The Use of Implicit Evidence for Relevance Feedback in Web Retrieval

Ryen W. White [1], Ian Ruthven[2], and Joemon M. Jose[1]

[1] Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ. Scotland
whiter, jj@dcs.gla.ac.uk

[2] Department of Computer and Information Sciences
University of Strathclyde, Glasgow G1 1XH. Scotland
Ian.Ruthven@cis.strath.ac.uk

Abstract. In this paper we report on the application of two contrasting types of relevance feedback for web retrieval. We compare two systems; one using explicit relevance feedback (where searchers explicitly have to mark documents relevant) and one using implicit relevance feedback (where the system endeavours to estimate relevance by mining the searcher's interaction). The feedback is used to update the display according to the user's interaction. Our research focuses on the degree to which implicit evidence of document relevance can be substituted for explicit evidence. We examine the two variations in terms of both user opinion and search effectiveness.

1 Introduction

The transformation of a user's information need into a search expression, or query, is known as *query formulation.* It is widely regarded as one of the most challenging activities in information seeking, [2]. In particular, problems arise if the user's need is vague [21], and if they lack knowledge about the collection make-up and retrieval environment [19].

The rapid expansion of the Web has led to an increase in the number of casual searchers using Web search engines. Many of these searchers are inexperienced and can have difficulty expressing their information need. For example, Jansen et al. [6] showed that 62% of queries submitted to the Excite web search engine contained only one or two terms. Such short queries can lack very useful search terms, which may detract from the effectiveness of the search [13].

A technique known as *relevance feedback* is designed to overcome the problem of translating an information need into a query. Relevance feedback is an iterative process where users assess the relevance of a number of documents returned in response to an initial, 'tentative' query, [19]. Users peruse the full-text of each document in this set, assess it for relevance and mark those that best meet their information need. The limitations in providing increasingly better ranked results

F. Crestani, M. Girolami, and C.J. van Rijsbergen (Eds.): ECIR 2002, LNCS 2291, pp. 93–109, 2002.

based solely on the initial query, and the resultant need for query modification have already been identified [22]. Relevance feedback systems automatically resubmit the initial query, expanding it using terms taken from the documents marked relevant by the user.

In practice, relevance feedback can be very effective but it relies on users assessing the relevance of documents and indicating to the system which documents contain relevant information. In real-life Internet searches, users may be unwilling to browse to web pages to gauge their relevance. Such a task imposes an increased burden and increased cognitive load [20]. Documents may be lengthy or complex, users may have time restrictions or the initial query may have retrieved a poor set of documents. An alternative strategy is to present a query-biased summary of each of the first *n* web pages returned in response to a user's query [23]. The summaries allow users to assess documents for relevance, and give feedback, more quickly.

However the problem of getting the users to *indicate* to the system which documents contain relevant information remains. In this paper, we examine the extent to which *implicit* feedback (where the system attempts to estimate what the user may be interested in) can act as a substitute for *explicit* feedback (where searchers explicitly mark documents relevant). Therefore, we attempt to side-step the problem of getting users to explicitly mark documents relevant by making predictions on relevance through analysing the user's interaction with the system.

Previously, many studies that endeavour through the use of various 'surrogate' measures (links clicked, mouseovers, scrollbar activity, etc.) [11], [7] to unobtrusively monitor user behaviour have been conducted. Through such means, other studies have sought to determine document relevance implicitly [4], [12], [9], [14]. These studies infer relevance from the time spent viewing a document. If a user 'examines' [10] a document for a long time, or if a document suffers a lot of 'read wear' [4] it is assumed to be relevant.

These studies only focus on newsgroup documents and rely on users interaction with the actual document. In this paper we extend these concepts onto web result lists, using document summaries instead of the actual document. Much can be gleaned from a user's ephemeral interactions during a single search session [15]. Our system seeks to capture these and predict relevance based on this interaction.

Specifically, we hypothesised that implicit and explicit feedback were interchangeable as sources of relevance information for relevance feedback. Through developing a system that utilised each type we were able to compare the two approaches from the user's perspective and in terms of search effectiveness.

This paper will describes the system and experiments used to test the viability of interchanging implicit and explicit relevance feedback. The experiments were carried out as part of the TREC-10 interactive track. In this paper we expand on our original analysis of our experiments and provide a deeper insight into our experimental results.

This paper describes the two systems used in section 2, the relevance feedback approaches in section 3, then outlines the experimental methodology employed in section 4. We present the initial results and analyse them in section 5, and conclude in section 6.

