
14 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Differ	53.08
9 Get Ready for Agile Methods with Care	51.80
33 Theoretical Reflections on Agile Development Methodologies	48.36
31 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two	46.31
44 A Survey of Controlled Experiments in Software Engineering	44.86
32 Agile Software Development: Adaptive Systems Principles and Best Practices	44.85
29 Iterative and Incremental Development: A Brief History	36.13
4 Web-Based Agile Software Development	35.48
23 Agile manufacturing: A framework for research and development	34.04
39 Explaining Software Developer Acceptance of Methodologies: A Comparison o	32.70
40 A review of agile manufacturing systems	29.75
22 How To Read A Paper	26.49
37 Lean Software Development: An Agile Toolkit	25.75
8 Extreme Programming Explained 2nd Ed.	23.38
36 A Practical Guide to Feature-Driven Development	21.17
20 Methodological Fit in Management Field Research	20.57
17 Are Two Heads Better than One? On the Effectiveness of Pair Programming	20.07
21 On the Effectiveness of the Test-First Approach to Programming	17.32
51 Millennials and the World of Work: The Impact of Obesity on Health and	**16.67**
52 Effect of Lean System 7 on metabolic rate and body composition	**15.18**
41 Agile Project Management with Scrum	14.79
54 Seville (sour) Orange Juice: Synephrine Content and Cardiovascular Eff	**13.81**
45 Extreme Programming Refactored: The Case Against XP	9.40
16 Improvisation in Small Software Organizations	9.18
53 Experimental Biology and Medicine	**8.62**
50 Prevalence and Trends in Obesity Among US Adults 1999-2008	**2.74**

Figure 4.20 The DurTeCs Report

The resources listed in the above table are ordered in descending order. Those with the higher metrics, at the top, offer the closest match to the Area of Research, and could be used in the creation of a reading list.

Phase 9: Results

The report will be used to consider the results of the DurTeC system processing. In the case of the Worked Example, and referring to the table above, a number of points were highlighted:

1. Seventeen of the papers had no abstract section, and so gained no weighting metric. The process could be run again using other texts to represent the Research Domain such as keywords, or even the whole of the paper text. These would create a different set of results, and may then include these seventeen papers.
2. The five non-related papers can be seen in bold text occupying five of the bottom eight positions.
3. Paper 17 was the paper which most closely matched Dyba's paper.

It is worth noting that there is no intrinsic value to the metrics other than to represent its paper's place in the hierarchy of the list. In other words, taking a single paper from the list in isolation such as the third top one, paper 43 with a metric of 75.23 in no way identifies this paper as a good match. The metric purely signifies that from this repository set, using the chosen algorithm, it is a slightly better match to Dyba's paper than paper 12, but not as good a match as paper 37. It may however still be a poor match.

The benefit of using the abstract as a representation of the papers is that the reasons behind anomalies such as why Papers 51 and 52 received better matches than Paper 41 can be identified. In the case of these two, their heightened position was a result of the strength of presence of the word 'study' within them.

This research does not consider whether the quantity of material given to DurTeC at this point affects the accuracy of the results. Another aspect is that it does not consider how many words used in the list found in the table would yield the best results. It may be that by using only words with more than one occurrence, for example, or words with more than a particular percentage may impact the results.

4.10 Discussion

The purpose of the DurBA model was to provide a means to address the various areas of concern highlighted in Chapter 2. This was illustrated through the use of a scenario. Having described the DurBA model, it became apparent that there were a number of concepts which were of specific interest to this research.

The subject of textual comparison systems was chosen for further research as it was considered closely relevant to the domain of academic research on the Web. From this, a second sub-system was proposed and implemented. This system was called DurTeC, standing for Durham Textual Comparison system.

The processing of the DurTeC system was broken down into 9 phases:

Phase 1: Select the Source Materials to Represent the Area of Research

Phase 2: Create the Source Resource Representation File (RRF)

Phase 3: Adjusting the Source RRF to More Accurately Represent the Area of Research

Phase 4: Select the Material Repository Files

Phase 5: Create the Comparator Resource Representation Files (RRF)

Phase 6: Algorithm Selection

Phase 7: Execution

Phase 7: Reporting

Phase 9: Results

These phases were described in Section 4.7. A small worked example was created Section 4.9, and the process was tested under experiment conditions with limited data in Section 4.10. The results showed that the processing worked as expected.

4.11 Conclusion

This Chapter began by listing a number of areas of concern using the Web for researching. It then went on to consider a scenario where two students studying similar topics, could without their direct knowledge make their combined research more effective by sharing research. After that, a system called DurBA was proposed which would make such a scenario plausible. It identified the architecture and the concepts which would be needed for it to work.

Following on from this, some specific areas of research were highlighted. It was decided that this research would concentrate on just one of these areas, progressing the understanding within it. That area would be the paradigm of the Textual Comparison System. This is the process by which two pieces of text are compared and given a metric to identify their level of similarity.

To further this research, a Textual Comparison System named DurTeC was proposed. The DurTeC system would be viewed not as a complete system, but as a sub-system or plug-in which could be used as part of a larger system such as part of DurBA. The DurTeC system was described in detail.

The Chapter then went on to present a Worked Example of the system processes, followed by a dry-run to confirm that it was producing valid results.

Chapter 5 The Design and Development of the Tool

The concept behind the DurTeC system was that it would be able to identify those documents from a set which most closely matched the requirements of the user. The user could supply one or more 'Source' articles to represent their Area of Research (known here as the Research Domain), and DurTeC would compare it against a corpus of 'Comparator' articles representing a Material Repository such as the Web. DurTeC would iterate through every article in the corpus comparing it with the Research Domain, and identify those which most closely match, and those which least closely match.

DurTeC could then produce a Reading List from the most suitable articles.

5.1 System Architecture

The system was executed as a single application with no need of Web access as all the data was available locally.

Source article(s) would be located within a specific file on the computer.

Comparator article(s) would be located within a specific file on the computer.

All options within the system were available from the Main screen.

5.2 Assumptions

The DurTeC system would make a number of assumptions:

1. The system would be a plug-in. DurTeC would not be a stand-alone system and would be designed with this in mind. Therefore:
    a. Constraints of the network and operating systems would not be considered.
    b. Any user management and security issues would not be

considered.

2. As it was a system designed for testing purposes, it would not be reliant upon on data quality and format such as data held on the Web.

5.3 Design

The design of the system would be as was described in Section 4.5 DurTeC – A Textual Comparison System. The image following is taken from that Section.

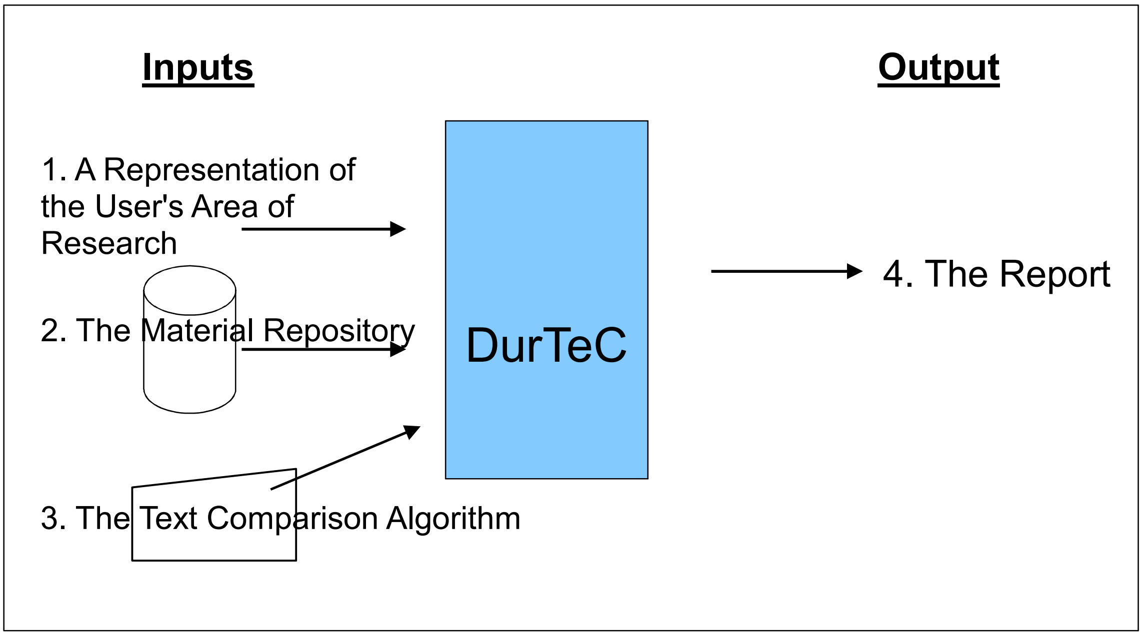

Figure 5.1 The DurTeC Overview

5.4 Inputs

DurTeC takes a number of inputs:

1. A Representation of the User's Area of Research; the Research Domain.

The user would be able to submit an individual, or a number of text documents to DurTeC to represent their Area of Research. These submitted documents were used by DurTeC to create a Domain Definition File.

2. The Material Repository. This would represent a collection of Comparator materials such as those that could be found on the Web. The Material Repository would take the form of a file containing a set of textual documents, each representing an individual Comparator item. These documents were formatted in XML so that DurTeC could understand the different parts of the document such as Abstract and Title. Either all or only a section of the document can be considered by DurTeC during processing.

3. The Text Comparison Algorithm. DurTeC would allow a number of algorithms be applied to the comparison process. These would be used to consider which offered the better potential.

5.5 Output

DurTeC produces one output:

1. The Report. The report will provide a list of every document in the Material Repository which was compared against the Research Domain. Each item in the list would have a metric created by DurTeC to represent how similar it judged it to be to the Research Domain.

5.6 Other Variants

DurTeC would allow a number of variables with the system.

1. The number of words considered. It is possible to limit the quantity of words that DurTeC uses from the Research Domain File during its

calculations. This can be done in two ways:

a. By entering a value in the 'Term Quantity' box on the Main Menu screen.

b. By editing the words in the Domain Definition File.

2. Domain Definition File term weightings. It is possible for the user to alter the values associated with each term in the Domain Definition File.

3. Use of a Thesaurus. Synonyms may be used to complement the Domain Definition File. Where this happens, additional relative words are added (with the same weighting value as the original) to the Domain Definition File.

4. The Domain Definition File can be saved and reused. This feature also allows new 'Source' articles to be added to an existing Domain Definition to ‘enrich’ it.

5.7 Screen Shots

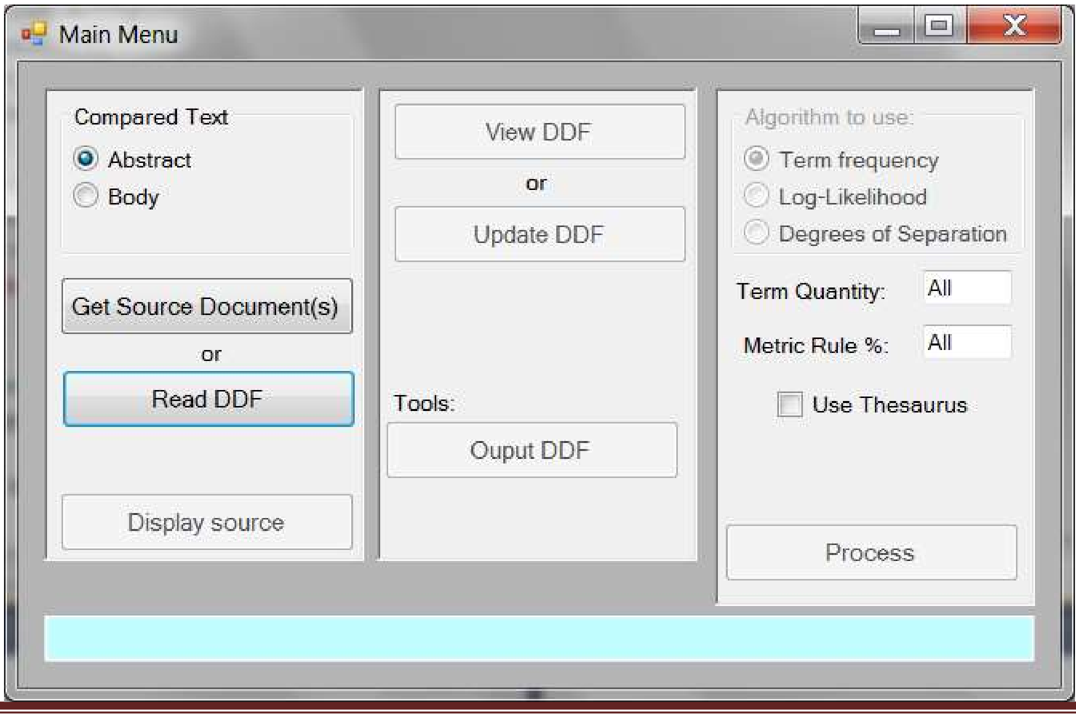

Figure 5.2 Main Menu Screenshot

The following images are of screen-shots taken of the DurTeC system.

In the Main Menu Screen, the user can:

1. Select Source documents.
2. Read in a previously saved Domain Definition File.
3. View the Source documents.
4. View the Domain Definition File. This can be used for editing also.
5. Update the saved Domain Definition File.
6. Select which algorithm should be used during the process.
7. Select the quantities of terms from the Domain Definition File which DurTeC should consider.
7. Select the minimum weighting value that words should hold for them to be included within the calculation
9. Select whether synonyms should be added to the Domain Definition File.
10. Start the processing.

When the user clicks the 'Get Source Document(s)' button, the screen following is displayed.

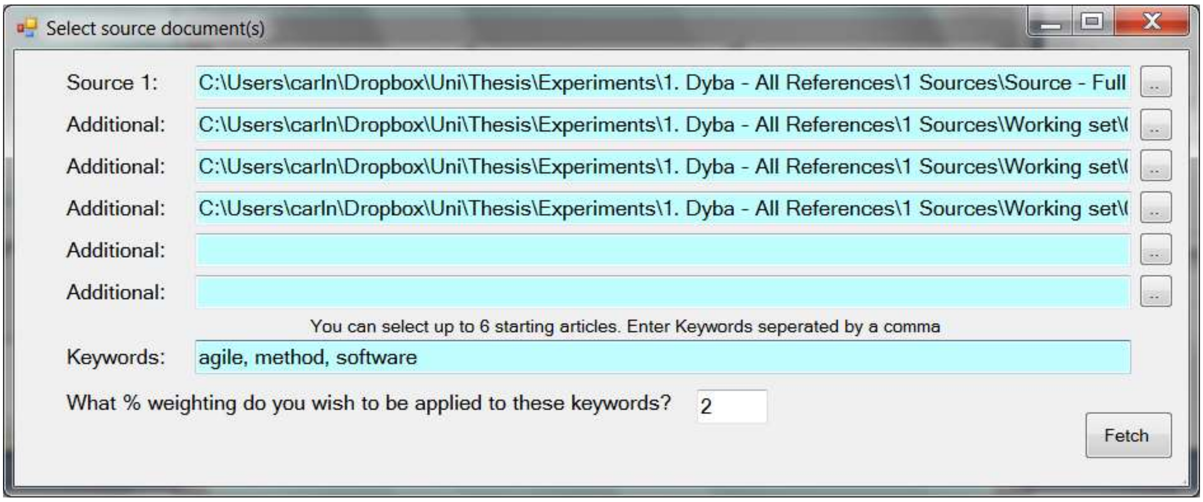

Figure 5.3 Source Document Selection Screen

In the above displayed screen, the user can:

1. Select one main Source document.
2. Select up to five additional Source documents.
3. Add a comma separated list of search-words.
4. Set weighting levels to words in the identified word-frequency list.

These options are used to create the Domain Definition File. After this has been created, it can be viewed using the 'View DDF' button in the main screen.

The Domain Definition File is presented on-screen as shown following:

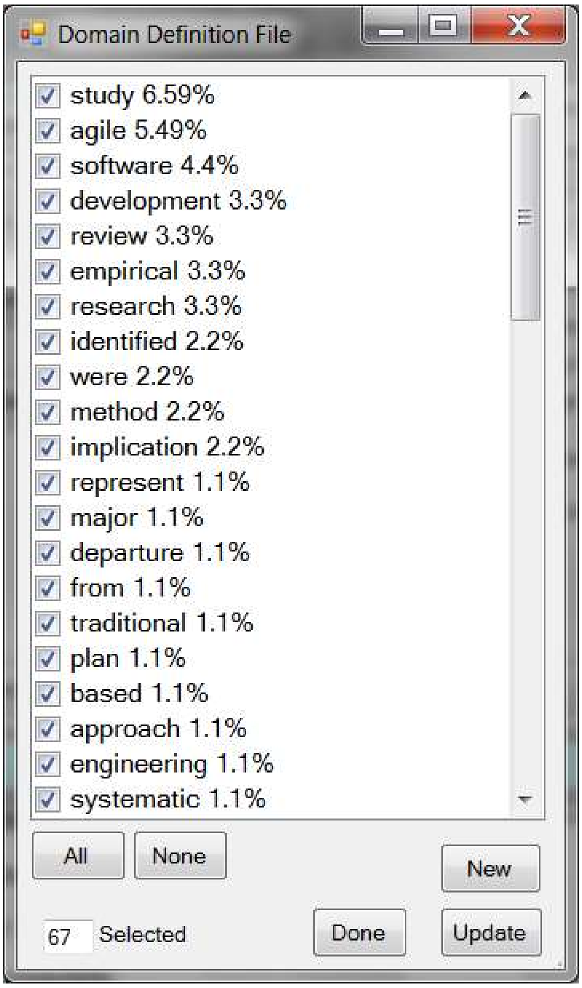

Figure 5.4 DurTeCs Domain Definition File

In this screen, the user can:

1. Remove those unwanted terms in the file thought inappropriate.
2. Edit individual metrics associated with each term. Where the user thinks a term is incorrectly rated, it can be given a different metric to move to a different position in the list.
3. Save or cancel the changes made.

The process is executed by the user clicking the 'Process' button on the main menu screen once all the required parameters have been set.

Paper #	Paper	Metric
15	15 Applying Systematic Reviews to Diverse Study Types: An Experience Report	66.99
2	2 An Introduction to Agile Methods	64.84
27	27 New Directions on Agile Methods: A Comparative Analysis	61.51
20	20 The Future of Empirical Methods in Software Engineering Research	55.09
4	4 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplin	53.08
1	1 Agile software development methods	52.11
29	29 Preliminary Guidelines for Empirical Research in Software Engineering	51.42
7	7 Agile Software Development: Its about Feedback and Change	46.44
22	22 Changing the Paradigm of Software Engineering	45.01
3	3 Get Ready for Agile Methods with Care	44.18
9	9 Assumptions Underlying Agile Software Development Processes	41.40
12	12 Web-Based Agile Software Development	40.24
26	26 Checklists for Software Engineering Case Study Research	39.29
25	25 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	36.80
10	10 Theoretical Reflections on Agile Development Methodologies	36.12
6	6 Agile Software Development: Adaptive Systems Principles and Best Practices	35.37
28	28 Guidelines for performing Systematic Literature Reviews in Software Engineering	29.83
24	24 Agile manufacturing: A framework for research and development	28.27
19	19 A Survey of Controlled Experiments in Software Engineering	25.42
8	8 A review of agile manufacturing systems	24.44
21	21 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theo	23.69
16	16 Methodological Fit in Management Field Research	20.57
32	32 Lean Software Development: An Agile Toolkit	19.19
30	30 Iterative and Incremental Development: A Brief History	19.11
5	5 Extreme Programming Explained 2nd Ed.	16.48
36	36 Effect of Lean System 7 on metabolic rate and body composition	15.18
33	33 A Practical Guide to Feature-Driven Development	13.14
14	14 Are Two Heads Better than One? On the Effectiveness of Pair Programming	12.78
38	38 Seville (sour) Orange Juice: Synephrine Content and Cardiovascular Effects in Normoten	10.33
13	13 Improvisation in Small Software Organizations	9.18
31	31 Agile Project Management with Scrum	8.44
17	17 On the Effectiveness of the Test-First Approach to Programming	8.01
23	23 How To Read A Paper	4.70
39	39 Millennials and the World of Work: The Impact of Obesity on Health and Productivity	4.31
11	11 Selecting a Projects Methodology	0.00
18	18 Extreme Programming Refactored: The Case Against XP	0.00
34	34 Challenges of Migrating to Agile Methodologies	0.00
35	35 Prevalence and Trends in Obesity Among US Adults	0.00
37	37 Experimental Biology and Medicine	0.00

Figure 5.5 The Result of DurTeC's Processing

The result of the processing is the automatic creation of a spread sheet report similar to that found in the table following:

5.8 Summary

Chapter 5 has continued on from Chapter 4 where DurTeC, the proposed textual comparison tool was described. This chapter introduced the specifications and assumptions made about this tool, and went on to list both the inputs and output to and from the tool.

