# Advantages of Query Biased Summaries in Information Retrieval

Anastasios Tombros
Computing Science Department
University of Glasgow
Glasgow G12 8RZ
Scotland
tombrosa@dcs.gla.ac.uk
www.dcs.gla.ac.uk/~tombrosa/

Mark Sanderson
CIIR, Computing Science Department
University of Massachusetts
Amherst, MA 01003
U.S.A.
sanderso@cs.umass.edu
www-ciir.cs.umass.edu/~sanderso/

**Abstract** This paper presents an investigation into the utility of document summarisation in the context of information retrieval, more specifically in the application of so called query biased (or user directed) summaries: summaries customised to reflect the information need expressed in a query. Employed in the retrieved document list displayed after a retrieval took place, the summaries' utility was evaluated in a task-based environment by measuring users' speed and accuracy in identifying relevant documents. This was compared to the performance achieved when users were presented with the more typical output of an IR system: a static predefined summary composed of the title and first few sentences of retrieved documents. The results from the evaluation indicate that the use of query biased summaries significantly improves both the accuracy and speed of user relevance judgements.

## 1 Introduction

In a typical interaction with an information retrieval (IR) system, the user enters a specific information need, expressed as a query. Figure 1 shows the typical response of a system in relation to the query "commercial aircraft manufacturers". For each of the documents presented in the *retrieved document list*, their title, first few sentences, and their location is shown to the user. This amounts to a form of *predefined static summary* of each document. A quantification of relevance to the query is also shown next to the title of each document. Utilising this information, users have to decide which of the retrieved documents are most likely to convey their information need. Ideally, it should be possible to make this decision without having to refer to the full document text. However, it is unlikely that the first few sentences of a document and its title will give a clear view of the way in which the document relates to a user's query. As a result, users frequently have to refer to the full text of the document, making the process of relevance judgement time-consuming. Even when users refer to the full text, its very nature may have a confounding effect: it may be large and difficult to manage, and relevant information may be widely scattered, and therefore hard for the user to extract.

Recognising the cognitive overhead this imposes, there have been attempts to concentrate users' attention on the parts of the text that possess a high density of relevant information. These methods, known as passage retrieval (Callan 1994; Knaus et al. 1995), identify and present to the user individual text passages that are more focussed towards particular information needs than the full document texts. The main advantage of these approaches is that they provide an intuitive overview of the distribution of the relevant pieces of information within the documents. As a result, it may be easier for users to decide on the relevance of the retrieved documents to their queries. However, even this approach does not alleviate the need to refer to the full text of the retrieved documents.

This paper investigates a novel approach to the presentation of clues on the relevance of retrieved documents to information needs. The approach aims to minimise users' need to refer to the full document text, while at the same time to provide enough information to support their retrieval decisions. It is proposed that an automatically generated summary of each document in a retrieved document list, biased to a user's query, can provide such a function.

A document summary conventionally refers to an abstract-like condensation of a full text document, that presents succinctly the objectives, scope, and findings of the document (Maizell et al., 1971). The minimal function that any useful summary should provide is being *indicative* of the source's content, thus helping a reader to decide whether looking at the whole document will be worthwhile. In this sense, summaries can serve as a preview format to support relevance assessments on the full text of documents (Rush et al., 1971). Some summaries may also contain *informative* material, in which case they can be used as stand-alone document surrogates.

Since its beginnings (Luhn, 1958; Edmundson, 1969), automatic text summarisation has been performed primarily by the selection of sentences from the original document; scores are assigned to sentences according to a set of extraction criteria (Paice, 1990), and the best-scoring sentences are presented in the summary. This approach can be better termed as *sentence extraction* rather than summarisation, and although it does not perform an in-depth analysis of the source text it can produce indicative summaries, which can help users in relevance judgements.


