Australian Government
Department of Defence
Defence Science and
Technology Organisation

The Use of a Context-Based Information Retrieval Technique

Kathryn Parsons, Agata McCormac, Marcus Butavicius,
Simon Dennis* and Lael Ferguson

Command, Control, Communications and Intelligence Division
Defence Science and Technology Organisation

*Ohio State University

ABSTRACT

Since users are faced with an ever increasing amount of data, fast and effective retrieval of required information is of vital importance. This study examined two methods of using Latent Semantic Analysis (LSA) to improve the results retrieved using a keyword-based technique using sentence or document context. Fifty participants retrieved information using a standard keyword technique and the two LSA techniques. Although the re-ranking provided by the LSA techniques ordered the documents in a significantly more efficient manner, no significant differences were found in user performance with regards to accuracy, time taken or documents accessed for the different techniques. However, individual differences did significantly influence results, most notably in regards to participants' scores on a comprehension test. This study therefore highlights the importance of examining the impact of individual differences in any information retrieval system.

RELEASE LIMITATION

Approved for public release

Report Documentation Page				*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUL 2009**	2. REPORT TYPE	3. DATES COVERED

4. TITLE AND SUBTITLE **The Use of a Context-Based Information Retrieval Technique**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **DSTO, , , , ,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited.**

13. SUPPLEMENTARY NOTES **The original document contains color images.**

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **57**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

APPROVED FOR PUBLIC RELEASE

The Use of a Context-Based Information Retrieval Technique

Executive Summary

The amount of information that users are required to process continues to rapidly grow, and this increases the requirement for an accurate and effective information retrieval tool. This is, however, a far from simple goal, and despite the extensive research in the area of information retrieval, an ideal tool remains elusive.

There are a number of complexities and ambiguities associated with the English language that result in difficulties associated with information retrieval. For instance, information retrieval tools must contend with obstacles such as polysemy, which refers to words with multiple meanings, and synonymy, which is used to describe multiple words with the same meaning.

Many of these problems can be minimised when the query is provided in context. Latent Semantic Analysis (LSA) is a statistical technique for inferring contextual and structural information, and previous studies have found promising correlations between LSA and human judgements of document similarity.

The aim of this study was to examine whether the results provided by a keyword based technique would be improved through the use of two LSA techniques. Participants were required to highlight query terms from within documents, and one LSA technique utilised the sentence of the query term, and the other LSA technique utilised the entire document. A baseline technique, in which results were not re-ranked, was also used.

Fifty participants were provided with a number of information retrieval questions, which involved retrieving the documents that would be useful if writing a hypothetical report on a specified topic. Using a counterbalanced repeated-measures design, participants utilised a customised interface, which retrieved and ranked documents using the three different techniques.

An analysis of the searches conducted by the users in the experiment revealed that, when utilising the LSA techniques, the relevant documents were significantly more likely to be placed towards the beginning of the retrieved list. Despite this, the LSA techniques were not associated with an advantage in terms of accuracy, time taken or documents accessed with respect to user performance. Instead, most participants accessed almost all of the documents in all retrieved lists, meaning that differences between the techniques had no impact on the participants' performance.

However, individual differences did influence results. Participants were required to complete a short comprehension test, and the participants who had higher scores on this test also tended to have better performance on the information retrieval task. The results

also indicated that LSA may compensate for the abilities of the participants who had lower comprehension scores, as there was far more variation across the techniques for the participants who did not perform well on the comprehension test, and very little variation across the techniques for the participants who performed well on the comprehension test.

This study therefore highlights the importance of testing the influence of individual differences on any IR system, and the importance of testing any IR tool on a population that closely reflects the intended users of the system. This study also suggests that tools such as LSA are unlikely to be necessary in relatively small document collections, as most participants are likely to use a brute force approach, in which all documents are accessed. It is hypothesised that such techniques will be far more useful in extremely large document collections, where it is impractical to access all documents.

Authors

Kathryn Parsons
Command, Control, Communications and
Intelligence

Kathryn Parsons is a research scientist with the Human Interaction Capabilities Discipline in C3ID where her work focuses on cognitive and perceptual psychology, information visualisation and interface design. She obtained a Graduate Industry Linked Entrepreneurial Scheme (GILES) Scholarship in 2005, with Land Operations Division, where she was involved in human factors research, in the Human Sciences Discipline, specifically in the area of Infantry Situation Awareness. She completed a Master of Psychology (Organisational and Human Factors) at the University of Adelaide in 2005.

———————————————   ————————————————————————————————

Agata McCormac
Command, Control, Communications and
Intelligence

Agata McCormac joined DSTO in 2006. She is a research scientist with the Human Interaction Capabilities Discipline in C3ID where her work focuses on cognitive and perceptual psychology, information visualisation and interface design. She was awarded a Master of Psychology (Organisational and Human Factors) at the University of Adelaide in 2005.

———————————————   ————————————————————————————————

Marcus Butavicius
Command, Control, Communications and
Intelligence

Marcus Butavicius is a research scientist with the Human Interaction Capabilities Discipline in C3ID. He joined LOD in 2001 where he investigated the role of simulation in training, theories of human reasoning and the analysis of biometric technologies. In 2002, he completed a PhD in Psychology at the University of Adelaide on mechanisms of visual object recognition. In 2003 he joined ISRD where his work focuses on data visualisation, decision-making and interface design. He is also a Visiting Research Fellow in the Psychology Department at the University of Adelaide.

———————————————   ————————————————————————————————

Simon Dennis
Ohio State University

Simon Dennis, PhD. is currently an Associate Professor at Ohio State University. Before moving to Columbus in 2007, he held positions at the University of Adelaide, University of Colorado and the University of Queensland. He was awarded his PhD in 1993 from the Department of Computer Science at the University of Queensland. Dr. Dennis has been awarded a series of grants as well as both defence and industry contracts in the areas of mathematical memory modelling, psycholinguistics and usability. In addition, he is published in many of the field's most prestigious journals including the Proceedings of the National Academy of Sciences, Neuropsychologia, Trends in Cogntiive Sciences and Psychological Review. In joint work with the Distributed Systems Technology Centre, he has also made a significant contribution in the area of information retrieval including papers in the Journal of the American Society for Information Science and Technology (JASIST) and the Special Interest Group on Information Retrieval (SIGIR).

Lael Ferguson
Command, Control, Communications and Intelligence

Lael Ferguson graduated from the University of South Australia in 1997 with a Bachelor of Applied Science (Mathematics and Computing) and began working for the Department of Defence in Canberra as a software developer. In 1999 she transferred to Geraldton and worked as a system administrator. In 2000 she transferred to the Defence Science Technology Organisation at Edinburgh as a system administrator/software developer, managing a computing research laboratory, and developing concept demonstrators and experimental software.

Contents

1. Introduction

Due to the expanse of information available, users are often overwhelmed by the challenge of finding relevant documents. Essentially, the amount of information readily available is so extensive and continues to grow at such a rate that it is often neither practical nor possible for a user to read the full text of all available documents (Carlson, 2004). Information retrieval tools can be used to ensure that users receive the most appropriate and relevant information, which can assist in reducing information overload. The overall goal is to increase the likelihood that the user will obtain relevant documents without having to search the whole collection.

Simply put, information retrieval refers to the process of finding material that satisfies an information need. Information retrieval does not inform or change the knowledge of a user, but rather, only reveals the existence (or non existence) of documents relating to a request (Lancaster, 1968). In other words, information retrieval will not answer a specific question, but will retrieve the documents that could be used to answer that question.

Typically, this will involve some form of keyword search over a document collection, and research suggests that keyword searching is one of the most effective retrieval techniques (Navigli & Velardi, 2003; Guo, Shao, Botev & Shanmugasundaram, 2003). However, in extremely large document collections, the number of documents retrieved by a keyword search can be unmanageably large, and the process of checking through the documents to find those that are relevant can be very time consuming.

Furthermore, keyword searches can suffer from the problem of synonymy, where words have multiple meanings and can take on multiple roles, only a subset of which may be relevant for the particular search (Ravin & Leacock, 2000). This means that a keyword search can result in the retrieval of many irrelevant documents.

Information retrieval can be further complicated by problems associated with developing an appropriate query. According to Ruthven, Tombros and Jose (2001) the formulation of a query can be very demanding, particularly in cases where the user is inexperienced with information retrieval or is unfamiliar with the document collection. Similarly, when the information need is vague, the formulation of a query can be extremely difficult (Ruthven et al., 2001). Users' queries often fail to fully describe the information need and they are commonly very short and ambiguous (Fonseca, Golgher, Pôssas, Ribeiro-Neto & Ziviani, 2005). According to Fonseca and colleagues (2005), to ensure better information retrieval, it is necessary to improve query formulation.

In order to improve query quality, the current study required participants to highlight their search terms from within the documents. Although this process limits the possible queries, it is thought to decrease problems associated with formulating query terms, as the user can select terms from within relevant documents. This process also eliminates problems associated with spelling or grammatical errors, as the selected term (or terms) come directly from the document, in the required context.

The aim of this experiment was to evaluate whether the results of a simple non-context search could be improved by re-ranking the results using the context provided by the surrounding terms. A non-context search was tested against two contextualised models, which used Latent Semantic Analysis (LSA) in different ways. One model involved using LSA on the sentence surrounding the search term, and the other involved using LSA on the context provided by the whole document (LSA will be described in more detail in a subsequent section).

1.1 Performance Measures

The performance of information retrieval tools is usually measured via recall and precision. As shown in the equations below, recall is defined as the proportion of retrieved relevant documents out of all relevant documents available, and precision is defined as the proportion of relevant items retrieved out of all retrieved items. In other words, recall is associated with the proportion of relevant items that are retrieved, and precision is associated with the proportion of retrieved items that are relevant.

$$Recall = \frac{\#\,(relevant\ items\ retrieved)}{\#\,(relevant\ items)}$$

$$Precision = \frac{\#\,(relevant\ items\ retrieved)}{\#\,(retrieved\ items)}$$

Information retrieval systems aim to maximise both of these measures, and generally, there is a trade off between the two, meaning that as the recall of a system increases, the precision tends to decrease. Although these measures are widely used, according to Spink and Wilson (1999), the measurement provided by precision and recall often fails to provide an accurate indication of the value of a retrieval tool.

An example may illustrate this point. Consider a document collection with 200 documents, 60 of which have been deemed to be relevant to a particular query. If two information retrieval systems retrieve 60 documents, including 30 relevant documents, this means that both of these systems have a recall and precision of 50%. However, presume that one system retrieved the 30 relevant documents first, and the other system retrieved 30 irrelevant documents before the 30 relevant ones. In this example, the retrieval tool that ranked the relevant documents first would be far more advantageous, but the recall and precision measures fail to provide this information.

There are a number of similar measurement techniques that attempt to add to the value of conventional precision and recall scores. For example, 'cut-off precision' involves the precision of a system being measured after a certain number of documents have been retrieved. For example, for web-based retrieval systems, precision is often measured after ten

documents have been retrieved. If this technique was used in the example above, the first information retrieval system would have a precision of 100% and the second would obtain a score of 0%. Therefore, although this technique reveals vital information regarding the initial performance of a system (which is often most important), it fails to provide information regarding the systems' eventual performance or overall performance.

Other measurement techniques include interpolated precision, in which the precision of a system is measured at various levels, such as every 10% of documents, and uninterpolated precision, in which the precision is measured after every document has been retrieved. Precision can also be measured via the mean average precision (MAP) score, which is based on the precision score after the retrieval of each relevant document. Typically, these scores are then averaged to obtain an overall precision score that is thought to more accurately reflect the systems' performance throughout.

Information retrieval performance can also be measured using the F-measure, which combines recall and precision with an equal weight (Yang & Liu, 1999). The F-measure arguably provides a more accurate quantification of a system, as it takes into account both recall and precision. Hence, if either the recall or precision scores are very low, the combined score will also be poor. This measure is a harmonic mean, and the formula is provided below.

$$\text{F-measure} = \frac{2 \times (\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$

Although these measures could be very useful in some situations, Spink and Wilson (1999) claim that current measures of information retrieval performance lack validity. This is particularly true in situations where the document collection is unfamiliar. Essentially, when dealing with an unfamiliar collection, there is no knowledge of the total number of relevant items, and therefore, it is difficult to quantify performance.

Moffat and Zobel (2008) attempt to resolve this issue with a new metric, referred to as rank-biased precision (RBP). This measurement technique is flexible, and can therefore be utilised in the presence of uncertainty, such as situations with partial relevance judgements or unjudged documents (Moffat & Zobel, 2008). Furthermore, the technique also takes into account user behaviour. The calculation includes a value referred to as persistence (or probability) p, which is a number between 0 and 1 that reflects the patience of the user. In situations where the user is highly impatient, a small p value (less than 0.5) can be used to place the emphasis on early ranked documents. In contrast, if the user is highly persistent, and is interested in finding every relevant document, a high p value (close to 1.0) would be suitable. The formula (where r_i is the relevance of the ith ranked document and d indicates depth) (Moffat & Zobel, 2008) is provided below. In this context, depth refers to the place in a list of ranked documents (i.e., a d value of 4 refers to the fourth document in a ranked list).

$$RBP = (1 - p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1}$$

Tague and Schultz (1989) suggest that 'informativeness' is the best method of measuring the performance of an information retrieval tool, and Su (1998) claims that information retrieval performance should be based on the value of the search results as a whole. However, it is unclear exactly how 'informativeness' or 'value' would be measured with these methods (Spink & Wilson, 1999), which limits their usefulness as actual measurement techniques.

Finally, the evaluation of an information retrieval system would be incomplete without information regarding the speed of the system and the time required for preparing the system for use. For example, a system could be exceptionally effective at retrieving the correct documents, but if it is extremely slow and takes a lot of processing power, then a system that retrieves documents with slightly lower accuracy, but with a faster processing time, could be more beneficial in a real-world environment.

1.2 Properties of Information Retrieval Systems

There are various information retrieval systems that work in different ways to assist users to satisfy their information need. Most of these systems are generally based upon a number of different hypotheses. For example, the relevance hypothesis is based on the idea that a document is more likely to be relevant to a request if it has more matching descriptive items (Sparck Jones, 1971). In other words, a document containing ten terms that match a query is more likely to be relevant than a document that only matches one of the query terms.

