Effectiveness of keyword-based display and selection of retrieval results for interactive searches

Ezio Berenci, Claudio Carpineto,* Vittorio Giannini,* and Stefano Mizzaro***

* Fondazione Ugo Bordoni, Via B. Castiglione 59, I-00142, Rome, Italy
{eberenci, carpinet, vgiannini}@fub.it

** Department of Mathematics and Computer Science, University of Udine,
Via delle Scienze, 206 - Loc. Rizzi, I 33100 Udine, Italy
mizzaro@dimi.uniud.it

Abstract

We present an approach to increasing the effectiveness of ranked-output retrieval systems that relies on graphical display and user manipulation of "views" of retrieval results, where a view is the subset of retrieved documents that contain a specified subset of query terms. This approach has been implemented in a system named VIEWER (VIEwing WEb Results), acting as an interface to available search engines. An experimental evaluation of the performance of VIEWER in contrast to AltaVista is the major focus of the paper. We first report the results of an experiment on single, short query searches where VIEWER, used as an interactive ranking system, markedly outperformed AltaVista. We then concentrate on a more realistic searching scenario, involving free query formulation, unconstrained selection of retrieval results, and possibility of query reformulation. We report the results of an experiment where the use of VIEWER, compared to AltaVista, seemed to shift the user effort from inspection to evaluation of results, increasing retrieval effectiveness and user satisfaction. In particular, we found that the VIEWER users retrieved half as many nonrelevant documents as the AltaVista users while retrieving a comparable number of relevant documents.

1. Introduction

Information retrieval, as experienced by most Web users, is an iterative and interactive process which consists of submitting a query, seeing the ranked document summaries returned in response to the query (which may possibly lead to download the associated full documents), and submitting a new query, until the sought information have been found or the search has been abandoned. Unfortunately, the unmanageably large response sets of Web search engines coupled with their low precision and ranked list presentation may make summary perusal hard, time-consuming, and costly for the user. Research in information retrieval is thus increasingly focusing on the lack of effectiveness of current retrieval engines' interfaces.

The need for concise display and user-oriented manipulation of retrieval results has been recognized before the advent of Web-based ranked-output search services. Among various other systems, Bead (Chalmer and Chitsons, 1992) and LyberWorld (Hemmje *et al*, 1994) depicted clustering patterns in a document space using three-dimensional visualization schemes, InfoCrystal (Spoerri, 1994) used a particular visual representation of a Venn diagram to suggest how to refine Boolean queries, TileBars (Hearst, 1995) displayed distribution of query terms within each document to locate its relevant parts, and Ulysses showed a lattice of terms and documents that can be searched in various and integrated ways (Carpineto and Romano, 1996). Most of these systems, however, cannot be applied to Web-based retrieval because they are either computationally expensive, or require sophisticated graphical facilities, or do not scale well, or rely on underlying retrieval models other than best-match ranking, or, more often, present a combination of these features.

The goal of our research is to facilitate inspection and utilization of Web retrieval results. Some of the more stringent requirements of Web searches such as presentation of results by summaries, interaction with nonexpert users, speed, and incrementality have been specifically addressed only very recently. Two examples are the works by Zamir and Etzioni (1998), in which they described a cluster-based method for reordering and labelling the first documents returned by Web engines, and by Tombros and Sanderson (1998), where they suggested using a query-biased document description to better support the users' need to refer to full documents. Instead of presenting the user with a different list of summaries or with a list of summaries with a different document description, we aim at giving the user more control over the set of summaries that can be selected for perusal.

We present a graphical interface to Web search engines that displays characteristics of documents which are significant in supporting the decision to peruse or not. This acts as an intermediate layer between the query specification stage and the actual display of the document summaries; the latter takes place only on user demand after interaction with the intermediate layer. Our approach is based on the notion of view, where a view is simply defined as the subset of retrieved documents that contain a specified subset of query terms. Similar to other recent exact matching retrieval systems that will be discussed below, the main rationale of our approach is that the selection of documents of interest can be facilitated by decomposing a query into its constituents and checking for their inclusion in a document individually.

A major part of this paper is then a study on the retrieval effectiveness of this kind of component for on-line interactive searches. We evaluate the performance of using the view mechanism to select document summaries in contrast to their conventional ranked presentation directly following a user query, as in Web search engines. We experiment over a large test collection with external subjects, considering both single-query searches

and multiple-queries searches. This kind of experiments has been regrettably rare in the literature on user-oriented visualization and manipulation of retrieval results, probably due to a combination of technological, organizational, and economical difficulties; we feel that this gap should be filled in order to assess the utility of these tools in a more realistic manner.