SIGIR'98, Melbourne, Australia

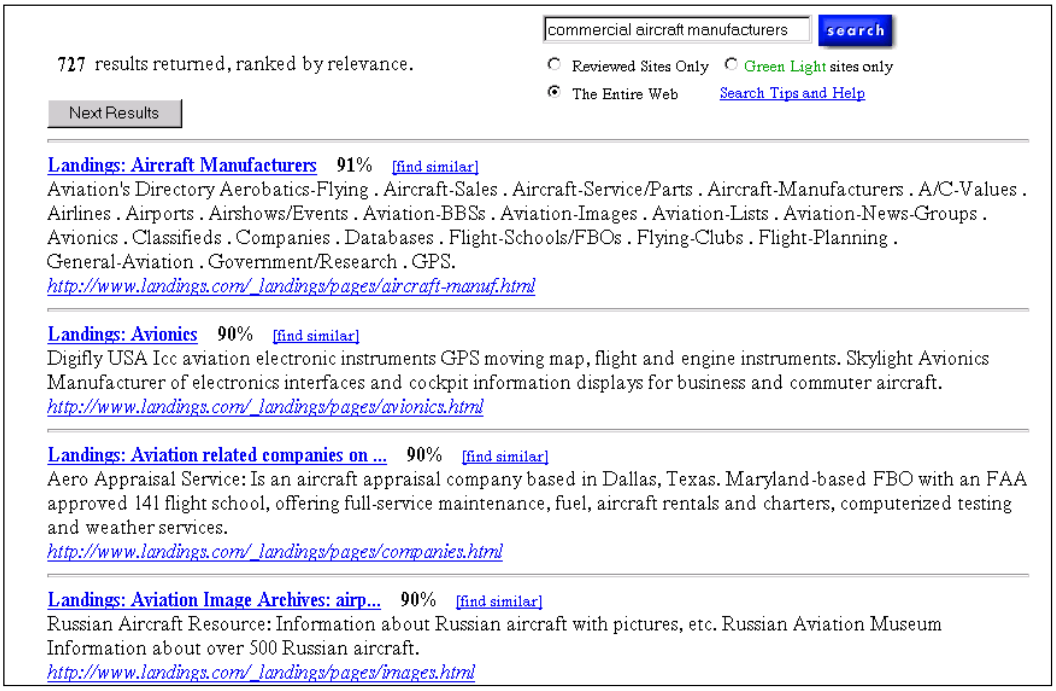



**Figure 1.** Typical output of an IR system

Although there have been attempts to produce coherent summaries by language generation, and artificial intelligence techniques (Jacobs & Rau, 1990; McKeown & Radev, 1995), they are capable of processing texts only within a narrow domain whose characteristics are predictable and well understood (e.g. news stories, financial and commercial reports). There is not enough evidence that such systems will be able to manipulate domain independent text in the foreseeable future.

In order to evaluate the effectiveness of the query biased summaries, a task-based evaluation scheme was developed. It involved the integration of the summarisation system with an IR system in order to perform the evaluation in an end-user, operational environment. Traditionally the evaluation of summarisation systems involves measuring quantitative attributes of the summaries (e.g. similarity between automatically generated summaries and human prepared ones) (Edmundson 1969; Kupiec et al. 1995; Salton et al. 1997). There have recently been proposals (Hand 1997) and attempts (Miike et al. 1994; Mani and Bloedorn 1997) to develop schemes that measure qualitative features of the systems in a task-based environment. However, these attempts have so far been limited as far as the experimental procedure is concerned. We believe that the work described in this paper offers a more complete approach to the task-based evaluation of summarisation systems.

In the remainder of this paper, we first describe the query biased summarisation system that was developed. Subsequently, the experimental procedure used to investigate the effectiveness of the system is detailed, and the results obtained from this procedure are presented. Finally some conclusions that can be drawn from the presented work are discussed, as well as some points for future research.

## 2 The Summarisation System

Before building a summarisation system one needs to establish the type of documents to be summarised and the purpose for which the summaries are required. The summaries generated by our system were aimed at providing users working on an interactive IR system with information on the relevance of documents retrieved in response to their query. It was assumed that users were willing to spend a limited amount of time to go through the list of retrieved documents and to decide on their relevance. Moreover, the summaries were to be user-directed: biased towards the user's query. With these factors in mind, we now describe the system. It was based on a number of sentence extraction methods (Paice 1990) that utilise information both from the documents of the collection and from the queries used.

The documents to be summarised were articles of the Wall Street Journal (WSJ) taken from the TREC collection (Text REtrieval Conferences) (Harman 1996). In order to decide which aspects of the articles would provide utility to generating a summary, their characteristics were examined in a small scale study. The methodology that was followed involved examining 50 randomly selected articles from the collection and attempting to extract conclusions about the distribution of important information within them. Their title, headings, leading paragraph, and their overall structural organisation were studied. This sample collection was used for experimentation with various system parameters, in order to approximate the best settings for the summarisation system. Although the sample of the documents was small, there was a strong uniformity in the characteristics of the sample that allowed for a generalisation of the conclusions to the entire collection. The sentence extraction methods that were employed are now described.

### 2.1 Title method

It is generally known that the titles (headlines) of news articles tend to reveal the major subject of the article; they usually act as a preview to the whole article. This belief was strengthened by the sample study of the document collection: titles in the WSJ collection tended to refer to the main subjects of the article. In order to exploit this feature of the collection, terms that occurred in the title section of the documents were assigned a positive weight (title score).

### 2.2 Location method

It was uniformly noted from the sample study, that the leading paragraph of each article provided much information on the article's content. This conclusion seemed to be in agreement with (Brandow et al. 1995), who suggested that "improvements (to the auto-summaries) can be achieved by weighting the sentences appearing in the beginning of the articles more heavily". In order to quantify this contribution, an ordinal weight was assigned to the first two sentences of each article.

Section headings within articles provide evidence about their division into meaningful semantic units. This was a uniform conclusion obtained from the sample study of the WSJ collection. In a similar way that the title of an article is indicative of its content, a section heading reveals its principal information. In order to exploit the evidence provided by section headings, a 'heading score' was assigned to each one of the sentences comprising a heading.

### 2.3 Term occurrence information

In addition to the evidence provided by the structural organisation of the documents, the system utilises the number of term occurrences (*TO*) within each document to further assign weights to sentences. Instead of merely assigning a weight to each term according to its frequency within the document, the system locates clusters of *significant words* (Luhn 1958) within sentences, and assigns scores to them accordingly.