This hypothesis must also take into account inverse document frequency and term weighting, which is related to the number of times that a word is used in a particular document, relative to the number of times that it is used in a corpus (Sparck Jones, 1972). This therefore involves discriminating the importance of a term, and allows for the fact that some words are more influential than others (Fang, Tao & Zhai, 2004). For example, there are many terms that occur regularly in all documents, and those terms are therefore of little use to a retrieval system. In order to resolve this, information retrieval systems often remove words on a stop list, such as 'a', 'the', 'of', 'by' and 'to'. It has been shown that by using a stop list between forty and fifty percent of the total number of words in a document can be removed (Salton & McGill, 1983). It is also important that these measures are normalised for length, as a short document that contains a query word five times could be far more useful than an extremely long document that uses the query word ten times.

Systems are also often based upon the associated hypothesis, which suggests that if one term is able to retrieve relevant documents then any other closely related term should also be able to retrieve relevant documents (van Rijsbergen, 1979). For example, if the term 'car' is useful, then it is highly likely that the associated term 'automobile' could also retrieve relevant information. A further hypothesis is the cluster hypothesis. This is based on the idea that if a relevant document is found, then any closely associated document is likely to be relevant to

the same queries (van Rijsbergen, 1971). Hence, two highly similar documents are likely to be relevant to similar requests.

1.2.1 Word Stemming

Another useful property of many information retrieval systems is referred to as 'word stemming'. This basically involves truncating related words to a common stem or root word, such that, for example, a search for the word 'kick' will also result in the retrieval of documents that use the words 'kicked', 'kicks' and 'kicking'. The aim is to ensure that the retrieval system will not miss relevant information.

Stemming can reduce vocabulary by between ten and fifty percent, leading to an increase in recall (van Rijsbergen, 1979). This increase in recall can be desired if the truncated words remain relevant to the original query. However, in some cases, word stemming can result in the retrieval of irrelevant information. For example, the word 'army' will be truncated into 'arm', but these words have very different meanings, and hence, in this case, the stemmer is unlikely to improve information retrieval, but rather, will result in an increase in recall and a decrease in precision (Krovetz & Croft, 1992).

The most common stemmer is the Porter Stemmer (Porter, 1980). Although some research (Manning, Raghavan & Schutze, 2007) has found strong empirical evidence for the effectiveness of this stemmer, other research (Singhal, 2001) opposes this, instead claiming that stemmers result in a very small increase in search effectiveness. This research indicates that stemmers result in a large increase in recall, which then decreases precision, meaning that the user is faced with a large increase in irrelevant (and therefore frustrating) documents.

1.2.2 Ranked Retrieval

A further property of many effective information retrieval systems is ranked retrieval. This tends to increase the value of the results, by altering the order or position, so that the results that most closely represent the query are at the top of the retrieved list.

Effective ranking is particularly important for Internet search engines, as research has indicated that over 80% of users view only one page of results (Beitzel, Jensen, Chowdhury, Grossman & Frieder, 2004), and hence, if the most crucial results are not retrieved promptly, they are unlikely to be viewed. Internet search engines produce positioning information using factors such as the HTML code, link popularity and domain name (Carlson, 2004). However, in order to reduce plagiarism and manipulation, the specific details of these algorithms are closely protected and changed frequently (Carlson, 2004).

In simple ranked retrieval, the results tend to be ordered based on aspects such as the frequency of the occurrence of query keywords. More complicated retrieval systems aim to take into account aspects such as the importance of the search terms. For example, in a search for 'Microsoft Corporation' the more specific term 'Microsoft' would be weighted more highly than the general term 'Corporation'. Essentially, in most ranking algorithms, weights are assigned to the words in the queries and to the words in the document, and then these weights are compared, to produce a ranking of the importance of a specific document.

Examples of models that use ranked retrieval include vector space models and probabilistic models, and these models will be described in more detail in a subsequent section.

1.3 Challenges for Information Retrieval Systems

Information retrieval systems must contend with many obstacles associated with complexities in language and information need. For example, the context of an information need can be highly influential, but it is often very difficult for this context to be expressed in a search query (Swanson, 1988). Associated with this, as already highlighted in this report, choosing appropriate search terms to meet an information need can be an extremely difficult task (Ruthven et al., 2001; Fonseca et al., 2005).

Information retrieval is also influenced by complexities associated with relevance. Most notably, the relevance or usefulness of a document can depend highly on the other documents that have been viewed (Swanson, 1988). Relevance can also be extremely subjective, and a document deemed to be relevant by one user may be judged quite differently by a different user (Parsons, McCormac & Butavicius, 2007). The evaluation of a retrieval system is also complicated as it is difficult to know how many relevant documents have been missed (Swanson, 1988). This is particularly true in large and unfamiliar document collections, as it is often practically impossible to review all documents to obtain an accurate and objective measure of the number of relevant documents.

The effectiveness of information retrieval systems is also complicated by the ambiguities that exist within language, which make it difficult to develop accurate queries (Ruthven, Lalmas & Rijsbergen, 2003). This problem is referred to as 'word mismatch', and it essentially occurs when users attempt to retrieve documents using words that do not match those used by the authors to describe the concepts (Xu & Croft, 2000). Studies by Furnas and colleagues (cited in Xu & Croft, 2000) examined word mismatch and discovered that, approximately 80% of the time, participants used a different term to describe the same object. Hence, there is a very high likelihood that a simple keyword search will fail to retrieve a large proportion of relevant documents.

This problem is associated with synonymy, which is used to describe multiple words with the same meaning (Ravin & Leacock, 2000). For example, a simple search using the word 'car' will fail to retrieve documents that use the term 'automobile'. Hence, relevant documents can be missed, which impacts on recall.

One of the most common word mismatch problems is referred to as polysemy, which refers to words with multiple meanings (Ravin & Leacock, 2000). For example, the word 'bank' can be used as a noun to refer to a river bank or a commercial bank, or it can be used as a phrasal verb, to refer to having confidence in or relying on someone. Hence, a keyword search using the word 'bank' will retrieve many irrelevant documents, which can then decrease precision (Krovetz, 1997).

The problem of word mismatch can be even more problematic in less structured and formal communication, such as emails or transcriptions of conversations. In these communication modes, there is a greater likelihood of spelling or grammatical errors, which could influence

the performance of an information retrieval system. Word mismatch can also occur when the user misspells a query term, and hence, the system will fail to retrieve relevant documents.

A system may also fail to retrieve relevant documents due to differences associated with American spelling and Australian spelling. For example, a search using the word 'organisation' is likely to miss documents that used 'organization'. These word mismatch problems are generally more severe for short queries as opposed to long queries (Xu & Croft, 2000).

Many of these problems can be minimised through the use of query expansion or clarification. The aim of query expansion is to reduce document mismatch by expanding the query, using words or phases with similar meaning (Xu & Croft, 1996). In contrast, query clarification determines which polysemous term is relevant to the search. An example of query clarification is provided in Figure 1.

Figure 1: An example of query clarification (Getty Images, Inc 2007)

This process increases the chances of finding relevant documents. However, although relevance is improved, the extra step can result in a large increase in time and an increase in cognitive load (Dennis, Bruza & McArthur, 2002). An increase in cognitive load is detrimental as it has been shown to increase fatigue and decrease learning and situation awareness (Dennis et al, 2002). This can potentially diminish the benefits gained through query expansion.

There are also a number of other complexities associated with word mismatch, which can increase the challenges facing information retrieval systems. For example, it is very difficult for a system to recognise the complexities associated with paraphrasing (Chang & Hsu, 1999). In other words, there are many different ways to say the same thing, and most systems do not have the necessary knowledge to recognise this.

The English language also has ambiguous sentences, which can further limit a system's ability to recognise meaning. For example, the sentence "I saw the man on the hill with the telescope" is ambiguous, as it is unclear who is holding the telescope (Simon, 1996, p.78). Hence, in a sentence such as this, the knowledge provided by the context is necessary to ascertain the full meaning.

Language is also complicated by anaphora, in which one expression refers to another expression (Ge, Hale & Eugene, 1998). For example, one sentence may describe an object, and the following sentence could then refer to the same object as 'it'. When both sentences are read together, the meaning is usually clear, but, when the second sentence is read independent of

the first, it can be difficult to interpret the meaning. This problem is particularly relevant for less formal communications. For instance, in an email communication, the actual topic of discussion may not be explicitly mentioned. Instead, it could be assumed knowledge, or the topic may have been mentioned in a previous communication. It is extremely difficult for an information retrieval tool to recognise relevance in such a situation.

Essentially, although many information retrieval systems are very advanced, they are still not advanced enough to recognise many of the complexities associated with language, and they do not have the same level of understanding as a human. Hence, these systems are almost always still less effective than human indexing (Swanson, 1988).

1.3.1 Challenges Associated with Individual Differences in Information Retrieval

It is also necessary for information retrieval tools to face challenges associated with the individual differences that exist between users. Users are likely to range widely in regards to factors such as technical knowledge, cognitive abilities, comprehension and personality (Dillon & Watson, 1996). Allen (1991) suggests that factors including previous knowledge, learning style and cognitive style can also influence users' search tactics. Furthermore, search effectiveness has been demonstrated to be influenced by logical reasoning ability (Allen, 1994). There is also evidence to indicate that age, academic background and gender can affect performance using information retrieval systems (Borgman, 1989).

These factors can result in dramatic differences in regards to user performance. For example, Chen and Dumais (2000) examined web search performance of 74 participants with intermediate experience and found an average reaction time of 52.3 seconds. However, there was an extremely large range in the results, with one participant taking only 22 seconds, and another taking 144 seconds. Hence, it can be challenging to develop a system that will result in effective performance for all users.

It is also necessary to note that studies assessing individual differences are generally highly dependent on context imposed by the specific system, and it is therefore difficult to generalise the findings from one study to others (Dillon & Watson, 1996). Consequently, for new systems, it is still necessary to assess the differences between users, as it is highly unlikely that stable individual differences will be found.

Despite this, there is evidence suggesting that certain design features or characteristics may optimise performance for some individuals (Allen, 2000). Stanney and Salvendy (1995) use two approaches referred to as 'capitalization' and 'compensatory', in which some features may capitalise on the skills of individuals with higher abilities, and other features may compensate for the lower levels of ability in other users.

This therefore highlights the importance of an adequate analysis of individual differences in the development of information retrieval systems. Appropriately designed systems and tailored training for users should increase the likelihood that a system will maximise users' skills (Dillon & Watson, 1996).

1.4 Types of Information Retrieval Systems

Various theoretical models have led to the creation of a number of different information retrieval tools, which match and rank documents in a variety of ways (Liddy, 2005). These range from simple keyword searches, to complicated algorithms that analyse syntax to retrieve information based on word meaning.

1.4.1 Keyword Search

Most information retrieval tools are based, at least in some respect, on keyword searches, where the user is required to enter a query term (or terms). With a simple keyword search, the system then analyses the collection, and any documents containing the query word are retrieved. Research suggests that keyword searching is one of the most effective retrieval techniques (Navigli & Velardi, 2003; Guo et al., 2003).

However, there are a number of problems associated with keyword searches, and most of these problems are related to the ambiguities that exist within language, which make it difficult to develop accurate queries (Ruthven et al., 2003). For example, as already indicated in the previous section, language is complicated by word mismatch problems such as synonymy and polysemy.

However, users generally have a good knowledge of language and the associated limitations that language can create for keyword searching, and therefore, many of these problems can often be overcome. For example, when searching for a document on riverbanks, users would generally not search with the word 'bank' alone, as they would know that this search would retrieve irrelevant information regarding financial institutions. Hence, the problem of synonymy can be reduced through the use of real-world knowledge.

A further problem associated with keyword searching is related to the ease with which such content or keyword based systems can be beaten. Many of the first Internet search engines were based on simple keyword searches, where Web pages with more occurrences of the search term were ranked higher. However, poor or irrelevant Web pages interested in improving their ranking could take advantage of this, by including a number of lines with popular keywords repeated many times. This is often referred to as 'spamdexing', which essentially involves actions that aim to provide an unwarranted increase in a Web pages' relevance (Gyöngyi & Garcia-Molina, 2005). Hence, relying on word occurrence alone will not necessarily result in the retrieval of the best documents.

1.4.2 Boolean Search

Some of these problems can be minimised by using a Boolean Search, which uses principles of Boolean logic. Boolean logic is made up of three logical operators: OR, AND, and NOT. OR logic is most commonly used to search for similar terms and concepts. AND retrieves documents which contain more than one search term, and NOT excludes terms from a search. Using these logical operators is not as simple as it first appears as there are a number of problems associated with using Boolean queries.

For instance, using AND does not guarantee that the selected terms are actually used together. The system can instead retrieve words that are used in different sentences or paragraphs, which means that the document is not necessarily relevant to both of the terms. It is also very difficult to choose the best terms for an OR search. For example, if an individual is searching using the word 'money', there are many synonymous terms that can be used, including cash, currency, capital and funds. A user also has to exhibit caution when using NOT, because the term that a user wants to avoid may appear in documents that also contains the query term. In this example by using NOT relevant documents may be excluded.

Boolean queries are very precise and to be applied effectively, a user has to understand the syntax and semantics of Boolean queries. By understanding how the queries are applied, it is easier to formulate a query specific to an information need. Therefore, Boolean queries tend to be ineffective if users are unfamiliar with Boolean search methods. If a user has difficulty in articulating their needs then the search is unlikely to accurately reflect their expectations. This reflects the main failing of most information retrieval tools, which, in essence, is the users' inability to choose effective search terms.

Furthermore, Boolean queries fail to account for relevance; this means that retrieval is based on a binary decision, with no partial match and no ranking provided. This is a major concern because information retrieval is most effective when highly relevant documents are retrieved first. Due to the problems associated with Boolean queries, many information retrieval systems use the Vector Space Model.

1.4.3 Vector Space Model

The Vector Space Model is often used in IR, and it consists of three stages; document indexing, term weighting and the similarity ranking of documents. In the first stage of document indexing, non significant words, such as *and, this* and *is,* are removed from the document vector (Salton & McGill, 1983). This is usually done by using a stop list, which is important because it allows the document vector to be represented primarily by content bearing words.

To enhance the retrieval of relevant documents, weightings are then assigned to the indexed terms. To obtain optimal results for both precision and recall it is suggested that the best term weight schemes are achieved by using term frequency, length normalisation and inverse document frequency (Lee, Chuang & Seanoms, 1997). Finally the documents are ranked according to similarity. This is achieved by applying a comparison function, usually the cosine coefficient, which measures the angle between a document vector and the query vector (Salton, 1988). Therefore the most relevant documents are the documents whose vectors are closest to the query vector.