It went on to offer a number of screen shots illustrating execution of the tool. It showed how the tool could be affected by variables, and by way of an illustration, showed what output could be expected in the form of a spread sheet report.

Chapter 6 The Two Experiments and their Results

This chapter employs the tool in two experiments to develop and examine the viability of a selection of algorithms.

6.1 Experiment 1

6.1.1 Overview

This experiment would take the following format:

Aim	To fine-tune the algorithm within DurTeC to increase comparison results. To examine whether the DurTeC can produce reliable results.
Hypothesis	DurTeC's processing algorithm could be developed to significantly improve its capability in identifying relevant documents given a Research Domain.
Variables and Sets	Control variables: DurTeC comparison algorithm. Response variables: The difference between DurTeC's opinion and the Researcher's opinion. Control group: The Researcher's rankings. Experimental group: DurTeC's rankings. Participants: 2
Inputs	39 Structured Abstracts representing a corpus of academic material.
Procedure	DurTeC would iteratively make a comparison using differing algorithm.
Format of Results	Metrics.

Figure 6.1 Experiment 1 Overview

6.1.2 Aim

The aim of the experiment was to improve how successfully the DurTeC system could compare a given Research Domain with a group of documents, and identify from that group, those documents which most closely match it.

The benefit of being able to undertake this process reliably lies in its potential in Web-based systems. In the scenario in Chapter 1, the students did not waste a great deal of their time searching for information on the Web. Through an automated process, they were presented with pertinent research materials. By using a successful comparison algorithm, a computer program could be employed to replicate this process to locate other materials on the Web pertinent to the research article, thus saving the student time. The computer system could also facilitate the sharing of those materials identified by both the student and the system to other students.

The experiment would concentrate on highlighting the algorithms which were most successful in identifying those articles most similar, and most dissimilar to the Research Domain. Therefore, the success of the algorithm was key to the comparison process.

For this to be quantified, the results from the DurTeC system would be compared with those from a human, in this case, the Researcher.

In order that comparisons could be made, DurTeC allowed for numerous algorithms to be applied in the comparison process. By doing this, these most successful could be identified. The algorithm used for the comparison process are described in 6.1.5 following.

6.1.3 Hypothesis

Figure 6.1 lists the hypotheses for this experiment.

Hypothesis ID	Hypothesis
H1	The basic processing algorithm can be manipulated to show significant improvement in its ability to achieve H2 and H3.
H2	The DurTeC system can identify more than 50% of the Researcher's Top 10 abstracts.
H3	The DurTeC system can identify more than 100% of the unrelated abstracts.

Figure 6.2 The Hypotheses for Experiment 1

6.1.4 Variables and Sets

Control variables. The control variables in this experiment would be those areas of the comparison process which could be manipulated in order to affect the results, see 6.1.5 for further information. These included:

1. Synonym lists.
2. Use of a percentage rule.
3. User defined word list.
4. Use of secondary Sources.

Response variables: These would be the metrics representing the difference between DurTeC's opinion and the Researcher's opinion.

Control group: The control group would be the set of Researcher's rankings.

Experimental group: This would be the DurTeC's rankings.

Participants: 2

6.1.5 Inputs

For this experiment, Structured Abstracts would be used. Where a Structured Abstract was not readily available, one was created by the Researcher using the abstract section from that article. The Source article was the same as in the previous Pilot Studies 1 and 2 (Dyba), and a sub-set of 34 articles referenced within that document were used. The Structured Abstracts would not contain the title of the article, and so, each would be allocated an identifier number.

The Experimental Group a ranked list, created by DurTeC, of its comparison results based on the applied algorithms.

The Control Group was a list, in the Researcher's opinion, were the ten most comparable Structured Abstracts from the set of 34. These were allocated identifier numbers from 1 to 10. These can be found in following:

Paper #	Title
1	Agile software development methods
2	An Introduction to Agile Methods
3	Get Ready for Agile Methods with Care
4	Toward a Conceptual Framework of Agile Methods: A Study of Agility in
5	Extreme Programming Explained 2nd Ed.
6	Agile Software Development: Adaptive Systems Principles and Best Prac
7	Agile Software Development: Its about Feedback and Change
8	A review of agile manufacturing systems
9	Assumptions Underlying Agile Software Development Processes
10	Theoretical Reflections on Agile Development Methodologies

Figure 6.3 The Control Group

To enhance the experiment, 5 completely unrelated Structured Abstracts were added. These 5 were given the numbers 34 to 39 and so brought the total of the Comparator Structured Abstracts to 39.

This selection of abstracts would allow DurTeC to be tested against:

1. How well it could identify those top 10 selected by the Researcher.
2. How well it could identify the bottom set of 5 unrelated abstracts.

Ideally, it would position them accordingly in its results.

Algorithm Variables

The DurTeC system processing algorithm uses a basic Term Frequency/Inverse Document Frequency mechanism. In addition to this, a set of variables can be applied to simulate user input and collaboration. These variables may be applied independently or simultaneously, and they provide DurTeC with further processing algorithms. The additional variable applied in the experiments within this thesis are listed following:

1. The Default Synonym List (**DSL**). This was a Standard English language synonym list of about 140,000 words. Where the DSL is used, words found in the comparison process are augmented with those matching synonyms from the list.
2. The CNSynonym List (**CNS**). This was a highly defined domain orientated synonym list based on the most popular words found in the subject Software Engineering. This consisted of about 70 words. As above, synonyms are used to augment the text in the comparison process.
3. The Percentage Rule (**PR(X)**). The percentage that a word occupied in the given text was an important factor in the process. DurTeC could be set to

only consider those words which occupied above a given figure, e.g. 1%. The words occupying less that that percentage of the given text would be ignored. The X represents the percentage rule applied, e.g. 1%

4. A User-Defined Word List (**DWL**). A number of comma-delimited words could be fed into DurTeC to augment the system's understanding of the Research Domain. This can be considered much like when a user enters words into a search engine.

5. Secondary Sources (**SS(X)**). DurTeC, at its most basic only considered one given Source text in order to create a Research Domain to represent the user's Area of Research. The system would also allow up to 6 other documents to be supplied to reinforce this understanding of the Research Domain. The X represents the number of Source texts used.

The above abbreviations shall be used in Figure 6.2 following.

6.1.6 Procedure

The procedure for the experiment would be as follows.

For each iteration of the experiment:

1. The algorithm variables for that individual execution would be set, using one or more of those listed above.

2. DurTeC would, using the selected algorithm, compare each of the 35 Comparator Structured Abstracts with the Source Structured Abstract. This resulted in a similarity metric being produced in each case. This would allow the abstracts to be ordered by the similarity metric to form a hierarchy descending in the level of relativity.

3. The results would be noted.

6.1.7 The Results

The results from Experiment 1 were substantial, and a complete listing of them can be found in Appendix 10. To simplify and clarify the results, Figure 6.2 shows a summary.

In Figure 6.2, the following columns can be seen:

Column 1. The Execution Identifier. Each iteration of processing that DurTeC undertook with a unique algorithm was given an identifier.

Column 2. The algorithm(s) applied

Column 3. The quantity of the Researcher's top 10 abstracts that the DurTeC System identified within its own top 10.

Column 4. The quantity of the 5 unrelated abstracts that the DurTeC System identified within its own bottom 10.

Column 5. A Relativity Total. This is the combined figure for the above to metrics. The optimum value here would be 15, i.e. all 10 of the Researcher's abstracts being present in DurTeC's top 10, and all 5 of the unrelated being present in DurTeC's bottom 5.

The Relativity Total figure was used as a measure of each algorithm's accuracy. Below is a summary of the Algorithm Results.

ID	Algorithm(s) Applied	Top 10	Bottom 5	Relativity Total
1	None	3	5	8
2	PR(2)	5	3	8
3	PR(3)	4	3	7
4	PR(5)	5	3	8
5	DSL	2	2	4
6	DSL PR(2)	3	2	5
7	DSL PR(3)	2	1	3
8	DSL PR(5)	4	1	5
9	DWL SS(2)	5	4	9
10	PR(2) DWL SS(2)	6	4	10
11	PR(3) DWL SS(2)	6	4	10
12	CNS	4	5	9
13	CNS PR(2)	4	5	9
14	CNS PR(3)	4	4	8
15	CNS PR(5)	4	4	8
16	DWL	5	4	9
17	PR(2) DWL	5	4	9
18	PR(3) DWL	5	4	9
19	PR(5) DWL	6	4	10
20	DSL DWL	1	2	3
21	DSL PR(2) DWL	2	2	4
22	DSL PR(3) DWL	2	3	5
23	DSL PR(5) DWL	3	3	6
24	CNS DWL	3	3	6
25	CSN PR(2) DWL	4	3	7
26	CNS PR(3) DWL	4	3	7
27	CNS PR(5) DWL	4	4	8
28	SS(2)	5	4	9
29	PR(2) SS(2)	7	4	11
30	PR(3) SS(2)	7	4	11
31	PR(5) DWL	6	4	10
32	DSL DWL SS(2)	2	4	6
33	DSL PR(2) DWL SS(2)	3	4	7
34	DSL PR(5) DWL SS(2)	4	4	8
35	CNS SS(2)	6	5	11
36	CNS PR(2) SS(2)	5	4	9
37	CNS PR(3) SS(2)	5	4	9
38	CNS PR(5) SS(2)	6	4	10
39	CNS DWL SS(2)	4	4	8
40	CNS PR(2) DWL SS(2)	3	4	7
41	CNS PW(3) DWL SS(2)	3	4	7
42	CNS PW(5) DWL SS(2)	3	4	7
43	DWL(Top 3 words)	5	5	10
44	DWL(Top 5 words)	5	5	10
45	DWL(Top 7 words)	5	5	10
46	DWL(Top 10 words)	6	5	11
47	DWL(Top 15 words)	7	5	12
48	PR(2) DWL(Top 15 words)	7	5	12
49	DWL(Top 20 words)	7	5	12
50	DLW(Top 20 words) SS(5)	6	5	11
51	SS(0)	2	5	7
52	SS(1)	5	3	8
53	SS(2)	5	4	9
54	SS(3)	6	4	10
55	SS(4)	6	5	11
56	SS(5)	8	5	13

Figure 6.4 The Algorithm's Results

6.1.7 Discussion

For the DurTeC system to be seen as viable, it should be able to identify at least half of the Researcher's top 10 abstracts in its top 10 list, and at least half of the 5 unrelated abstracts in its bottom 10 list. The results of the algorithms can be seen in the table above.

Observations

There are a number of notable observations which can be found within these figures:

1. The Use of Secondary Sources. Secondary Sources improved results considerably. This is illustrated in Executions 51 to 56. It can be seen that when additional secondary Sources are submitted, improvements are made. The use of Secondary Sources appeared to work optimally when not used in conjunction with another rule.

2. The use of Defined Word Lists. The use of the Defined Word Lists Rule when used alongside other rules showed both increases and decreases in accuracy depending upon the partnered rule.

    a. Execution 5 used only the Default Synonym List. By adding the Defined Word List Rule, a reduction in accuracy could be seen in Execution 20.

    b. Execution 7 used the Default Synonym List and the 5% Percentage Rule. This gave a Relativity Total of 5. By adding the Defined Word List Rule in Execution 23, there was a slight increase in accuracy.

    c. However, using the Defined Word List on its own resulted in an Improvement through its use, as can be seen in numbers 43 to 47.

3. The Use of the Default Synonym List. The Default Synonym List had a negative effect in most cases. Consider the following:

    a. On Execution 1, where no rules were applied, the Relativity Total returned was 7. When the Default Synonym list was applied in Execution 5, this figure dropped to 4.

    b. On Execution 16, where the Defined Word List was applied, the Relativity Total returned was 9. When the Default Synonym List Rule was applied to this in Execution 20, this figure was reduced to 3.

4. The use of the CNSynonym list. This was essentially a refined synonym list. It had varied effect

    a. On Execution 1, where no rules were applied, the Relativity Total returned was 7. When the CNSynonym list was applied in Execution 12, there was a slight increase.

    b. On Execution 16, where the Defined Word List was applied, the Relativity Total returned was 9. When the CNSynonym List Rule was applied to this in Execution 24, this figure was reduced to 6.

5. The Use of a Percentage Rule. Generally, the use of the percentage rule had a variable effect, sometimes improving, but at other occasions not so. As examples, consider:

    a. Increases such as Execution 24 and its following 3 executions.

    b. Decreases such as Execution 12 and its following 3 executions.

Notable Algorithms

One of the aims to the experiment was to identify which of the applied algorithms produced results that most closely matched those given by the Researcher. Referring to Figure 6.2, the 6 algorithms listed in offered the best results.

Algorithm identifier	Rule(s)	Relativity Total
10	2% Percentage Rule (PR(2)) + Defined Word List + 2 Secondary Sources (DWL + SS(2))	10
19	5% Percentage Rule + Defined Word List (PR(5) + DWL)	10
29	2% Percentage Rule + 2 Secondary Sources (PR(2) + SS(2))	11
35	CNSynonym + 2 Secondary Sources (CNS + SS(2))	11
47	Defined Word List (DWL(Top 15 words))	12
56	5 Secondary Sources (SS(5))	13

Figure 6.5 The 6 Most Successful Algorithms

6.1.9 Conclusion

The research has shown that the algorithm that the DurTeC system uses can be improved considerably to reflect the opinions of the Researcher. At this point, there could be a discussion about the validity of some of the study parameters. It could be argued that the opinions of the Researcher's top 10 may be incorrect. We must however, bear in mind that it is just an opinion, and like most pedagogical processes, opinions vary. There must be, however, some starting point, some set of base comparison data, and this is what the Researcher's opinions provided.

It could also be argued that, if the baseline for the DurTeC is incorrect, i.e. the results when no algorithm is applied, then the results may be inaccurate. The problem with using DurTeC's top 10 is that they do not truly represent the meaning of the abstract, just the words. Compare this with the Researcher's top 10 which consider the meaning behind the text. However, we must consider the domain that this is being applied to; the Web. In this domain there is, for the foreseeable future anyway, going to be a certain amount of vagary between what humans are searching for, and what computers *understand* they are searching for. Conversely, it could be argued that although this baseline is only a nominal guide, that improvement are still just that, improvements, and should stand as such.

It is the Researchers opinion that the following hypotheses from Figure 6.2 have been proven:

Hypothesis ID	**Result**
H1	Out of 56 executions, 31 algorithms were capable of identifying at least 50% of the Researchers Top 10 abstracts. One algorithm achieved 70%, and 5 algorithms achieved 70%
H2	Out of 56 executions, 15 algorithms were capable of identifying 100% of the unrelated abstracts.
H3	Algorithm 1, which was the original processing algorithm with no variables applied, returned a Relativity Total of 7. Out of the 56 executions, 29 algorithms were capable of improving this figure. If we consider a Relativity Total of more than 11 to be a significant increase on the original figure of 7, then 10 of the algorithms achieved this.

Figure 6.6 The Confirmed Hypotheses for Experiment 1

6.2 Experiment 2

6.2.1 Overview

This experiment would take the following format:

Aim	To assess how well the DurTeC system using the 6 Notable Algorithms could match the opinions of the human results from Pilots Study 2.
Hypothesis	The results returned from DurTeC should significantly match those from Pilot Study 2.
Variables and Sets	Control variables: The applied algorithm variables Response variables: The variance between DurTeC's opinion and participant's opinions from Pilot Study 2. Control group: Pilot Study 2 set Experimental group: DurTeC set Participants: 2
Inputs	Structured Abstracts representing a set of academic articles. The 6 Notable Algorithms highlighted in Figure 6.3.
Procedure	Compare the human results returned from Pilot Study 2 with results from the DurTeC system using each of the Notable Algorithms as highlighted in the previous section.
Format of Results	Metrics.

Figure 6.7 Experiment 1 Overview

6.2.2 Aim

The aim of the experiment had two foci:

1. To determine how accurately DurTeC could mimic the human results from Pilot Study 2.
2. To determine which of the 6 Notable Algorithms offered the best comparison.

6.2.3 Hypothesis

There are a number of hypothesis for this experiment. Figure 6.8 lists these.

Hypothesis ID	Hypothesis
H1	The DurTeC system could reliably match the humans' opinions in Pilot Study 2.
H2	That the most accurate comparison algorithms would be identified.

Figure 6.8 Hypotheses for Experiment 2

6.2.4 Variables and Sets

Control variables. The control variables are the variables set when using the 6 Notable Algorithms.

Response variables. The difference between DurTeC's opinion and the participant's opinions from Pilot Study 2.

Control group. The Pilot Study 2 set will remain unaffected by the experiment.

Experimental group. The algorithms applied by DurTeC will produce a set of data which can be compared with the participants' opinions in Pilot Study 2.

Participants: 2

6.2.5 Inputs

For this experiment, the DurTeC system would use the same data as used in Pilot Study 2. The results from the DurTeC system would be compared against the results from Pilot Study 2. These results found in Figure 6.9 are reproduced from those found in Figure 3.4.

Abstract #	Avg.	Rank
Abstract 01	2.92	8
Abstract 02	2.25	4
Abstract 03	2.17	2
Abstract 04	1.83	1
Abstract 05	2.25	5
Abstract 06	2.33	6
Abstract 07	2.17	3
Abstract 08	2.75	7
Abstract 09	3.42	10
Abstract 10	2.92	9

Figure 6.9 Results from Pilot Study 2

Figure 6.9 consists of the following columns:

Column 1. Abstract identifier.

Column 2. The average ranks from the participants (better matches are lower numbers).

Column 3. The position based on the average ranks (in ascending order).

For this experiment, it is assumed that the data collected from Pilot Study 2 was accurate. It could be argued quite easily that this is an incorrect assumption because:

1. The participants may have made erroneous selections.
2. The participants, asked to do the study again would give completely different answers.
3. Were a different set of participants asked to undertake the study, the results would be different.
4. Given different data, the study may prove to be more or less accurate.

However, for its purpose, the data will have to be accepted as 'as is' and as correct. Remember that the study is asking for opinions, not facts.

The DurTeC system would in turn, use each of the 6 algorithms identified in Figure 6.3.

6.2.6 Procedure

The procedure for the experiment would be as follows; for each iteration of the experiment:

1. One of the 6 Notable Algorithms would be set.
2. DurTeC would, using the algorithm, compare each of the 10 Comparator Structured Abstracts with the Source Structured Abstract. This resulted in a similarity metric being produced in each case. This would allow the abstracts to be ordered by the similarity metric to form a hierarchy descending in the level relativity.
3. The results would be noted.

6.2.7 The Results

The following table shows the results from Experiment 2.

Algorithm name	Pilot Study 2 Rank	DurTeC Basic	PR(2) DWL SS(2)	PR(5) DWL	PR(2) SS(2)	CNS SS(2)	DWL(15)	SS(5)
Algorithm ID	**N/A**	**1**	**10**	**19**	**29**	**35**	**47**	**56**
Abstract #								
Abstract 01	8	10	8	10	7	10	10	5
Abstract 02	4	4	3	3	1	6	5	3
Abstract 03	2	6	7	7	8	3	6	2
Abstract 04	1	7	5	5	6	8	7	4
Abstract 05	5	5	6	6	4	5	4	1
Abstract 06	6	9	10	9	9	4	9	9
Abstract 07	3	3	2	2	2	2	3	6
Abstract 08	7	2	4	4	5	7	2	7
Abstract 09	10	8	9	8	10	9	8	10
Abstract 10	9	1	1	1	3	1	1	8

Figure 6.10 Results from Experiment 2

The table is made up of the following columns:

Column 1. The abstract ID.

Column 2. The Pilot Study 2 position - taken from Figure 6.5.