Based on the sample study of the WSJ collection, we concluded that a reasonable *TO* value for establishing the significance of a term was 7, and that this value should be adjusted according to the length of the document. The value of 7 was applied to medium-sized documents of the collection (between 25 - 40 sentences). These numbers were also obtained through the sample analysis of the document corpus. For documents that contain more than 40 sentences, the *TO* value is augmented by 10% of the increase in document size. The increase is calculated in respect to the upper limit of the medium document size, i.e. 40. For example, for a document that is 50 sentences long, the increase in size is 10, and therefore the *TO* limit is set to: $7 + [0.1 \bullet (10)] = 8$. For documents smaller than 25 sentences the same procedure was applied, calculating the decrease in document size in respect to the lower limit of the medium document size (i.e. 25).

In order to extend the notion of significance from single terms to clusters of terms, we define two terms as being *significantly related* if both of them are significant, and between them are no more than 4 non-significant words. If in that way a sentence contains two or more clusters, the one with the highest significance factors is taken as the measure for that sentence. This approach is in agreement with Luhn's suggestions (Luhn 1958), as well as with more recent studies that show that in the English language 98% of the lexical relations occur between words within a span of 5 words in a sentence (Abracos and Lopes 1997). The scheme that is used for computing the significance factor for a sentence was originally proposed by (Luhn 1958). This scheme consists of defining the extent of a cluster of related words (i.e. the actual number of words in the cluster), and dividing the square of this number by the total number of words within this cluster.

### 2.4 Biasing summaries towards queries

The long standing motivation for this work was a belief that if, in the retrieved document list, users of IR systems could see the sentences in which their query words appeared, they could better judge the relevance of documents. Therefore, a 'query score' was calculated for each of the sentences of a document. The computation of that score was based on the distribution of query terms in each sentence. This was based on the belief that the larger the number of query terms in a sentence, the more likely that sentence conveyed a significant amount of the information need expressed in the query. The actual measure of significance of a sentence in relation to a specific query, was derived by dividing the square of the number of query terms included in that sentence by the total number of the terms of the specific query.

For each sentence, that score was added to the overall score obtained by the sentence extraction methods, and the result constituted the sentence's final score. The summary for each document was then generated by outputting the top-scoring sentences, until a desired summary length was reached. This was defined to be 15% of the document's length up to a maximum of five sentences. Such a value seems to be in general agreement with suggestions made by (Edmundson 1964), and (Brandow et al. 1995).

## 3 Experimental Design

The aim of the specific experiment was to establish that the use of query biased summaries in a retrieved document list would have a positive effect on the process of relevance judgement by users. However, throughout the discussion below, the reader should bear in mind that this hypothesis is related to the task-based evaluation scheme for summarisation systems which is proposed in this paper.

The proposed evaluation scheme judges the utility of a summarisation system in the context in which it will eventually be used, and for the purposes for which it has been built. According to this rationale, the *indicative* function (Rush et al. 1971) of a summary is the one which should be primarily evaluated. By integrating the summarisation system into an existing IR system, we both define its operational context, and its primary function: the query biased summaries are used as a preview format in order to support a relevance decision by users. Therefore, the proposed evaluation scheme aims at measuring the effectiveness of the summaries in supporting user's relevance decisions. This principal aim of the evaluation process can now be clearly mapped to the research hypothesis that we propose to establish: by proving this hypothesis a positive indication for the effectiveness of query biased summaries is gained.

### 3.1 Design considerations

Having established the actual hypothesis to be examined, we can introduce the basic design settings upon which the testing of the hypothesis was conducted.

**Experimental conditions.** We are interested in two levels of an independent variable in our experimental design: the use of query-biased summaries in a ranked list of retrieved documents; and the use of static pre-defined summaries (the title and first few lines of a document) in such a list. In this way, the design comprises two tasks that a group of subjects will have to perform: to judge the relevance of the documents in a ranked list, with either query biased or predefined summaries. The performance of the users in these tasks constitutes the dependent variable of the experiment, and we shall attempt to prove that any variation of the performance between the two groups is attributed only to the change in the level of the independent variable.

**Groups of subjects.** In the experiment described in this paper two groups consisting of 10 subjects each were employed. Subjects were randomly assigned to a group (by means of a draw), and each group was assigned to one experimental condition only (*independent groups design* (Miller 1984)). It is believed that the number of 20 subjects is sufficient for attributing significance to any results obtained. The subjects comprised mainly of postgraduate students doing a conversion course in computer science. Clearly, this population is not representative of that which we wish to generalise the conclusions to. However, relevant studies have shown that although there are 'risks' in generalising experimental results in such cases, an investigator may feel safe in doing so since the statistical differences introduced are generally of a small scale (Keppel 1973).

**Situational variables.** Such variables are associated with the experimental situation itself (e.g. background noise, equipment settings, experimenter's behaviour, etc.). Such factors can easily confound the effects of the independent variable if they change systematically from one condition to another. There was an effort to hold the situational variables constant throughout the experimental procedure: computers used were identically set up (both from a hardware and software point of view); there was only one experimenter present throughout the experimental procedure; and the experimental sessions took place at similar times in a short space of days in an effort to ensure that the external conditions of the room (especially the external noise) would be as similar as possible.