2 Systems

In this section we introduce the systems used during our experiments. Our basic experimental system is a generic interface that can connect to any web search engine. In our experiments we use the interface to connect to the Google search engine. The interface is based on a summarisation interface developed for investigating web search behaviour, [23], [24]. The system developed for the experiments in this paper also incorporates a component that displays sentences from the retrieved set of web pages. These sentences are ones that have a high degree of match with the user's query. The set of sentences and the ranking of the sentences automatically updates in the presence of relevance information from the user (relevance feedback). We shall discuss these components in more detail in the following sections. Here, we simply note that the effect of relevance information is presented to the user by the changes in the top-ranking sentence list.

Two interfaces were developed; one which uses explicit feedback and one which uses implicit feedback, Fig. 1. We shall discuss the differences in the two interfaces in section 3.

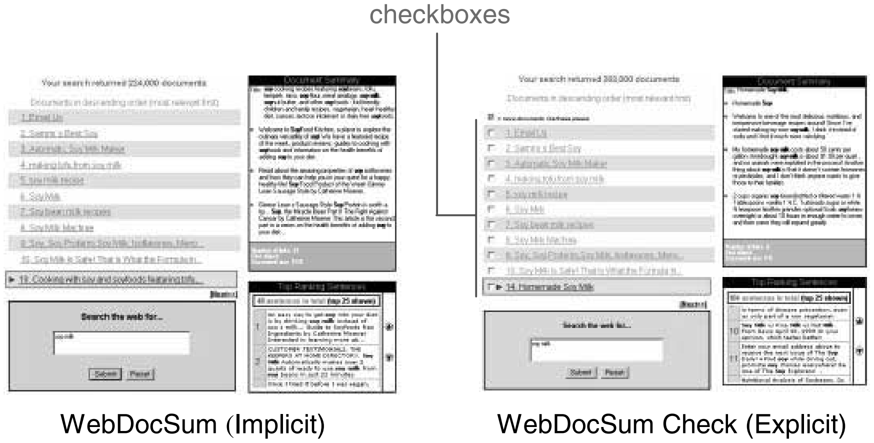

checkboxes
Your search returned 224,000 documents
Document Summary
Top Ranking Sentences
Search the web for...
Submit
Reset
Your search returned 383,000 documents
WebDocSum (Implicit)
WebDocSum Check (Explicit)

Fig. 1. System interfaces

Both versions of our basic interface contain four components; query input window (bottom left, Fig. 1), summary (top right, Fig. 1), results list (top left, Fig. 1), top-ranking sentences (bottom right, Fig. 1). We shall discuss each of these in turn.

2.1 Query Input

The query input component displays the current query for quick and easy reformulation by the user. The system supports any syntax that the underlying search engine, Google, supports. However, no participants in the study used anything other than keywords separated by spaces.

Upon submission, the query is sent to Google, where the first 3 result pages are parsed, extracting the titles, web addresses and abstracts of the first 30 documents. A thread is dispatched to each of these pages, the source for each is downloaded and a summary created. All 30 documents are summarised in parallel. The entire process, from query submission to result presentation takes around 7 seconds.

2.2 Summary Window

In [23], [24] it was shown that the presence of document summaries could lead to more interaction with the results of a web search. That is, searchers would assess more documents if they could access document summaries than if they had access to the full-text alone. In this research document summaries were used to facilitate the relevance assessment process.

The summaries are created through a sentence extraction model, presented in [24], in which the web pages are broken up into their component sentences and scored according to how useful they will be in a summary. A number of the highest-scoring sentences are then chosen to compose the summary.

Sentences are scored according to their position (initial introductory sentences are preferred), the words they contain (words that are emphasised by the web page author, e.g. emboldened terms, or words in the document title are treated as important), and the proportion of query terms they contain. The latter component – scoring by query terms – biases the summaries towards the query.

The summary window displays a summary if the mouse passes over either the document title in the results list *or* a sentence in the top ranking sentences list. The window displays the document title, the top four highest ranking sentences and extra document information such as document size and number of outlinks.

The summary window can also display graphical output and provide feedback should a web error occur. Such an error would occur if a web page was unavailable or taking too long to retrieve. In such circumstances the summary window will show the abstract offered by the underlying search engine and an error message detailing the reason for the web error.

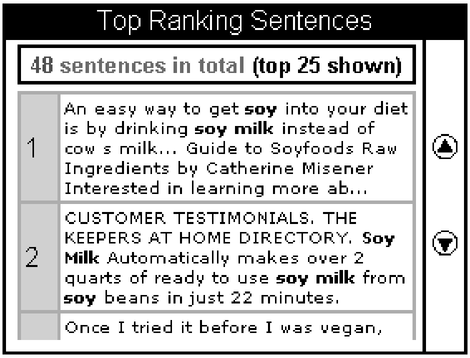
2.3 Top Ranking Sentences

Relevance feedback techniques typically either modify a query automatically, [19], or present the user with a set of new terms to allow the user to interactively modify their query, [1]. Automatic relevance feedback techniques can suffer from the fact that users are often not willing to use relevance feedback techniques. In particular this can be because the user does not understand the relation between the relevance assessment and the effect of relevance feedback. Interactive query modification has often been preferred on the grounds that it allows the user more control over their search, [8]. However, searches often do not fully utilise interactive query modification techniques; either because they do not know how to choose good new query terms, [1] of because they do not understand the effect the new terms will have on their search [18].