Essentially, the vector space model involves a degree of similarity between a query and a document, meaning that partial matches are taken into account, and documents are ranked by relevance. The two major challenges to consider when using the vector space model are selecting an appropriate set of base vectors and choosing an appropriate scheme for terms.

Examples of systems that use vector space modelling include SMART and Wide Area Information Servers (WAIS).

1.4.4 Latent Semantic Analysis

LSA, which is also known as latent semantic indexing (LSI), uses a statistical and mathematical technique for inferring contextual and structural information within words and sentences (Landauer, Foltz & Lahan, 1988; Deerwester, Dumais, Landauer, Fernass & Harshman, 1990). LSA is a fully automated process, and is predominantly used in information retrieval and document similarity.

LSA involves the creation of a term by document matrix, and weighting functions are then applied to this matrix. Essentially, the words of a corpus are represented in columns and the documents are represented in rows, creating a matrix of the document collection, showing the frequency with which each word occurs (Kintsch, 2001). LSA then applies Singular Value Decomposition (SVD) to the matrix to find the semantic dimensions in the document set. Basically, SVD is used to discard redundant information and focus only on essential semantic information (Kintsch, 2001).

LSA has many advantages over other techniques. It has been shown to outperform vector-based methods, in regards to precision and recall, and is often able to successfully address the problems associated with polysemy and synonymy (Papadimitriou, Raghavan & Tamaki, 1998). LSA has also been demonstrated to have comparable findings to some aspects of human performance, including judgements of essay quality, word recognition, word categorisation, sentence to word semantic priming and speech comprehension (Laudauer et al., 1998).

However, this technique does have a number of limitations. Most importantly, LSA uses a 'bag of words' approach, which means that it does not take into account the order of words (Wallach, 2006). Word order can be extremely useful and can reveal important information regarding the context of a sentence (Wallach, 2006). Although LSA has been shown to correlate well with human judgements, these correlations tend to be highly variable depending on differences between individuals, and they are greatly influenced by the selection of weight functions, factors retained, stopping and backgrounding (Pincombe, 2004).

1.4.5 Probabilistic Models

Probabilistic models of retrieval estimate the probability that a document will be relevant to a given query. Examples of probabilistic based systems include Cheshire II, Inktomi and INQUERY. The underlying assumption is that the terms in a relevant document are distributed differently to the terms in a non-relevant document (Fuhr, 1992).

Although probabilistic models are able to rank documents in order of their probability of being relevant, the model has three major disadvantages. First, the initial definition of what is and is not relevant is, of course, highly subjective. Second, the method ignores the frequency of the index term within a document, and finally, the model assumes that the index terms are independent.

1.4.6 Language Models

In contrast, natural language models apply algorithms that combine statistical information with semantic information. Semantic information is gathered by processing the language rather than treating each term independently; this enhances the indexing method and improves search precision, by reducing the retrieval of non-relevant items (Kowalski, 1997). Essentially, natural language processing is able to add another level of disambiguation by indexing phrases rather than individual terms (Kowalski, 1997). Fagan (1987) was able to show that by using phrases, retrieval improved between 2 to 23 per cent, with variation being query dependent.

The ability to use natural language in information retrieval has great potential because it means that retrieval would no longer have to rely on only keywords, but rather, would be based on meaning. However, a major problem for language processing models is effectively dealing with the problem of lexical ambiguity (Krovetz, 1997).

Although the potential certainly exists, the current consensus within the information retrieval community is that using semantic information alone does not significantly increase the performance of information retrieval tools (Gonzola, Verdejo, Chugur, Cigarran, 1998). Further research into natural language processing is still necessary.

1.5 The Current Study

As highlighted in this report, there are a vast range of information retrieval tools that work in different ways. Although aspects of the various tools have potential, evidence suggests that standard keyword-based systems are often the most effective, particularly when they have been enhanced via techniques such as query expansion (Navigli & Velardi, 2003).

Although keyword-based techniques are generally successful, the success can be limited by poor queries. Often short and ambiguous queries are used, which can result in the retrieval of many irrelevant responses. This can be particularly problematic in extremely large collections, as a search may result in the retrieval of many thousands of items, and reviewing each item to determine its relevance would be impractical.

This study aimed to determine whether this problem could be reduced by utilising the context of the surrounding terms to re-rank the results. Essentially, rather than increasing cognitive load by requiring participants to expand their queries or provide semantically related terms, the current study used a keyword-based technique.

Participants were required to highlight terms from within documents, and the results were then automatically re-ranked based on the context of the surrounding sentence for one condition, and the context of the whole document for another condition. In a baseline condition the results were not re-ranked, and were instead based on term occurrence, with normalisation for document length. In order to ascertain whether user performance was influenced by individual differences, participants were also asked to complete a demographic questionnaire and two short cognitive tests. The methodology and results will now be described in the following sections.

2. Methodology

2.1 Participants

The sample consisted of 50 university students, from the University of Adelaide. To ensure that the participants were more likely to represent the anticipated customer employees, in relation to age, gender and academic qualifications, a large proportion of the sample were recruited from a third year level or higher.

2.2 Materials

2.2.1 Demographic Questionnaire

Participants were asked to complete a demographic questionnaire. This included questions regarding their age, gender, education level, area of study, and other experience using visualisation tools.

2.2.2 Comprehension and Information Test

In this experiment, the ability to read and understand passages of text was extremely important. Therefore, participants were required to complete a short test of English comprehension and a short information test. In the comprehension test, participants were required to read short passages of text and then answer multiple choice questions that referred to information covered in the passage. The information test included 30 multiple choice questions, which assessed the participants' general knowledge.

The results of these tests should provide an indication of whether participants had difficulties understanding the experiment, or whether the participants' performance on the task was influenced by their performance on the cognitive tests.

2.2.3 Document Collections

The documents used were newspaper articles from the TREC-8 document collection, containing three comparable sets of documents, with 150 documents in each collection. The documents were from the following sources: Foreign Broadcast Information Service (FBIS); LA Times; Financial Times; and the Federal Register (see Voorhees and Harman (2000) for more information). Examples of the documents are shown in Appendix A.

For each of the document sets, 30 of the documents (or 20% of each collection), were from one of three different research topics, adapted from the TREC-8 Ad-hoc Retrieval topics. These research topics were used to produce the questions, and the questions used are provided in Appendix B. The number of relevant documents for each of the three research topics varied from eight to twelve to prevent participants from perceiving a pattern in the document collections. The documents contained a maximum of approximately 400 words each, which ensured that the documents fit within the experimental interface.

In each of the document collections, 40% (60 documents) were irrelevant to act as noise. 40% (60 documents) were distractor documents, which contained at least one topic related word from one of the provided research questions (or a direct synonym of a topic related word), but were chosen to be irrelevant to the queries. For example, one question required participants to find the documents on cosmic events. One of the distractor documents for the question was a review of a play, referring to "cosmic boredom". Hence, this document contains a topic related keyword (cosmic), but was irrelevant to the question.

It is also necessary to note that, although the collections were designed so that 40% of the documents were distractor documents, the actual percentage of distractors was higher. The collections also included documents that were 'accidentally' distractors – for example, a document on cosmic events may have also included the word 'technology', meaning that it could be considered a distractor document for the question on robotic technology. Also, the documents designed to be 'noise' may have conceivably contained synonyms of topic related words that were not considered by the authors. Hence, the actual proportion of distractor documents is likely to be at least 50%.

A practice set with 30 documents was also utilised, which allowed participants to familiarise themselves with the interface. This practice set contained a 'walkthrough' question, which was demonstrated by the experimenter to provide the participants with a practical example of how to use the interface. The 'walkthrough' question contained four relevant documents. The practice set also contained a question that was completed by the participants, which contained three relevant documents and three distractor documents.

2.2.4 The Information Retrieval Techniques

Three independent information retrieval techniques were utilised. These were:
- A non-context search;
- LSA using the sentence of the query term;
- LSA using the whole document,

The non-context search has similarities to a keyword search, with the main difference that users were required to highlight the search term from within the document rather than typing it into a keyword box. The open source text search engine, Lucene (The Apache Software Foundation, 2007), was utilised to search the collection using the highlighted term. The search was normalised for length and the retrieved documents were ranked, so that the documents with more occurrences of the highlighted term were ordered before those with fewer occurrences of the term. This retrieval technique does not use any contextual information.

In this experiment, two forms of LSA were used; LSA using the sentence of the document (this will be referred to as *LSA Sentence*), and LSA using the whole document (this will be referred to as *LSA Document*). As detailed previously in this report, LSA assesses the co-occurrence of words and uses singular value decomposition (SVD) to discard redundant information, meaning that it only focuses on essential semantic information (Kintsch, 2001).

In the same manner as the non-context search, participants were required to highlight a search term, and the Lucene algorithm was then used to retrieve a list of the documents containing

that query term. The results of the non-context search were then re-ordered based on contextual information. When utilising LSA Sentence, the documents were re-ranked based on the similarity of the sentence in which the search term was located to the retrieved documents. When utilising LSA Document, the documents were re-ordered using the similarity between the document containing the search term and the retrieved documents. Further details of this method of calculating similarity values using LSA can be found in Pincombe (2004).

LSA was trained on the Touchstone Applied Science Associates (TASA) corpus, which contains 10 million words from a variety of areas, including science, social studies, language, arts, health, business and home economics. It was important to utilise a corpus that contained information from a wide variety of areas to ensure that the effectiveness of LSA did not differ between the different questions.

Participants were not provided with any information regarding the specific techniques, and were only told that the study was assessing different information retrieval techniques that work in different ways.

2.3 Method

The research assistant was provided with detailed instructions to follow for carrying out the experiment, which ensured that all participants received the same information. Before the experiment began, participants were given an information sheet and consent form, explaining their participation in the study. They were then asked to complete a demographic questionnaire, followed by a short test of comprehension and a short general knowledge test.

The study consisted of a practice question followed by three conditions, with each condition testing a different document retrieval technique. A repeated measures design was used, which means that all participants completed the experiment with each of the information retrieval techniques. Hence, all participants used the non-context search, and both the LSA sentence technique and the LSA document technique.

To control for possible learning effects or fatigue, all conditions were completed in a counterbalanced order. In addition, the allocation of document set to search condition was balanced such that all different combinations of set and search type occurred with the same frequency across all of the participants.

A custom interface, created by the Defence Science and Technology Organisation, was provided on a computer monitor (see Appendix C for a screenshot from the interface). Participants were given a basic user guide, which specified all necessary functions and terms.

The task of the experiment involved participants searching for and compiling all of the documents that would be used to write a report on a specified topic. For each of the three conditions, participants were asked to complete three questions based on research topics. For example, for the topic 'robotic technology', participants were provided with the following guidelines:

"Imagine that you are writing a report on robotic technology. Please indicate the documents that you would refer to in writing such a report."

The interface displayed the current question, a document window for displaying the full text of a document, and a summary list of the documents that were relevant to a search. In the summary list, the documents were labelled with an identification number. At the start of each question the document window contained a document that was deemed to be 'relevant' to the first question and the summary list was blank. This starting document was the same for each participant.

Participants were required to make queries by clicking and dragging the mouse over the text to highlight a word or series of consecutive words in the document, and they were then required to press the 'search' button to initiate the search. The retrieval tool then searched the collection and the documents containing the query term (or terms) were displayed in the summary list.

The interface also contained a box for the 'marked' documents (the documents considered relevant to the question), which appeared above the summary list. Participants were required to 'mark' or 'unmark' documents by clicking on an up or down arrow to move the document into or out of the 'marked' box.

When a document was selected, the whole text was viewed in the document window. The background of the currently viewed document was coloured to indicate whether it was 'unread', 'read' or 'marked'. At any point, participants could choose to highlight and search with another query, could choose to 'mark' the document, or could choose to move to any other documents in the summary list. Participants were also able to access and search from the 'marked' documents. This search process continued until the participant decided that all of the documents relevant to the question had been 'marked'.

The program not only recorded the documents that participants selected, it also recorded all other user interaction with the interface, including the documents viewed, the order in which documents were viewed, and all timing information.

3. Results

3.1 Summary of Results

Fifty participants answered three questions for each of the three techniques. In contrast to expectation, there were no significant differences in terms of the participants' average accuracy, number of search terms and time taken across the three techniques.

However, an analysis of the documents retrieved using the different techniques indicated that the LSA techniques did retrieve documents in a more efficient manner. It is theorised that the document collection was not large enough for the effective re-ranking to significantly influence the participants' performance. Despite this, the results do suggest that LSA may assist in information retrieval.

Furthermore, findings indicate that individual differences had a large influence on the results. Characteristics such as the participants' level of education and performance on the comprehension test tended to be better predictors of success. Hence, the participants who obtained higher comprehension scores and the participants who had a higher level of education tended to perform more successfully on the information retrieval task.

The results for each of the performance measures and the other demographic information will now be analysed in more detail.

3.2 Efficiency of the Re-Ranking Techniques

In order to determine the efficiency of the different re-ranking techniques, the documents retrieved for each search were analysed in detail. The aim was to examine the placement of the predetermined relevant documents within the retrieved list. This is based on the assumption that a more efficient technique will retrieve the relevant documents first.

As indicated earlier (see Section 1.1) there is little agreement in regards to the best performance measure, and most measures have both benefits and disadvantages. Therefore, the documents retrieved by every search were examined in detail, and a number of measurement techniques were utilised to assess performance.

3.2.1 Average Rank

The three techniques were assessed by examining the average rank of the relevant documents retrieved by each search. A one-way Repeated Measures Analysis of Variance (RMANOVA) was conducted and there was a significant effect for technique, Wilks' Lambda = .852, $F(2, 48)$ = 4.16, $p < 0.05$. The placement of relevant documents was closest to the beginning of the retrieved list when utilising LSA Document (M = 2.71, SD = 0.73), followed by LSA Sentence (M = 2.76, SD = 0.87), then the Word Search technique (M = 3.05, SD = 0.63). The effect size, calculated using multivariate η_P^2, was 0.15, meaning that approximately 15% of the variance in the placement of relevant documents was associated with the technique. This indicates that

the re-ranking provided by the LSA techniques successfully places the documents in a more efficient position.

3.2.2 Placement of the Final Relevant Document

In order to further assess this, the placement of the final relevant document in each retrieved list was assessed. This is based on the assumption that a more efficient technique will rank the relevant documents towards the beginning of the retrieved list. Hence, this indicates the proportion of documents in a retrieved list that the user would need to view in order to access all relevant documents.