**Retrieval task.** In choosing the exact form of retrieval task to be performed by the subjects, it was decided to opt for a task that involved subjects in only performing relevance judgements. Therefore subjects were presented with a retrieved document list and told that this list was the result of a retrieval based on a particular query, which they were also shown. The only actions they could perform was to move through the list or to fetch the full text of the documents listed within it. Their task, therefore, was to identify, in a limited amount of time, as many relevant documents as possible.

**Queries used.** The queries used (50) were randomly selected from the queries of the TREC test collection. Test collection queries were chosen because a list of documents manually judged to be relevant to each query was already available. This list was used as the standard against which the subjects' relevance judgements were compared. In Figure 2 a sample TREC query, also known as a topic, is shown. As can be seen, TREC topics are long and detailed and this raised the issue of what part of the topic should be used when generating the retrieved document list to present to the subjects. The 'title' section of the queries was felt to be typical of the queries entered by users in an interactive IR system, while the other sections were regarded as a detailed description of the information need. Therefore, to generate the ranked document list, the title was submitted as a query to an IR system and the rest of the topic was shown to the subjects so that they could better understand the information need of the query.

<top>
<num> Number: 033

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description: Document must report on how the importation of foreign textiles or textile products has influenced or impacted on the U.S. textile industry.

<narr> Narrative: The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. "Textile industry" includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.

</top>

**Figure 2.** A sample TREC topic

**IR system used.** The retrieval system used to generate the retrieved document lists was a classic document ranking system employing a tf•idf term weighting scheme with stop word removal and word stemming using the Porter stemmer (Porter 1980)

### 3.2 Operationalising the experiment

The actual steps of the experimental procedure are as follows:

- Each subject was randomly assigned to one of the two levels of the independent variable in the way that was previously explained. In that way the task that each subject should perform was defined.
- In order to perform the relevance judgements, each subject was presented with 5 queries which were randomly assigned to subjects (by means of a draw) from the set of 50 TREC queries.
- As soon as the subject was placed in front of the assigned computer, instructions about the experiment were handed to him/her. Subjects could then go through the instructions in their own time. Any questions about the instructions were answered by the experimenter. Subjects were otherwise not told of the hypothesis being tested.

You have just typed in the following query:

**"Alternatives to Postscript"**

**A brief explanation about the query:**

To be relevant, a document must identify a need for, or the existence of, an alternative to Postscript, a page description language.

---

**The following articles have been retrieved in relation to that query:**

**Documents 1-10 (of 50) matching the query.**

**[1]**

**Publishing: Papers Take Alternative Path to Success ---- By Andrew Patner Staff Reporter of The Wall Street Journal**

As publisher of the Chicago Reader, an alternative weekly newspaper, Robert A. Roth refuses to read daily papers. He's afraid their stodgy ways might be infectious. The 68 papers that belong to the Association of Alternative Newsweeklies now reap $100 million-plus in combined revenue, and their 3.5 million total circulation includes an enviable share of the coveted 18- to 35-year-old market, the very group that daily newspapers are having the most trouble attracting. The alternative weeklies group admitted seven new papers this year and had applications from seven more. Cartoonists Lynda J. Barry, whose work appears in Esquire, Mother Jones and other magazines, and Matt Groening, of "The Simpsons," started with alternative papers and still do weekly strips for them.

[ Click here to get the whole article ]