An alternative, one which we follow in this paper, is not to use relevance feedback to modify the user's query, or to suggest query terms, but to use relevance feedback to suggest new documents to the user. Specifically we use relevance feedback to recommend documents that have been retrieved but have not yet been viewed by the user. We do this by the notion of top-ranking sentences, Fig. 2.

Our summarisation model is basically a sentence extraction model: all retrieved pages are split into their component sentences, ranked and a number are selected to compose the summary. As well as being used to create summaries, these sentences can also be used to indicate to the searcher the sentences in the retrieved set of documents that have the closest match to the query.

The system pools all of the sentences from the documents that it was able to summarise from the first 30 returned by the underlying search engine. It then ranks these initially based on the score assigned by the query-biasing algorithm, section 2.2, and presents the top 25 sentences to the user. Fig. 2 shows the top ranking sentences window.



Top Ranking Sentences

48 sentences in total **(top 25 shown)**

1 An easy way to get **soy** into your diet is by drinking **soy milk** instead of cow s milk... Guide to Soyfoods Raw Ingredients by Catherine Misener Interested in learning more ab...

2 CUSTOMER TESTIMONIALS. THE KEEPERS AT HOME DIRECTORY. **Soy Milk** Automatically makes over 2 quarts of ready to use **soy milk** from **soy** beans in just 22 minutes.

Once I tried it before I was vegan,

Fig. 2. Top Ranking Sentences window

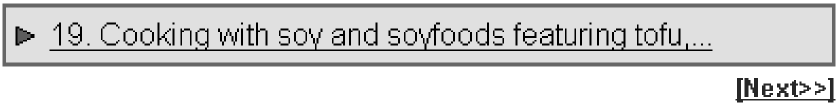
The top ranking list only contains sentences from summaries that the user has not yet seen. As the user passes over a document title, or in the explicit case, marks a document as relevant, the sentences from the summary associated with that document are removed from the list.

Relevance feedback is used to re-rank the list of sentences. As the user interacts with the system – either through explicit or implicit relevance feedback, section 3, the list of top-ranking sentences is re-ranked. The re-ranking is intended to display the most similar sentences to the user's original query combined with the user's interaction which may show the user's changing information needs. We shall discuss how this is achieved in section 3.

2.4 Results List

Only the title of the document is shown in the result list. When the user moves the mouse over an entry in the result list, the summary window will change to show a summary for that page. If a title is *clicked*, the page will open in a new window. The query form retains the current query for quick and easy reformulation. If the user goes over a sentence in the top ranking sentences list, the document's title to which that sentence belongs is highlighted.

A small window below the main results list displays the title of a document should it fall outside the top 10 documents (i.e. in the range 11 to 30). This is shown in Fig. 3 and comes from Fig. 1, (left hand side).



▶ 19. Cooking with soy and soyfoods featuring tofu,...

Fig. 3. Document title window

3 Explicit and Implicit Feedback

In the previous section we outlined the details of our interface. We developed two interfaces; one which uses explicit relevance feedback, in the form of check-boxes before the document titles, Fig. 1 (right), and one which uses implicit feedback and has no means for the user to explicitly indicate the relevance of a document, Fig. 1 (left). The main component is the top-ranking sentence list; a list of the sentences that best-match the system's representation of the user's information need. This list updates in the presence of relevance information from the user.

The main difference, therefore, between the two systems lies in how the call is made to re-rank the sentences. In the explicit case, this is made by clicking on a checkbox and hence marking a document relevant. In contrast, the implicit system re-ranks the list when the user moves the mouse over a document title. Through these means the user no longer has to be concerned with marking a document as relevant, the system has mined their interaction and made an educated assumption based on this. That is we assume that viewing a document summary is an indication of a user's interest in the document's contents. Both systems use the same method of re-ranking sentences. That is, the sentences from that document's summary are removed from the top ranking list and the remaining sentences will be re-ranked immediately, in real-time, based on the occurrence of the query expansion terms and query terms within them.

Each time the list of sentences are to be updated, the summaries from the assessed relevant documents (explicit system) or assumed relevant documents (implicit system) are used to generate a list of possible expansion terms. From this listing, the top 6 expansion terms are added to the user's original query. These terms are taken from *all* assumed relevant summaries (i.e. all those that the user had viewed so far).