A RMANOVA was conducted on the proportion of the retrieved lists that contained relevant documents for each technique. A significant effect was found, Wilks' Lambda = .523, $F(2, 48)$ = 21.85, $p < 0.001$. When utilising LSA Document ($M = 0.85$, $SD = 0.06$) and LSA Sentence ($M = 0.85$, $SD = 0.07$), the relevant documents were ranked in a significantly more efficient position than when utilising the Word Search technique ($M = 0.90$, $SD = 0.04$). This means that in order to view all relevant documents in a list, it was necessary to view approximately 85% of the list for the LSA techniques versus approximately 90% of the list (an extra 5%) when using the Word Search technique. The effect size, calculated using multivariate η_p^2, was 0.48 meaning that approximately 48% of the variance in the position of the final relevant document was associated with the technique.

3.2.3 Rank-Biased Precision (RBP)

RBP is a new method that provides a robust, flexible and user oriented measure of the effectiveness of an information retrieval tool (Moffat & Zobel, 2008). As mentioned previously, the technique includes a value, p, which represents the persistence or patience of the user. Since this experiment utilised a reasonably small document collection and aimed to assess whether a keyword search could be improved through re-ranking the results, p was set to 0.5. This places a 50:50 chance on the user continuing from one document to another, and therefore places a large amount of emphasis on documents that are near the beginning of a retrieved list (Zhang, Park & Moffat, 2008).

A RMANOVA found a significant effect, Wilks' Lambda = .360, $F(2, 48)$ = 13.53, $p < 0.001$. RBP was significantly lower for the Word Search technique ($M = 0.61$, $SD = 0.07$) than for LSA Document ($M = 0.68$, $SD = 0.07$) or LSA Sentence ($M = 0.66$, $SD = 0.05$). The effect size was .036 indicating that 36% of the variance in RBP score was accounted for by the different technique. This finding is depicted in Figure 1, below, clearly demonstrating that precision was significantly worse for the Word Search technique.
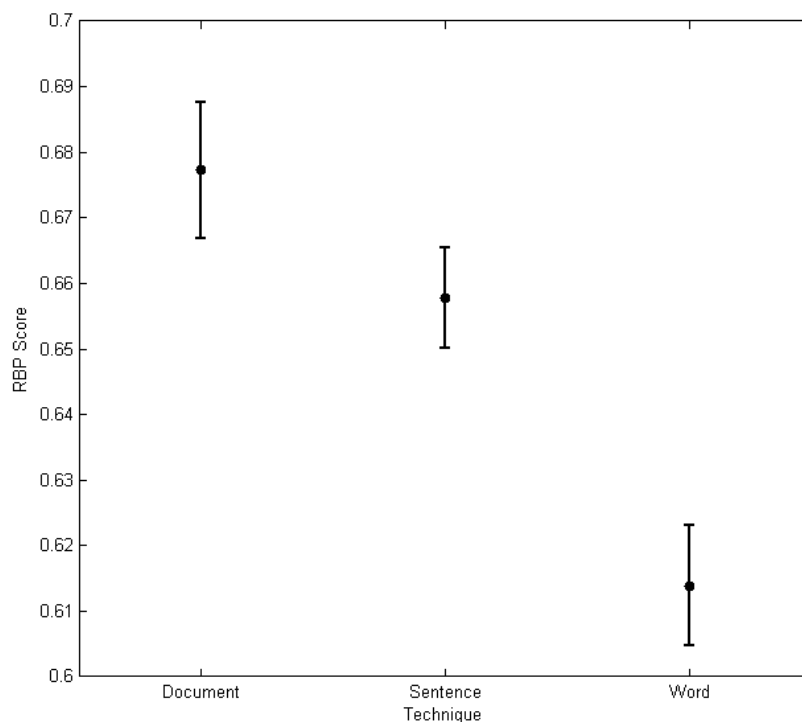
Figure 1: Rank-biased precision by technique used (One standard error about the mean is shown)

3.2.4 Precision and Recall

Although overall precision and recall scores are imperfect measures that reveal little regarding the initial performance of a technique, a precision-recall graph can be used to depict the trade-off between precision and recall. In order to comprehensively illustrate the performance of the different techniques, precision and recall scores were obtained for each search after the retrieval of every document up to a depth of 15. The precision and recall scores for each technique were averaged for each participant, and overall average scores were then obtained.

The precision-recall graph is shown in Figure 2. This graph clearly shows that initial performance was best for LSA Document, followed by LSA Sentence, which means that the first few documents were far more likely to be relevant when the participant was using the LSA techniques. The Word Search technique had the worst initial performance. However, an analysis of the graph reveals that the performance had essentially equalised by the retrieval of approximately eight documents (recall of ~0.28).
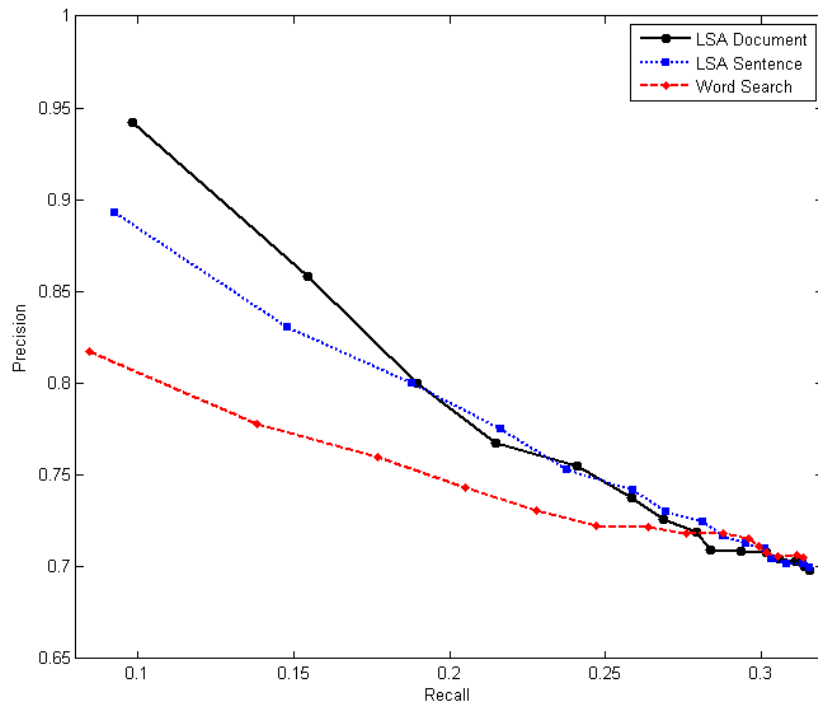
Figure 2: Precision-Recall Curve for each technique

3.3 Initial Analysis of Participants' Responses

The results were then examined in regards to the participants' performance. Performance was measured based on accuracy, time taken and the number of search terms used.

Accuracy for the experiment was defined as the percentage of predetermined 'relevant' documents that were marked by participants. Overall, the accuracy for the experiment was quite high, with participants marking an average of 81% (SD = 10.97) of the predetermined relevant documents across all questions. However, there was a large range in the accuracy obtained. In total for the nine questions there were 90 relevant documents, and the participant with the highest accuracy marked 87 of those documents (97% correct). In contrast, the lowest score was only 34, which equates to only 38% of the relevant documents.

There was a very large range in the amount of time taken to complete questions. Overall, the average time was 422 seconds, which equates to approximately 7 minutes to complete one question (SD = 209.68). There was also a large range in the number of search terms used, with a minimum of only 2 searches, and a maximum of 51. Generally, participants required approximately 11 search terms to answer a question (M = 11.31, SD = 7.90, $Mode$ = 10).

3.4 The Influence of Technique

Despite the results in Section 3.2, which indicated that the LSA techniques did rank the documents more efficiently, there were no significant differences between the techniques in regards to accuracy, time taken or search terms used.

3.4.1 Accuracy by Technique

As shown in Table 1, there was very little difference in relation to the accuracy scores across the three techniques. The highest accuracy was obtained using LSA Sentence, and the lowest accuracy was obtained when utilising LSA Document.

Table 1: Accuracy by technique

Technique	Accuracy (%)	Standard Error	SD
LSA Document	80.21	1.96	13.89
LSA Sentence	82.13	1.89	13.36
Word Search	81.61	1.77	12.51
Total	*81.31*	*1.52*	*10.97*

In order to obtain a thorough assessment of performance, participants' were also measured based on their precision, recall, f-measure and mean average precision (MAP) scores. Figure 3 displays these scores for each of the techniques, clearly indicating that there was very little variation in any of these measures across the three techniques.
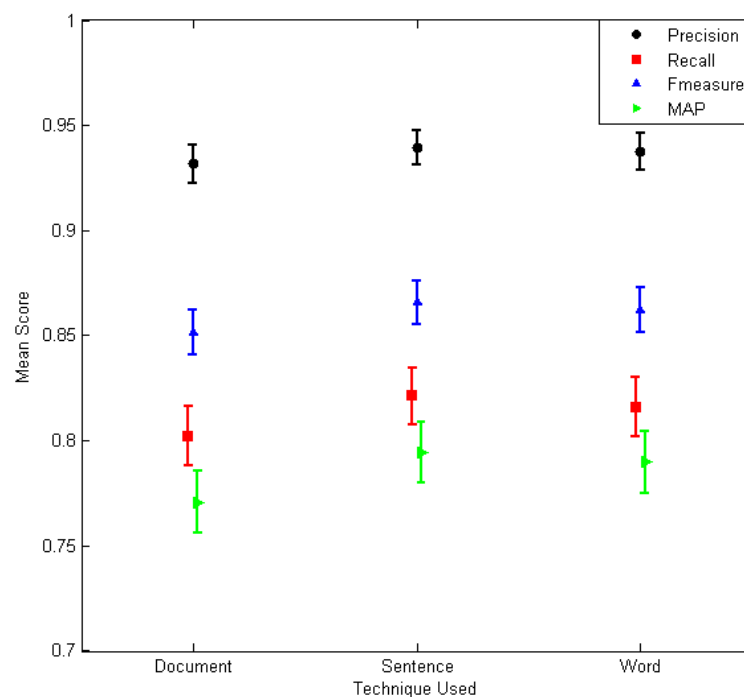
Figure 3: Accuracy by technique used (One standard error about the mean is shown)

Participants tended to obtain very high precision ($M = 0.94$, $SD = 0.10$), indicating that the documents that were marked tended to be correct, and that they were unlikely to mark many additional documents. The recall score ($M = 0.81$, $SD = 0.17$) indicates that participants were very unlikely to mark all of the relevant documents. This could indicate a disagreement between the participants' assessment of relevance and the predetermined relevance

judgements, or it could be caused by an inability to find the documents using the provided interface.

3.4.2 Time by Technique

As shown in Figure 4, there were no significant differences in the amount of time taken when using the three techniques [LSA Document (M = 426.84, SD = 220.96), LSA Sentence (M = 426.74, SD = 198.25), Word Search (M = 413.57, SD = 210.34)].
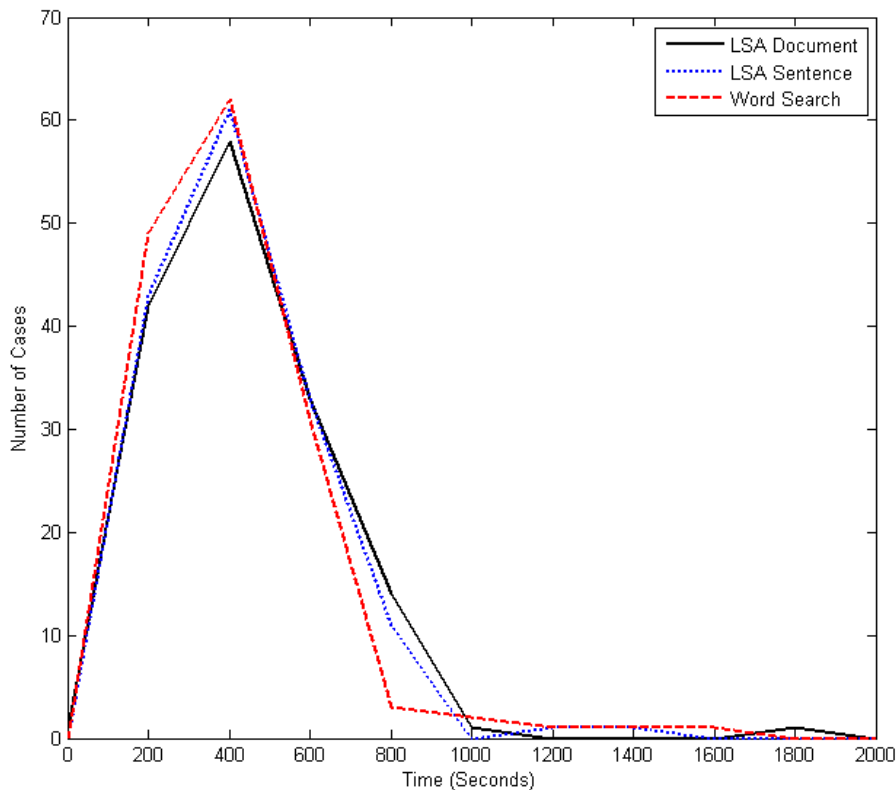

Figure 4: Histogram showing time by technique

The average time was slightly lower for the Word Search technique; however, this is a reflection of the 'wait' time rather than any difference in the participants' performance. Essentially, the LSA techniques required more processing time, and therefore, for each question, the participants waited an average of approximately 10 seconds [LSA Document (M = 11. 15, SD = 9.78), LSA Sentence (M = 9.11, SD = 7.86)] compared to less than half a second (M = 0.44, SD = 1.33) when using the Word Search technique. This wait time could have decreased participants' confidence in the system, and hence, this may have reduced the success of the LSA techniques.