Column 3. Results from DurTeC with no algorithm applied.

Columns 4-9. Results from DurTeC with each of the 6 algorithms applied.

We can see, for example, that Abstract 9, the *Indicator Abstract* is in position 10 in the results from Pilot Test 2 (PS2 Position column). The same abstract has been similarly placed when DurTeC used algorithms 29 and 56. However, Abstract 4, which was ranked top in Pilot Study 2, only managed to be ranked 6th

and 4th respectively.

To make these results more understandable, a Degrees of Separation Matrix as introduced in 3.2.7, and used again in 3.3.7 was created from this data. That Matrix can be found following.

Algorithm name	DurTeC Basic	PR(2) DWL SS(2)	PR(5) DWL	PR(2) SS(2)	CNS SS(2)	DWL(15)	SS(5)	Total DOS
Algorithm ID	**1**	**10**	**19**	**29**	**35**	**47**	**56**	
Abstract ID								
Abstract 01	2	0	2	1	2	2	3	12
Abstract 02	0	1	1	3	2	1	1	9
Abstract 03	4	5	5	6	1	4	0	25
Abstract 04	6	4	4	5	7	6	3	35
Abstract 05	0	1	1	1	0	1	4	8
Abstract 06	3	4	3	3	2	3	3	21
Abstract 07	0	1	1	1	1	0	3	7
Abstract 08	5	3	3	2	0	5	0	18
Abstract 09	2	1	2	0	1	2	0	8
Abstract 10	8	8	8	6	8	8	1	47
Degrees of Separation (DOS)	**30**	**28**	**30**	**28**	**24**	**32**	**18**	**190**
Average DOS	**3**	**3**	**3**	**2.8**	**2.5**	**3.2**	**1.8**	**1.8**
Rank	**5**	**3**	**5**	**3**	**2**	**7**	**1**	**1**

Figure 6.11 Degrees of Separation Matrix

The above table is made up of the following columns and rows:

Column 1. The abstract ID.

Column 3. The Degrees of Separation (DOS) from the basic DurTeC calculation with no algorithm applied.

Columns 3-7. The DOS from DurTeC with each of the 6 algorithms applied.

Row 15. The total DOS for that column.

Row 16. The average DOS for that column.

Row 17. The rank based on Row 16.

The important row above is the one second from the bottom, which holds the Average DOS metric for each algorithm. The lower the DOS, the better the result.

6.2.7 Discussion

If we look at Algorithm 1 which is called DurTeC Basic, it uses TF/IDF without any other variables applied, it can be seen that there was a total of 30 Degrees Of Separation between it and the results from Pilot Study 2. This meant that on average, over the 10 Abstracts, Algorithm 1 was 3 DOS away from those supplied by the participants in Pilot Study 2.

This figure can be used as a Comparator against other algorithms, seeing where improvements were made when applying the variables.

If we refer to the best ranked algorithm, Algorithm 56, it can be seen that there was a total of 17 DOS between it and the results from Pilot Study 2. This meant that on average, Algorithm 56 was 1.7 DOS away from those supplied by the participants in Study 2. This was the best improvement experienced over the Notable Algorithms.

As can be seen, Algorithms 56 and 35 offered the best results. These algorithms were:

Algorithm 56: 5 Secondary Sources

Algorithm 35: CNSynonym + 2 Secondary Sources

Algorithm 56 was particularly successful, as on average it had less than two ranks difference for each comparison, when compared with the human result found in Pilot Study 2. This two rank difference can also be presented as the algorithm only being 17% inaccurate, or in other words, 72% accurate.

6.2.9 Conclusion

It is the Researchers opinion that the hypotheses from Figure 6.4 have been proven:

Hypothesis ID	Hypothesis
H1	The results showed that using Degrees of Separation as a comparison process, DurTeC's best algorithm could achieve 72% accuracy of the human results.
H2	The results have shown that algorithms 56 and 35 were most successful in mimicking the participants' choices in Pilot Study 2

Figure 6.12 The Confirmed Hypotheses for Experiment 2

Chapter 7 Discussion

Chapter 1 highlighted the problems and issues associated with the size of the Web; how it affects the viability of academic research. This research considered what technologies are currently available which could be used to address this problem and went on to propose a system which would utilise them.

7.1 The Potential Offered by the Web

Chapter 2, the Literature Review sought to identify those most potentially useful technologies which could be made use of to address the concerns highlighted in Chapter 1.

It suggested that the areas which appear to offer the most potential for academic research were:

1. Academic Learning Environments.
2. Knowledge sharing.
3. Agents.
4. Document classification and text comparison.

The research showed that some of the most important aspects of using the Web as a learning resource were:

1. The importance of collaboration.
2. The importance of sharing.
3. The potential for information overload.
4. The development of new skills for digital age learners.

A number of these concepts were to be borne in mind as this research progressed, namely collaboration, sharing and the control of information overload.

7.2 Findings from the 2 Pilot Studies

The research began by undertaking 2 pilot studies aimed at understanding how students perceived information in given texts, in this case academic articles. On the Web, in many cases, the student may have only the article's title to make a judgement on its relativity to their sought subject, and its quality.

Pilot Study 1

Pilot Study 1 asked participants to consider academic titles. They were asked to make judgements, based on the titles alone, on how well several articles matched a single given Source article. It was found that, in many cases, the titles did not adequately describe the article contents. In some cases, the titles were quite misleading. The students' opinions mainly matched one another, but there were anomalies.

Those anomalies included:

1. The title which participants deemed the best match to the Source title was actually a book covering a sub-topic of the Source article; therefore had they relied solely on this book for research, they would have missed a considerable amount of information.
2. The participant ranked *Chrysler Goes to Extreme* alongside *Extreme Programming Explained 1st Edition*. This does not seem entirely logical as the former paper's title is much more vague than the latter's, which sounded like a highly targeted book.
3. The participants ranked *Extreme Programming Explained 2nd Edition* 14

places above the 1st edition which again was not expected.

It was however, reassuring to see that the participants had generally identified the most relevant reading materials given the Source article fairly well in the top half of the table.

Pilot Study 2

Pilot Study 2 took the experiment to a further stage. Instead of giving participants just the article titles to use for their comparison, they were issues with Structured Abstracts on which to base their opinions. This resulted in an improvement in cohesion between the participants’ opinions.

The pilot study produced a number of observations:

1. Structured Abstract 4 most closely matched the Source article's Structured Abstract. Out of the 12 participants, 11 of them judged it as either very or fairly similar, giving it a '1' or a '2'.
2. Participant 9 'bucked' the general trend of the other ratings, and may well have affected the overall accuracy of the study.
3. It had been expected that Abstract 9, the unrelated Indicator Abstract, would be ranked as 'not similar' by all participants. This was not the case as only half of them rated did so.

Pilot Study 2 furnished us with some data which would be called upon for the later experiments.

7.3 Proposed System

The research then went on to consider a scenario where two students, undertaking similar research may, with the use of a software tool, achieve considerably more productive researching. This researching was assisted by the creation of a model called the Durham Browsing Assistant (DurBA). This model was presented and it was explained how areas of it made use of some of the technology highlighted in earlier research. During this presentation of the DurBA model, a number of key areas were highlighted as listed following:

1. Autonomous Browsing.
2. The Textual Comparison System.
3. Search-word Selection.
4. Recommendation System.
5. Recommendation Rewards.

It was decided that this research should concentrate on just one of these areas, and that chosen area was the Textual Comparison System.

A second system, considered a 'plug-in' to the DurBA model was proposed as part of this concentrated research. This sub-system was called DurTeC for Durham Textual Comparison system and would concentrate on the part of the DurBA model which would make judgements on how similar two given documents were. If, through the use of the DurTeC system, DurBA could autonomously mimic how students compared documents, it would be an important tool in reducing the amount of academic materials that student would need to review to get to the more valid materials.

The DurTeC system was created, not as a plug-in for DurBA, but as a stand-alone testing framework for the technology that it encapsulated. It would allow numerous comparison algorithms to be implemented and adjusted within its framework for testing purposes.

7.4 Findings from Experiment 1

Experiment 1 involved using the DurTeC system together with a set of differing algorithms, in order to identify which offered the greatest potential in comparing documents. In order that a comparison could be made against DurTeC's results from this, a second set of data was required. For this, the Researcher's opinion was used as a Comparator set. The Researcher's opinion set would be compared with results from the DurTeC system.

The experiment included the testing of 56 different algorithms. From the results of these tests, a set of 6 of the most Notable Algorithms were identified. The Notable Algorithms are listed following.

Identifier	Algorithm
10	2% Percentage Rule + Defined Word List + 2 Secondary Sources
19	5% Percentage Rule + Defined Word List
29	2% Percentage Rule + 2 Secondary Sources
35	CNSynonym + 2 Secondary Sources
47	Defined Word List (Top 15 words)
56	5 Secondary Sources

Figure 6.13 The 6 of the Most Notable Algorithms

These algorithms would be used again in Experiment 2.

It was observed that some of the algorithms worked better together where others suffered a negative effect when used together.

For example, The CNSynonym list worked well on its own, being able to identify all of the unrelated abstracts in its bottom 10. However, when the CNSynonym list was combined with Secondary Sources, for example, its accuracy dropped across the figures. Also, combining the CNSynomym list with the Defined Word List did not work well. At best, this combination managed to identify four out of the top ten and four out of the bottom ten, where purely on its own, it achieved five and four respectively.

To make observation more easily identified, a graph was created using the data from DurTec.

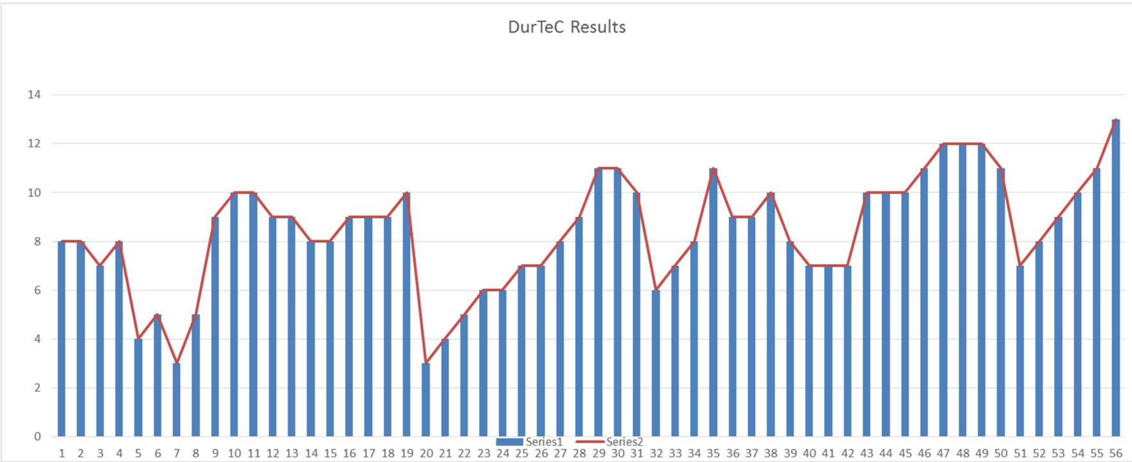

Figure 6.14 The Results of DurTeC Algorithms

The above graph which is a representation of DurTeC's results lists along the base, the numbered columns representing the algorithm identifiers.

The numbered rows along the left hand edge show the number of top 10 and bottom 5 abstracts identified using each algorithm.

It can be seen in rows 5 to 9, 20 to 24 and 32 to 34, that the introduction of the Default Synonym List always caused a drop in accuracy.

It can also be seen that the introduction of the Secondary Sources in columns 9 to 11, and 50 to 56 caused a distinct improvement in accuracy.

Finally, there was a marked positive surge in the accuracy whenever the top (n) occurring words were from the Defined Word List were used. This can be seen in columns 43 to 49 where the top 3, 5, 7, 10, 15 and 20 from the DWL were used.

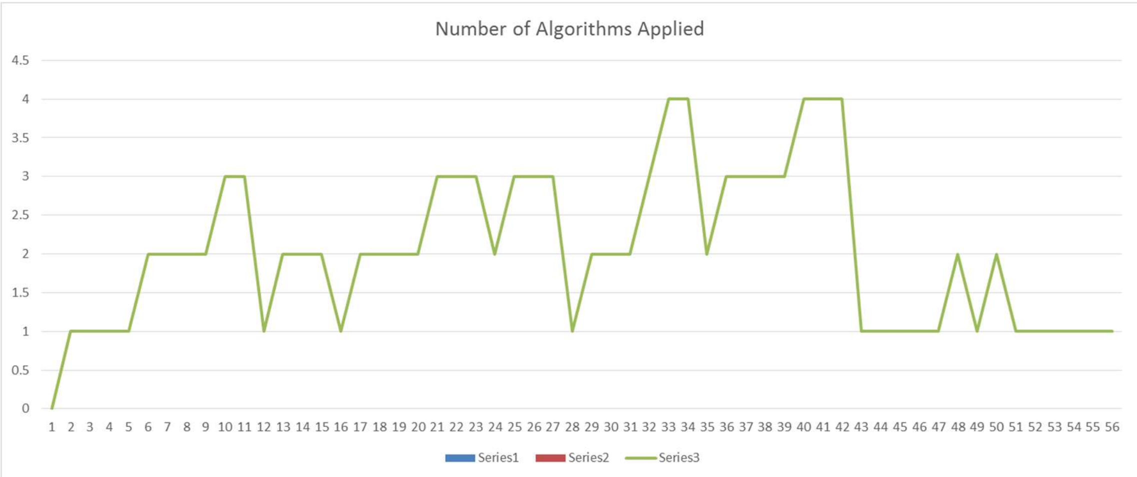

Figure 6.15 The Number of Algorithms Applied

The above graph illustrates the quantity of variables applied to each algorithm. It can be seen that Algorithm 40 utilised 4 variables.

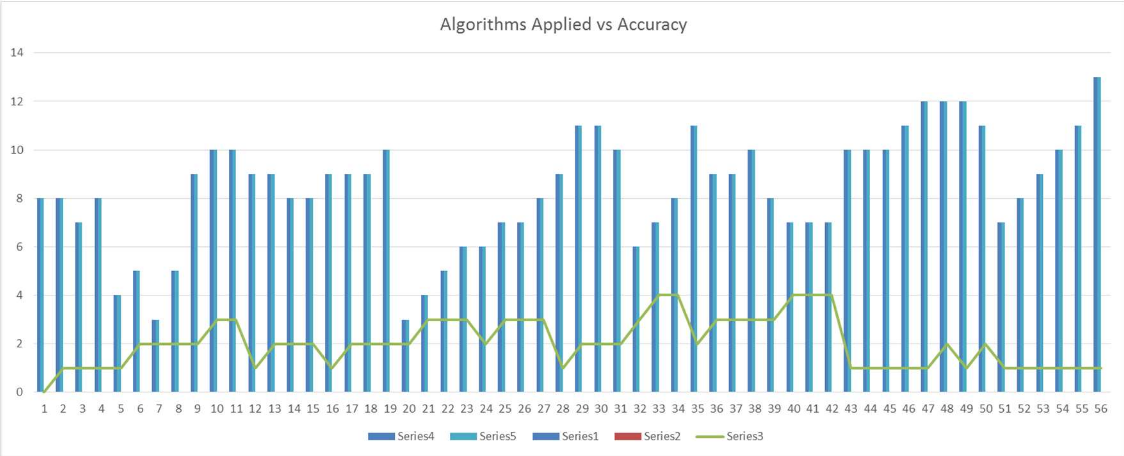

Figure 6.16 The Number of Algorithms Applied vs DurTeC's Results

The above graph combines both of the above graphs in order to see if there is any relationship between the number of variables applied and accuracy of the results.

As can be seen, the relationship is variable.

From Algorithms 7 to 11, an increase in accuracy is observed. However, from Algorithms 27 through to 30, a reduction in variables shows an increase in accuracy. From Algorithm 43 onwards where in most cases, only one variable was used, there was a definite maintaining of accuracy. Compare this with Algorithms 39 to 42 where those same variables were augmented with others and a drop in accuracy was observed. It seemed that an increase in variable complexity did not have a positive effect on accuracy in most cases.

7.5 Findings from Experiment 2

The Notable Algorithms as highlighted previously were used in Experiment 2. On this occasion, they were compared against the data collected in Pilot Study 2. Pilot Study 2 had used Structured Abstracts as Comparator documents, and a number of participants had given their opinions.

In Experiment 2, the DurTeC system would utilise each of the 6 identified algorithms, and the results would be compared against the participant results from Pilot Study 2. The experiment aimed to see which of these 6 algorithms proved to be most successfully in mimicking the participants' opinions.

The results have shown that algorithms 56 and 35 were most successful in mimicking the participants' choices in Pilot Study 2:

Algorithm 56: 5 Secondary Sources

Algorithm 35: CNSynonym + 2 Secondary Sources

The results demonstrated that by using Degree of Separation as a comparison process, DurTeC's best algorithm, Algorithm 56, could achieve 72% accuracy of the human results.

7.6 Consequences on the DurTeC System

The results from experimenting with the DurTeC system suggested that by combining simulated collaboration and an autonomous comparison process, it could reliably identify documents of a similar nature. This meant that it could also identify non related documents.

The benefits of this within an academic research background would be impressive. Even if the system were only capable of identifying those least related articles that students came across whilst browsing the Web, it would reduce their work-load substantially.

Consider the two algorithms identified as most suitable; algorithms 35 and 56.

Algorithm 35

Algorithm 35 used a Defined Synonym List which related to words used within the sought Research Domain, in this case empirical reviews of Agile programming methods. The benefit of employing a Defined Synonym List rather than a standard synonym list was that only words considered appropriate to the Domain were considered. For example, if a standard synonym list was used during the processing, certain words would be added to the Research Domain to complement it, and some of these words might not necessarily be valid. The effect would be a watering down of the Research Domain.

A good example here is synonyms for the term 'agile'. In a clearly defined synonym list, the term agile might be associated with other terms such as *planning*, *iterative*, *review*, *collaboration*, *redefine*, *increment*, *adaptive*, *rapid*, *team work*, *change* and *practice*. All of these terms might be used when describing agile methods with this domain. However, a standard synonym list might include words such as *nimble*, *supple*, *lithe*, *sprightly*, *swift* and *lively* which, when incorporated into the Research Domain, and used in a comparison process would most likely dilute the results with lower quality resources.

Algorithm 35 also made use of 2 additional Secondary Sources. Secondary Sources are used by DurTeC in addition to the Primary Source document when creating the Research Domain. In effect, they strengthen DurTeC's understanding of the Area of Research. In this case, it has increased DurTeC's accuracy.

Algorithm 56

Algorithm 56 relied on additional Source documents to create the Research Domain. The experiments showed that by adding further Source documents, DurTeC could create a better understanding of the Research Domain and achieve a higher standard of comparison. This algorithm was tested with up to 6 additional Sources. At some point, the benefits of adding additional Sources could be expect to be somewhat neutralised by the effect of dilution.

Both of these algorithms would fit quite snugly into the DurBA model. To recap, DurBA is designed to identify suitable research materials given a Research Domain. The Research Domain is the system's understanding of the research topic. Researches of that topic are members of that Research Domain group. Once a Research Domain is defined, DurBA can search the Web autonomously for relative materials, reporting back to users those which it finds most suitable.

The two algorithms highlighted would be useful in this process, consider:

1. Algorithm 35 - CNSynonym + 2 Additional Sources.

   A bespoke synonym list could be created by the users of the system over a period of time. Words could be added, removed, and re-weighted so their effects could be manipulated. Each user would have the same rights to alter the list of words. In effect, a group of students researching highly defined topics could collaborate on creating the synonym list for the Research Domain. Secondary Sources could be submitted to DurBA by users to strengthen the Research Domain. Secondary Sources could be identified by tutors or by students who come across relative material as they undertake their Web research. The secondary Sources located by users, also in effect become part of the research reading list distributed to all members of the Research Domain group.

2. Algorithm 56 - 5 Secondary Sources.

   As above, Secondary Sources could be submitted to DurBA by the group members to augment the Research Domain.

Both algorithm mechanisms lend themselves collaboration. By making recommendations, users are in effect providing Secondary Sources to strengthen the Research Domain. By providing Search-words which can be seen as extending the synonym list, the users are also strengthening the Research Domain.