To rank the possible expansion terms we used the *wpq* algorithm [17] shown in Equation 1. For any term *t*, where *N* is the total number of summaries, *n* is the number of summaries containing *t*, *R* is total number of relevant summaries and *r* is the total number of relevant summaries that contain *t*. *R* and *r* are based on relevance assessments and are therefore prone to increase as users interact with the systems.

Possible expansion terms only come from the document summary generated by the system, and it is assumed that N (i.e. the total number of summaries) is 30.

$$wpq_t = \log\left(\frac{(r+0.5)(N-n-R+0.5)}{(n-r+0.5)(R-r+0.5)}\right)\times\left(\frac{r}{R}-\frac{n-r}{N-R}\right) \qquad \textbf{(1)}$$

The new query representation – the original terms plus the new expansion terms – are then used to re-rank the top-ranking sentence list, assigning a score to each remaining sentence based on term occurrence. The score is calculated based on the term occurrence of the top six query expansion terms. If a sentence contains a query expansion term, its score is incremented by the wpq score for that term. If a sentence contains multiple occurrences of the same term, its score is incremented for each occurrence.

Traditional relevance feedback techniques would reset the sentence scores each feedback iteration, i.e. the sentence score would be recalculated from the wpq values each time. In our application we do not do this: the new score for a sentence is added to the score from the previous iteration. The effect of this is that the *order* in which a user views summaries is important; the same summaries viewed in a different order can give a different ranking of sentences.

Fig. 4 shows how the scores of 5 sentences (s1…s5) change through three iterations of re-ranking. The first sentence shown in the top ranking sentences list after each iteration is the most relevant sentence taken from summaries the user has yet to see, based on the summaries they have viewed so far. Sentences that might have started low in the initial ranking can ‘bubble up’ (as is the case with s1 in Fig. 4) to the top of the list and present the user with a means of accessing a potentially relevant document based on all previous interaction. A sentence will be removed from the list if it is judged relevant (either implicitly or explicitly). Only sentences from document summaries *not yet viewed* are shown in the list.

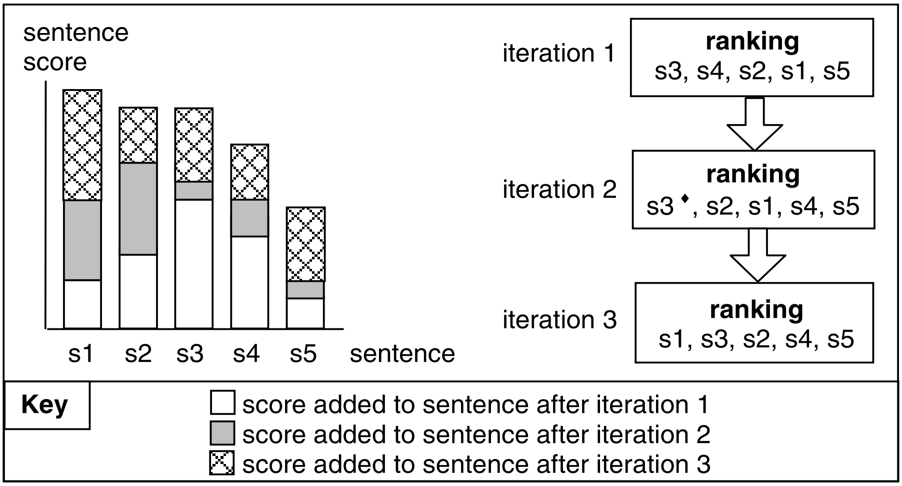


Fig. 4. The cumulative affects of sentence scoring, over three iterations

3.1 Summary

The research question under investigation in this paper is whether implicit indications of relevance can be substituted for explicit assessments of relevance. To test this we developed two interfaces; one that allows users to explicitly mark documents relevant, the other attempts to estimate what the user finds relevant through the user's interaction. The relevance information from both systems is used to alter the user's query using relevance feedback techniques. The modified query is then used to dynamically re-rank a list of sentences that match the modified query. This list of sentences are intended to act as a means of recommending unseen documents to the user. The relevance information is not used to initiate new retrievals, as in standard relevance feedback, but is used to update the information presented to the user. The degree to which the implicitly updated information is as useful as the explicitly updated information is how we measure the degree to which implicit relevance evidence can be substituted for explicit relevance evidence.

4 Experimental Details

The experiments we report in this paper were carried out as part of the TREC-10 interactive track [5].