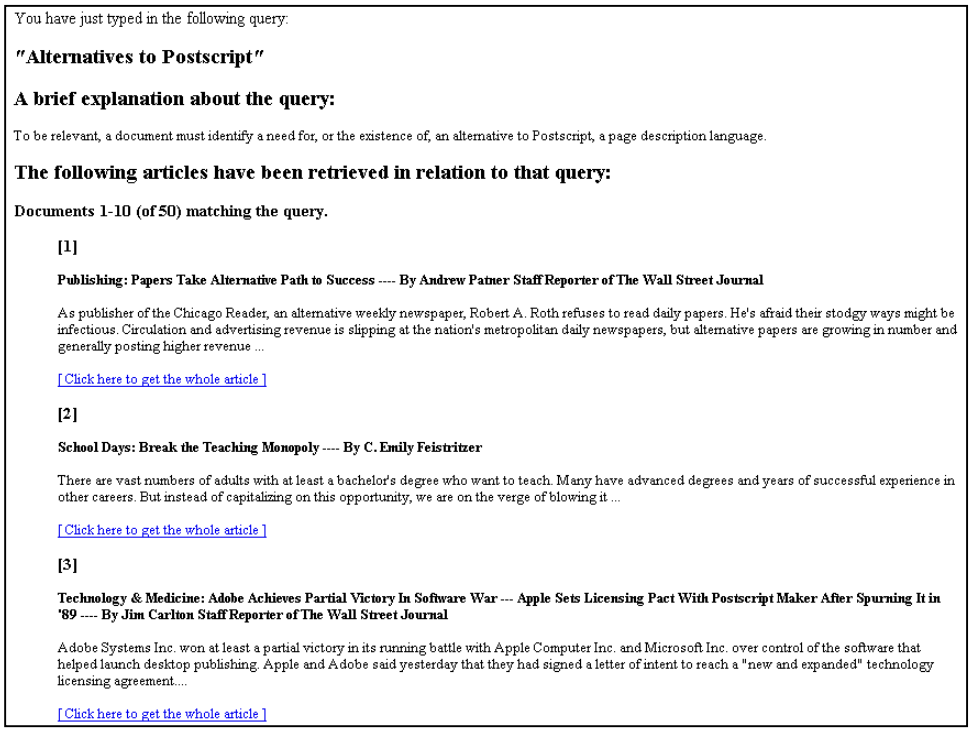
**[2]**

**School Days: Break the Teaching Monopoly ---- By C. Emily Feistritzer**

There are vast numbers of adults with at least a bachelor's degree who want to teach. Many have advanced degrees and years of successful experience in other careers. That means one has to go to college and complete a series of education courses approved by the state department of education. Alternative teacher certification programs are not wanting for detractors. Since last week, after an article about our study of alternative teacher certification appeared in the New York Times, we've gotten scores of phone calls from people who want to know where the alternative certification programs are -- they included a dentist who wants to teach biology, an international businessman with a master's degree in physics who wants to teach high school physics and math, and a corporate executive who wants to teach elementary school.

[ Click here to get the whole article ]

**Figure 3a.** A ranked document list with summaries biased to the query “Alternatives to Postscript”. The summaries contain 15% of the original document's sentences, up to a maximum of 5 sentences.



You have just typed in the following query:

**"Alternatives to Postscript"**

**A brief explanation about the query:**

To be relevant, a document must identify a need for, or the existence of, an alternative to Postscript, a page description language.

**The following articles have been retrieved in relation to that query:**

**Documents 1-10 (of 50) matching the query.**

**[1]**

**Publishing: Papers Take Alternative Path to Success ---- By Andrew Patner Staff Reporter of The Wall Street Journal**

As publisher of the Chicago Reader, an alternative weekly newspaper, Robert A. Roth refuses to read daily papers. He's afraid their stodgy ways might be infectious. Circulation and advertising revenue is slipping at the nation's metropolitan daily newspapers, but alternative papers are growing in number and generally posting higher revenue ...

[ Click here to get the whole article ]

**[2]**

**School Days: Break the Teaching Monopoly ---- By C. Emily Feistritzer**

There are vast numbers of adults with at least a bachelor's degree who want to teach. Many have advanced degrees and years of successful experience in other careers. But instead of capitalizing on this opportunity, we are on the verge of blowing it ...

[ Click here to get the whole article ]

**[3]**

**Technology & Medicine: Adobe Achieves Partial Victory In Software War --- Apple Sets Licensing Pact With Postscript Maker After Spurning It in '89 ---- By Jim Carlton Staff Reporter of The Wall Street Journal**

Adobe Systems Inc. won at least a partial victory in its running battle with Apple Computer Inc. and Microsoft Inc. over control of the software that helped launch desktop publishing. Apple and Adobe said yesterday that they had signed a letter of intent to reach a "new and expanded" technology licensing agreement....

[ Click here to get the whole article ]

**Figure 3b.** A ranked document list presenting the titles and first three sentences of the documents. This representation is assumed to be the one of a typical IR system.

Subsequently, the retrieved document list, composed of the 50 highest ranked documents, was presented to each subject. In Figures 3.a and 3.b sample screenshots from the two different types of ranked lists used in the experiment are presented.

- Subjects then had a time limit of 5 minutes to identify the relevant documents to each query that was assigned to him/her. The timing was performed by the experimenter. The relevant documents were marked by the subjects on an answer sheet prepared for each query. Subjects were instructed to examine documents in the order presented in the retrieved document list and to mark the document they were last examining when the 5 minute time period expired. If a subject managed to examine all the retrieved documents for a query before the specified time ended, the experimenter was notified and this fact was noted on the answer sheet. The sheet was returned to the experimenter after the 5 minutes were up.
- Once the subject had completed the assigned task, a questionnaire was presented to him/her. Once completed, the questionnaire was returned to the experimenter.
- A brief discussion was subsequently held with those subjects that were further interested. At that point, the nature of the experiment was presented to them. They were encouraged to express their opinions about the experimental procedure and the overall reasoning of the experiment.

The data that was collected from each subject comprised the completed questionnaires and answer sheets for each query. The next section will present an analysis of the results derived from them.

## 4 Experimental Results

The variable we wished to examine through experimentation (the dependent variable), was the performance of the users in the process of relevance judgements on documents retrieved by specific queries. In order to do so, a set of criteria that provided a satisfactory coverage of the aspects of the dependent variable had to be defined. Such criteria for the experiment conducted were:

- The recall and precision of the relevance judgements performed by the subjects.
- The speed with which these judgements were performed.
- The need of the subjects to seek assistance from the full text of the retrieved documents.
- The subjective opinion of the users about the assistance provided by the information that was accompanying each retrieved document.

In the following paragraphs the results obtained through the experimental procedure are presented and analysed.

### 4.1 Recall and Precision

The effectiveness of the relevance judgements can be quantified by two measures: the recall (number of relevant documents correctly identified by a subject for a query divided by the total number of relevant documents, within the examined ones, for that query), and the precision of the judgements (relevant documents correctly identified divided by the total number of indicated relevant documents for a query).