3.4.3 Search Terms by Technique

The number of search terms did not differ significantly across the three techniques. The most search terms were used for LSA Sentence (M = 11.49, SD = 8.47), and this was followed closely

by LSA Document (M = 11.39, SD = 7.50), and the Word Search technique (M = 11.05, SD = 7.74). However, although there were no significant differences, as shown in Figure 5, participants were less likely to require a small number of searches when utilising the LSA Document technique. It is also necessary to note that although LSA Sentence had the highest mean number of search terms, this average was increased because the three most extreme scores (42, 43 and 51) all occurred when participants were utilising LSA Sentence.
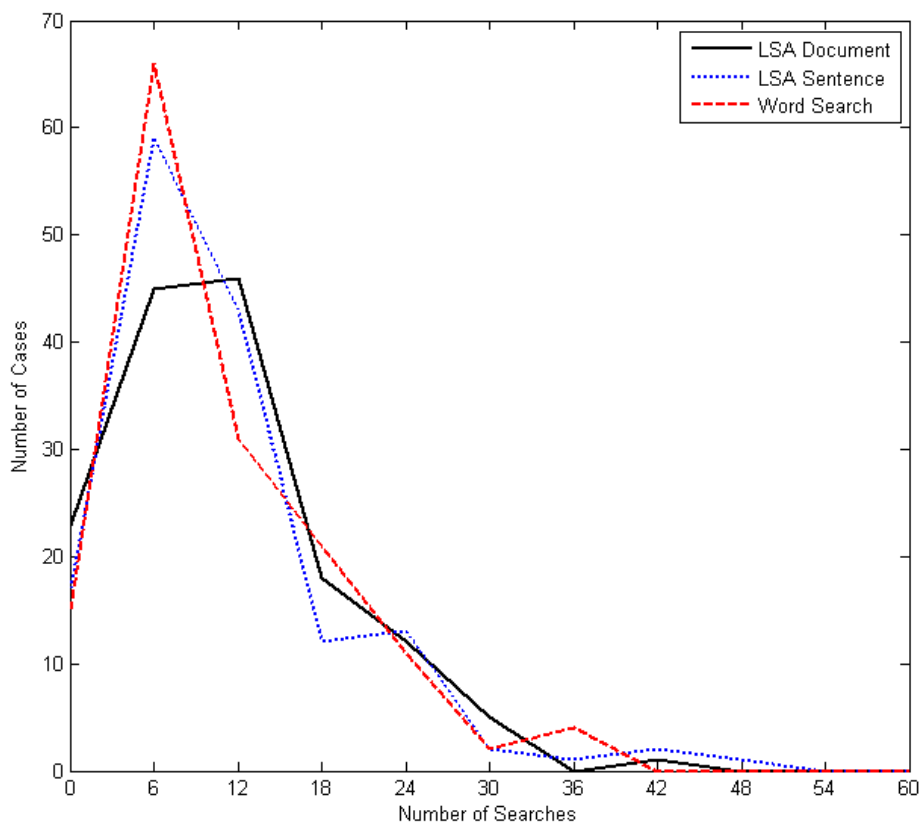

Figure 5: Histogram showing search terms by technique

3.5 Documents Accessed and Relevance Assessments

Since the participants' performance was not influenced by the efficiency of the re-ranking technique demonstrated in Section 3.2, the documents accessed and marked were examined in more detail. There were two hypotheses to explain the observed results, (1) participants may have accessed all of the retrieved documents, and therefore the efficient re-ranking would not have made a difference, and (2) the subjectivity associated with the relevance assessments may have influenced the results.

3.5.1 The Proportion of Documents Accessed

In order to test the first hypothesis, the documents accessed were examined in detail. On average, participants accessed approximately 94% of the retrieved documents that had not been previously viewed. There was little difference between the three techniques. Participants

were least likely to access all retrieved documents for LSA Document (93%) followed by LSA Sentence (94%) and the Word Search techniques (94%).

An analysis of the order in which documents were accessed indicated that participants regularly started at the top of the list, and accessed all documents that they had not previously viewed in a sequential manner. This behaviour can be explained via Zipf's (1949) *principle of least effort*, as the users have defaulted to a brute force technique that minimises effort, whilst guaranteeing that their search will find all relevant documents in the list. Hence, regardless of the efficiency of the re-ranking, since the participants were likely to continue to the end of the list, any benefit would be lost.

3.5.2 Relevance Assessments

It is also possible that the results were influenced by the subjectivity associated with relevance assessments. Disagreements regarding relevance assessments would decrease the accuracy of participants' performance, and any time spent deliberating potential relevance would also impact upon the results.

3.5.2.1 Documents Missed

As alluded to earlier, a number of participants failed to mark some of the documents that were predetermined to be relevant. In fact, there were a number of documents where a high proportion of the participants disagreed with the predetermined relevance assessments. For instance, for the question on wildlife preservation and poaching, 42 participants (84%) did not mark Document 62. In total, of the 90 documents that had been predetermined to be relevant, an average of 17.3 (SD = 9.8) documents were missed by participants, and there were 36 documents that were missed by 10 or more participants.

Since relevance is such a subjective concept, it is quite likely that different individuals may not have agreed on what constituted 'relevant'. For example, the documents varied in regards to the proportion of the text that was 'on topic', and it is possible that participants may have differed as to their threshold to determine whether a document was relevant. Some participants may have felt that any mention of the topic was enough to make it relevant, whereas other participants may have only marked the document if the topic was mentioned in detail.

3.5.2.2 Documents Added

There were also a number of documents that a large percentage of the participants added, despite the fact that the documents were not in the predetermined relevance list. In total, 266 documents were added, and there were 7 documents that were added by ten or more participants. That equates to 0.59 (SD = 0.99) documents per question. An analysis of the added documents discovered that the majority contained topic related words, or direct synonyms of topic related words, but yet were not relevant to the queries.

This suggests that participants may not have read the documents properly, or may have had a different threshold for determining whether a document was relevant to a given query. For instance, the question on robotic technology contained non-relevant documents that involved business dealings of companies involved in robotic technology. Hence, although the

24

documents may have contained the words 'robotic technology', since there was no mention of technology of a robotic nature these documents were deemed to be non-relevant. However, it is very conceivable that some participants may have still considered those documents to be relevant.

3.5.2.3 *Reassessment of Contentious Documents*

Since these relevance judgements may have influenced the participants' performance, the documents where there was a high level of disagreement between the participants' responses and the predetermined ratings were reassessed by two additional judges. This included the relevant documents that participants missed and the irrelevant documents that were added by participants. The additional judges were not told whether the documents had been added or missed by participants, and were only provided with the question, and asked to make a relevance decision for each of the documents.

In total, 36 documents were missed by 10 or more participants, and 7 documents were added by 10 or more participants, making a total of 43 documents that were reassessed by the additional judges. In the vast majority of instances, the additional judges both supported the predetermined relevance assessments. Judge 1 agreed with the predetermined responses in 39 of the 43 cases (91%), and Judge 2 agreed with the predetermined responses in 30 of the 43 cases (70%).

There were 15 instances where one of the additional judges disagreed with the predetermined assessments, and there was only one case where both of the judges disagreed. This means that in 98% of the reassessed cases at least one of the additional judges supported the predetermined decisions.

In order to determine whether the contentious documents influenced participants' performance, the results were re-analysed, with the decisions for 16 contentious documents reversed; that is, the contentious documents that were added by participants were included in the relevance list, and contentious documents that were missed by participants were removed from the relevance list. However, the re-analysis showed that these changes did not significantly influence results.

3.6 The Influence of Question

Although there were no significant differences in regards to the performance across the three techniques, there were significant differences based on the performance across the different questions. In total, participants answered nine questions, and there was a large range in performance, suggesting that information retrieval could be influenced by the specific information to be retrieved.

3.6.1 Accuracy by Question

Although the questions were designed to be of equal difficulty, the results indicate that participants' accuracy was strongly influenced by the question. The highest accuracy was obtained for the question on Cosmic Events (M = 91.0, SD = 16.76) and the lowest accuracy was obtained for the question on Heroic Acts (M = 71.17, SD = 15.63). A RMANOVA on

participants' accuracy for the nine questions indicated that this was a significant effect, Wilks Lambda = 0.28, $F(8, 42)$ = 13.76, $p < 0.001$, multivariate partial eta squared = 0.72.
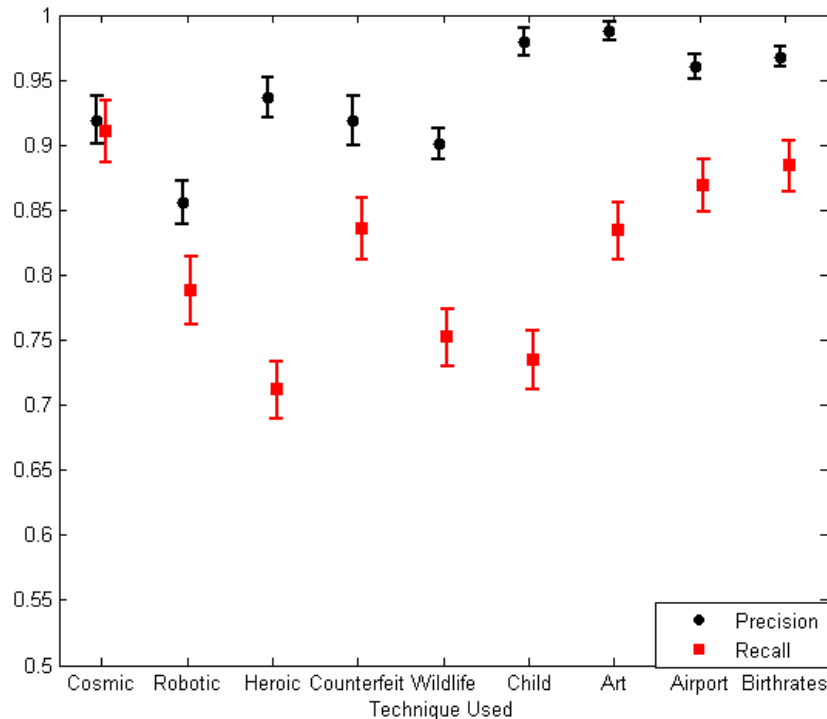

Figure 6: Precision and recall by question (One standard error about the mean is shown)

The graph in Figure 6 shows participants' precision and recall scores, and clearly demonstrates that participants performed differently depending on the question. For the questions on Child Labour, Art Theft, Airport Security and Birth Rates, participants obtained extremely high precision, signifying that non-relevant documents were very rarely marked. In contrast, precision was far lower for the question on Robotic Technology, suggesting that many participants marked documents that were not relevant to that question. Participants also varied greatly in regards to the recall score, indicating that participants were less likely to mark all relevant documents for some questions.

Interestingly, a stable pattern was not found in the relationship between recall and precision. For the question on Cosmic Events, participants tended to obtain high scores in both precision (M = 0.92, SD = 0.13) and recall (M = 0.91, SD = 0.17), and using Pearson's product-moment correlation coefficient, there was a strong positive relationship between the two variables [r = 0.72, n = 50, $p < 0.001$]. In contrast, for the question on Child Labour, participants tended to obtain high precision scores (M = 0.98, SD = 0.73), but the recall scores were generally far lower (M = 0.73, SD = 0.16). Although there was still a positive correlation between the two variables, it was far less significant [r = 0.39, n = 50, $p < 0.01$].

There was also no stable pattern found when the results were analysed based on the accuracy obtained for each question when using each of the different techniques. When participants

used LSA Document, they tended to obtain lower accuracy on the questions involving Art Theft, Airport Security and Declining Birth Rates. In contrast, when participants were utilising the Word Search technique, accuracy was far lower for the questions involving Child Labour and Counterfeit Money. The LSA Sentence technique was least effective for the question involving Robotic Technology. These results suggest that the success of the techniques could be more influenced by the specific documents used, and it is possible that each of the techniques could have benefits when searching for certain documents. Unfortunately, it is difficult to pinpoint any consistent differences in the documents or questions that may explain why certain techniques were more successful for some questions, and less successful for other questions.

3.6.2 Time by Question

Although there were no significant differences in the time taken for the three techniques, there were significant differences in the time taken for the various questions. A RMANOVA was conducted to compare the time in seconds on the nine questions completed, and there was a significant effect, Wilks Lambda = 0.23, $F(8, 42) = 18.02$, $p < 0.001$, multivariate partial eta squared = 0.77. The significance is clearly depicted in Figure 7, showing that the question on Cosmic Events took longest, and the question on Birth Rates was completed quickest. The means and standard deviations are presented in Table 2.
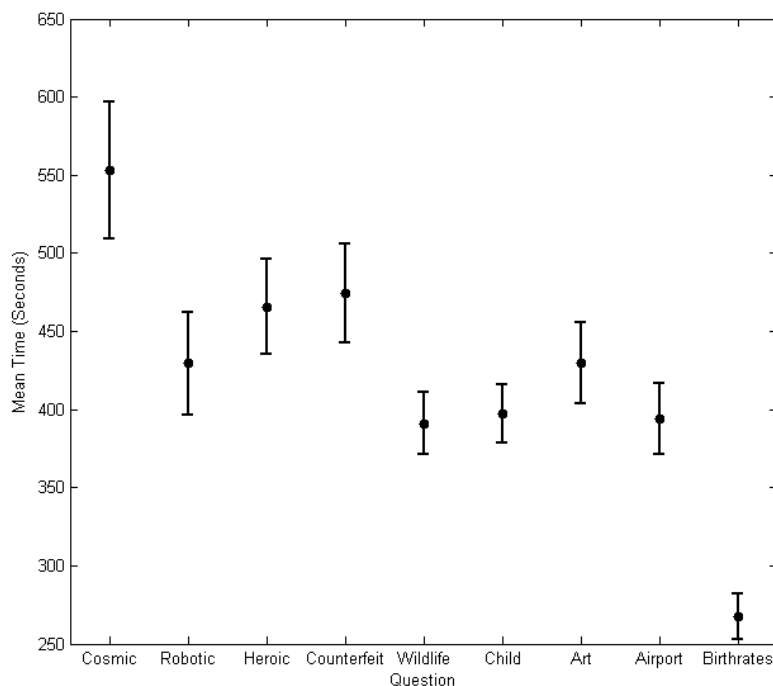

Figure 7: Time taken for each of the questions (One standard error about the mean is shown)

Table 2: Description statistics for time and search terms by question

Question	Time (M)	Time (SD)
Cosmic Events	553.04	309.51
Robotic Technology	429.53	232.42
Heroic Acts	465.63	215.55
Counterfeit Money	474.14	224.14
Wildlife Poaching	391.05	141.16
Child Labour	397.27	129.71
Art Theft	429.83	184.92
Airport Security	393.63	160.71
Birth Rates	267.32	101.92

3.6.3 Search Terms by Question

As shown in Table 3 and Figure 8, there was a large range in the average number of search terms required for each of the questions. Generally, participants required far more search terms when answering the question on Cosmic Events, and far fewer search terms for the question on Birth Rates. Interestingly, as shown previously (see Section 3.3.2) the question on Birth Rates was also quickest, and the question on Cosmic Events was the slowest.