We used a within-subjects experimental design and in total, 16 subjects participated. Each user attempted 4 web search tasks and used 2 systems. A Greco-Latin square experimental design [16] was used to randomise the tasks and guard against learning effects. Prior to starting the experiment, subjects were given a 15 minute introduction to the system, during which the experimenter walked them through a demonstration task.

4.1 Subjects

In total, 16 subjects participated in our experiments. All subjects were educated to graduate level in a non-computing, non-LIS discipline, with three exceptions, all our subjects were recruited from the Information Technology course at the University of Glasgow. All users, with one exception, used the Internet on a regular basis. Through using a diverse mixture of participants, with a diverse mixture of skills and experiences, the heterogeneous nature of the web populace was better represented.

The average age of the subjects was 24.75 with a range of 11 years. Most users used computers and the Internet frequently – the average time spent online per week was 14 hours. With three exceptions, all users cited Google amongst their favourite search engines.

4.2 Tasks

The tasks were split into four categories: Medical, Buying, Travel and Project. There were 4 tasks in each category – users attempted one task from each. The tasks allocated were randomised to reduce potential learning effects and task bias. Fig. 5 shows the tasks used.

Medical

- Find a website likely to contain reliable information on the effect of second-hand smoke.
- Tell me three categories of people who should or should not get a flu shot and why.
- List two of the generally recommended treatments for stomach ulcers.
- Identify two pros or cons of taking large doses of Vitamin A.

Buying

- Get two price quotes for a new digital camera (3 or more megapixels and 2x or more zoom).
- Find two websites that allow people to buy soy milk online.
- Name three features to consider in buying a new yacht.
- Find two websites that will let me buy a personal CD player online.

Travel

- I want to visit Antarctica. Find a website with information on organized tours/trips there.
- Identify three interesting things to do during a weekend in Kyoto, Japan.
- Identify three interesting places to visit in Thailand.
- I'd like to go on a sailing vacation in Australia, but I don't know how to sail. Tell me where can I get some information about organized sailing cruises in that area.

Project

- Find three articles that a high school student could use in writing a report on the Titanic.
- Tell me the name of a website where I can find material on global warming.
- Find three different information sources that may be useful to a high school student in writing a biography of John F. Kennedy.
- Locate a site with lots of information for a high school report on the history of the Napoleonic wars.

Fig. 5. Tasks used in TREC-10 interactive track experiments

Users were allowed a maximum of 10 minutes for each task. They were asked to use the system presented to them (either implicit or explicit, depending on the particular Greco-Latin square allocation) to search the Internet and attempt to find an answer to the task set. Users were allowed to browse away from the result list to any degree.

4.3 Data Capture

We utilised two different means of collecting data for post-experimental analysis; questionnaires and background system logging. Through these means we could collect data that would allow us to thoroughly test the experimental hypothesis.

We administered two main types of questionnaire during the course of the experiment, five in total. The first gathered both demographic and Internet usage information, and was completed at the start of the experiment. The second made use of Likert Scales and Semantic Differentials to assess the systems and to some extent the tasks, from the perspective of the user. This was completed after each of the four tasks.

The background system logging sought to capture certain aspects of the user's interaction with the system. Information such as the task time, calls for an update of the top ranking sentences list and the query expansion terms used were all logged. Through a detailed analysis of these logs, it was hoped that a complete comparison of the implicit and explicit methods could be carried out.

5 Results & Analysis

In this section, we discuss the results obtained from the experiments. We look at the effectiveness of the searches (number of result pages viewed, task completion and task times) and the users' perceptions of the systems. We also present a subjective evaluation of the usefulness of both the automatic query expansion terms generated by the system and the feedback components.

5.1 Search Effectiveness

Most of the data used to assess search effectiveness came from the logs generated by the system during the experiments.

5.1.1 Number of Result Pages Viewed

The total number of result pages viewed and queries submitted during all the experiments was recorded. Table 1 shows the average results per user obtained.

Table 1. Average result page views and query iterations per user

Variation	Number of result pages	Number of query iterations
Implicit	3.375 *	3.5625
Explicit	2.5 *	2.625

* users occasionally refined query before result page appeared, so result pages ≠ query iterations

These differences are not significant using a Mann-Whitney Test at $p \leq 0.05$ ($p = 0.234$). Our system gave access to the first 30 documents retrieved by the underlying search engine, and in many cases this was sufficient to complete the tasks. This meant that there was no real need for users to browse to the next 30 results (i.e. results 30 to 60 in standard search engines). The lack of a significant difference between the implicit and explicit systems shows that the type of system used does not affect the number of result pages viewed or query iterations needed.