7.7 Additional Comparison

Comparing all of the 56 Algorithms with the Results from Pilot Study 2.

Experiment 2 compared the six Notable Algorithms with the human results from Pilot Study 2. As an additional study, all 56 of the algorithms were executed against the same data to see how they compared. It was of interest to see if any of those algorithms identified as the most accurate in Experiment 2 would also percolate to the top when comparing all of the algorithms. Using the same criteria as in Experiment 2, the top 6 algorithms were highlighted and can be seen in the following table. Note that the complete results can be found in Appendix 11.

Algorithm name	DurTeC Basic	SS(2) PR(2) DWL	SS(2) PR(3) CNS	SS(2) PR(5) CNS	CNS SS(2)	SS(4)	SS(5)	Total DOS
Algorithm ID	1	10	37	38	35	55	56	
Abstract ID								
Abstract 01	2	2	2	2	5	5	3	21
Abstract 02	4	3	3	1	1	0	1	13
Abstract 03	5	0	1	5	0	0	0	11
Abstract 04	1	7	1	1	3	0	3	16
Abstract 05	0	1	1	0	4	5	4	15
Abstract 06	3	1	1	2	4	2	1	14
Abstract 07	1	0	5	2	4	4	5	21
Abstract 08	2	6	3	3	1	2	0	17
Abstract 09	4	1	9	0	4	4	4	26
Abstract 10	8	3	0	0	0	0	0	11
Degrees of Separation (DOS)	30	24	26	16	26	22	21	
Average DOS	3	2.4	2.6	1.6	2.6	2.2	2.1	
Rank		4	5	1	5	3	2	

Figure 6.17 The Top 6 algorithms from Pilot Study 2

It was interesting to see that of the six Notable Algorithms, identified from Experiment 1, three of them where again identified in this list. Algorithms 10, 35 and 56 occupied positions 4, 5 and 2 respectively. The remaining of the Six Notable Algorithms ranked as 7th, 10th and 14th.

7.8 Conclusion

Chapter 2 highlighted a number of areas of concern with using the Web as an academic resource. The DurBA model, using a plug-such similar to the described DurTeC system, showed how each of these issues could be overcome. By combining collaboration between students, and providing a platform for the reporting and reviewing of suitable materials, the time taken to locate materials could be reduced. The concept of a Reading List was suggested as a collection of materials suitable to a given Area of Research. It would mean that users have a readily available set of highly targeted resources 'at hand' which would reduce their searching time. This would prevent the need for multiple students undertaking synchronous or repeated research. Where a user locates an item of research on the Web which has previously been reviewed by a peer, they will be notified of this fact; this will reduce the risk of users having to re-read unsuitable materials.

The concept of a Research Domain to represent a user's, or a collection of users' Area of Research was introduced. This could be used by any automated process where the comparing of materials was necessary.

A recommendation system is suggested which encourages peers to propose and rate materials they come across on the Web suitable to their areas of research. This method would go some way in addressing the concern over the identification and quality of academic materials on the Web. Where suitable materials were identified, a copy of the materials would be stored in the database. Future users of that Research Domain would have immediate access to it, via the 'clickable' Reading List. It will also address the problem of ephemeral resources; where items on the Web seem to just 'disappear' when a URL link is broken.

The DurBA model was designed using Ausserhofer's guidelines as listed in Figure 2.2; these are briefly listed again following:

1. Has overall educational goals.
2. Offers material management.
3. Allows collaboration.
4. Keeps information overload to a minimum.
5. Offers a user friendly interface.
6. Adapts to a variety of subjects.

7. Is extensible.

The DurBA model has addressed all 5 of the Areas of Concern which were highlighted in Figure 2.1 (Items 6 and 7 having been discounted in Section 5.2). It is acknowledged that the system has ignored a number of key issues which would be encountered if implemented such as the identified subjectivity and bias of document classification algorithms, and the negative effect of placing such algorithms in the real world.

Chapter 8 Conclusion

8.1 Research Contribution

Chapter 1 highlighted a number of concerns with allowing students to rely on the Web as their main academic resource. Those four areas of concern were addressed in a hypothetical scenario as highlighted in Section 1.5. This Thesis aimed to examine current technologies and identify those which offered the best potential for implementation in such a scenario.

Using the Scenario as a target, the Durham Browsing Assistant model (DurBA) was created and described which utilised several of the identified technologies. This model enabled students to collaboratively locate academic materials pertinent to their area of research. An implementation of such a model was expected to reduce the students' researching time and increase the quality of their study time.

The creation of the DurBA system highlighted a number of key areas of research which were integral to it. It was argued that improvement in any of these key areas would result in a substantial increase in the system's viability.

From those key areas of research, the topic of Textual Comparison Methods was chosen as one that would be investigated further.

For this investigation to take place, a computer based system called DurTeC was proposed and implemented. This Durham Textual Comparison system could be considered as a testing 'harness' allowing multiple comparison algorithms to be compared alongside one another.

Prior to this, a number of participant-based pilot studies were used to investigate how closely the human participants considered selected texts. The data from some of these pilot studies would be used in the final experiments where they would be compared with the data returned by the DurTeC system to see how accurate it was.

The aim of the first experiment was to compare the 56 algorithms with each other and identify those which most closely matched the Author's opinion of the compared texts. This process highlighted 6 Notable Algorithms which were used in Experiment 2.

In Experiment 2, the 6 Notable Algorithms were employed against the data from Pilot Study 2 to see how closely they would match the human participants' opinions.

These algorithms were tasked with matching the top 10 human selected texts and the bottom 5 human selected texts. This gave them a maximum achievable value of fifteen if they could identify all of the texts. The 6 notable algorithms managed to match between 10 and 13. This displayed an improvement from the baseline of a standard TF/IDF algorithm of 8 matches.

It can be seen how by implementing one of these improvement algorithms into the DurBA model, an improvement could be expected in its ability to identify materials, given the topic a student may be researching.

8.2 Criteria of Success

The Criteria for Success for this Thesis was listed in Chapter 1, and has been reproduced following in bold. Each criterion is followed by an explanation as to how it has been met.

1. **To propose a model which can provide a viable solution for the identified issues.**

   In Chapter 4, this research described the Durham Browsing Assistant (DurBA) as a model for addressing the problems identified in Chapter 1. It combined a collaborative researching environment with a recommendation system which allowed peers in working groups to review and recommend materials.

It utilised the concept of a Research Domain to represent a user's or a group of users' Area of Research. It provided to the users, a Reading List identifying the most suitable materials for their Area of Research.

It used an Agent to proactively seek out similar materials which matched the Research Domain, and copies of any materials found will be recorded.

2. **Highlight any pertinent areas of research identified in the proposed model, and undertake further research into it.**

   As a result of the investigation and description of DurBA, the following areas of research were highlighted in Section 4.6: Autonomous Web Searching, Collaboration, Textual Comparison Systems, Keyword Selection, Recommendation Systems and Recommendation Rewards.

   It was decided that this Thesis should further research Textual Comparison Systems. As part of this, 4 pilot studies and two experiments were created to identify comparison algorithms which showed substantial improvement on a standard TF/IDF algorithm. 6 such Notable Algorithms were identified through the use of the DurTeC testing harness.

8.3 Threats

We must bear in mind however that the results from this research depend on the participants' and the Researcher's opinions in some of the experiments, rather than fact. Performing the experiments again with different data, and differed participants might produce considerably different results. However, as mentioned earlier, at this point in time, and on this occasion, this is all we have to draw on, and so should be accepted as such.

8.4 Further Work

Having encountered such a large quantity of information during this research, identifying further work was not a difficult task; however, identifying the further work with the most potential was considerably more difficult. The Researcher has listed the following as further work:

1. Textual comparison methods. The mechanism by which the computer system can identify material similarities.
2. An accepted universal Web-based document classification and identification system. Standards for marking Web based document with meta-data in order that they can be more easily identified and categorised. Although this already exists with RDFs and the Semantic Web.
3. Research on the values of collaborative learning. Although collaborative learning is seen in a positive light, concepts such as over-exposure, academic standards, and its effect on different learner style students need to be considered.
4. Research on the values of peer reviewing. Although peer reviewing had shown that it can increase the quality of work being reviewed, is it, in general, an effective method of increasing quality?
5. Research on the values of Recommender systems. How effective and how necessary are recommendation systems within an academic environment?
6. Research on the effects of just-in-time and on-demand learning. Is too much learning information being delayed through the use of this paradigm?
7. Research on how much the ‘bright lights’ effect of the Web impacts on the quality of student output. Qualitative research methods should be identified and taught.

7. The need for students to have life-long-learning skills as part of their education in order to allow them to compete throughout their working life.

9. Universal Research Database. The development of a universal research database based containing all mainstream learning materials, indexed using a universal classification and identification system.

10. Research into the outcomes of students using the Web as their main learning tool compared with taking part in purely class-based courses.

11. Research into Agents and their abilities. Can developers instil sufficient intelligence into the Agent? Does the cost of deployment outweigh the benefit? How accepting are users to a foreign Agents accessing personal information such as dates of travel for booking holidays?

12. Automatic Text Summarisation. Without accurate automatic text summarisation, the mechanisation of comparison processing would not be successful.

13. I could also include in this section the potential to expand the undertaken experiments by further algorithms and variables. One example of this would be to restrict the quantity of synonyms that DurTeC could use, for example to 2, 3 and 5 synonyms, to see what effect this would have on accuracy.

Reference List

[ADHO79] Ad-hoc Working Group for Critical Appraisal of the Medical Literature, A proposal for more informative abstracts of clinical articles, Annals of Internal Medicine, 106(4), pp.597–604, 1977.

[AMBR90] Ambron S., A Short History of Hypermedia, The Interactive Learning Revolution: Multimedia in Education and Training, 1990.

[ARAN14] Arancibia, J.D., Rusu, C., Ubiquitous Computer-Supported Collaborative Learning: A Literature Review. 11th International Conference on Information Technology: New Generations, 2014.

[ARMS91] Armstrong, J.S., Hubbard, R., Does the need for agreement among reviewers inhibit the publication of controversial findings? Behavioural and Brain Sciences, 14, pp. 136-137, 1991.

[ARMS00] Armstrong, S. J., The Influence of Individual Cognitive Style on Performance in Management Education, Educational Psychology, 20, pp. 323-339, 2000.

[AROY04] Aroyo L., Dicheva D., The New Challenges for E-learning: The Educational Semantic Web, International Forum of Educational Technology & Society, 7, pp. 59-69, 2004.

[ASTI93] Astin, A. W., *What matters in college? Four critical years revisited.* San Francisco: Jossey-Bass. 1993.

[ATTE05] Attewell, J., Mobile technologies and learning: a technology update and earning project summary [online], Available: www.lsda.org.uk/cims/order.aspx?code= 052232&XOWEB (Accessed: 16th Oct 2006), 2005.

[ATTL07] Attle, S., Baker, B., Cooperative Learning in a Competitive Environment: Classroom Applications, International Journal of Teaching and

Learning in Higher Education, 19(1), pp. 77-73, 2007.

[AUSS97] Ausserhofer A., Teaching Computing using an Interactive Multi-paradigm Programming Environment. Proceedings of the 15th International Conference on Technology and Education, 1997.

[BAHI10] Bahiraey, M.H., Quality of Collaborative and Individual Learning in Virtual Learning Environments, The 2nd International Conference on E-Learning and E-Teaching, 2010.

[BALA97] Balabanovic M., An Adaptive Web Page Recommendation Service, 1st International Conference on Automated Agents, 1997.

[BARO09] Baron, A., Rayson, P., Archer, D., Word frequency and key word statistics in corpus linguistics, Anglistik, pp. 41-67. 2009.

[BAYL03] Bayley, L., Eldredge, J., 2003. The Structured Abstract: An Essential Tool for Researchers, The Journal of the Research Section of MLA, 17(1), Spring 2003

[BAYL03a] Bayley, L., Maragno, J., Wyndham, L., Structured Abstracts – They Really Work!, The Journal of the Research Section of MLA, 17(1), 2003.

[BECT03] Becta, Virtual and managed learning environments. [Online] Available at www.becta.org.uk (Accessed: 17th December 2011), 2003.

[BERG02] Bergstrom C., Bergstrom T., Electronic Subscriptions to Scientific Journals: A Boon for Whom? [Online], Available at www.octavia.zoology.washington.edu/publishing/ boon.pdf (Accessed: 2nd February 2007), 2002.

[BERN01] Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web, Scientific American, 2001.

[BHAT10] Bhatti, R., Internet Use among Faculty Members in the Changing Higher Education Environment, Library Philosophy and Practice, 2010.

[BISC12] Bischof, J.M., Airoldi, E.M., Summarizing topical content with word frequency and exclusivity, Proceedings of the 29th International Conference on Machine Learning, 2012.

[BLAC05] Blackhall, L., Die LMS! You too PLE! [Online] Available at www.teachandlearnonline .blogspot.com/2005/11/die-lms-die-you-too-ple.html (Accessed: 15th August 2011), 2005.

[BOLL99] Bollacker K., Lawrence S., Giles C., A System for Automatic Personalised Tracking of Scientific Literature on the Web, Proceedings of the 4th ACM Conference on Digital Libraries, 1999.

[BRIT00] Britain, S. & Liber, O., A framework for pedagogical evaluation of virtual learning environments. JTAP, 41, 2000.

[BRUS01] Brusilovsky P., Weber G., ELM-ART: An Adaptive Versatile System for Web-Based Instruction, Proceedings of International Journal of Artificial Intelligence in Education, 12, 2001.

[BRUS03] Brusilovsky P., Peylo C., Adaptive and Intelligent Web-based Educational Systems, International Journal of Artificial Intelligence in Education, 13, 2003.

[BUDG07] Budgen, D., Kitchenham, B., Charters, S., Turner, M., Brereton, P., Linkman, S., Preliminary Results of a Study of the Completeness and Clarity of Structured Abstracts, 11th International Conference on Evaluation and Assessment in Software Engineering, 2007.

[CAME07] Cameron R., Emamy K., Citeulike: A Researcher's Social Bookmarking Service, Ariadne, 51, 2007.

[CAPR10] Capra, R., Marchionini, G., Velasco-Martin, J., Muller, K., Tools-at-Hand and Learning in Multi-Session, Collaborative Search. CHI 2010: Sharing Content and Searches, 2010.

[CARL05] Carlson, S., The Net Generation Goes To College, Chronicle of Higher Education, 2005, 52(7), 2005

[CAVA02] Cavaglia, G., Measuring corpus homogeneity using a range of measures for inter-document distance, Proceedings of Language Resources and Evaluation, 2002.

[CHAK06] Chakravarty, R., Randhawa, S., Academic Search Engines: Librarian's Friend, Researcher's Delight, 4th Convention - Planner, 2006.

[CHAN09] Chang, J, Boyd-graber, J., Gerrish, S., Wang, C., Blei, D.M., Reading Tea Leaves: How Humans Interpret Topic Models, *Neural Information Processing Systems*, 2009.

[CHEV04] Cheverst, K., Byun, H.E., Utilising Context History to Provide Dynamic Adaptions, Applied Artificial Intelligence. 2004.

[CHIC96] Chickering, A., Ehrmann, S. C., Implementing the seven principles: Technology as lever, AAHE Bulletin. pp. 36, Available at www.tltgroup.org/programs /seven.html (Accessed: 19th July 2011), 1996.

[CHOU05] Chou, S.W., Liu, C.H., Learning effectiveness in a Web-based virtual learning environment: a learner control perspective. Journal of Computer Assisted Learning, 21, pp. 65-76. 2005.

[CHUR95] Church K.W., One Term or Two? Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 310-317, 1995.

[CLAR94] Clark R.E., Media will never Influence Learning. Educational Technology Research and Development, 42, 1994.

[CMAL03] Models of teaching and learning, Council for Museums, Archives and Libraries, Available at www.resource.gov.uk/action/learnacc/muslearn/models.asp# behaviourist (Accessed: 11th November 2012), 2003.

[COLL99] Collis B., Applications of Computer Communications in Education: An overview, IEEE Communications Magazine, 37(3), 1999.

[COLL09] Collins, A., Halverson, R., Rethinking Education in the Age of Technology: The Digital Revolution and the Schools, New York: Teachers College Press, 2009.

[CONL05] Conlan O., Lawless S., Wade V., Dynamic Contextual eLearning – Dynamic Content Discovery, Capture and Learning Object Generation from Open Corpus Sources, Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2005.

[CONN05] Connolly, F. W., It's Not the Change, It's the Difference: Evaluating Technology on Campus, EDUCAUSE, 27(4), pp. 5-6. 2005.

[CONO06] Conole, G. et al., JISC LXP – Student experiences of technologies [online], Available at /www.jisc.ac.uk/media/documents/programmes/elearningpedagogy/lxpprojectfinalreportdec06.pdf (Accessed: 7th March 2011), 2006.

[CORM01] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., Introduction to Algorithms, 2nd ed., MIT Press (McGraw-Hill), 2001.

[COYL07] Coyle K., The library catalogue in a 2.0 world, Journal of Academic Librarianship, 33(2), pp. 279–291, 2007.

[CRON57] Cronbach L.J., The Two Disciplines of scientific psychology, Journal American Psychologist, 12, 1957.

[CROS07] Cross, J., All or nothing, Informal learning blog [online], Available: www.informl.com/2007/02/09/all-or-nothing/ (Accessed: 7th March 2011), 2007.

[DAS07] Das D., Martins A., A Survey of Automatic Text Summarization, Literature Survey for the Language and Statistics II course at CMU, available at http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf (Accessed: 5th June 2011), 2007.

[DAWS07] Dawson, R., Information Research, 13(3). 2007.

[DEGA07] DeGagne, M., Wolk R.M., Unwired: Student Use of Technology in the Ubiquitous Computing World, Information Systems Educational Journal, 5(13), 2007.

[DERO00] De Roure D., Hall W., Reich S., Pikrakis A., Stairmand M., Hill G., MEMOIR – An Open Framework for Enhanced Navigation of Distributed Information, Information & Processing Management, 37, 2000.

[DIET13] Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H.Q., Giordano, D., Marenzi, I. & Nunes, B.P., Interlinking Educational Resources and the Web of Data: A Survey of Challenges and Approaches. Program: Electronic Library and Information Systems, 47(1), pp. 60-91. 2013.

[DRAC10] Drachsler, H., Pecceu, D., Arts, T., Hutten, E., Rutledge, L., Van Rosmalen, P., Hummel, H., and Koper, R. ReMashed – A Usability Study of a Recommender System for Mash-Ups for Learning, In Proceedings of iJET, pp. 7-11, 2010.

[DROR07] Dror, E., 2007, Technology enhanced learning: The good, the bad, and the ugly, Pragmatics & Cognition, 16(2), pp215-223

[DUMA02] Dumais S.T., Lewis D.D, Sebastiani F., Report on the workshop on operational text classification system, Available at: http://www.sigir.org/forum/F2002/ sebastiani.pdf (Accessed: 27th June 2007), 2002.

[DUNN93] Dunning, T., Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), pp.61-74. 1993.

[EDIT04] Editors, Addressing the limitations of Structured Abstracts. Annals of Internal Medicine, 140, pp.470−471, 2004.

[EKLU97] Eklund J., Brusilovsky P., Schwarz E., Web-based Education for All: A Tool for Development Adaptive Courseware, Proceedings of Computer Networks and ISDN Systems, 1997.

[ERDE02] Erdelez, S., Students Rely on the Web for Hard to find Information, Educause Quarterly, 1, 2002.

[EVEL04] Eveland W.P., Cortese J., Park H., Dunwoody S., How Web Site Organization Influences Free Recall, Factual Knowledge, and Knowledge Structure Density,' Human Communication Research 30(2), 2004.

[FELD05] Felder, R.M., Spurlin, J.E., Applications, Reliability, and Validity of the Index of Learning Styles, International Journal of Engineering Education, 21(1), pp. 103-112, 2005.

[FIDE99] Fidel R., Davies R., Douglass M., Holder J., Hopkins C., Kushner E., Miyagishima B., Toney C., A Visit to the Information Mall: Web Searching Behaviour of High School Students, Journal of the American Society for Information Science, 50, 1999.

[FISH25] Fisher, R.A., 1925. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd. 1925.