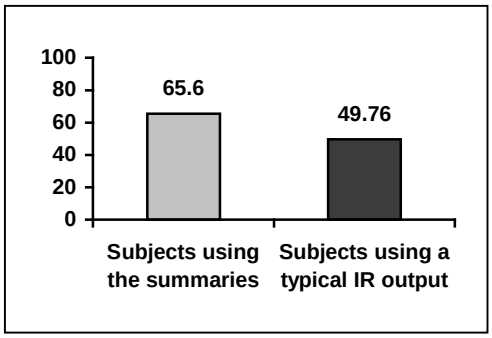



**Figure 4.** Recall values for the two groups

The recall values for the group of subjects using the query biased summaries is considerably larger than that of the group using a typical IR output: the difference in performance is 15.84%. The interpretation of this result is that users in the 'summary group' managed to successfully identify a larger number of relevant documents than the other group.

Nevertheless, in order to have an overall view of the effectiveness of the relevance judgements, we need to examine the performance of the two groups in the precision of the judgements. The precision values obtained for the two experimental groups are presented in Figure 5. The data presented in Figures 4 and 5, were acquired by averaging the results for each query over the total number of queries, thus producing the average recall and precision values per query. In order to establish the statistical significance of these results, *t-tests* (Miller 1984) were performed on both these measures, indicating that, with a probability of error 0.05, the results are attributed to the change of level of the independent variable and not to chance factors.

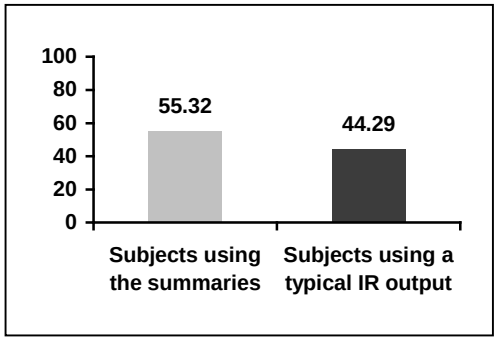



**Figure 5.** Precision of the judgements

The use of the TREC relevance assessments in obtaining these results is a factor that has to be examined. Initially, it can be argued that the acceptance of the TREC assessments as the 'correct answers' for the relevant documents to each query is not fully justified: different users may judge different documents as relevant, and furthermore it is possible that the TREC assessments are not totally accurate. This is possibly a significant factor for the rather low values obtained in the recall and the precision of the relevance judgements in both experimental conditions.

A further point of discussion is that TREC documents do not have to be wholly, or even primarily, about a request topic

in order to be deemed relevant. A couple of sentences relevant to query content are an adequate criterion for the relevance of document in TREC. The summarisation system selects information from each document based on, among other criteria, the distribution of query terms in its constituent sentences. In this way, the generated summaries aim to help users to more easily identify the relevant pieces of information that are contained in each document. Thus, the rationale of the summarisation system approximates the rationale of the TREC assessments: high distribution of query terms in a sentence, can possibly be evidence of its relevance to the query. The results reported in this section strengthen this belief: subjects using the summaries have been significantly assisted by the conveyed information in relation to the specific query.

A final point of discussion is the effectiveness of the summaries in 'warning' users on the non-relevance of documents. A first indication is provided by the highest precision values that the group of subjects using the summaries showed. They tended to erroneously judge less documents than the other group, therefore, it could be said that summaries help them to identify non-relevant documents more effectively. In the case where no relevant documents where present in the part of the retrieved document list that users examined - there were 8 such cases in the group using the summaries, 9 in the other group - the former group marked irrelevant documents as relevant 4 out of 8 times, while the latter group did this 8 out of 9 times.

Therefore, we conclude that subjects using query biased summaries in a retrieved document list, performed their relevance judgements significantly better than those using the classic IR standard: the title and first few lines of a document. In essence this means that query biased summaries allow users to identify more relevant documents, and identify them more accurately.

### 4.2 Speed

The actual number of documents that each subject managed to examine within the specified 5 minute period was known, since for each query the last document examined was indicated by each subject. This number is used as an indication of the speed with which the relevance judgements were performed by each subject.

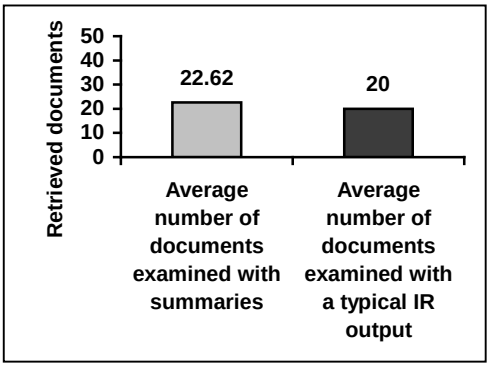



**Figure 6.** Speed results

In Figure 6 the results obtained for the speed of the relevance judgements are presented. These results have been obtained by averaging the number of examined documents for the two experimental conditions over the total number of queries, i.e. 50. Thus, subjects using the query biased summaries examined on average 22.62 documents per query, while subjects using a typical IR output examined on average 20 documents. However small this difference may seem, it amounts to a 13% increase in the average number of documents examined. Taking also in consideration the specific time limits of the experiment, we conclude that there is a definite tendency for users presented with the query biased summaries to perform relevance judgements quicker than users presented with a standard static system output.