5.1.2 Task Completion

As part of the post-task questionnaire users were asked whether they felt they had successfully completed the task just attempted, it is these results that are presented in Table 2. The choice of whether a task was complete was left up to the user. It was thought that this best reflected real-world retrieval situations. However, the experimenter was occasionally asked to verify the correctness of the results obtained. Table 2 shows these results (out of 64).

Table 2. Number of tasks completed

Variation	Number of tasks completed
Implicit	61
Explicit	57

Again these results are not significant using a Mann-Whitney Test at $p \leq 0.05$ ($p = 0.361$). There is no significant difference between the number of tasks that users completed on the implicit and the explicit systems.

5.1.3 Task Times

The time taken to complete tasks on both systems was measured. When a task was incomplete, a figure of 600 seconds (10 minutes) would be recorded by the system. This was the time limit imposed on each task and users were not allowed to work past this. On no occasion did a user stop before the 10 minute limit had been reached unless they had completed their current task. In Table 3 we can see these results.

Table 3. Average time per task

Variation	Average time per task (secs)
Implicit	372.29
Explicit	437.43

Again these are not significant using a Mann-Whitney Test at $p \leq 0.05$ ($p = 0.228$).

From an analysis of the log files we were able to establish that no significant difference existed between the two variations. This appears to add a little weight to our claim that perhaps the implicit and explicit feedback are at least to some degree substitutable, although factors such as the similarity of the interface design may be important to. If the results obtained were significant we could suggest that one type of system promotes search effectiveness more than the other. In this case, there is no significant difference, and it is safe to assume that some degree of substitutability does indeed exist.

5.2 User Perceptions

As well as assessing the systems in terms of search effectiveness, we also attempted to assess both the systems used and the tasks allocated, from the perspective of the user. By using Likert Scales and Semantic Differentials, subjects could express their feelings with relative ease.

5.2.1 Tasks

Subjects had the opportunity to rate the tasks allocated to them in terms of clarity and difficulty. Each task, or category, was attempted four times in total (twice on each system). Table 4 shows the results obtained.

Table 4. Semantic differential values obtained from user task assessment (lower = better, range = 1 – 5)

Differential	Implicit	Explicit	Significance [Mann-Whitney]
clear	1.42	1.83	0.5160
easy	2.48	2.42	0.6134

As is apparent from the last column in the table, neither of the two differentials are significant (at $p \leq 0.05$). This test is across all tasks, so any significant value would have pointed to a serious error in our experimental methodology and/or task design. A similar result (i.e. no significance) was obtained when we analysed the fully and partially defined tasks separately.

5.2.2 Systems

Subjects had a chance to rate the systems using both Likert Scales and Semantic Differentials in a similar way as in section 4.2.1. They answered questions relating to the way the top ranking sentences list was updated and its respective usefulness. Table 5 provides an overview of the results obtained.

Table 5. Semantic differential/Likert scale values obtained from system assessment (lower = better, range = 1 – 5)

Differential/Scale	Implicit	Explicit	Significance [Mann-Whitney]
useful (summaries)	2.46	2.16	0.3173
useful (top ranking list)	2.63	2.29	0.4394
see ♦	1.71	1.75	0.244
see o	1.54	1.96	0.00606

♦ potentially relevant sentences from unseen documents helped user decide which to view

o in the user's opinion the top ranking sentences list was updated often enough

There is no significant difference between the first three rows in Table 5 (above) when measuring at $p \leq 0.05$. The way the summaries are produced, the content of the top ranking sentences list and the ability presented to the user to view sentences from unseen documents does not differ from system to system. The averages are less than 2.64 in all cases, with 3 being the middle ('undecided') value. Based on this, we can assume a generally positive response from users to the system in general. Although the significance of this was not tested, it is interesting to note nonetheless and may provide us with some scope for future research.

When users assessed the frequency with which the top ranking sentences list updates, there is a significant difference between the implicit and explicit systems ($p = 0.00606$ using Mann-Whitney at $p \leq 0.05$). It appeared that subjects preferred the top ranking sentences list to update automatically, rather than under their instruction. Users were unwilling to 'instruct' the list to update. This is in accordance with previous research [3, Kons97] that also shows users' apparent unwillingness to provide relevance information explicitly.

5.3 Query Expansion

The query expansion terms were generated automatically by the system each time the list was re-ranked, and were used by the system in this re-ranking only; *the terms were to used to generate a new result set.* The nature and 'usefulness' of query expansion terms is subjective, but a brief evaluation based on how correct they 'appear' to be is possible. The systems we developed utilised automatic query expansion techniques to re-rank the list of top-ranking sentences. Such systems (or variations in our case), by their very nature, hide query expansion terms from the user. We evaluated the terms based on what we thought were 'reasonable', no input from users was utilised.