[FLET97] Fletcher, R.H., Fletcher, S.W., Evidence for the effectiveness of peer review. Science and Engineering Ethics, 3(1), pp.35-50, 1997.

[FLIC04] Flick, U., von Kardof, E., Steinke, I., A Companion to Qualitative Research, Sage Publications, London. 2004.

[FORE03] Foreman J., NEXT-Generation: Educational Technology versus the Lecture, EDUCAUSE Review, 2003.

[FRAS05] Fraser M., Virtual Research Environments: Overview and Activity, Ariadne, 44, 2005.

[GARB09] Garbarino, S., and Holland, J., Quantitative and Qualitative Methods in Impact Evaluation and Measuring Results, Governance and Social Development Resource Centre, available online at http://www.gsdrc.org/docs/open/ EIRS4.pdf (Accessed: 9th May 2013), 2009.

[GARC04] Garca-Gallego, A., Georgantzs, N., Pereira, P., Pernas-Cerrillo, J., Risk Attitudes and Internet Search Engines: Theory and Experimental Evidence, NET Institute Working paper, 4(3), 2004.

[GARC05] Garcia E., Information Retrieval Tutorials [Online], Available at: www.miislita.com/information-retrieval-tutorial/indexing.html (Accessed: 5th June 2007), 2005.

[GILB96] Gilbert, S.W., Welcome to the Internet: Nightmare or paradise?, AAHE Bulletin, 46(7), pp. 3-4. 1996.

[GLAS77] Glaser R., Adaptive education: Individual, diversity and learning, New York, H.B. Holt, 1977.

[GORD01] Gorder Pam F., Research Communications News, Ohio State University, 2001

[GOSL11] Gosling, P., Noordam, B., Mastering Your Ph.D.: Survival and Success in the Doctoral Years and Beyond 2nd Ed., Springer, pp. 27-35. 2011.

[GRAN07] Granic A., Glavinic V., Stankov S., Usability Evaluation Methodology for Web-based Educational Systems, Universal Access in the Information Society, 2007.

[GRAV73] Graves D.H., Writing: Teachers and Children at Work, Portsmouth, NH: Heinemann, 1973.

[GREE09] Greenhow, C., Robelia, B., and Hughes, J.E., Learning, Teaching, and Scholarship in a Digital Age, Educational Researcher, 37(4), pp. 246-259, 2009.

[GRIM01] Grimes D., Boening C., Worries with the Web: A look at Student Use of Web Resources, College & Research Libraries, 62(1), pp. 11-23, 2001.

[GROF07] Groff, J., Mouza, C., A Framework for Addressing Challenges to Classroom Technology Use, AACE Journal, 16(1), pp. 21-46. 2007.

[HAEN10] Kaplan, A.M., Haenlein, M., Users of the world, unite! The challenges and opportunities of Social Media, Business Horizons, 53(1), pp. 59-67. 2010.

[HARA93] Harasim L., Collaborating in Cyberspace: Using Computer

Conferences as a Group Learning Environment. Interactive Learning Environments, 3(2), pp. 119-130, 1993.

[HARL04] Harley, J., Current Findings from Research on Structured Abstracts, Journal of the Medical Library Association, 2004, **92**(3), pp. 367-371, 2004.

[HASH09] Hashim, N., Che Rusuli, M.S., and Zainuddin Z., The Usage of Electronic Information Resources in Information Searching: A Case Study of Students of Faculty of Engineering, International Conference on Teaching and Learning in Higher Education, 2009.

[HEAD07] Head, A.J., Beyond Google: How Do Students Conduct Academic Research, First Monday, 12(7), 2007.

[HEAR97] Hearst M., Texttiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), pp. 33–64, 1997.

[HEID09] Heidi, J., Barker, S., 2009. How high-school students find and evaluate scientific information: A basis for information literacy skills development, Library and Information Science Research, 31(1), pp. 12-17, 2009.

[HITC02] Hitchcock, S, Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C., Harnad, S., 2002. Open Citation Linking - The Way Forward, D-Lib Magazine, 7(10). 2002.

[HOFL72] Hofland, K. Johansson, S. Word frequencies in British and American English. The Norwegian Computing Centre for the Humanities, 1972.

[HONG03] Hong, K.S., Ridzuan, A.A., and Kuek, M.K., Student's attitudes toward the use of the Internet for learning: A study at a university in Malaysia, Technology & Society, 6(2), pp.45-49. 2003.

[HORN26] Horn, E., A Basic Writing Vocabulary: 10,000 Frequently Used Words in Writing, [Monograph First Series No. 4]. 1926.

[IVAN10] Ivaniin, M., Mujacic, S., Debevc, M., Introducing e-learning to institutions and classrooms: Contributions to quality assurance of e-education and e-contents, Proceedings of the 9th International Conference on Information Technology Based Higher Education and Training, 2010.

[JACO00] Jacobsen, D. M., Kremer, R., 2000. Online testing and grading using WebCT in computer science. Paper presented at WebNet 2000: World conference on the WW and the Internet, 2000.

[JAIN07] Jain, R., Introduction to Experimental Design, [online] Available at http://www.cse.wustl.edu/~jain/cse567-07/ftp/k_16ied.pdf (Accessed: 25/04/2013), 2007.

[JEDL07] Jedlitschka, A., Ciolkowski, M., 2007. Reporting Experiments in Software Engineering, Guide to Advanced Empirical Software Engineering, 2, pp. 201-227. 2007.

[JENH07] Jen-Hwa Hu P., Hui W., Clark T., Yan Tam K., Technology - Assisted Learning and Learning Style: A Longitudinal Experiment, IEEE Transactions on Systems Management and Cybernetics, 37(6), 2007.

[JISC09] JISC, Higher education in a Web 2.0 world [Online]. Available at: www.jisc.ac.uk/publications/documents/heweb2.aspx (Accessed: 5 May 2011), 2009.

[JONE02] Jones S., The Internet Goes to College, Pew Internet & American Life, 2002.

[KARE06] Kareal F., Klema J., Adaptivity in e-learning, Current Developments in Technology Assisted Education, Czech Technical University Report, 2006.

[KEMP02] Hall W., Kemp M., Davis H., Roche W., From Classroom Tutor to Hypertext Advisor: An Evaluation, Association for Learning Technology Journal (ALT-J), 10, 2002.

[KEND37] Kendall, M., A New Measure of Rank Correlation, Biometrika, 30(1–2), pp.71-79. 1937.

[KEND90] Kendall, M.G., Gobbons, J.D., Rank Correlation Methods (5th edition). London: Arnold, 1990.

[KENN09] Kennedy D.J., Virtual Learning Environments (VLEs): here to stay, or on the brink of demise?, The Plymouth Student Educator, 1(1), pp. 57-66, 2009.

[KILG97] Kilgarriff, A., Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora, Language Engineering for Document Analysis and Recognition. Proceedings, AISB Workshop, Falmer, 1997.

[KILG97] Kilgarriff, A., Salkie, R., Measures for corpus similarity and homogeneity Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-3), June, pp.46-52. 1997.

[KLEB10] Klebl, M., Krämer, B.J., Zobel, A., From Content to Practice: Sharing Educational Practice, in Edu-Sharing. British Journal of Educational Technology, 41(6), pp. 936–951. 2010.

[KNIG11] Knight, L.V., Steinbach, T.A., Adapting Peer Review to an Online Course: An Exploratory Case Study, Journal of Information Technology Education, 10, 2011.

[KOLB90] Kolb D.A., Rubin I.M., Osland J., Organisational Behavior: An Experimental Approach, Englewood Cliffs, NJ: Prentice Hall, 1990.

[KOPE04] Koper R., Use of the Semantic Web to Solve Some Basic Problems in Education, Journal of International Media in Education, 6, 2004.

[KOSC96] Koschmann T.D., CSCL: Theory and Practice of an Emerging Paradigm, Routledge, 4, 1996.

[KOST02] Kostoff, N., Hartley, J., Open letter to technical journal editors. Journal of Information Science, 27(3), pp. 57–61, 2002.

[KUH001] Kuh, G.D., Hu, S., The Relationship between Computer and Information Technology Use, Selected Learning and Personal Development

Outcomes, and Other College Experiences, Journal of College Student Development, 42(3), pp. 217-32, 2001.

[KVAV04] Kvavik, R. B., Caruso, J. B. &Morgan, G., ECAR study of students and information technology 2004: Convenience, Connection, and Control, EDUCAUSE Centre for Applied Research, 2004.

[LEHM04] Lehman, J.S., Santner, T.J., Notz, W.I., Designing computer experiments to determine robust control variables, Statistica Sinica, pp. 571-590. 2004.

[LENN94] Lennon J., Maurer H., Lecturing Technology: A Future with Hypermedia, Educational Technology, 1994.

[LEWI01] Lewis D.D, Sebastiani F., Report on the workshop on operational text classification systems, [Online] Available at: http://www.sigir.org/forum/F2001/ textClassification.pdf (Accessed: 27th June 2007), 2001.

[LEWI04] Lewis, D.D., Yang, Y., Rose, T.G., Li, F., RCV1: A New Benchmark Collection for Text Categorisation Research, Journal of Machine Learning Research, 5, pp. 361-397, 2004.

[LIEB95] Lieberman, H., Letizia: An Agent That Assists Web Browsing, International Conference on Artificial Intelligence. 1995

[LIGH90] Light, R. J., Singer, J. D., & Willett, J. B., By Design: Planning Research on Higher Education, MA: Harvard, 1990.

[LIU006] Liu, Z., Jin, H., Fang, Z., Collaborative Learning in E-Learning based on Multi-Agent Systems, Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design, 2006.

[LIU012] Liu, M., Yang, J., An improvement of TFIDF weighting in text categorization, 2012 International Conference on Computer Technology and Science (ICCTS 2012), 47, 2012.

[LONG96] Longworth, N., Davies, W. K., Lifelong Learning. New vision, new implications, new roles for people, organizations, nations and communities in the 21st century, London: Kogan, pp. 179, 1996.

[LORE02] Lorenzen, M., The Land of Confusion? High School Students and Their Use of the World Wide Web for Research, Research Strategies, 17(2), pp. 151-63, 2002.

[LOZA06] Lozano-Nieto, A., Guijarro, and E., Berjano E.J., Critical Assessment of the World Wide Web as an Information Resource in Higher Education: Benefits, Threats, and Recommendations, Journal of Online Teaching and Learning, 2(1), pp. 22-29, 2006.

[LUHN57] Luhn H.P., A Statistical Approach to the Mechanised Encoding and Searching of Literary Information, IBM Journal of Research and Development, pp. 309-317, 1957.

[MACC02] MacColl J., Pinfield S., Climbing the Scholarly Publishing Mountain with SHERPA, Ariadne, 33, 2002.

[MACG99] MacGregor S.K., Hypermedia Navigation Profiles: Cognitive Characteristics and Information Processing Strategies, The Journal of Educational Computer Research, 20(2), 1999.

[MACH07] Machado, M., Tao, E., Blackboard vs. Moodle: Comparing User Experience of Learning Management Systems, 37th ASEE/IEEE Frontiers in Education Conference, 2007.

[MAES94] Maes P., Agents that Reduce Work and Information Overload, Communications of the ACM, 37(7), 1994.

[MANU02] Manuel, K., Teaching Information Literacy to Generation Y. Journal of Library Administration, 36, pp. 195-217. 2002.

[MARS03] Marshall S., Mitchell G., Potential Indicators of e-Learning Process Capability, Proceedings of UDUCAUSE, 2003.

[MCCA97] McCallum, A., Rosenfeld, R., Mitchell, T.M., Ng, A.Y., Improving Text Classification by Shrinkage in a Hierarchy of Classes, International Conference on Machine Learning, 1997.

[MCCO97] McCormack C., Jones D., Building a Web Based Educational System, Wiley Computer Publishing, 1997.

[MCGO11] McGowan, H.M., Planning a Comparative Experiment in Educational Settings, Journal of Statistics Education, 19(2). 2011.

[MENC02] Menczer F., Complementing Search Engines with Online Web Mining Agents, Decision Support Systems, 2002.

[MIMN11] Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., Optimizing Semantic Coherence in Topic Models, EMNLP, Edinburgh, Scotland, 2011.

[MIOD02] Mioduser D., Nachmias R., Web in Education: An Overview, Handbook on Information Technologies for Education and Training, Springer, pp. 23-43, 2002.

[MITR97] M., Singhal A., Buckley C., Automatic Text Summarization by Paragraph

Extraction, Workshop On Intelligent Scalable Text Summarization, 1997.

[MODR04] Modritscher F., Garcia-Barrios V., Gutl C., The Past, the Present, the Future

of Adaptive E-Learning, Proceedings of International Conference of Learning, 2004.

[MOLE00] Molebash P., Technology and education: current and future trends,

Information Technology Journal, 7, 2000.

[MONT05] Monthienvichienchai R., Seyedarabi F., Conlan O., Discrepancies Between Reality and Expectation: Can Adaptive Hypermedia Meet the Expectations of Teachers, CELDA2005, Cognition and Exploratory Learning in

Digital Age, 2005.

[MORG03] Morgan, B., Cooperative learning in higher education: Undergraduate student reflections on group examinations for group grades. College Student Journal, 37(1), pp. 40-50, 2003.

[MORG13] Morgan, T., and Bullen M., Crossing Boundaries: Exploring Social and Academic Uses of Technology in Higher Education. International Perspectives on Technology Enhanced Learning (IPTEL13), 2013

[MOXL01] Moxley, J.M., Universities Should Require Electronic Theses and Dissertations, EDUCAUSE Quarterly, 24(3), pp.61-63, 2001.

[MOYO05] Moyo, L., The Virtual Patron, Science and Technology Libraries, 25, pp.175-209, 2005.

[MUHL04] Muhlhauser M., E-learning After Four Decades: What About Sustainability, [Online], Available at www.citeseer.ist.psu.edu/726924.html (Accessed: 07/01/07), 2004.

[NAHL96] Nahl D., Harada V.H., 1996. Composing Boolean Search Statements: Self Confidence, Content Analysis, Search Logic and Errors, School Library Media Quarterly, 24, 1996.

[NAIS04] Naismith, L., Lonsdale, P., Vavoula, G., and Sharples, M., Literature Review in Mobile Technologies in Learning, Futurelab Series, 11, 2004.

[NANA03] Nanas N., Uren V., De Roeck A., A Comparative Study of Term Weighting Methods for Information Filtering, The Knowledge Media Institute – Open University, 2003.

[OAKS07] Oakes, M. and Farrow, M., Use of the Chi Squared test to examine vocabulary differences in English-language corpora representing seven different countries. Literary and Linguistic Computing, 22(1), pp. 75-100. 2007.

[OBLI04] Oblinger D., The Next Generation of Educational Engagement, Journal of Interactive Media in Education, 2004.

[OEHL00] Oehlert, G.W., A First Course in Design and Analysis of Experiments, W. H. Freeman and Company, 2004.

[OLIV13] Oliver, M., Learning Technology: Theorising the Tools We Study. International Perspectives on Technology Enhanced Learning (IPTEL13), 2013.

[OMAL79] O'Malley, C., Scanlon, E., Computer-supported collaborative learning: Problem solving and distance education (Report No. 75), Open University, Centre for Information Technology in Education. 1979.

[OWST97] Owston R., The World Wide Web: A Technology to Enhance Teaching and Learning, Educational Researcher, 26(2), 1997.

[PANK04] Pankratius V., Sandel O., Stucky W., Retrieving Content with Agents in Web Service E-learning Systems, Proceedings of the Symposium on Professional Practice in Artificial Intelligence, IFIP WG12.5, First IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI), 2004.

[PARR93] Parrenas, C. S., Parrenas, F. Y., Cooperative learning, multicultural functioning, and student achievement. In L. M. Malave (Ed.), Proceedings of the National Association for Bilingual Education Conferences, pp. 171-179. 1993.

[PEDE96] Pederson, T., Fishing for Exactness. In Proceedings South Central SAS User Group, Texas, 1996.

[PETR05] Petroski, H., The Internet has become the Research Tool of Choice for Today's Students, Who Have All But Abandoned Books, ASEE Prism, 2005.

[PETZ10] Petzold, J., Winterman, B., and Montooth, K., Science Seeker: A New Model for Teaching Information Literacy to Entry-Level Biology Undergraduates, Issues in Science and Technology Librarianship, 2010

[PFLE95] Pfleeger, S.L., ACM Sigsoft Software Engineering Notes, 20(1). 1995.

[PICC01] Piccoli G., Ahmad R., Ives B., Web Based Virtual Learning

Environments: A Research Framework and Preliminary Assessment of Effectiveness in Basic IT Skills Training, MIS Q, 25(4), 2001.

[PISH10] Pishva, D., Nishantha, G.G.D., Dang, H.A., A Survey on How Blackboard is Assisting Educational Institutions around the World and the Future Trends, ICACT, 2010.

[POPE10] Popescu, E., Badica, C., Moraret, L., Accommodating Learning Styles in an Adaptive Educational System, Informatica, 35, pp. 451-462, 2010.

[PREN05] Prensky, M., Listen to the Natives, Learning in the Digital Age, 63(4), pp.7-13, 2005.

[PRES26] Pressey S.L., A Simple Apparatus which Gives Tests and Scores and Teaches, Journal of School and Society, 27, 1926.

[PUMI07] Pumipuntu, S., Phromchanthuek, S., Collaborative Learning, the Learning Method through Internet, Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology, 2007.

[RADE02] Radev D. R., Hovy E., McKeown K. Introduction to the special issue on summarization. Computational Linguistics, 27(4), pp. 399-407, 2002.

[RAGH01] Raghavan S., Garcia-Molina H., Crawling the Hidden Web, Proceedings of the 27th International Conference on Very Large Databases, pp. 129 -137, 2001.

[RAYS00] Rayson, P., Garside, R., Comparing Corpora using Frequency Profiling, Proceedings of the Workshop on Comparing Corpora, 9, pp.1-6, 2000.

REIB06] Reiber, L. J. Using peer review to improve student writing in business courses. Journal of Education for Business, 71(6), pp. 322-326. 2006.

[REID06] Peer Review: Successful from the Start, Reid, E.S., Mason, G., The Teaching Professor, 20(9), 2006.

[REST07] Resta, P., and Laferrire, T., Technology in Support of Collaborative Learning, Educational Psychology Review, 19(1), pp.65-73, 2007.

[ROCK00] Rockland, R.H., Reducing the Information Overload: A Method on Helping Students Research Engineering Topics Using the Internet, IEEE Transactions On Education, 43(4), 2000.

[ROMI04] Romiszowski, A., How's the E-learning Baby? Factors Leading to Success or Failure of an Educational, Technology Innovation. Educational Technology, 44 (1), pp. 5–27. 2004.

[RUMB01] Rumbough, T., Controversial uses of the Internet by college students [online]. Available: www.educause.edu/ir/library/pdf/csd1617.pdf (Accessed: 13th April 2011), 2001.

[SACK79] Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., Design and analysis of computer experiments, Statistical Science, 4(4), pp. 409-423. 1979.

[SALT75] Salton G., Yang C.S., Yu C.T., A Theory of Term Importance in Automatic Text Analysis, Journal of the American Society for Information Science, pp. 33-44, 1975.

[SALT77] Salton G., Buckley C., Term-Weighting Approaches in Automatic Text Retrieval, Information Process Management, 2(5), pp. 513-523, 1977.

[SAPP06] Sapp, M.R., Van Epps, A., Equal Access: What does the digital revolution mean for library websites, Science and Technology Libraries, 26(3/4), pp. 157-173. 2006.

[SCHA07] Schaffert, S., Hilzensauer, W., On the way towards Personal Learning Environments: Seven crucial aspects, eLearning Papers, 9, 2007.

[SCOT97] Scott-Armstrong, J., Peer Review for Journals: Evidence on Quality Control, Science and Engineering Ethics, 3, pp. 63-74, 1997.

[SEBA02] Sebastiani F., Machine learning in automated text categorization, ACM Computing surveys, 34(1), pp. 1–47, 2002.

[SELT12] Seltman H.J., Experimental Design and Analysis, D(2), SAGE Publications, Inc., pp. 1-427, 2012.

[SHAH14] Shah, C., Advances in Information Science: Collaborative Information Seeking, Journal of the Association for Information Science and Technology, 65(2), pp. 215-236. 2014.