Table 3: Description statistics for time and search terms by question

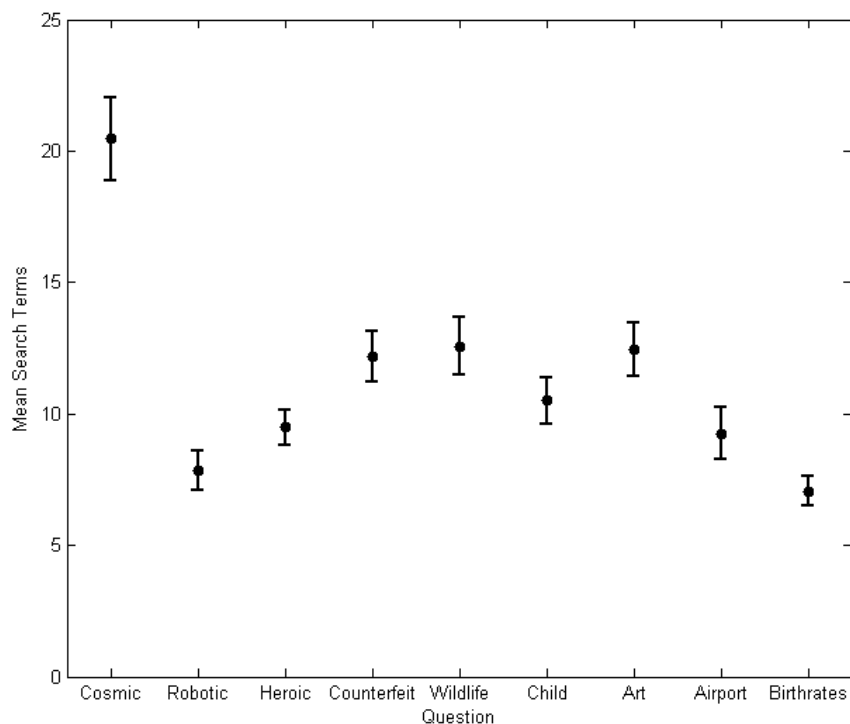| Question	Search Terms (M)	Search Terms (SD)
Cosmic Events	20.46	11.23
Robotic Technology	7.84	5.30
Heroic Acts	9.48	4.58
Counterfeit Money	12.18	6.93
Wildlife Poaching	12.56	7.73
Child Labour	10.50	6.28
Art Theft	12.46	7.28
Airport Security	9.26	7.15
Birth Rates	7.06	4.10

Figure 8: Number of search terms by question (One standard error about the mean is shown)

3.6.3.1 Examination of the Search Terms Used

The specific search terms used were also analysed in detail. Across the nine questions, there were 14 occasions where 10 or more participants used the same initial search term, and each of the questions had at least one occurrence of these common search terms. The most common initial search term was heroic, which was the first term used by 28 participants. This was the second search term used by a further 11 participants, meaning that 78% of participants used this term first or second.

3.6.3.2 Repeated Search Terms

There were also a number of occasions where participants repeated the same search. In total, there were 547 repeats, with an average of 1.22 searches repeated for each question (*SD* = 0.84). Participants were far more likely to repeat searches when answering the question on Cosmic Events, with a total of 139 repeats (*M* = 2.78, *SD* = 4.2). In contrast, only 26 terms were repeated for the question on Birth Rates (*M* = 0.52, *SD* = 0.89). There were only two participants who did not have any repeated search terms, and the highest number of terms repeated was 30.

3.6.3.3 Incorrect or Invalid Search Terms

It is also possible that some participants may have had difficulties with the interface or the search tool. It total, across the 50 participants, there were 85 occasions where participants made searches that were invalid or incorrect, meaning that they did not retrieve any documents. These searches, and possible reasons for their occurrence, will be explained in more detail in the Discussion section.

3.7 Individual Differences

The results also indicated that the participants' performance was significantly influenced by individual differences. Aspects such as the participants' education level and score on the comprehension test influenced results, with the more educated participants and the participants who scored highly on the comprehension test more likely to perform well in the information retrieval experiment.

3.7.1 Comprehension Score

Before the commencement of the experiment, participants were asked to read a number of short passages of text, and were then provided with multiple choice questions, where the answers required a level of comprehension of the passages. The maximum score for the comprehension test was 6, and the average score obtained was 4.3 (SD = 1.36), suggesting that most participants tended to perform well on the task.

Participants who had higher scores on the comprehension test also tended to obtain higher accuracy on the information retrieval experiment, and tended to complete questions quicker, with fewer searches required. The relationship between accuracy (as measured by the percentage of predetermined relevant documents marked) and comprehension score was investigated using a Pearson product-moment correlation coefficient. There was a moderate positive correlation between the two variables [r = .39, n = 50, p < .001], with higher accuracy scores associated with higher comprehension scores. Thus, comprehension scores accounted for 15.21% of the variance in accuracy. The participants' comprehension scores were not significantly correlated with the amount of time taken (r (50) = -.11, p = .46) or the number of searches used (r (50) = .10, p = .47).

A RMANOVA was also conducted to determine whether participants' comprehension score influenced their performance on the different information retrieval techniques. As shown in Figure 9, it appeared as though participants' performance when utilising the different techniques was more varied for the participants with low scores on the comprehension test. In contrast, the participants with high comprehension scores tended to perform equally well when utilising each of the techniques. However, this was not a significant effect (Wilks' Lambda = .96, F(2, 46) = 0.95, p = .393, multivariate partial eta squared = .04).
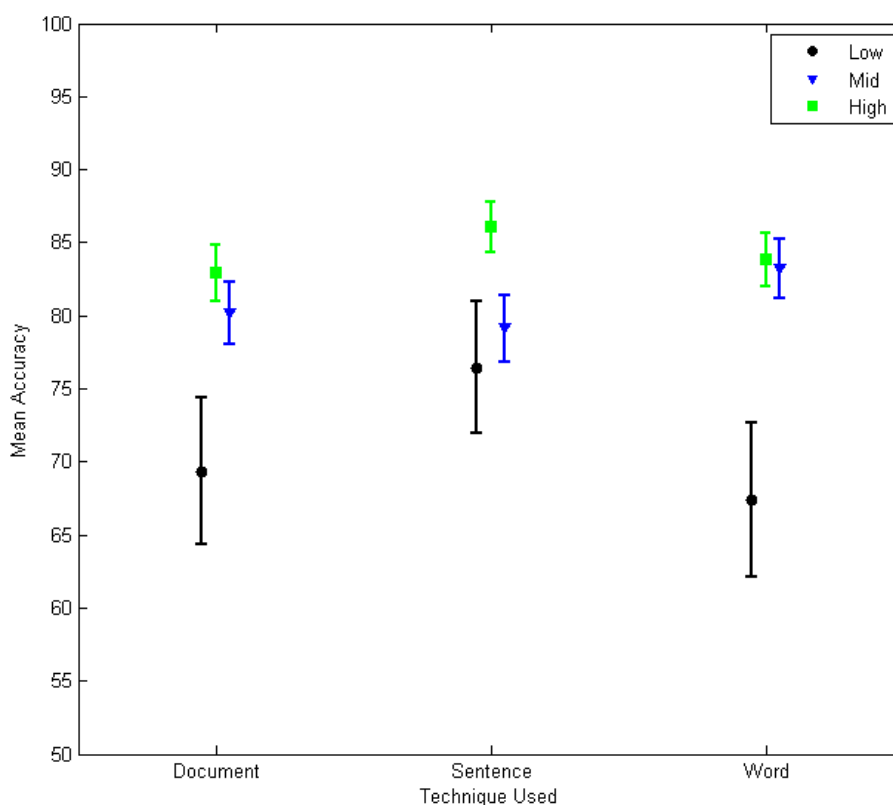
Figure 9: Mean accuracy by technique and comprehension score (One standard error about the mean is shown)

3.7.2 Information Test

Participants were also required to complete a short information test, which consisted of 30 multiple choice questions, and the average score obtained was 21.32 (SD = 3.5). The relationship between the performance on the information test and the performance in the information retrieval task were tested in a series of correlations. Scores on the information test were not significantly correlated with either the number of search terms used (r (50) = .22, p = .13) or the amount of time taken (r (50) = -.05, p = .74). However, scores were moderately, positively correlated with accuracy, r (50) = .42, p < 0.01, with the information test scores accounting for approximately 17% of the variation in accuracy.

Another RMANOVA was conducted to determine whether participants' information test scores influenced their performance on the different information retrieval techniques. Although there was not a significant effect (Wilks' Lambda = .94, F(2, 46) = 1.56, p = .222, multivariate partial eta squared = .06), there was a similar pattern to the relationship observed for the comprehension test, where participants' performance was more varied for the participants with lower scores, and less varied for the participants with higher scores, who performed well regardless of the technique utilised.

3.7.3 Education Level

In the demographic questionnaire, participants were asked to provide an indication of their highest level of education. These scores were then used to analyse the results, with participants divided into four groups, based on their level of education (1st year, 2nd year, 3rd year and post grad).

A one-way between groups analysis of variance was conducted to determine whether participants' education level had a significant influence on their performance in the information retrieval task. The mean accuracy score obtained by the first year participants (M = 75.45, SD = 13.59) was lower than the average score obtained by the second year participants (M = 81.93, SD = 10.16) and the third year participants (M = 83.07, SD = 8.85). The average score for the participants with post graduate qualifications was highest (M = 84.33, SD = 10.41). However, this was not a significant difference [$F(3, 46)$ = 1.66, p = 0.19], suggesting that although there was a tendency for education to influence performance, the finding was not true for all individuals. Despite this, the histograms in Figure 10 clearly indicate that first year participants were far less likely to obtain very high accuracy scores.
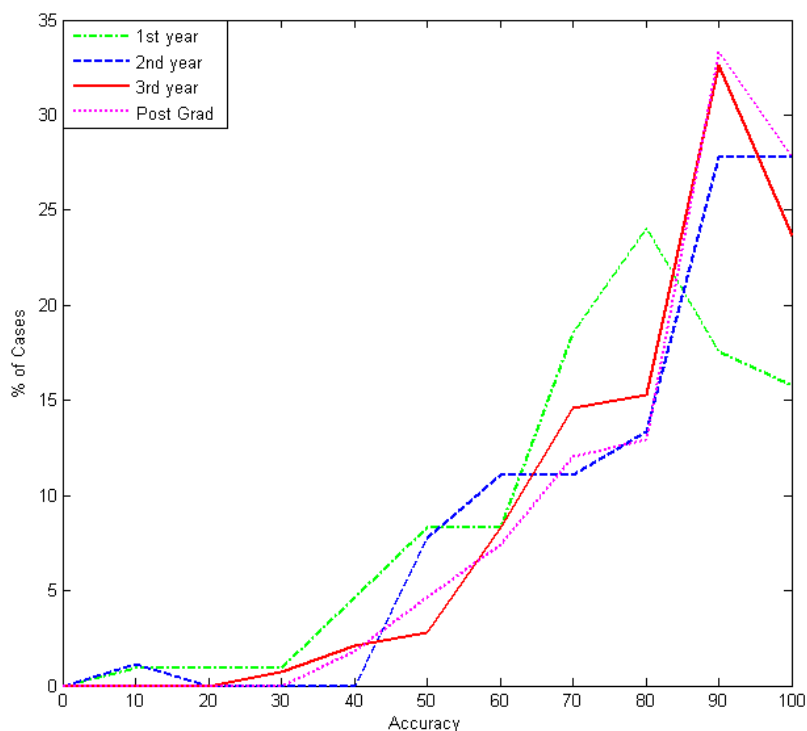
Figure 10: Histograms showing accuracy by education level

3.7.4 The Effect of Other Individual Differences

The participants also provided information regarding a number of other individual differences, including age, gender, experience with visualisations and area of study. Those additional factors did not significantly influence results in terms of the time taken, searches required, accuracy obtained or the performance when utilising the specific techniques.

However, participants' performance was significantly influenced by whether English was their primary language. A series of independent samples t-tests were conducted to compare the performance obtained by the participants for whom English was the primary language (Group 1), and the participants for whom English was not the primary language (Group 2). Participants' accuracy was significantly influenced by language, with the participants in Group 1 obtaining significantly higher scores (M = 82.91, SD = 8.30) than the participants in Group 2 [M = 72.94, SD = 18.54; t(50) = 2.47, p < .05]. The magnitude of the differences was quite large (eta squared = .113), which means that approximately 11% of the variance in accuracy was related to the participants' primary language. This is clearly shown in Figure 11, which shows the histograms for the accuracy obtained by both groups of participants, indicating higher accuracy for the participants who have English as their primary language.
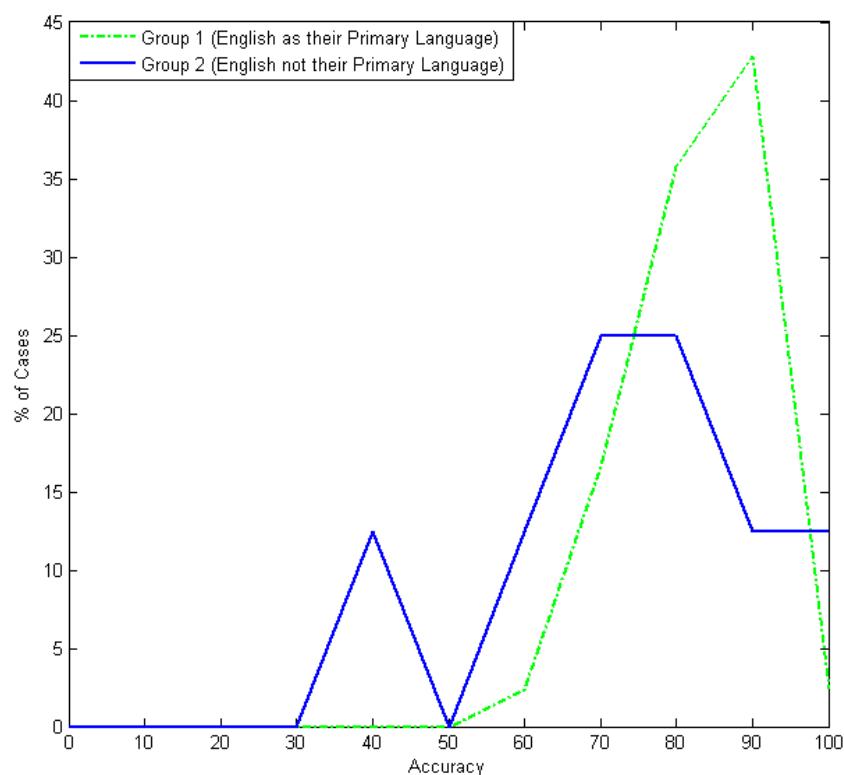
Figure 11: Histograms showing accuracy by education level

The results also indicate that the amount of time taken for participants in Group 1 (M = 405.26, SD = 111.31) was significantly less than the amount of time taken by participants in Group 2 [M = 512.15, SD = 105.10; t(50) = -2.47, p < .05]. The magnitude of the differences was also quite large (eta squared = .113), indicating that 11% of the variance in time taken was related to whether the participants' primary language was English. This suggests that there may have been a language barrier for some participants.