Overall, the choice of automatic query expansion appeared good, the terms appeared to be a natural expansion of the initial query submitted by the user. Fig. 6 shows an example of how the query expansion terms for the query "effects smoke" changed through three re-ranking iterations. This was taken from the experiments in response to the task: *Find a website likely to contain reliable information on the effect of second-hand smoke*.

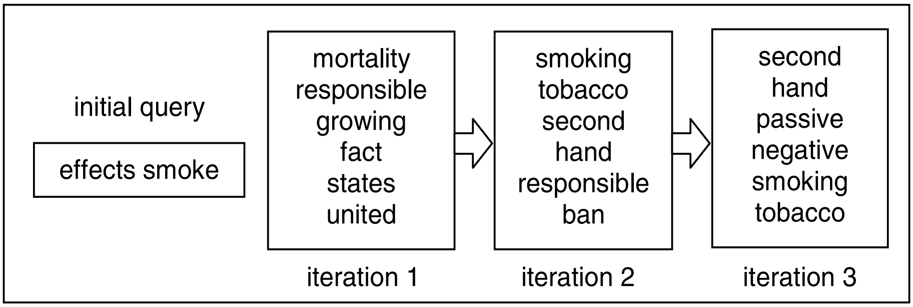


Fig. 6. Changes in automatic query expansion terms (taken from experiments)

As the user views summaries, the terms are modified based on the summaries they view or 'mark' relevant. Fig. 6 shows how this can be an effective means of reformulating a query for web retrieval.

5.4 Use of Feedback Components

In a similar way to the previous subsection, the results reported here are mainly based on observation. One of our main concerns prior to starting the experiments was that users would not make use of the feedback components, namely the top ranking sentences list, the highlighting of the document titles and the checkboxes. Previous research [18] has shown that users often have a reluctance to use such features. Our concerns proved unfounded, as users made use of all components, especially the top ranking sentences (39% of all 'Summary Window' updates came from this source).

The inclusion of checkboxes appealed to some users and not to others; some saw them as a means by which they could retain control over the system's operation whereas others saw them as a hindrance. Overall, just as many users preferred the implicit system as did the explicit one.

5.5 Technical Drawbacks

It is worth mentioning a technical problem observed during our experiments. The implicit system suffered from the effects of accidental 'mouseovers', with users passing over document titles en route to those that interested them. This meant from time to time that sentences were removed from the top ranking list by accident (i.e. the system was interpreting the mouseover as an indication of relevance). A possible solution to this problem would be the introduction of a timing mechanism so that only mouseovers actually meant by the user are taken into account in re-ranking the list.

6 Conclusions

The aim of the research reported in this paper was to test whether implicit relevance feedback was a viable substitute for its explicit counterpart in web retrieval situations. To test this, we developed two interfaces – one that used implicit feedback and another that used explicit feedback. The two systems were then compared in terms of user perception and search effectiveness in a real-life information search environment. We hypothesised that there would be no significant differences in the results of this comparison.

We endeavoured to use techniques previously restricted to the full-text of newsgroup documents and apply them to the result list of a web search interface, using query-biased web page summaries as the means of assessing relevance, instead of the actual document. A list of potentially relevant sentences that evolved, in real-time, to reflect the user's interactions provided users with a 'gateway' to potentially relevant summaries (and hence documents) that they have not yet viewed.

The experiments undertaken to test our hypothesis were as part of the TREC-10 interactive track [5]. The tasks allocated to users appeared to be sufficiently random and exhibited no correlation between search system used and the user's assessment of task difficulty.

Users found the summaries and top ranking sentences useful. There were no significant differences in the comparison of systems, with one exception. The implicit system updated the top ranking sentences list automatically. This was preferred in favour of the explicit system which gives them full control over when the list updates. Users performed equally well on the implicit and explicit systems, which leads us to conclude that perhaps substituting the former for the latter may indeed be feasible. This initial result will be exploited in future research to investigate how implicit evidence is best collected, how it should be used and what are good implicit indicators of relevance.

We assumed that the viewing of a document's summary was an indication of an interest in the relevance of the document's contents. There are several grounds on which this can be criticised; users will view non-relevant summaries, the title rather than the summary was what the user expressed an interest in, and the user may look at all retrieved documents before making real relevance decisions. Nevertheless we felt that our assumption was fair enough to allow an initial investigation into the use of implicit feedback. A future alternative could be to introduce a timing mechanism to eliminate the problems caused by the accidental 'mouseover' of document titles and the unwanted removal of sentences from the top ranking sentences list that follows.

The top ranking sentences proved a useful aid to users, but they were only the means we chose to test the research hypothesis and other alternatives are equally viable. This subject area obviously needs a lot more research, but we hope that we have shown that implicit and explicit relevance feedback can be interchanged in web retrieval.