[SHAW03] Shaw, M., Writing Good Software Engineering Research Papers-Mini-tutorial; In Proceedings of the 25th International Conference on Software Engineering (ICSE’03), pp. 726-736, 2003.

[SIEG57] Siegel, S., Non-parametric Statistics, The American Statistician, 11(3), pp. 13-19. 1957.

[SINH01] Sinha, R., Swearingen, K., Comparing Recommendations Made by Online Systems and Friends, DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, 2001.

[SKIE07] Skiena, S., The Algorithm Design Manual, 2nd ed., Springer, New York, 2007.

[SLAU09] Slauson, G.J., Snyder, J., Carpenter, D., Encouraging Students to Learn on the Fly in CIS courses, Information Systems Educational Journal, 7(6), 2009.

[SLAV95] Slavin, R.E., Cooperative learning: Theory, research, and practice. Allyn and Bacon, 2, 1995.

[SLAV74] Slavin, R. E., Research Methods in Education: A Practical Guide, J:Prentice-Hall, 1974.

[SMIT09] Smith, S.D., Salaway, G., The ECAR study of Undergraduate Students and Information Technology [online], Available: www.net.educause.edu/ir/library/pdf/ERS0607/ ERS0607w.pdf (Accessed: 29th June 2010), 2009.

[SONW05] Sonwalkar, N., Adaptive Learning Technologies: From One Size Fits All to Individualisation, EDUCAUSE, 7, 2005

[SOTO13] Spearman's rank correlation coefficient, Southampton University [online], Available: http://www.soes.soton.ac.uk/teaching/courses/oa132/module10/m10.html (Accessed: 17th July 2013), 2013.

[SPEA04] Spearman, C., The proof and measurement of association between two things. American Journal of Psychology, 15, pp.72–101. 1904.

[SPIR07] Spires, H. A., Lee, J. K., Turner, K. A., and Johnson, J., Having our say: Middle grade student perspectives on school, technologies, and academic engagement. Journal of Research on Technology in Education, 2007, 40, pp. 49-75, 2007.

[SPSS04] Hinton, P.R., Brownlow, C., McMurray, I., Cozens, B., SPSS Explained, Routledge, 2004.

[SPUR04] Spurlock-Johnson J., Zhang W., Allen-Hayes L., Can E-learning Replace the Traditional Classroom? A Case Study at a Private High School, Proceedings of ISECON, 21, 2004.

[STAH04] Stahl G., Groupware Goes to School: Adapting BSCW to the Classroom, International Journal of Computer Applications in Technology, 19(3/4), pp. 162-174, 2004.

[STAH95] Stahl G., Sumner T., Repenning A., Internet Repositories for Collaborative Learning: Supporting both Students & Teachers, Proceedings of the CSCW, 1995.

[STEV09] Stevenson, H.A., Harrison, J.E., Structured Abstracts: Do they improve citation retrieval from dental Journals? Journal of Orthodontics, 36(1), pp.52-60, 2009.

[STIL07] Stiles, M., Death of the VLE?: a challenge to a new orthodoxy. Serials, [online], Available: www.serials.uksg.org/openurl.asp?genre=article&issn=0953-0460&volume=20&issue=1&spage=31 (Accessed: 17th August 2011), 2007.

[SU2007] Su, H., Zhang H., Hashemi, R., Distributed, Integrated, and

Collaborative Research Environment (DiCore): An Architectural Design, 7th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel Distributed Computing, 2007.

[TARA07] Taraborelli, D. Soft peer review: social software and distributed scientific evaluation, Proceedings of the 7th International Conference on the Design of Cooperative Systems, pp. 99-110. 2007

[THOM03] Thompson, C., Information Illiterate or Lazy: How College Students Use the Web for Research, Libraries and the Academy, 3(2), pp. 259-267, 2003.

[TIAN07] Tian, Y., Wu, D., Kam-Wing, N., On Distributed Rating Systems for Peer-to-Peer Networks, The Computer Journal, 51(2), 2007.

[TRIB97] Tribble, C. and Jones, G., Concordances in the classroom. Athelstan, 1997.

[USUN03] Usun, S., Undergraduate Students Attitudes towards Educational Uses of Internet Interactive educational multimedia: IEM, 7, 2003.

[VARL06] Varlamis I., Apostolakis I., The Present and Future of Standards for E-learning Technologies, Interdisciplinary Journal of Knowledge and Learning Objects, 2, 2006.

[VEEN06] Veen, W., Vrakking, B., Homo Zappiens: Growing up in a digital age. London: Continuum International Publishing Group, 2006.

[VYGO77] Vygotsky, L.S., Mind and society: The development of higher mental

processes, Harvard University Press, 1977.

[WANG07] Wang, F., Rabsch, C., Kling, P., Liu, P., Pearson, J., Web-based Collaborative Information Integration for Scientific Research, International Conference on Data Engineering, 2007.

[WANG07] Wang, H., Hu, P., Hu, L., Collaborative Research-Based Learning under the Network Environment, International Conference on Computer Science and Software Engineering, 2007.

[WELL06] Weller, M., VLE 2.0 and future directions in learning environments. Proceedings of the First International LAMS Conference 2006: Designing the Future of Learning, pp. 99-106, 2006.

[WELL10] Wells, J.D., Campbell, D.E., Valacich, J.S., and Featherman, M., The Effect of Perceived Novelty on the Adoption of Information Technology Innovations: A Risk/Reward Perspective, Decision Sciences, 2010, 41(4), pp. 713-743, 2010.

[WESS02] Wessner, M., Haake, J.M., Tietze, D.A., An Infrastructure for Collaborative Lifelong Learning, Proceedings of the 35th International Conference on System Sciences. 2002.

[WIEZ97] Wiezel, A., Measuring the success of virtual tutoring, FIE Conference, 1997.

[WILL07] Williams, B.C., Brian, B., Elger, D., and Schumacher, B.E., Peer Evaluation as a Motivator for Improved Team Performance in Bio/Ag Engineering Design Classes, International Journal of Engineering Education, 23(4), pp. 697-704, 2007.

[WILS05] Wilson, S., 2005. Architecture of virtual spaces & the future of VLEs. [Online] Available at: www.zope.cetis.ac.uk/members/scott (Accessed: 22nd December 2011), 2005.

[WILS06] Wilson, S., Liber, O., Johnson, M., Beauvoir, P., Sharples, P., and Milligan, C. 2006. Personal Learning Environments: Challenging the dominant design of educational systems, EC-TEL Workshops, 3(2), pp. 27-37, 2006.

[WIND97] Windschitl, M., The Web and classroom research: What path should we take? Educational Researcher, 27(1), pp. 27-33, 1997.

[WU2004] Wu, C. F. J. & Hamada, M. 2000. Experiments: Planning, Analysis, and Parameter Design Optimization, New York: John Wiley and Sons. 2000.

[WU2012] Wu, P.H., Hwang, G.J., Tsai, W.H., An Expert System-based Context-Aware Ubiquitous Learning Approach for Conducting Science Learning Activities. Educational Technology & Society, 16 (4), pp. 217–230, 2012.

[YEW00] Yew, L.C., Hovy, E., Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, Proceedings of HLT-NAACL, Canada, pp. 71-77, 2000.

[YULE44] Yule, G. U., The statistical study of literary vocabulary. Cambridge: Cambridge University Press, 1944.

[ZAFA04] Zafar, A., Ahmad, N., Towards Adaptive e-learning: Technological Challenges and Enabling Technologies, e-Learning Research Agenda Forum, 2004.

[ZWYN03] Zwyno M.S., Student Learning Styles, Web Use Patterns, and Attitudes Toward Hypermedia-Enhanced Instruction, Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, 2003.

Appendices

Appendix 1: Confirming the Results for the 'Degrees of Separation' Algorithm

Processing the *Degrees of Separation* Algorithm

Each Domain file from the Material Repository would be compared against the user's RD file to test for how closely it matched.

Where a word was found in both the RD file and the compared Material Repository file, the 'Degrees of separation' metric would be calculated. The combined metrics from all the matched words would be added together, and added to the running total for that comparison. In the case of Degrees of Separation comparisons, the lower the metric, the better the match between two documents. We shall take each comparison in turn.

Comparing the RD File with File 1

The words that were present in both the RD file and 'File 1' were 'agile' and 'software'. The column 'Position' in the following table represents the position the word holds in the respective files; i.e. the word at the top of the list would be position one, and the word at the bottom of the list would be position 5.

RDF word	Position
study	1
agile	**2**
software	**3**
development	4
review	5

File 1 word	Position
lifecycle	1
agile	**2**
method	3
software	**4**
paradigm	5

The word 'agile' is in position 2 in the RDF, and Position 2 in File 1, therefore, the degree of separation between these two words is 0. The word 'software' is in position 3 in the RDF, and Position 4 in File 1, therefore, the degree of separation between the two words is 1.

The other three words, study, development, and review, in the RDF are not present in File 1, therefore they each gain the maximum degrees of separation metric which in this case is 5 (the number of terms used). Therefore a comparison between the two files results in the calculation as found following.

RDF word	Metric
study	5
agile	0
software	1
development	5
review	5
Total	**16**

Comparing the RD File with File 2

The words that were present in both the RD file and 'File 2' were 'agile', 'software' and 'development'.

RDF word	Position
study	1
agile	2
software	3
development	4
review	5

File 2 word	Position
software	1
development	2
agile	3
system	4
formal	5

The word 'agile' is in position 2 in the RDF, and Position 3 in File 2, therefore, the degree of separation between these two words is 1.

The word 'software' is in position 3 in the RDF, and Position 1 in File 2, therefore, the degree of separation between the two words is 2.

The word 'development' is in position 4 in the RDF, and Position 2 in File 2, therefore, the degree of separation between the two words is 2.

The other 2 words, 'study' and 'review', in the RDF are not present in File 2, therefore they each gain the maximum degrees of separation metric which in this case is 5.

Therefore a comparison between the two files results in the calculation as found following.

RDF word	Metric
study	5
agile	1
software	2
development	2
review	5
Total	**15**

Comparing the RD File with File 3

The only word that was present in both the RD file and 'File 3' was 'development'.

RDF word	Position
study	1
agile	2
software	3
development	4
review	5

File 3 word	Position
java	1
basic	2
development	3
python	4
language	5

The word 'development' is in position 4 in the RDF, and Position 3 in File 2, therefore, the degree of separation between the two words is 1.

The other 2 words, 'study', 'agile', 'software', and 'review', in the RDF are not present in File 3, therefore they each gain the maximum degrees of separation metric which in this case is 5.Therefore a comparison between the two files results in the calculation as found following.

RDF word	**Metric**
study	5
agile	5
software	5
development	1
review	5
Total	**21**

Comparing the RD File with File 4

The words that were present in both the RD file and 'File 4' were 'agile', 'software' and 'development'.

RDF word	**Position**
study	1
agile	2
software	3
development	4
review	5

File 4 word	**Position**
apples	1
oranges	2
pears	3
grapes	4
peaches	5

There were no words in the RDF which were present in File 4. All words therefore gained the maximum degree of separation metric which in this case is 5. Therefore a comparison between the two files results in the calculation as found following.

RDF word	Metric
study	5
agile	5
software	5
development	5
review	5
Total	**25**

Summary of the Manual Results

The summary of the results from this illustrative experiment can be seen following.

File #	Metric
1	16
2	15
3	21
4	25

Above, it can be seen that File 2 offered the closest match to the user's RD file, followed closely by 'File 1'.

The DurTeC System's Results

By submitting the same data to the DurTeC system, we could see if it was providing the same results as the manually created dry-run results.

The results from the DurTeC system can be found following.

File #	Metric
Paper 1	16
Paper 2	15
Paper 3	21
Paper 4	25

This shows that the returned report matches the manual run results thus confirming the results as correct.

Appendix 2: Confirming the Results for the 'Log Likelihood' Algorithm

Processing the *Log Likelihood* Algorithm

Each Domain file from the Material Repository would be compared against the user's RD file to test for how closely it matched.

Where a term was found in both the RD file and the compared Material Repository file, the 'Log Likelihood' metric would be calculated. The combined metrics from all the matched terms from those two texts would be added together to create a comparison metric.

The Log Likelihood equation used for this purpose was as displayed following:

SizeOfSource: the count of all terms in the source text

SizeOfComparator: the count of all terms in the Comparator text

FrequencyOfTermInSource: the count of the individual term occurrence in source text

FrequencyOfTermInComparator: the count of the individual term occurrence in Comparator text

FrequencyOfTermsTotal: FrequencyOfTermInSource + FrequencyOfTermInComparator

ExpectedFrequencyOfTermInSource: SizeOfSource * FrequencyOfTermsTotal / (SizeOfSource + SizeOfComparator)

ExpectedFrequencyOfTermInComparator = SizeOfComparator * FrequencyOfTermsTotal / (SizeOfSource + SizeOfComparator)

Log-likelihood calculation

Metric = 2 * ((FrequencyOfTermInSource * Math.Log(FrequencyOfTermInSource / ExpectedFrequencyOfTermInSource)) + (FrequencyOfTermInComparator * Math.Log(FrequencyOfTermInComparator / ExpectedFrequencyOfTermInComparator)))

Comparing the RD File with File 1

The words that were present in both the RD file and 'File 1' were 'agile' and 'software'. The column 'Position' represent the position the word holds in the respective files; i.e. the word at the top of the list would be position one, and the word at the bottom of the list would be position 5.

RDF term	File 1 term	LL Calculation
study	no match	1
agile	agile	0.00017
software	software	0.00604
development	no match	1
review	no match	1
Total		**3.00623**

Comparing the RD File with File 2

The words that were present in both the RD file and 'File 2' were 'agile', 'software' and 'development'.

RDF term	File 1 term	LL Calculation
study	no match	1
agile	agile	0.00017
software	software	0.00604
development	development	1
review	no match	0.25252
Total		**2.41027**

Comparing the RD File with File 3

The only word that was present in both the RD file and 'File 3' was 'development'.

RDF term	File 1 term	LL Calculation
study	no match	1
agile	no match	1
software	no match	1
development	development	0.09439
review	no match	1
Total		**4.09439**

Comparing the RD File with File 4

There were no words present in both the RD file and 'File 4'.

RDF term	File 1 term	LL Calculation
study	no match	1
agile	no match	1
software	no match	1
development	development	1
review	no match	1
Total		**5**

Summary of the Results

The summary of the results from this illustrative experiment can be seen following.

File #	Metric
1	3.00623
2	2.41027
3	4.09439
4	5

From the above table, it can be seen that File 2 offered the closest match to the user's RD file, followed closely by 'File 1'.

The DurTeC System's Results

By submitting the same data to the DurTeC system, we could see if it was providing the same results as the manually created dry-run results.

The results from the DurTeC system can be found following.

File #	Metric
1	3.00624
2	2.4103
3	4.0944
4	5

The above table shows that the returned report matches the dry-run results found in the previous table thus confirming the results as correct.

Appendix 3: Pilot Study 1 - Questionnaire

Pilot Study 1 – Comparing Abstract Titles

This is an experiment I am doing in the relationships between the titles of academic paper and their perceived contents.

Assuming the following title of a paper:

"Empirical studies of agile software development: A systematic review"

Following is a list of academic article titles. Which of these articles, based solely on their titles, would you expect to closely match the above title?

Use the following rating system:

A tick = closely match

A cross = a poor match

A blank = unsure

There is no correct way of doing this – it is a personal opinion based on the wording used.

Thanks for your help – Carl Nattrass

Page 1

Article Number/Title	Answer
02 Agile software development methods	
03 New Directions on Agile Methods: A Comparative Analysis	
05 Chrysler goes to extreme	
06 Web-Based Agile Software Development	
09 Extreme Programming Explained 1st Ed.	
10 Extreme Programming Explained 2nd Ed.	
11 Get Ready for Agile Methods, with Care	
15 Selecting a Project's Methodology	
17 An Introduction to Agile Methods	
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	
24 Evidence-Based Software Engineering for Practitioners	
26 On the Effectiveness of the Test-First Approach to Programming	
29 How To Read A Paper	
30 Agile manufacturing: A framework for research and development	
33 Checklists for Software Engineering Case Study Research	
34 Extreme Programming Considered Harmful for Reliable Software Development	
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	
37 Preliminary Guidelines for Empirical Research in Software Engineering	
40 Iterative and Incremental Development: A Brief History	
42 Questioning Extreme Programming. Should we optimize our software development process?	
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	
44 Agile Software Development: Adaptive Systems Principles and Best Practices	
46 Theoretical Reflections on Agile Development Methodologies	
47 Challenges of Migrating to Agile Methodologies	
50 A Practical Guide to Feature-Driven Development	
52 Lean Software Development: An Agile Toolkit	
53 Changing the Paradigm of Software Engineering	
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Theoretical	
56 Agile Project Management with Scrum	
59 A Survey of Controlled Experiments in Software Engineering	
61 Extreme Programming Refactored: The Case Against XP	
64 Assumptions Underlying Agile Software Development Processes	
66 Agile Software Development: It's about Feedback and Change	
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	

Appendix 4: Pilot Study 1 – Results

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	X
05 Chrysler goes to extreme	✓
06 Web-Based Agile Software Development	
09 Extreme Programming Explained 1st Ed.	
10 Extreme Programming Explained 2nd Ed.	
11 Get Ready for Agile Methods, with Care	
15 Selecting a Project's Methodology	
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	X
30 Agile manufacturing: A framework for research and development	
33 Checklists for Software Engineering Case Study Research	X
34 Extreme Programming Considered Harmful for Reliable Software Development	
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	
42 Questioning Extreme Programming. Should we optimize our software development process?	
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	
47 Challenges of Migrating to Agile Methodologies	X
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	
53 Changing the Paradigm of Software Engineering	
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	✓
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✓
61 Extreme Programming Refactored: The Case Against XP	
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	

Page 1

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	X
05 Chrysler goes to extreme	
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	X
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	
15 Selecting a Project's Methodology	
17 An Introduction to Agile Methods	
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	X
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	X ✓
29 How To Read A Paper	✓
30 Agile manufacturing: A framework for research and development	✓
33 Checklists for Software Engineering Case Study Research	
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	X
42 Questioning Extreme Programming. Should we optimize our software development process?	
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	
46 Theoretical Reflections on Agile Development Methodologies	X
47 Challenges of Migrating to Agile Methodologies	
50 A Practical Guide to Feature-Driven Development	
52 Lean Software Development: An Agile Toolkit	
53 Changing the Paradigm of Software Engineering	X
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	X
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	X
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	✓
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✓
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	✓
33 Checklists for Software Engineering Case Study Research	✓
34 Extreme Programming Considered Harmful for Reliable Software Development	
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	
42 Questioning Extreme Programming. Should we optimize our software development process?	✓
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✓
47 Challenges of Migrating to Agile Methodologies	✓
50 A Practical Guide to Feature-Driven Development	
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	✗
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	✓
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✓
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✓
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	✓
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	✓
42 Questioning Extreme Programming. Should we optimize our software development process?	✓
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✓
47 Challenges of Migrating to Agile Methodologies	✗
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	✓
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	✓
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✓
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✓

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	✓
06 Web-Based Agile Software Development	✗
09 Extreme Programming Explained 1st Ed.	✓
10 Extreme Programming Explained 2nd Ed.	
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✓
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✗
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✗
24 Evidence-Based Software Engineering for Practitioners	✗
26 On the Effectiveness of the Test-First Approach to Programming	
29 How To Read A Paper	✓
30 Agile manufacturing: A framework for research and development	
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	~~✗~~ ✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	✓
42 Questioning Extreme Programming. Should we optimize our software development process?	✓
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✗
47 Challenges of Migrating to Agile Methodologies	✓
50 A Practical Guide to Feature-Driven Development	✗
52 Lean Software Development: An Agile Toolkit	✗
53 Changing the Paradigm of Software Engineering	✓
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	
56 Agile Project Management with Scrum	
59 A Survey of Controlled Experiments in Software Engineering	
61 Extreme Programming Refactored: The Case Against XP	
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✗