### 4.3 Reference to the full text of the documents

The data collected on the users' reference to the full text of documents showed that subjects using the query biased summaries had to refer to 0.3 full texts per query, whereas subjects from the other experimental group had to refer to 4.74 on average. If we normalise these values to the average number of documents that each experimental group examined for each query, we obtain the results shown in Figure 7. This figure shows that each subject using the summaries had to refer to the full text of 1.32% of the documents for each query, while subjects in the other experimental condition had to refer to 23.7% of the documents.

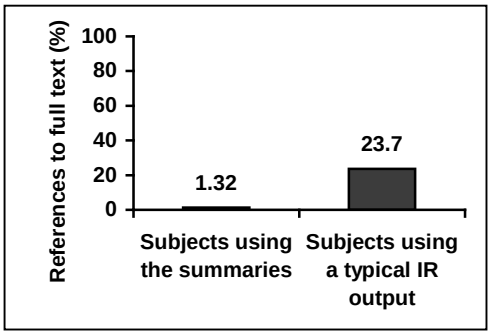



**Figure 7.** Average number of references to the full text of the documents (per query)

This difference can be clearly attributed to the summary information that the subjects were presented with for each retrieved document. The result verifies the initial assumption that the approach adopted by the majority of IR systems for presenting the user with static summaries based on the first few lines of a document is inadequate. Users need more clues to establish the relevance of documents, and especially they need clues about the context in which the query terms are used in these documents. If these clues are not provided from the accompanying information, users refer to the full text of the documents. It is the case in the specific experimental situation, that the query biased summaries provided the subjects with enough evidence to support their relevance judgements. Furthermore, bearing in mind the results pertaining to the accuracy of the relevance judgements, we conclude that the summaries also provided the subjects with the necessary information to adequately decide on the relevance of the documents.

### 4.4 Opinions of the users

As a form of confirmation of the results obtained in the previous categories, the subjective opinions of the users,

gathered from the questionnaire they were asked to fill in after their session, rated the utility of the auto-summaries higher than that of the typical IR output. This result is depicted in Figure 8 where the scale ranges from 1 (most helpful) to 5 (least helpful). The data shown in this figure indicates that subjects using summaries rated on average the utility of the accompanying information at 1.5, while subjects assigned in the other experimental condition indicated a rating of 2.5.

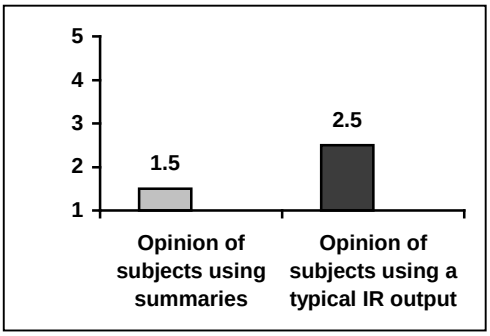



**Figure 8.** Subjective opinion of the users

During the post-experimental discussions, users presented with a typical IR output expressed their dissatisfaction regarding the information they were presented with. More specifically, they emphasised on the fact that they had to refer to the full text for almost every document they were examining. Some subjects even mentioned that it would be helpful for their relevance judgements if they could somehow see how the query terms are used within each retrieved document. Hence, the outcome of the post-experimental discussions is yet another indication in favour of the assumption made, that users require more clues about the relevance of the retrieved documents than they are usually presented by typical IR systems. The automatically generated query biased summaries have focused on capturing that requirement.

## 5 Further Work and Conclusions

From the experimental results, we can draw the following conclusions on the effect of query biased summaries on the process of user relevance judgements:

- They assist users in performing relevance judgements more accurately and more quickly. Users can identify more relevant documents for each query, while at the same time make fewer mistakes. This effect of the summaries is attributed mainly to their indicative nature, and especially to the fact that they adequately indicate the context within which potentially ambiguous query terms are used in the retrieved documents.
- They almost alleviate the users' need to refer to the full text of the documents. Users rely almost solely on the information conveyed in the query-biased summaries in order to perform their relevance judgements. If we examine this result in relation to the increase in the accuracy of the relevance judgements, we can conclude that these summaries successfully provide users with clues about the relevance of the retrieved documents.

One clear application for query biased summaries is in the context of the ubiquitous 'web search engines'. Although these services are generally reliable and take a short time to return a retrieved document list in response to a user's query, access to the full text of documents can be poor as they are stored on other web servers that are potentially slower, less reliable, and more remote. Therefore users of such search engines are likely to want information, like the output of query biased summaries, to help them reduce the number of full documents they try to access. An issue that would have to be addressed in such a case, is the generation of an index file that would allow the re-construction of the query biased summary without having to retrieve the entire web page to the server running the search or to the client applet presenting the search results.

One possible extension of the work reported here is to repeat the experiments, but this time simulate the conditions encountered when searching for information on the web. It is anticipated that the benefits of summaries shown here would be amplified under such conditions. We also intend to examine different summarisation techniques and to apply our system to alternative test collections. In addition we will compare the accuracy of user relevance judgement when using our system to the accuracy when other techniques such as passage retrieval are applied.