The rest of the paper is organized as follows. We first discuss the renewed interest for exact matching retrieval and describe the system prototype in which our approach has been implemented, named VIEWER (VIEwing WEb Results). Then we present the experimental study, which consists of two distinct experiments. In the first experiment we compare the ranking performance of AltaVista to that of VIEWER, used as an interactive ranking system, on single, short query searches. In the second experiment we compare the performance of VIEWER to that of AltaVista on a more realistic subject searching task, involving free query formulation, free inspection and selection of results by users, and possibility of query reformulation. For this task, we use evaluation measures related to document relevance, as well as to time and subjective opinions of the users. A discussion of some general lessons learned from these experiments along with directions for future work conclude the paper.

2. Exact matching retrieval

Consistently with earlier results on the effectiveness of information retrieval systems, Web search services have heavily favored best matching over other types of document retrieval. However, some specific requirements of Web-based retrieval challenge this view. First of all, while we are often primarily interested in precision rather than recall, there is evidence that best matching retrieval achieves lower precision ratios than exact matching retrieval for large databases, and that this difference increases as databases grows (Blair and Maron, 1990). Secondly, while best matching retrieval is designed to take advantage of the presence of many query terms to describe a user's information need, the average number of user-supplied query terms in Web searches is usually very small, often less than 2. Given this scenario, exact matching retrieval is seen with renewed interest, both as an alternative or as a complementary technique to traditional best matching retrieval.

One of the most interesting recent findings that supports the use of exact matching retrieval was presented by Clarke *et al.* (1997), who showed that a variant of the well known coordination level-based retrieval method may achieve not only better precision but also better recall-precision performance than best match ranking when the user queries are short.[1] This result is particularly impressive considering that since

[1] Using coordination level-based retrieval the documents are ranked according to the number of distinct query terms that they contain, which is referred to as their coordination level (see for instance Van Rijsbergen (1979)). If the user query contains *n* terms, the documents that contains *n* query terms are

coordination level-based retrieval uses a purely syntactical ranking criterion, it fails to recognize all the situations in which documents containing fewer query terms are more relevant than documents containing more query terms. This usually happens when a short (exact) partial match between query and documents is found that closely corresponds to the query concept while there are other longer (exact) partial matches that are less "about" the query concept. As a result, coordination level-based retrieval may easily favour spurious or irrelevant matches over relevant ones, thus lowering precision with many false drops. In order to improve the effectiveness of (syntactic) keyword retrieval we must therefore deal with the (semantic) problem of discriminating between proper and improper partial matches between query and documents.[2]

To deal with this problem we may use contextual information on the query terms. Bodoff and Kambil (1998), for instance, suggested using cataloger-provided dependencies between the subject terms of the documents being searched. Another way of obtaining contextual information is to have the user express Boolean constraints over the set of query terms to indicate which terms cover which aspects of the query, e.g., the constraint (A OR B) AND (C OR D) specifies two aspects of the query, each represented by two keywords (Hearst, 1996). These approaches may be useful to solve specific aspects of the exact partial match problem but the formulation of the additional information that they require lends itself to improper partial matches. Furthermore, the specification of query filters takes place before searching the database, while the relevance of a partial match depends also on the content of the database being searched.

A semi-automatic solution to the partial match problem is to present the user with information that highlight the distribution of the possible various meanings (arising from partial matches between the query and the documents) in the documents themselves, and then let the user select the documents containing the meanings of interest. This approach has been taken by Veerasamy and Heikes (1997), with the main goal of clarifying the role played by query terms in the result of ranked output systems. They visualize the weights of the query terms contained in all documents retrieved in response to a query and then let the user choose the documents that contain the most relevant combinations of weighted terms. Our approach shares a similar concern but employs a radically different visualization and interaction scheme. Instead of visualizing the weights of the query terms of each retrieved document we concentrate on all the possible subsets of query terms (i.e., subqueries) that can be generated from the user query, showing their distribution in the set of retrieved documents and allowing the user to select the set of documents associated to each of them. We speak of view, because in

---

ranked before those containing *n-1* terms, which, in turn, are ranked before those containing *n-2* terms, and so on. This method is also termed quorum-level searches (Salton, 1989).

this way the user may see parts of results without seeing the whole list. Views are defined in a precise way from the retrieved documents through a simple and comprehensible characteristic of their content, i.e., the subset of distinct query terms that they contain.