Despite this, there were no significant differences for Group 1 (M = 11.23, SD = 4.55) and Group 2 [M = 11.75, SD = 4.62; t(50) = -.30, p = .77] in regards to the number of search terms used.

4. Discussion

The aim of this experiment was to ascertain whether participants' performance on an information retrieval task could be improved by re-ranking results retrieved from a keyword search using the context provided by the words surrounding a user's chosen search terms. Two LSA techniques (based on the context provided by the sentence and the context provided by the document) were tested against a context free Word Search technique.

An analysis of the documents retrieved by each technique revealed that the re-ranking provided by the LSA techniques significantly improved the efficiency of the retrieved list. However, contrary to expectations, this re-ranking did not improve participants' performance. Instead, the question and individual differences in regards to aspects such as participants' comprehension and knowledge of the English language were far more influential.

There are a number of plausible reasons for this unusual pattern of results, and these reasons will be analysed in detail in this section. Mechanisms to reduce these problems in future experiments and other suggestions for future research will also be examined.

4.1 Issues Associated with the Document Collection

It is possible that the document collection utilised may have influenced results. An analysis of the documents accessed indicated that participants regularly viewed all of the retrieved documents. Therefore, despite the fact that the relevant documents were placed in a significantly more efficient position when utilising the LSA based techniques, the participants' performance was not altered.

This is likely to be associated with the size of the collection. The current experiment utilised three documents sets, with 150 documents in each set. A real-life document collection could consist of many thousands or hundreds of thousands of documents, where it would be impractical for participants to use the default strategy of examining the whole list. Hence, a system that ranked documents in an effective manner could be extremely beneficial.

The document collection utilised was designed with a large proportion of noise and distractor documents, which were expected to increase the difficulty of the exercise, and decrease participants' ability to easily locate all documents (more accurately reflecting a real-life task). However, most participants searched with terms that were unlikely to retrieve more than 10 documents. In fact, only 14% of all searches retrieved more than 10 documents, and only 2% of all searches retrieved more than 20 documents. Hence, when such a small number of documents were commonly retrieved, participants generally examined the entire list. This is interesting as it suggests that in cases where the collection is not very large, participants may conclude that the risk associated with missing relevant documents was not worth the benefit associated with stopping their examination of the retrieved list.

4.2 The Subjectivity of Relevance

Participants' performance in this experiment was assessed based on whether the participants 'marked' the documents that were predetermined to be relevant to the various questions. However, since relevance is such a subjective concept, it is possible that results may have been influenced by this subjectivity. For example, it is possible that participants may have spent a great deal of time agonising over whether to 'mark' specific documents, resulting in an artificial increase in the timing information.

It is also possible that participants who did not mark certain documents may have had valid reasons for their decisions. Section 3.5.2.3 described the results of the re-analysis of the contentious documents, and this assessment highlighted the subjectivity of relevance. For instance, for the question on declining birth rates, there was one document where both additional judges disagreed with the predetermined relevance decision. The document in question referred to an increase in birth rates, but the additional judges claimed that, despite quoting an increase, the document could still be a useful resource for a report on declining birth rates.

In contrast, for the question on the forgery of currency, one of the additional judges claimed that two of the documents were not relevant, as they only referred to counterfeit money that had been found, rather than the act of forging the currency. It can be argued that the documents referring to counterfeit money could still be useful when writing a report on the forgery of currency, but it is equally valid to argue against this. Hence, it is possible that the subjectivity of the documents may have influenced the participants' results.

4.3 Human Factors Issues Associated with the Interface

As mentioned in the results, there were a number of occasions where participants made incorrect or invalid searches. Some of those were searches that appeared to reflect errors in the highlighting process. For example, one participant searched with 'oreign museums' rather than 'foreign museums' and another searched with 'hieves stole 20 paintings' whereas the first word in that search should (presumably) be 'thieves'. The highlighting process used in this interface would have been unfamiliar to most participants, and therefore, they may have required some practice to become accustomed to the technique.

There were also a number of other searches that may have been instances where the participants were using search techniques that were not available in this tool. Many information retrieval tools include stemming, which allows users to search with parts of words, and any word containing that segment of text will be retrieved. For example, during the 'Heroic Acts' question, one participant searched with a part of word, 'sav', which could have been a highlighting problem, or could have been an attempt to retrieve any documents containing words such as 'save', 'saved' and 'saving'.

Similarly, another participant searched with 'poach', but rather than retrieving documents containing poachers, poaching and poacher, no documents were retrieved. There were also occasions where participants removed a plural from a term. For instance, a number of participants highlighted 'highjack' from the term 'highjackers'. However, as 'highjack' did not

appear as a term in any document, this search was ineffective. The information retrieval tool used for this experiment did not use stemming and did not allow participants to search in this manner. Hence, such searches only retrieved documents if that specific part of a word was present in a document. Problems such as this may have decreased confidence in the retrieval system, which may have impacted upon future performance.

4.4 An Analysis of User Behaviour with the Interface

An examination of the search terms used found that almost all of the most common searches were from the first or second sentence of the document. This behaviour could be explained via Zipf's (1949) *principle of least effort*, as participants may have been looking for the first possible word, and hence, may have chosen less than ideal search terms simply because those terms were easy to locate in the first sentence of the document. For instance, for the question on Cosmic Events, the most common initial search term was 'universe', which was located in the first sentence, whereas 'cosmic', which first occurred in the fifth sentence, was the first search term in only six cases.

However, it should be noted that the document collection consisted of newspaper articles, which tend to be written with a first sentence that summarises the whole document. Hence, it is likely that the first sentence may have contained words that were particularly pertinent to the question, and therefore this was not necessarily an ineffective technique.

An examination of the searches also found that many participants used two word searches. For this experiment participants were informed that there were only 150 documents in the collection, and hence, two word searches were often too specific to retrieve relevant documents. This is likely to be a reflection of peoples' experience with Internet search engines, where (due to the massive number of pages indexed) a search with one word is unlikely to be specific enough to retrieve the desired information.

There were also a number of occasions when participants repeated a search term. There are a number of plausible reasons for this behaviour. It is possible that participants may have repeated searches because they had forgotten previous searches. However, it is equally possible that the participant may have intentionally repeated a previously effective search, to verify that all documents had been marked, or to return to a previously viewed document. A number of participants also repeated ineffective searches, where only the source document was retrieved. This could be a reflection of a lack of confidence in the system, or users' may have lacked confidence in their own ability to correctly utilise the system. It is likely that such actions would be reduced with further instruction and practice with the interface.

4.5 Individual Differences

The results of this study indicate that individual differences, including aspects such as education level, primary language, comprehension and general knowledge, tended to influence participants' abilities on the information retrieval task. These findings therefore emphasise the importance of measuring individual differences in the development of an information retrieval system.

As mentioned in Section 1.3.1, there are certain features or characteristics of systems that may increase the performance of some individuals, and decrease the performance of others (Allen, 2000). Hence, it is extremely important that systems are designed with the users in mind, and are designed to maximise those users' skills (Dillon & Watson, 1996).

The findings of this study indicate that the participants who had higher scores on the comprehension test also tended to have extremely little variation in their performance utilising the three information retrieval techniques. In contrast, the participants with lower scores in the test had far more variation, and tended to perform worst when utilising the Word Search technique, and best with LSA Sentence. It is therefore possible that LSA tends to be a 'compensatory' approach, as it increased the performance of the participants with lower abilities.

4.6 Suggestions for Future Research

Although the re-ranking provided by the LSA techniques significantly improved the efficiency of the retrieved lists, the different techniques did not alter participants' performance. This study therefore raises a number of important issues, which should be examined in more detail in future experiments.

The results of this study suggest that participants faced with a reasonably short list will tend to examine all documents in a systematic and sequential manner. This experiment therefore provides a useful insight into the strategies utilised. In future experiments, it would be useful to provide participants with a far larger corpus, to determine the point at which most participants will no longer access all retrieved documents. If a document collection contained many thousands or hundreds of thousands of documents, participants would be unlikely to access all documents, and therefore a technique that effectively re-ranks the retrieved list would be far more useful. However, it is necessary to repeat the experiment with a larger corpus to determine whether this assumption is confirmed.

Furthermore, the results of this study suggest that some features of the interface may have been counterintuitive. For example, most information retrieval tools utilise word stemming, allowing users to search with part of a word and retrieve all documents that contain that word stem. The interface used in this study did not act in this manner, which may have reduced confidence in the system, and may have resulted in an assumption that certain words were not present in documents (e.g., if a search for 'poach' did not retrieve documents, participants may have assumed that there were no documents containing words such as poachers, poaching and poacher). It is necessary to note that participants may have produced more effective queries if they were not limited to only consecutive words.

Additionally, some participants may have found it difficult to identify the relevant portion of documents. Users might be accustomed to information retrieval tools that highlight the searched terms within the documents, allowing participants to quickly and efficiently scan through the document to determine whether the highlighted word occurs in a useful context. Since some aspects of the interface were counterintuitive, it is likely that participants may have required more training to become familiar with the unaccustomed techniques.

It would also be beneficial to use a more varied corpus rather than only newspaper articles. This would reflect a more realistic search and may also be more appropriate to use in conjunction with an LSA technique. Newspaper articles are generally written in a certain formulaic and stylised way, with many similar words and phrases. LSA applies a 'bag of words' approach and therefore may perform better with a more varied corpus.

Finally, since this study indicated the importance of individual differences in information retrieval, it could be useful to include more tests of individual abilities. This could include measures of visual perception or cognitive abilities, to determine exactly which individual traits tended to predict participants' performance. This information could be extremely useful for future training and system development.

5. Conclusions

Achieving efficient and accurate information retrieval is a challenging task. The aim of this study was to determine if the results of a keyword-based information retrieval technique could be improved by re-ranking the results based on the context provided by the surrounding terms. A baseline technique was compared against two LSA techniques, and an analysis of the retrieved documents indicated that the re-ranking provided by the LSA techniques significantly improved the efficiency of the retrieved list.

However, the participants' performance was not altered by the different techniques. Instead, the findings suggest that, when dealing with a small number of documents, participants will generally access all documents retrieved in a systematic manner. It is therefore hypothesised that the re-ranking technique would be more useful in a significantly larger document collection, where a thorough assessment of all documents is impractical.

This study has also emphasised the importance of assessing the impact of individual differences in any information retrieval system. For example, it was found that LSA did improve performance for participants with lower scores on the comprehension test. Results also suggested that education level, primary language and general knowledge tended to influence participants' abilities on the information retrieval task.

Individual abilities may play a very important role in information retrieval and more research is required to determine what aspects of individual differences will most significantly affect outcomes. This suggests that a greater focus needs to be placed on designing information retrieval systems that can be tailored to maximise each user's skills to optimise performance.

6. References

Allen, B. (2000). Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the American Society for Information Science, 51*, 508-520.

Allen, B. (1994). Cognitive abilities and information system usability. *Information Processing and Mangement, 30*, 177-191.

Allen, B. (1991). Cognitive research in information science: Implications for design. *Annual Review of Information Science and Technology, 26*, 3-37.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O. (2004). Efficiency and scaling: hourly analysis of a very large topically categorized web query log. *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, Sheffield, 321-328.

Borgman, C. (1986). The user's mental model of an information retrieval system: an experiment prototype online catalog. *International Journal of Man-Machine Studies, 24*, 47–64.

Carlson, C.N. (2004). Information overload, retrieval strategies and Internet user empowerment. In L. Haddon (Eds), *Proceedings of The Good, the Bad and the Irrelevant (COST 269) 1*, 169-173, Helsinki, Finland.

Chen, H. & Dumais, S.T. (2000). Bringing order to the web: Automatically categorizing search results. *Proceedings of CHI'00, Human Factors in Computing Systems*, pp. 145-152.

Chang, C. & Hsu, C., (1999). Enabling concept-based relevance feedback for information retrieval on the WWW, *IEEE Transactions on Knowledge and Data Engineering, 11(4)*, 595-609.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnass, G.W., Harshman, R.A., (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41(6)*, 391-407.

Dennis, S., Bruza, P. & McArthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology, 53(2)*, 120-133.

Dillon, A. & Watson, C. (1996). User analysis in HCI – the historical lessons from individual differences research. *International Journal of Human-Computer Studies, 45*, 619–637.

Fagan, J.L. (1987). Automatic phrase indexing for document retrieval: an examination of syntactic and non-syntactic methods. *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '87)*, New

Orleans, 97-101.

Fang, H., Tao, T. & Zhai, C. (2004). A formal study of IR heuristics. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in IR (SIGIR'04)*, Sheffield, United Kingdom, 49-56.

Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N. (2005). Concept-based interactive query expansion, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany.

Fuhr, N. (1992). Probabilistic models in information retrieval. *Computer Journal, 35*, 243-55.

Ge, N., Hale, J., & Eugene, C. (1998). A statistical approach to anaphora resolution, *Proceedings of the 6th Workshop on Very Large Corpora*, Montreal, Canada, 161–170.

Getty Images (2007). Retrieved December 2007 from http://www.gettyimages.com/

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, August.

Guo, L. Shao, F., Botev, C., Shanmugasundaram, J. (2003). XRANK: ranked keyword search over XML documents, *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, San Diego, California.

Gyöngyi, Z. & Garcia-Molina, H. (2005). Web spam taxonomy, *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005 in The 14th International World Wide Web Conference (WWW 2005), Nippon Convention Center (Makuhari Messe)*, Chiba, Japan., New York, NY: ACM Press

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.

Kowalski, G (1997). *Information Retrieval Systems Theory and Implementation*. Boston: Kluwer Academic Publishers.

Krovetz, R. (1997). Homonymy and polysemy in information retrieval. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, Madrid, Spain, 72-79.

Krovetz, R. & Croft, W.B. (1992). Lexical ambiguity and information retrieval. ACM *Transactions on Information Systems (TOIS), 10*, 115-141.

Lancaster, F. W. (1968). *IR Systems: Characteristics Testing, and Evaluation*. New York: John Wiley and Sons, Inc.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Lee, D. L., Chuang, H and Seamons, K. (1997). Document ranking and the vector-space model. *IEEE Software, 14(2)*, 67-75.