## 6 Acknowledgements

The work reported in this paper was carried out by Anastasios Tombros as partial fulfilment of his M.Sc. in Advanced Information Systems at the Department of Computing Science at the University of Glasgow. Supervision of the project was by Mark Sanderson and Phil Gray. Mark Sanderson was funded by the VIPIR project.

## References

Abracos, J., and Lopes, G.P. 1997. Statistical methods for retrieving most significant paragraphs in newspaper articles. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarisation (ISTS '97), 51-57. Madrid, Spain, July 11 1997.

Brandow, R., Mitze, K., and Rau, L.F. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management* 31(5): 675-685, September 1995.

Callan, J.P. 1994. Passage-level evidence in document retrieval. In Croft, W.B., and van Rijbergen, C.J. eds. Proceedings of the Seventeenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 302-310. ACM Press, July 1994.

Edmundson, H.P. 1964. Problems in automatic abstracting. *Communications of the ACM* 7(4):259-263, April 1964.

Edmundson, H.P. 1969. New methods in automatic abstracting. *Journal of the ACM* 16(2):264-285, April 1969.

Hand, T.F. 1997. A proposal for a task-based evaluation of text summarisation systems. In Proceedings of the

ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarisation (ISTS '97), 31-38. Madrid, Spain, July 11 1997.

Harman, D. 1996. Overview of the Fifth Text REtrieval Conference (TREC-5). In Proceedings of the Text Retrieval Conference (TREC-5), National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

Jacobs, P.S., and Rau, L.F. 1990. Scisor: Extracting information from on-line news. *Communications of the ACM* 33(11):88-97, November 1990.

Keppel, G. 1973. *Design and analysis: A researcher's handbook*. New Jersey: Prentice Hall.

Knaus, D., Mittendorf, E., Schauble, P., and Sheridan, P. 1995. Highlighting relevant passages for users of the interactive SPIDER retrieval system. In Proceedings of the Text Retrieval Conference (TREC-4), National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

Kupiec, J., Pedersen, J., and Chen, F. 1995. A trainable document summariser. In Fox, E.A., Ingwersen, P., and Fidel, R. eds. Proceedings of the Eighteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 68-73. ACM Press, July 1995.

Luhn, H.P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2): 159-165, April 1958.

Maizell, R.E., Smith, J.F., and Singer, T.E.R. 1971. *Abstracting scientific and technical literature: An introductory guide and text for scientists, abstractors and management*. New York: Willey-Interscience, John Willey & Sons Inc.

Mani, I., and Bloedorn, E. 1997. Multi-document summarisation by graph search and matching. In Proceedings of AAAI-97, Providence Rhode Island.

McKeown, K., and Radev, D.R. 1995. Generating summaries from multiple news articles. In Fox, E.A., Ingwersen, P., and Fidel, R. eds. Proceedings of the Eighteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 74-82. ACM Press, July 1995.

Miike, S., Itoh, E., Ono, K., and Sumita, K. 1994. A full-text retrieval system with a dynamic abstract generation function. In Croft, W.B., and van Rijbergen, C.J. eds. Proceedings of the Seventeenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 152-161. ACM Press, July 1994.

Miller, S. 1984. *Experimental design and statistics* (2nd edition). New York: Routledge.

Paice, C.D. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing & Management* 26(1):171-186.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program - automated library and information systems* 14(3):130-137.

Rush, J.E., Salvador, R., and Zamora, A. 1971. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science* 22(4):260-274.

Salton, G., Singhal, A., Mitra, M., and Buckley, C. 1997. Automatic text structuring and summarisation. *Information Processing & Management* 33(2):193-20.

## 7 Appendix

In our experimental design, the amount of text shown in the two experimental conditions was not equal (see Figure 3.a and 3.b). While presenting this paper at the *AAAI'98 Spring Symposium on Intelligent Text Summarisation*, it was suggested that this design decision might have distorted the experimental results.

In order to determine the effect on the results presented above, we reran the experiment for the 'typical IR system' condition. In the new design, the number of sentences shown in the two experimental conditions was always the same (i.e. 15% of the original text, up to a maximum of 5 sentences). All the other design parameters were kept as they were.

The results that were obtained through the new experimental design were:

- For Section 4.1 (Recall and Precision): Recall: 48.01% and Precision 44.61%
- For Section 4.2 (Speed): The average number of documents examined was 23.24
- For Section 4.3 (Reference to the full text of the documents): The average number of references (per query) was 15.75%
- Finally, for Section 4.4 (Opinion of the users): The users rated the system with 2.3

The new results show that the accuracy of the judgements and the opinion of the users about the system were not significantly affected by the amount of text shown. The number of times that users had to refer to the full text of the documents was decreased (by approximately 8%), but it still remained significantly higher than the other group's figure (14.42% higher). Finally, users examined more documents per query with the new settings (3.24 more documents on average), but just 0.62 documents more than the group using the summaries.

Based on the new results, we can conclude that the amount of text shown was not a significant factor, and that the difference in performance in the two experimental groups can be attributed to the presence of the query biased summaries in the retrieved document list.