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	
10 Extreme Programming Explained 2nd Ed.	
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✗
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✗
24 Evidence-Based Software Engineering for Practitioners	✗
26 On the Effectiveness of the Test-First Approach to Programming	
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✗
37 Preliminary Guidelines for Empirical Research in Software Engineering	✗
40 Iterative and Incremental Development: A Brief History	✗
42 Questioning Extreme Programming. Should we optimize our software development process?	
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✓
47 Challenges of Migrating to Agile Methodologies	✓
50 A Practical Guide to Feature-Driven Development	
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✗
61 Extreme Programming Refactored: The Case Against XP	
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✗

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✗
05 Chrysler goes to extreme	✓
06 Web-Based Agile Software Development	✗
09 Extreme Programming Explained 1st Ed.	✓
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	✗
15 Selecting a Project's Methodology	✗
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✗
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✗
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	✓
33 Checklists for Software Engineering Case Study Research	✓
34 Extreme Programming Considered Harmful for Reliable Software Development	✗
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	✗
42 Questioning Extreme Programming. Should we optimize our software development process?	✗
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✗
47 Challenges of Migrating to Agile Methodologies	✗
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	✗
53 Changing the Paradigm of Software Engineering	✗
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	✓
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✗
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✗
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✗

Page 1

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	✗✓
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	
10 Extreme Programming Explained 2nd Ed.	
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✓
17 An Introduction to Agile Methods	
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✗
24 Evidence-Based Software Engineering for Practitioners	
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	✗
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✓
37 Preliminary Guidelines for Empirical Research in Software Engineering	
40 Iterative and Incremental Development: A Brief History	
42 Questioning Extreme Programming. Should we optimize our software development process?	✗✓
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	
46 Theoretical Reflections on Agile Development Methodologies	✓
47 Challenges of Migrating to Agile Methodologies	✗✓
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	✓
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✗✓
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	

Page 1

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	✓
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	✓
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	✗
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✓
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✓
24 Evidence-Based Software Engineering for Practitioners	✓
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	✗
30 Agile manufacturing: A framework for research and development	✗
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	✗
37 Preliminary Guidelines for Empirical Research in Software Engineering	✗
40 Iterative and Incremental Development: A Brief History	✓
42 Questioning Extreme Programming. Should we optimize our software development process?	✓
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	✓
47 Challenges of Migrating to Agile Methodologies	✓
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	✓
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	✓
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✗
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✓

Experiment 1 - Answer Sheet

Article Number/Title	Answer
02 Agile software development methods	
03 New Directions on Agile Methods: A Comparative Analysis	✓
05 Chrysler goes to extreme	✓
06 Web-Based Agile Software Development	✓
09 Extreme Programming Explained 1st Ed.	✓
10 Extreme Programming Explained 2nd Ed.	✓
11 Get Ready for Agile Methods, with Care	✓
15 Selecting a Project's Methodology	✓
17 An Introduction to Agile Methods	✓
19 Toward a Conceptual Framework of Agile Methods: A Study of Agility in Different Disciplines	✓
22 Are Two Heads Better than One? On the Effectiveness of Pair Programming	✗
24 Evidence-Based Software Engineering for Practitioners	✗
26 On the Effectiveness of the Test-First Approach to Programming	✓
29 How To Read A Paper	
30 Agile manufacturing: A framework for research and development	✓
33 Checklists for Software Engineering Case Study Research	✗
34 Extreme Programming Considered Harmful for Reliable Software Development	✓
36 Guidelines for performing Systematic Literature Reviews in Software Engineering	
37 Preliminary Guidelines for Empirical Research in Software Engineering	✓
40 Iterative and Incremental Development: A Brief History	✗
42 Questioning Extreme Programming. Should we optimize our software development process?	✗
43 Is Extreme Programming Just Old Wine in New Bottles: A Comparison of Two Cases	✓
44 Agile Software Development: Adaptive Systems Principles and Best Practices	✓
46 Theoretical Reflections on Agile Development Methodologies	
47 Challenges of Migrating to Agile Methodologies	✓
50 A Practical Guide to Feature-Driven Development	✓
52 Lean Software Development: An Agile Toolkit	✓
53 Changing the Paradigm of Software Engineering	✓
54 Explaining Software Developer Acceptance of Methodologies: A Comparison of Five Models	✓
56 Agile Project Management with Scrum	✓
59 A Survey of Controlled Experiments in Software Engineering	✓
61 Extreme Programming Refactored: The Case Against XP	✓
64 Assumptions Underlying Agile Software Development Processes	✓
66 Agile Software Development: It's about Feedback and Change	✓
68 Flexible and Distributed Software Processes: Old Petunias In New Bowl	✗

Appendix 5: Pilot Study 1.1 - Questionnaire

Pilot Study 1.1

Topic

Z. 80s Music - Boy George. Need I say more?

Comparators

A. Eighties on the airwaves
B. The birth of the Metrosexual
C. Why I miss eighties music
D. Have I got Spitting Image for you?
E. The path to rock mediocrity
1. A decade of decades
2. Bring back the eighties
3. The ascendancy of Hip-Hop
4. Social Media and its Social effects
5. The science of music

Rating System

1 Very relevant
2 Fairly relevant
3 Vaguely relevant
4 Not relevant

Appendix 6: Pilot Study 1.1 - Results

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	4	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	1	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	2	4	3	1

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	4	3	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	2	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	2	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	1	2	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	1	2	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	2	3	4

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	3	2	3	4

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	4	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	3	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	2	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	3	1	3	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	1	4	4

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	3	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	~~3~~ 2	~~4~~ 1	~~3~~ 2	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	3	2	3	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	1	3	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	3	1	3	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	3	1	3	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	4	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	4	2	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	1	2	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	1	2	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	2	4	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	4	3	4

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	1	~~4~~ 3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	4	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	1	2	4	2

Appendix 7: Pilot Study 2 – Questionnaire

Using the following rating system, compare each of the ten Comparator Structured Abstracts in turn to the Master Abstract. In each case, make a judgement on how similar you perceive the contents of each pair of papers would be based on the abstract text.

Rating System

1 very relevant

2 fairly relevant

3 vaguely relevant

4 not relevant

Please enter your ratings into the following table:

Abstract #	Rating
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Master abstract

Background:

Agile software development represents a major departure from traditional, plan-based approaches to software engineering. This paper conducts a systematic review of empirical studies of agile software development up to and including 2005.

Objectives:

To investigate what is currently known about the benefits and limitations of, and the strength of evidence for, agile methods. The implications for research and practice are presented.

Methods:

The studies were grouped into four themes: introduction and adoption, human and social factors, perceptions on agile methods, and comparative studies.

Results:

This research found that most studies (73%) dealt with professional software developers, the remaining 27% were conducted in a university setting. Whilst there was a range of research methods used, 76% of it was done on XP, whilst another 15% was done on Agile Development in general.

Most projects were of short duration and were completed in small teams. It found no empirical studies of agile software development prior to 2001, but since then, there has been a steady increase.

Conclusion:

The research concludes that there is a need for more and better empirical studies of agile software development within a common research agenda.

Comparator Structured Abstract 1

Background:

Academics are generally agreed that qualitative approaches to information systems research are finally gaining acceptance; they now accepted as equal in value to quantitative approaches when used appropriately.

Objectives:

This paper celebrates and recommends Action Research, because this particular qualitative research method is unique in the way it associates research and practice. It argues that researchers should try out their theories with practitioners in real situations and real organisations.

Methods:

By highlighting several key contributions of Action Research, this paper demonstrates positive observable effects on practice.

Results:

This paper shows situations where the use of Action Research has beneficial consequences and makes comparisons with more sedentary approaches.

Conclusion:

The paper highlights the benefits of Action Research and suggests a need for an Action Research monograph, to serve as a comprehensive framework and a guide for the larger community.

Comparator Structured Abstract 2

Background:

Agile Methods are creating a buzz in the software development community, drawing their fair share of advocates and opponents. While some people con-sider agile methods the best thing that has happened to software development in recent years, others view them as a backlash to software engineering and compare them to hacking.

Objectives:

The aim of this paper is to introduce the reader to agile methods allowing a judgement upon whether or not agile methods could be useful in modern software development.

Methods:

The paper discusses the history behind agile methods as well as the agile manifesto, a statement from the leaders of the agile movement. It uses a number of Agile methods as examples to illustrate points.

Results:

This paper serve as a discussion and forum for further consideration of Agile methods.

Conclusion:

Agile Methods are here to stay, no doubt about it. Agile Methods will probably not “win” over traditional methods but live in symbiosis with them. While many Agile proponents see a gap between Agile and traditional methods, many practitioners believe this narrow gap can be bridged.

Comparator Structured Abstract 3

Background:

Ever since the first major software systems were developed, a chronic ‘software crisis’ has been seen either looming ahead or haunting the community. Solutions have been sought mostly from raising the productivity of programmers, making systems less defective, or developing systems by methods that treat the end-users as equals to the designers in the development process

Objectives:

This paper explores the concept of Extreme Programming (XP) as a new approach to Information System development.

Methods:

Using an empirical study of two companies that apply an XP style development approach throughout the information systems development life cycle we could study the social processes and try to understand users at the local level.

Results:

The paper argues that XP is more of a new bag of old tricks than a totally new way of doing things. It also demonstrates that XP is actually a new name for an old way of working when developing a tailored system in-house.

Conclusion:

We found that XP is hindered by its reliance on talented individuals, which makes its large scale deployment as a general purpose method difficult. We claim that XP can be useful for small teams of domain experts, who are physically close to and able to communicate well with the end-users and who are good designers and implementers. However, these skilled and motivated individuals with high working moral can exhibit high productivity regardless of the methods used, if they are not overly constrained by bureaucracy.

Comparator Structured Abstract 4

Background:

The accelerated pace of software development and the geographically distributed nature of many development teams demand new process models. Reducing the development cycle is now one of software engineering's most important missions.

Objectives:

The author describes the Agile Software Process, a model that tackles these challenges and that is already in use at Fujitsu.

Methods:

This paper takes the form of a discussion about the option open for addressing this paradigm. It proposes the use of the Agile Software Process.

Results:

Since 1993, the Agile Software Engineering environment has been used in a large-scale communications software project at Fujitsu. This paper highlights 6 areas which have been improved in the process.

Conclusion:

The paper demonstrates how Agile Software Process employed by Fujitsu provides a just-in-time management of the software development process by providing the right process and product information to the right people at the right time which is still in use today.

Comparator Structured Abstract 5

Background:

Agile software development methods have evoked a substantial amount of literature and debates. However, academic research on the subject is still scarce, as most existing publications are written by practitioners or consultants.

Objectives:

The aim of this paper is to attempt to make sense out of the jungle of emerged agile software development methods.

Methods:

This paper uses comparative analysis from five perspectives: software development life-cycle including the process aspect, project management, abstract principles vs. concrete guidance, universally predefined vs. situation appropriate, and empirical evidence.

Results:

The majority of the methods observed did not provide true support for project management, and abstract principles appeared to dominate the current method literature and developers' minds. While universal solutions have a strong support in the respective literature, empirical evidence is currently very limited.

Conclusion:

The paper offers new directions; it was suggested that emerging new agile methods need to clarify their range of applicability and explain the interfaces to those parts of the software development life-cycle which are not a part of the chosen focus. In addition it was suggested that emphasis should rather be placed on method specialization rather than generalisation. It also noted that the project management perspective cannot be neglected.

Comparator Structured Abstract 6

Background:

The Chrysler Comprehensive Compensation scheme was launched in May 1997. A little over a year before that, the project had been declared a failure and all code thrown away. This paper reports on how using Kent Beck's Extreme Programming methodology, Chrysler delivered a very successful result.

Objectives:

The aim of this paper is to highlight how simplicity, testing, and aggressiveness were key in this process and how looking primarily at communication, the basis of the planning and tracking processes were undertaken.

Methods:

Through the use of a case study, this paper highlights elements which should be priority when implementing Extreme Programming methods when developing computer systems.

Results:

The paper found that Extreme Programming is a great approach for teams implementing object-based applications. Object-orientation lends itself well to evolutionary development of systems.

Conclusion:

This paper illustrates that by using process described, the team was able to start over on a very difficult problem and deliver a high-quality application on time and within budget. The combination of simplicity, communication, testing, and aggressiveness, applied by a disciplined team, gave the best results.

Comparator Structured Abstract 7

Background:

Software development methods are attempting to offer answer to the eager business community asking for lighter weight along with faster and nimbler software development processes. These Agile Methods have evoked a substantial amount of literature; however academic research is still scarce.

Objectives:

The aim of the publication is three fold; it synthesises existing literature on agile programming, it describes it as a set of paradigm methods, and finally, it offers a selection of criterion for comparing those methods.

Methods:

Using a systematic literature review, this paper is able to identify and analyse key areas of the Agile Programming paradigm.

Results:

This paper provides a structured review of the Agile Methods together with some comments and suggestions.

Conclusion:

This paper concludes that Agile Software development offers a people-centric novel approach to software engineering problems, and that those methods are by no means exhaustive or capable of solving all the associated problems.

Comparator Structured Abstract 8

Background:

Almost twenty years have passed since the first qualitative research study in software engineering was published. However, research publications using qualitative methods are still scarce.

Objectives:

This paper attempts to make existing qualitative research more visible and further the understanding of qualitative research and its importance in the software engineering community.

Methods:

A collection of eleven articles are used to represent a diverse set of theoretical frameworks and methods, while focusing on a wide range of software engineering activities from requirements engineering, project management to software process improvement.

Results:

This paper offers an introduction to qualitative research relating it to the tradition of software engineering. It then gives an overview of the different discourses in which qualitative research on software engineering is published and finally concludes by giving a discussion of potential quality criteria for qualitative research.

Conclusion:

This paper highlights the fact that, although there is substantive research into software engineering, there is a lack of research into qualitative methods. This paper offers review of a selection of such papers.

Comparator Structured Abstract 9

Background:

Healthcare workers using clinical practice guidelines and other recommendations need to know how much confidence they can place in the recommendations. Systematic and explicit methods of making judgments can reduce errors and improve communication.

Objectives:

This paper proposes a system for grading the quality of evidence and the strength of recommendations that can be applied across a wide range of interventions and contexts. It presents a summary of the approach from the perspective of users of guidelines.

Methods:

The paper offers an example review of an application of the proposed system within a scenario. It goes on to offer summaries, and suggestions for future studies.

Results:

This paper finds that the GRADE approach make it easier for users to assess the judgements behind recommendations.

Conclusion:

The study found that the proposed framework for structured reflection helped to ensure that appropriate judgement were made, but it did not remove the need for judgement.

Comparator Structured Abstract 10

Background:

Systematic reviews are one of the key building blocks of evidence-based software engineering. Current guidelines for such reviews are, for a large part, based on standard meta-analytic techniques. However, such quantitative techniques have only limited applicability to software engineering research.

Objectives:

This paper describes a study into an approach to combine diverse study types in a systematic review of empirical research of agile software development.

Methods:

The study was undertaken by reporting on the steps of the research methods used and the experience gained in performing these steps.

Results:

The research found that the general procedures for systematic reviews worked well and were relevant. A key challenge however, was to include evidence from a variety of perspectives and research methods.

Conclusion:

In conclusion, the study recommends that software engineering researchers increase the rigour with which they design, conduct, analyse, and report software engineering interventions. It also highlighted the limited facilities currently offered by the software engineering specific bibliographic databases.

Appendix 8: Pilot Study 2.1 – Questionnaire

Pilot Study 2.1

Topic

Why is the Eighties music so popular? New wave music rocks them all, that's why. It was also when rap was introduced and Rap in the Eighties made sense; it was not all about sex and violence like now. Also, ONLY Eighties music makes me dance. What more reason do you want? Oh, it also makes me reminisce over my youth; about how I got it all wrong, even up to now. My wasted youth as opposed to my wasted life.

Two words: Boy George. Need I say more?

Well, I guess we also need to include Michael Jackson, Duran Duran and George Michael's music in Wham!, and of course, Queen. I still believe that Madonna's music in the Eighties is the best of all time in her music career.

But, I have many other reasons behind my thinking. The main one is maybe because I have never appreciated music again the way I did in the Eighties. It may be because I'm older now and music is not such a big deal to me now. We all know that appreciation of music is at its peak in your teenage years. Why do you think the teenagers are ALWAYS the marketing target? They are the age group that almost always buys/downloads new music.

Well, that was true then, I'm not so sure now. I see a lot of middle-aged people (I now class myself, reluctantly as middle aged) listening to music on their music players.

Rating System

1 Very relevant

2 Fairly relevant

3 Vaguely relevant

4 Not relevant

Text A

It’s been almost 15 years since the idea of an all - Eighties format radio station generated any excitement among broadcasters. The format went through a quick boom/bust around the turn of the century – one entirely expected by programmers who had seen the short-lived all - Seventies format seven years earlier. Now, according to Radioinsight’s Lance Venta, an all- Eighties format exists only on four U.S. broadcast stations—two of which are translator/FM combos. Only one, KRKE Albuquerque, N.M., http://www.krke.fm/ is in a top 100 market.

The Eighties as a format was rife with challenges. Hot stations hadn’t relinquished the biggest Eighties hits. The 10-year window was narrow chronologically, but too broad stylistically. Christopher Cross’ “Sailing” and Guns N’ Roses' “Welcome to the Jungle” were nominally from the same decade, but nobody had ever needed to put them on the radio together before. Other radio stations overcompensated, doubling down on pop/rock and Eighties alternative, occasionally including some new wave classics that weren’t radio hits as currents. It was also easy to overindulge in novelties, or take them out altogether and reduce the era’s fun factor.

Text B

New wave's bleached roots sprouted from Seventies glam rock, another British subgenre featuring grown men playing dress-up, one that, unlike Sixties Beatlemania and Eighties synth pop, never caught on in any significant way in the United States. But for all their eyeliner, glitter and platform boots, there was something unmistakably straight about glam bands like Roxy Music, T. Rex and The Sweet as well as solo superstar David Bowie, despite his declaration of gayness in Melody Maker in 1972.

Duran Duran, Depeche Mode, Spandau Ballet and their new-wave peers were an entirely different story. Yes, at its most flash-in-the-pan, new wave's style was purely about being bright and shocking (see A Flock of Seagulls in all their sexless short-lived glory), but when stars like Boy George and Adam Ant got dolled up, it was more than just performance art. They were teen idols (unlike the glam rockers) and, in Ant's case, a bona fide sex symbol, setting fashion trends while challenging the strict definition of masculinity and what "gay" and "straight" were supposed to look like. It was, in a sense, the first metrosexual movement.

Text C

The Eighties was probably one of the best times to be alive. It was a creative revolution with music being the voice of people who had no words. The song lyrics had meaning that transcended race. Songs could express what people were afraid to say in their interactions. Guys could also find a song that could say what he had no words to say to the woman he was trying to win over.

When I hear an Eighties song on the radio, I often go look it up, see who sung it, the year it was in the charts and the position it got to. Time and time again, I think that I've just heard a former number 1, but when I check it, I find that it only got to #9 in the charts. It should really have got to number 1. Man, many of those which didn't make number 1, I think, should have. Was there really that much great music around all at the same time?

My kids actually love Eighties music and they know the lyrics as well as I do. Those were the days!!! :)

I just loved the synth styles music, the big hairdos, the bright clothes, and of course, the 'we don't actually look stupid doing this, do we?' attitude.

My only regret is that my children will never have the opportunity to experience such a unique, transient, influential and fun time. We may never see the likes again.

Text D

Do today's politicians supply plenty of colourful material for satire, asks Anoosh Chakelian, or are they all too dull to send up?

“He was just so boring” - This is the stark conclusion of one member of Ed Miliband’s team regarding their boss’ performance in the media eye at the very beginning of his opposition leadership.

But what does this mean for Britain’s satirists and sketchwriters, shouldering the peculiar burden of deriving humour from an emerging, and seemingly increasingly uniform, political class? Is their ink beginning to dry out?

Is this because we lack great frontline characters in the image of Margaret Thatcher, Willie Whitelaw, Michael Heseltine, Denis Healey, Winston Churchill, Benjamin Disraeli, and even Tony Blair, who, arguably, pioneered this era of slick parliamentary blandness? Big personalities like Dennis Skinner and Peter Tapsell merely rumble from the depths of the backbenches.

So, these times of increasingly slippery and subtle dramatic personae on the Parliament channel seem to be making room for a certain more cynical type of comedy, both in print and onscreen, rather than the traditional impersonations and flamboyant parodies of the past. The humour isn’t dead yet.