Liddy, E.D. (2005). Automatic document retrieval. In K. Brown (Eds) *Encyclopedia of Language and Linguistics*, 2nd Edition, Oxford: Elsevier Limited.

Manning, C.D., Raghavan, P. & Schutze, H., (2007). *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Moffat, A. & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems, 27(1)*, 2:1–2:27.

Navigli, R. & Velardi, P. (2003). An analysis of ontology-based query expansion strategies. *Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM 2003) at the 14th European Conference on Machine Learning (ECML 2003)*, Dubrovnik, Croatia.

Papadimitriou, C. H., Raghavan, P., and Tamaki, H. (1998) Latent semantic indexing: A probabilistic analysis. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 159-168.

Parsons, K., McCormac, A. & Butavicius, M. (2007). *An Analysis of Topicality and Relevance*, DSTO Technical Report, DSTO-TR-2078.

Pincombe, B. (2004). *Comparisons of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus*, DSTO Research Report, DSTO-RR-0278.

Porter, M.F. (1980). An algorithm for suffix stripping, *Program, 14(3)*, 130-137

Ravin, Y. & Leacock, C. (2000). Polysemy: An overview. In Y. Ravin & C. Leacock, *Polysemy: Theoretical and Computational Approaches*. Oxford University Press: New York, New York.

Ruthven, I., Lalmas, M. & van Rijsbergen, C.J. (2003). Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology, 54(6)*, 528-548.

Ruthven, I., Tombros, A. & Jose, J. (2001). A study on the use of summaries and summary-based query expansion for a question-answering task. *Proceedings of the 23rd BCS European Annual Colloquium on IR Research (ECIR 2001)*. Darmstadt.

Salton, G. (1988). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Readings Massachusetts: Addison-Wesley Publishing Company.

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company.

Simon, H.A. (1996). *The Sciences of the Artificial*. Third Edition, MIT Press: Cambridge, Massachusetts.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4)*, 35–43.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Sparck Jones, K. (1971). *Automatic keyword classification for retrieval*. London: Butterworth.

Spink, A., & Wilson, T.D. (1999). Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context. *Proceedings of MIRA 99: Evaluation Frameworks for Multimedia Information Retrieval Applications*, Department of Computing Sciences, University of Glasgow. Scotland, 75-92.

Stanney, K.M., & Salvendy, G. (1995). Information visualization: Assisting low spatial individuals with information access tasks through the use of visual mediators. *Ergonomics, 36*, 1184-1198

Su, L. (1998). Value of search results as a whole as the best single measure of IR performance. *Information Processing and Management, 34(5)*, 57-579.

Swanson, D.R. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39*, 92-98.

Tague, J., & Schultz, R. (1989). Evaluation of the user interface in an IR system: A model. *Information Processing and Management, 25(4)*, 377-389.

The Apache Software Foundation (2007). Apache Lucene, Retrieved June 2007 from http://lucene.apache.org/index.html

van Rijsbergen, C.J. (1971). An Algorithm for Information Structuring and Retrieval. *Comput. J. 14 (4)*, 407-412.

van Rijsbergen, C. J. (1979). *Information Retrieval.* London: Butterworths.

Voorhees, E.M. & Harman, D. (2000). Overview of the Eighth Text REtrieval Conference (TREC-8), In E.M. Voorhees & D.K. Harman (Eds.), *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*. Gaithersburg, Maryland, USA: National Institute of Standards and Technology.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine learning, Pittsburgh, Pennsylvania, 977-984.*

Xu, J & Croft, W.B (2000) Improving the Effectiveness of Information Retrieval with Local Context Analysis, *ACM Transactions on Information Systems, 18(1)*, 79-112.

Xu, J & Croft, W.B. (1996). Query Expansion Using Local and Global Document Analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 4-11. Zurich, Switzerland.

Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhang, Y., Park, L.A.F. & Moffat, A. (2008). Parameter sensitivity in rank-biased precision. *Proceedings of the 13th Australasian Document Computing Symposium (ADCS '08)*, Hobart, pp. 61-68.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort.* Addison Wesley, Cambridge, MA.

Appendix A: Example Documents

An international team of astronomers last week reported that it may have witnessed the birth of a pair of quasars. Dr. Georges Meylan of NASA's Space Telescope Institute said his team has detected an object in the constellation Pisces that appears to be a pair of quasars separated by less than 100,000 light years -- a small distance in cosmic terms. One light year equals about 6 trillion miles.

Quasi-stellar objects, or quasars, are thought to be the very bright cores of certain galaxies, and some astronomers believe that their intense energy may be caused by their material being pulled inward to giant black holes.

The international team, which also includes scientists from Caltech in Pasadena and from West Germany, said it has discovered what it believes to be the first true set of quasar "twins" using a 40-inch telescope near Le Serena, Chile.

In findings presented at the American Astronomical Society's national meeting in Alexandria, Va., the researchers said the quasar twins lie so close together in space that astronomers speculate their gravity fields may be interacting with one another and spurring formation of the intensely bright objects.

A fleeing suspect bit off more than he could chew -- or perhaps it was the other way around -- when he tried to fight off a police dog working with California Highway Patrol officers Thursday night.

The 26-year-old suspect was taken to San Clemente General Hospital for treatment of dog bites on both thighs.

The incident began at 6:47 p.m. when the CHP pursued a suspected drunk driver south on Interstate 5 in the Mission Viejo area, CHP Sgt. Dennis Dyer said. A check of the license plate indicated that the car was stolen.

The suspect suddenly stopped in the freeway's center divider near Cristianitos Road and ran across the freeway lanes, almost getting hit, and into a large field near Camp Pendleton, he said.

The CHP searched the field, assisted by helicopters from the Orange County Sheriff's Department. A second helicopter from the Costa Mesa Police Department located the suspect by using infrared sensors that detected the suspect's body heat in the tall, dense brush, Dyer said.

That's when Nick, a German shepherd police dog with the Sheriff's Department, was sent in. The suspect, identified as Fernandez Hernandez of Modesto, tried to fight off the dog. "He lost," Dyer said.

Hernandez was arrested on suspicion of car theft and driving while intoxicated, Dyer said.

Four North Hollywood residents who helped authorities capture an arsonist responsible for several fires in their neighborhood were honored Thursday by the Los Angeles Fire Department.

Stephanie Sheppard-Tapper, Linda Ornelas, Jim Khavarian and Sam Silberscheim, all residents of the 5400 block of Bellingham Avenue, were awarded certificates of appreciation by the Fire Commission.

The certificates cited the foursome for "unselfish courage and acts of heroism" in the Feb. 5 capture of Reunald Parker, 31, after a fire was intentionally started in a parking garage on the block.

Authorities said the four residents were members of a fledgling Neighborhood Watch group that had begun patrolling the area after several arson fires had occurred. On Feb. 5 they saw Parker leaving the scene of the garage fire and followed him. They later led fire investigators to Parker and he was arrested on suspicion of arson. Parker pleaded guilty to 17 counts of arson on April 18 and was sentenced to 14 years in prison.

A woman suffered minor injuries today when she was struck by a train as it approached the Fullerton Amtrak station.

Guadalupe Magdalena, 21, was apparently walking near the railroad track bordering Orangethorpe Avenue about a mile from the station when the 7:30 a.m. northbound train struck her, Fullerton Police Lt. Al Burks said.

Engineers told police that the train was traveling about 40 m.p.h. when the woman was spotted walking north on a pedestrian walkway. The train's horn was sounded, and the woman tried to move aside when the train hit her, Burks said.

The train "clipped her on the left shoulder," Burks said.

Magdalena could not remember how she was struck, he said.

She was taken to Anaheim Memorial Hospital's emergency room, where she was treated for multiple scrapes and bruises, a hospital representative said.

The accident is under investigation.

An international art dealer suspected in the theft of millions of dollars worth of oil paintings was found dead in his London, Ontario, apartment in an apparent murder-suicide, Canadian police said Tuesday. Superintendent Don Andrews told the Associated Press that the bodies of Peter Nixon, 58, and his wife Evelyn, 55, were found Monday in the bedroom of their apartment and were believed to have been dead about a week. A handgun was also found. Andrews said Peter Nixon was at the center of an international investigation into the theft of 20 oil paintings stolen from two British homes -- one in England and one in Scotland -- in 1981 and 1982. He said both the Canadian and British police, as well as the FBI, had been involved in the investigation. Twelve of the 20 paintings, some of which date back several centuries and are valued at more than $10 million, have been recovered in Europe, the United States and Canada.

Appendix B: Questions

Document Set A

Imagine that you are writing a report on cosmic events. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on robotic technology. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on heroic acts. Please indicate the documents that you would refer to in writing such a report.

Document Set B

Imagine that you are writing a report on the forgery of currency. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on the theft and forgery of art. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on the exploitation of children in the labour market. Please indicate the documents that you would refer to in writing such a report.
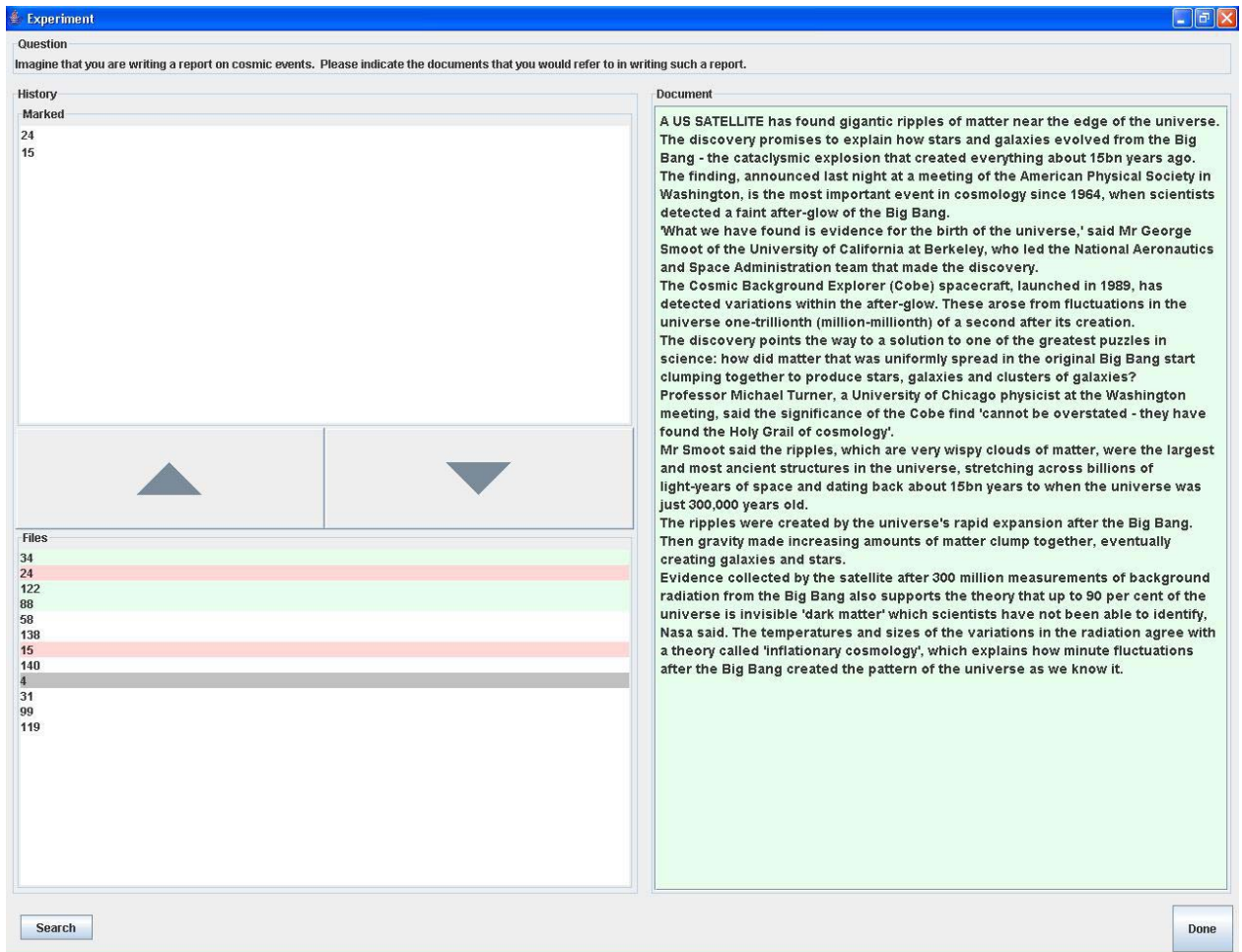
Document Set C

Imagine that you are writing a report on wildlife preservation and wildlife poaching. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on airport security. Please indicate the documents that you would refer to in writing such a report.

Imagine that you are writing a report on the decrease in birth rates. Please indicate the documents that you would refer to in writing such a report.

Appendix C: A screenshot from the interface

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA		1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)
2. TITLE The Use of a Context-Based Information Retrieval Technique	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)	
4. AUTHOR(S) Kathryn Parsons, Agata McCormac, Marcus Butavicius, Simon Dennis and Lael Ferguson	5. CORPORATE AUTHOR DSTO Defence Science and Technology Organisation PO Box 1500 Edinburgh South Australia 5111 Australia	

6a. DSTO NUMBER DSTO-TR-2322	6b. AR NUMBER AR- 014-585	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE July 2009

8. FILE NUMBER 2008/1020495/1	9. TASK NUMBER INT 007/020	10. TASK SPONSOR Intelligence	11. NO. OF PAGES 50	12. NO. OF REFERENCES 57

13. URL on the World Wide Web http://www.dsto.defence.gov.au/corporate/reports/DSTO-TR-2322.pdf	14. RELEASE AUTHORITY Chief, Command, Control, Communications and Intelligence Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS	Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.shtml

Information retrieval
Psychology
Latent Semantic Analysis
Empirical methods
Individual differences

19. ABSTRACT
Since users are faced with an ever increasing amount of data, fast and effective retrieval of required information is of vital importance. This study examined two methods of using Latent Semantic Analysis (LSA) to improve the results retrieved using a keyword-based technique using sentence or document context. Fifty participants retrieved information using a standard keyword technique and the two LSA techniques. Although the re-ranking provided by the LSA techniques ordered the documents in a significantly more efficient manner, no significant differences were found in user performance with regards to accuracy, time taken or documents accessed for the different techniques. However, individual differences did significantly influence results, most notably in regards to participants' scores on a comprehension test. This study therefore highlights the importance of examining the impact of individual differences in any information retrieval system.