Acknowledgements

We would like to thank fellow members of the Information Retrieval group for their thoughts and assistance. We would also like to thank those who participated in the experiments, their comments and enthusiasm were greatly appreciated.

References

1. M. Beaulieu. *Experiments on interfaces to support query expansion*. Journal of Documentation 53. 1. (1997) 8-19
2. Cool, C., Park, S., Belkin, N.J., Koenemann, J. and Ng, K.B. *Information seeking behaviour in new searching environment.* CoLIS 2. Copenhagen. (1996) 403-416
3. Grundin, J. *GroupWare and Social Dynamics: Eight Challenges for Developers.* Communications of the ACM 35. (1994) 92-104
4. Hill, W.C., Hollan, J.D., Wrobelwski, D. and McCandless, T. *Read wear and edit wear.* Proceedings of ACM Conference on Human Factors in Computing Systems, (CHI '92) Monterey, California, USA, 3-7 May (1992)
5. Hersh, W. and Over, P. *TREC-10 Interactive Track Report.* NIST Special Publication: 10th Text REtrieval Conference, Gaithersburg, Maryland, USA. 13-16 November (2001)
6. Jansen, B.J., Spink, A. and Saracevic, T. *Real life, real users, and real needs: A study and analysis of users on the web*. Information Processing & Management 36. 2. (2000) 207-227
7. Joachims, T., Freitag, D. and Mitchell, T. *WebWatcher: A Tour Guide for the World Wide Web* Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '97) Nagoya, Aichi, Japan, 23-29 August (1997)
8. J. Koenemann and N. Belkin. *A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness*. Proceedings of the ACM Conference on Computer-Human Interaction (CHI '96). Vancouver. (1996)
9. Konstan, J.A., Miller, B.N, Maltz, D., Herlocker, J.L, Gordon, L.R. and Riedl, J. *GroupLens: Applying Collaborative Filtering to Usenet News*. Communications of the ACM 40. 3. March (1997) 77-87
10. Kim, J., Oard, D.W. and Romanik, K. *Using implicit feedback for user modeling in Internet and Intranet searching*. Technical Report, College of Library and Information Services, University of Maryland at College Park. (2000)
11. Lieberman, H. *Letizia: An Agent That Assists Web Browsing* Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95) Montreal, Canada, 20-25 August (1995)
12. Morita, M. and Shinoda, Y. *Information filtering based on user behavior analysis and best match text retrieval*. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94) Dublin, Ireland. 3-6 July (1994)
13. Mitra, M., Singhal, A. and Buckley, C. *Improving Automatic Query Expansion* Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98) Melbourne, Australia, 24-28 August (1998)
14. Nichols, D.M. *Implicit ratings and filtering*. Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering (DELOS '97) Budapest, Hungary, 10-12 November (1997)

15. Oard, D. and Kim, J. *Implicit Feedback for Recommender Systems*. Proceedings of the AAAI Workshop on Recommender Systems (AAAI '98) Madison, Wisconsin, 26-30 July (1998)
16. Maxwell, S. E., & Delaney, H. D. *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum Associates. (2000)
17. Robertson, S.E., *On Term Selection for Query Expansion*. Journal of Documentation 46. 4. (1990) 359-364
18. Ruthven, I., Tombros A. and Jose, J. *A study on the use of summaries and summary-based query expansion for a question-answering task*. 23rd BCS European Annual Colloquium on Information Retrieval Research (ECIR 2001). Darmstadt. (2001)
19. Salton, G., and Buckley, C. *Improving retrieval performance by relevance feedback.* Journal of the American Society for Information Science. 41. 4. (1990) 288-297
20. Shavlik, J. and Goecks, J. *Learning users' interests by unobtrusively observing their normal behavior* Proceedings of the 2000 International Conference on Intelligent User Interfaces (IUI '00) New Orleans, USA, 9-12 January (2000)
21. Spink, A., Greisdorf, H., and Bateman, J. *From highly relevant to not relevant: examining different regions of relevance*. Information Processing and Management. 34. 5. (1998) 599-621
22. van Rijsbergen, C.J. *A New Theoretical Framework For Information Retrieval* Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '86) Pisa, Italy, 8-10 September (1986)
23. White, R., Jose, J.M. and Ruthven, I. *Query-Biased Web Page Summarisation: A Task-Oriented Evaluation*. Poster Paper. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01) New Orleans, USA, 9-13 September (2001)
24. White, R., Ruthven, I. and Jose, J.M. *Web document summarisation: a task-oriented evaluation.* International Workshop on Digital Libraries. Proceedings of the 12th International Database and Expert Systems Applications Conference (DEXA 2001) Munich, Germany, 3-7 September (2001)