Text E

Rock criticism has two schools of thought regarding the Eighties. One complains that it was all crass, commercial rubbish, breathing a sigh of relief that we made it through that Seventies (thanks to IRS, SST, jangle pop, college rock, and hardcore punk, of course). The other celebrates the music as "cheesy" fun, full of naïve, silly singles; bad haircuts; big synthesizers. It's a school intent on reducing it all as nostalgic fodder -- and whenever ' Eighties music is written about in this fashion, it's always given ironic adjectives, straight out of the height of valley girl speak.

All this ghettoizes an era in pop music that was rich in innovation, great one-hit wonders, oddities, and inexplicable flukes that make it a wonderful cross between the first days of the British Invasion and the peak of AM pop in the early Seventies. It was the last great era for pop singles -- the last time that singles really mattered, the last time that something totally unexpected could capture the minds of the public, before radio consolidation meant hits couldn't build in a region, before MTV turned to non-music programming and cut off a national outlet for new music.

Appendix 9: Pilot Study 2.1 – Results

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	3 2	3	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	3

D

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	3	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	3	2	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	2

c

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	3	1	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	2	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	3	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	2	1	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	3	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	2	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	2	1	3	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	2	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	3	4	4

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	3	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	3	2	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	3	3	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	4	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	4	3	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	2	2	4	2

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	2	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	2	3	4	1

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	1	4	2	4	1

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	2	3	4	3

	Text 1	Text 2	Text 3	Text 4	Text 5
Rating (1 to 4)	2	3	2	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	1	1	3	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	2	1	4	3

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	1	2	1	4	2

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	2	1	4	2

D

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	2	2	4	4	~~4~~ 1

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	3	4	2	4	3

L

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	3	3	1	4	2

5

	Text A	Text B	Text C	Text D	Text E
Rating (1 to 4)	3	2	2	3	4

Appendix 10: Experiment 1 - Results

Explanation of the Results

Each column (except the first column) represents an iteration of the DurTeC system with a unique algorithm applied.

The left hand column identified what each row represents. These rows are:

1. Default Synonym List. A coloured bar is shown in each column where this is applied.
2. The CNSynonym List. A coloured bar is shown in each column where this is applied.
3. The Percentage Rule. A coloured bar is shown in each column where this is applied.
4. The use of a User-Defined Word List. A coloured bar is shown in each column where this is applied.
5. Secondary Sources. The default is using two additional secondary sources. A coloured bar is shown in each column where this is applied.
6. The Algorithm Identifier. This is an identification number given to each iteration DurTeC undertook.

7-16. The Top 10. These are the top 10 abstracts that the DurTeC system identified based on their similarity metrics.

17. Top 10 Present. This is a metric identifying how many of the Researcher's top 10 abstracts where present in DurTeC's top 10.

17. Blank.

19-27. The Bottom 10. These are the bottom 10 abstracts that the DurTeC system identified based on their similarity metrics.

29. Bottom 5 Present. This is a metric identifying how many of the 5 unrelated abstracts where present in DurTeC's bottom 10.

As an example, 'Execution 3' employed the Default Synonym List, indicated by a grey bar, and a 2% Percentage Rule, indicated by a yellow bar, and the term '2%'.

Using 'Execution 3' again, it highlighted that the top 10 abstracts were 2, 27, 29, 26, 15, 20, 31, 25, 4 and 6. It can be seen that abstracts 2, 4 and 6 have been highlighted. This is because these three were all selected by the Researcher as those most comparable with the source Abstract. The figure 3 can be seen on the 'Total top 10 present' row to signify this.

In the Bottom 10, only 2 of the unrelated abstracts have been identified, namely Abstracts 35 and 37.The figure 2 can be found at the bottom of this column to signify this.

Default Synonym list									
CNSynonym List									
Percentage Rule		2%	3%	5%		2%	3%	5%	
Defined Word List									
Secondary Sources									
Algorithm Identifier	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	15	15	15	2	27	2	2	2	2
	22	2	2	22	2	27	27	27	1
	20	27	27	15	20	29	29	1	15
	2	20	29	7	31	26	15	10	20
Top 10	27	4	4	27	15	15	26	22	7
	7	1	7	9	29	20	20	15	27
	29	29	20	12	26	31	31	29	22
	26	7	22	26	28	25	28	7	4
	28	22	1	4	19	4	36	19	9
	1	3	12	1	7	6	4	28	26
Total top 10 present	**3**	**5**	**4**	**5**	**2**	**3**	**2**	**4**	**5**
	39	13	13	5	35	13	13	30	31
	36	31	31	11	30	16	30	25	39
	31	17	17	16	14	35	23	14	23
	38	23	23	17	16	30	34	5	17
Bottom 10	18	39	11	18	17	23	16	11	13
	13	11	18	23	34	17	37	16	38
	37	18	34	34	13	14	14	17	18
	35	34	35	35	18	37	18	23	34
	11	35	37	37	37	18	17	34	35
	34	37	39	39	11	11	11	37	37
Total bottom 5 present	**5**	**3**	**3**	**3**	**2**	**2**	**1**	**1**	**4**

2%	3%		2%	3%	5%		2%	3%
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
2	2	2	2	2	2	15	15	15
15	15	15	27	27	4	2	2	2
22	22	27	4	29	6	20	20	20
4	4	20	29	15	1	22	27	27
27	27	4	20	4	27	27	4	4
3	3	29	1	20	22	7	29	29
20	20	22	15	25	36	29	22	22
1	1	1	6	6	15	4	1	1
9	9	7	25	1	29	1	3	3
7	7	25	24	24	20	9	9	9
6	**6**	**4**	**4**	**4**	**4**	**5**	**5**	**5**
38	38	18	18	18	18	16	13	13
31	31	11	35	35	38	39	38	38
11	11	39	34	34	13	31	31	31
16	16	17	11	38	11	38	23	23
18	18	23	17	13	16	13	39	11
23	23	35	38	17	17	18	11	18
34	34	34	39	39	23	37	18	34
35	35	38	13	23	35	35	34	35
37	37	13	23	11	37	11	35	37
39	39	37	37	37	39	34	37	39
4	**4**	**5**	**5**	**4**	**4**	**4**	**4**	**4**

5%		2%	3%	5%		2%	3%	5%
19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
2	31	2	2	31	2	2	2	2
15	27	31	27	2	15	4	4	4
22	2	27	31	27	20	29	29	1
4	20	20	20	6	27	27	27	27
27	15	29	29	24	4	20	20	6
3	29	15	15	32	29	15	15	29
20	24	26	26	8	1	1	1	15
1	32	24	24	15	22	24	24	24
7	26	32	6	19	24	6	6	20
9	19	6	32	20	28	25	25	32
6	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**4**
38	30	30	39	13	16	17	17	35
31	35	13	13	17	35	16	16	18
11	14	17	17	38	11	18	18	11
16	17	23	23	14	18	11	11	34
18	34	14	14	35	39	34	34	38
23	16	35	35	23	23	38	38	16
34	13	16	16	37	38	39	13	13
35	37	37	37	18	34	13	39	39
37	18	18	18	16	13	23	23	23
39	11	11	11	11	37	37	37	37
4	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**4**

	2%	3%	5%		2%	5%		2%
28	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
2	2	2	2	2	31	31	2	4
1	27	27	3	1	2	24	1	6
7	3	3	1	31	15	8	4	2
15	7	7	27	26	32	27	7	12
22	1	1	4	27	27	2	22	9
27	22	22	22	15	6	6	20	24
9	9	4	9	29	8	32	6	25
20	4	9	7	32	24	15	9	27
4	12	12	10	28	29	20	24	3
26	10	10	12	20	28	4	27	32
5	**7**	**7**	**6**	**2**	**3**	**4**	**6**	**5**
36	38	38	16	23	23	17	36	18
14	31	31	17	11	11	38	17	13
31	11	11	18	17	17	14	18	16
17	16	16	23	14	14	23	39	28
13	18	18	31	35	13	37	14	38
18	23	23	34	13	35	35	23	11
34	34	34	35	39	39	39	35	23
35	35	35	37	18	18	18	13	35
38	37	37	38	37	37	16	38	37
37	39	39	39	38	38	11	37	39
4	**4**	**4**	**4**	**4**	**4**	**4**	**5**	**4**

2%	3%	5%		2%	3%	5%			
							Top 3 occ	Top 5 occ	Top 7 occ
36	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
4	4	2	2	4	4	4	15	15	15
6	6	3	1	2	2	2	22	7	7
2	2	1	20	15	15	15	2	2	2
12	12	27	4	20	20	20	20	20	20
9	9	4	15	6	6	6	27	22	22
24	24	22	24	24	24	24	7	27	27
25	25	9	29	25	25	25	1	1	4
27	27	7	27	27	27	27	9	4	1
3	3	10	6	32	32	32	29	10	10
32	32	12	26	29	29	29	4	26	28
5	**5**	**6**	**4**	**3**	**3**	**3**	**5**	**5**	**5**
18	18	16	14	18	18	18	39	36	17
13	13	17	18	34	34	34	36	16	36
16	16	18	11	38	38	38	31	39	16
28	28	23	34	35	35	35	38	31	39
38	38	31	35	16	16	16	13	38	38
11	11	34	39	13	13	13	18	13	13
23	23	35	38	11	11	11	37	18	18
35	35	37	23	23	23	23	35	37	37
37	37	38	13	37	37	37	11	35	35
39	39	39	37	39	39	39	34	34	34
4	**4**	**4**	**4**	**4**	**4**	**4**	**5**	**5**	**5**

		2%				
Top 10 occ	Top 15 occ	Top 15 occ	Top 20 occ	Top 20 occ		
				+5 additional	Source only	+1 additional
46	**47**	**48**	**49**	**50**	**51**	**52**
7	15	7	15	2	15	1
15	7	15	7	1	22	15
10	10	4	4	5	20	7
27	4	10	9	4	2	22
2	27	3	10	3	27	2
4	3	27	20	15	7	20
20	20	1	27	7	29	27
1	1	2	1	20	26	9
3	9	9	6	27	28	26
22	2	20	3	22	1	4
6	**7**	**7**	**7**	**6**	**2**	**5**
14	16	14	16	14	39	14
11	11	11	36	23	36	11
39	39	34	11	39	31	17
36	34	36	39	18	38	31
38	36	39	34	34	18	34
13	18	13	18	13	13	13
18	38	18	38	36	37	38
37	13	38	13	38	35	37
35	37	35	37	35	11	18
34	35	37	35	37	34	35
5	**5**	**5**	**5**	**5**	**5**	**3**

+2 additional	+3 additional	+4 additional	+5 additional
53	**54**	**55**	**56**
2	2	2	5
1	1	1	2
7	3	4	1
15	7	3	4
22	15	7	3
27	22	9	7
9	9	15	15
20	27	27	27
4	20	22	9
26	4	20	6
5	**6**	**6**	**8**
36	5	39	14
14	36	23	18
31	31	31	23
17	17	17	39
13	13	36	34
18	18	18	13
34	34	13	36
35	35	35	35
38	38	38	37
37	37	37	38
4	**4**	**5**	**5**

Appendix 11: Additional Comparison – Results

Default Synonym list									
CNSynonym List									
Percentage Rule		2%	3%	5%		2%	3%	5%	
Defined Word List									
Secondary Sources									
Algorithm Identifier	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Position 1	10	10	10	10	10	10	10	10	2
Position 2	8	2	2	2	8	8	8	8	1
Position 3	7	7	7	7	5	7	7	7	10
Position 4	2	5	5	4	7	5	5	5	8
Position 5	5	8	8	8	3	3	3	2	7
Position 6	3	4	4	5	1	2	2	4	5
Position 7	4	3	3	3	9	1	1	1	3
Position 8	9	6	1	1	4	4	4	3	4
Position 9	6	1	6	6	2	6	6	6	9
Position 10	1	9	9	9	6	9	9	9	6
Abstract 1	10	9	8	8	6	7	7	7	2
Abstract 2	4	2	2	2	9	6	6	5	1
Abstract 3	6	7	7	7	5	5	5	8	7
Abstract 4	7	6	6	4	8	8	8	6	8
Abstract 5	5	4	4	6	3	4	4	4	6
Abstract 6	9	8	9	9	10	9	9	9	10
Abstract 7	3	3	3	3	4	3	3	3	5
Abstract 8	2	5	5	5	2	2	2	2	4
Abstract 9	8	10	10	10	7	10	10	10	9
Abstract 10	1	1	1	1	1	1	1	1	3
Student's Opinion									
Abstract 8	2	2	2	2	2	2	2	2	6
Abstract 4	4	2	2	2	4	4	4	4	3
Abstract 2	5	5	5	5	3	5	5	5	8
Abstract 1	1	5	5	4	6	4	4	4	7
Abstract 5	0	3	3	3	2	2	2	3	2
Abstract 6	3	2	2	1	5	4	4	2	1
abstract 3	1	0	0	0	6	2	2	2	0
Abstract 7	2	1	6	6	3	3	3	4	3
Abstract 10	4	9	4	4	8	4	4	4	1
Abstract 9	8	0	0	0	3	0	0	0	3
Degrees of Separation	30	29	29	27	42	30	30	30	34

2%	3%		2%	3%	5%		2%	3%	5%
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
10	10	10	10	10	10	10	10	10	10
7	7	7	7	7	7	8	7	7	7
2	8	6	3	3	3	7	2	2	2
8	2	3	6	6	6	2	8	8	8
4	4	8	5	5	8	5	5	5	4
5	5	5	2	2	5	4	4	4	5
3	3	9	8	8	2	3	3	3	3
1	9	4	4	4	4	9	9	9	9
9	6	2	9	9	9	6	6	6	6
6	1	1	1	1	1	1	1	1	1
8	10	10	10	10	10	10	10	10	10
3	4	9	6	6	7	4	3	3	3
7	7	4	3	3	3	7	7	7	7
5	5	8	8	8	8	6	6	6	5
6	6	6	5	5	6	5	5	5	6
10	9	3	4	4	4	9	9	9	9
2	2	2	2	2	2	3	2	2	2
4	3	5	7	7	5	2	8	4	4
9	8	7	9	9	9	8	4	8	8
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	4	3	3	3
0	6	4	1	1	1	5	1	1	1
7	1	2	5	5	5	1	7	7	7
1	1	3	0	0	3	0	0	0	1
1	1	1	4	4	1	2	2	2	1
0	0	6	5	5	1	0	0	0	0
6	2	3	3	3	3	2	2	2	2
1	4	8	1	1	1	4	4	4	4
3	8	8	8	8	8	8	8	8	8
24	28	40	32	32	28	28	29	29	29

	2%	3%	5%		2%	3%	5%		2%
20	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**
10	10	10	10	10	10	10	10	2	2
8	7	7	7	7	7	7	7	1	7
7	8	8	8	6	6	6	6	10	10
5	5	5	5	3	3	3	8	8	5
3	9	9	9	8	8	8	3	7	8
9	3	3	3	9	2	2	2	5	4
1	6	1	1	5	5	5	4	3	1
6	1	2	2	2	9	4	5	4	3
4	2	6	6	4	4	9	9	9	6
2	4	4	4	1	1	1	1	6	9
7	8	7	7	10	10	10	10	2	7
10	9	8	8	8	6	6	6	1	1
5	6	6	6	4	4	4	5	7	8
9	10	10	10	9	9	8	7	8	6
4	4	4	4	7	7	7	8	6	4
8	7	9	9	3	3	3	3	10	9
3	2	2	2	2	2	2	2	5	2
2	3	3	3	5	5	5	4	4	5
6	5	5	5	6	8	9	9	9	10
1	1	1	1	1	1	1	1	3	3
2	2	2	2	2	2	2	2	6	6
4	3	3	3	3	3	3	3	3	3
5	6	6	6	4	4	4	4	8	8
4	4	4	4	2	2	2	7	7	4
2	4	4	4	3	3	3	2	2	3
3	3	3	3	3	4	4	4	1	2
2	3	2	2	3	3	3	1	0	2
1	6	5	5	5	2	3	2	3	4
6	8	4	4	6	6	1	1	1	4
7	5	5	5	8	8	8	8	3	0
36	44	38	38	39	37	33	34	34	36

3%	5%		2%	5%		2%	3%	5%	
30	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**
2	2	3	10	10	10	7	6	6	1
7	7	10	7	7	7	6	7	3	2
5	5	2	8	8	3	3	3	7	10
10	10	7	9	9	6	10	2	2	7
8	8	8	5	5	5	2	4	5	6
4	1	6	1	3	2	4	5	8	8
1	4	9	3	2	8	5	8	1	3
3	3	1	6	1	4	8	10	4	9
6	9	5	2	6	9	1	1	10	4
9	6	4	4	4	1	9	9	9	5
7	6	8	6	8	10	9	9	7	1
1	1	3	9	7	6	5	4	4	2
8	8	1	7	6	3	3	3	2	7
6	7	10	10	10	8	6	5	8	9
3	3	9	5	5	5	7	6	5	10
9	10	6	8	9	4	2	1	1	5
2	2	4	2	2	2	1	2	3	4
5	5	5	3	3	7	8	7	6	6
10	9	7	4	4	9	10	10	10	8
4	4	2	1	1	1	4	8	9	3
6	6	5	2	2	2	1	2	2	7
3	3	6	3	3	3	2	3	1	2
3	3	0	6	6	1	1	1	5	8
9	9	6	8	8	5	9	1	1	6
3	3	3	0	0	0	3	1	0	1
2	5	0	5	3	4	2	1	2	2
2	1	6	0	1	5	2	5	2	0
4	4	6	1	6	3	1	3	3	2
4	1	5	8	4	1	9	9	0	6
0	3	5	5	5	8	0	0	0	3
36	38	42	38	38	32	30	26	16	37

2%	3%	5%						2%
			Top 3 occ	Top 5 occ	Top 7 occ	Top 10 occ	Top 15 occ	Top 15 occ
40	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
10	10	10	10	10	10	10	10	10
7	7	7	8	8	8	8	8	7
6	6	6	7	7	5	5	7	8
8	8	8	5	5	7	7	5	2
3	3	3	2	2	2	2	2	5
2	2	2	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4
9	9	5	9	9	9	9	9	9
5	5	9	6	6	6	6	6	6
1	1	1	1	1	1	1	1	1
10	10	10	10	10	10	10	10	10
6	6	6	5	5	5	5	5	4
5	5	5	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7
9	9	8	4	4	3	3	4	5
3	3	3	9	9	9	9	9	9
2	2	2	3	3	4	4	3	2
4	4	4	2	2	2	2	2	3
8	8	9	8	8	8	8	8	8
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
3	3	3	4	4	4	4	4	3
4	4	4	5	5	3	3	5	6
7	7	7	4	4	6	6	4	1
2	2	2	3	3	3	3	3	0
4	4	4	3	3	3	3	3	3
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
5	5	1	4	4	4	4	4	4
8	8	8	8	8	8	8	8	8
38	38	34	36	36	36	36	36	30

Top 20 occ	Top 20 occ						
	+5 additional	Source only	+1 additional	+2 additional	+3 additional	+4 additional	+5 additional
49	50	51	52	53	54	55	56
10	3	10	1	2	3	3	5
8	5	8	10	1	1	4	3
7	2	7	8	10	2	2	2
2	4	5	7	8	10	1	4
5	1	2	2	7	8	10	1
3	10	3	5	5	7	8	7
4	7	4	3	3	4	7	8
9	8	9	4	4	5	5	7
6	6	6	9	9	6	6	6
1	9	1	6	6	9	9	9
10	5	10	1	2	2	4	5
4	3	5	5	1	3	3	3
6	1	6	7	7	1	1	2
7	4	7	8	8	7	2	4
5	2	4	6	6	8	8	1
9	9	9	10	10	9	9	9
3	7	3	4	5	6	7	6
2	8	2	3	4	5	6	7
8	10	8	9	9	10	10	10
1	6	1	2	3	4	5	8
2	5	2	7	6	5	5	3
4	1	4	6	3	3	0	1
5	0	5	6	8	0	0	0
1	3	4	6	7	9	0	3
0	4	3	3	2	3	5	4
3	4	2	1	1	1	2	1
1	4	1	0	0	1	4	5
2	1	2	3	3	2	2	0
4	4	4	1	1	4	4	4
8	0	8	3	3	0	0	0
30	26	35	36	34	28	22	21