3. Description of VIEWER

VIEWER is built around available "primary" Web search services, presenting users with a single unified interface. Users enter a query, which VIEWER forwards to a selected search engine (AltaVista, in the current implementation). VIEWER then collects the query results and shows, in a scrollable window, a subset of the document summaries, in the same ranked order as returned by the search engine; in addition, it shows, in the rest of the screen, a graphical visualization of results. The visualization consists of an aligned sequence of horizontal bars, one for each of the nonempty subqueries that can be formed with the n query terms (at most 2^n-1). Subqueries are displayed in the order of increasing number of terms, with the longest subqueries at the bottom; the length of each bar is proportional to the number of document summaries containing that subquery, which is also explicitly displayed next to the bar. By clicking on a bar the user may select the corresponding view, bringing up the associated summaries into the document window. The number of retrieved document summaries considered by VIEWER is currently set at 40; the maximum number of query terms used for visualization is set at 4, because the display of the subquery distribution in the retrieved documents would become cumbersome for longer queries. In practice, however, the latter is not a serious limitations due to the paucity of query terms in real searches.

As an example session with VIEWER, consider searching the following subject over the Web: "scientific accuracy of Bible predictions". Figure 1 shows the response of VIEWER to the user query: *scientific accuracy Bible predictions*, as of 25 March, 1999. The document summaries returned by AltaVista are shown on the right window. The graphical display quickly shows that the results produced by AltaVista were, in general, dissatisfying, because most retrieved documents did not deal with the Bible. What happened was that some subqueries (e.g., "predictions accuracy", "scientific accuracy", or just "predictions") matched out of the context of the primary topic subquery (i.e., Bible), and retrieved documents about such diverse domains as currencies, weather, physics, and astrology. What is more important, the non-relevant retrieved documents were ranked by AltaVista well ahead of the relevant ones. In fact, the documents about "scientific accuracy of the Bible" were ranked by AltaVista from

[2] Bodoff and Kambill (1998) identified several types of out-of-context matches (i.e., when some query terms match out of context of their relationships to the other terms), including polysemy, out of phrase

30 to 40, and thus a user would have probably completely missed them in a normal search. Using VIEWER, instead, the user may immediately select those few summaries that appear to be relevant, without perusing the others. In addition, the results displayed by VIEWER suggest that the user might profitably reformulate the query by emphasizing the primary subject of the search (e.g., adding such terms as biblical, christian, religious) and by not using the subqueries that matched out of context (e.g., replacing predictions by prophecies).

The scope of VIEWER encompasses a number of situations where the retrieval results can be usefully related to a query's constituents. The questions that can be quickly answered with VIEWER include: how many documents contain a certain subquery *s*?, which documents contain *s*?, which terms can be added to *s* in such a way that the resulting set of documents is more manageable (when *s* is contained in too many documents)?, which terms can be deleted from *s* in such a way that the resulting set of documents is still manageable (when *s* is contained in few documents)? We might also be interested in the relationships between different subqueries. So we might ask: what is the contribution of subquery *q* compared to subquery *s*?, does subquery *q* occur more frequently alone or in conjunction with *s*?, and so on. In addition to enriching inspection of retrieval results with facilities for selection, comparison and refinement involving groups of query terms, VIEWER has also potentials for facilitating query reformulation, as seen in the example. In particular, VIEWER may suggest that in order to focus on the intended meaning of the query, the user should reformulate the query by adding (deleting) terms to (from) it or by replacing some terms with narrower/broader terms. VIEWER may also help detect failure of intended senses of words, i.e., when two terms used in the query to identify one particular meaning do not occur together in the retrieval results, or, symmetrically, discover unwanted senses of words (Cooper and Byrd, 1997).

VIEWER has been implemented as a client-server system; its user interface is a Java applet which can be downloaded on a Web browser from*: http://www.fub.it/viewer/*. Thus, VIEWER copes with most computational constraints of Web-based retrieval (e.g., efficiency, portability, adaptability) that are not usually addressed in other document visualization systems. A more detailed description of VIEWER including architectural aspects and more Web session examples can be found in (Berenci *et al.*, 1998).

---

terms, secondary topic keyword, and non-categorical terms.

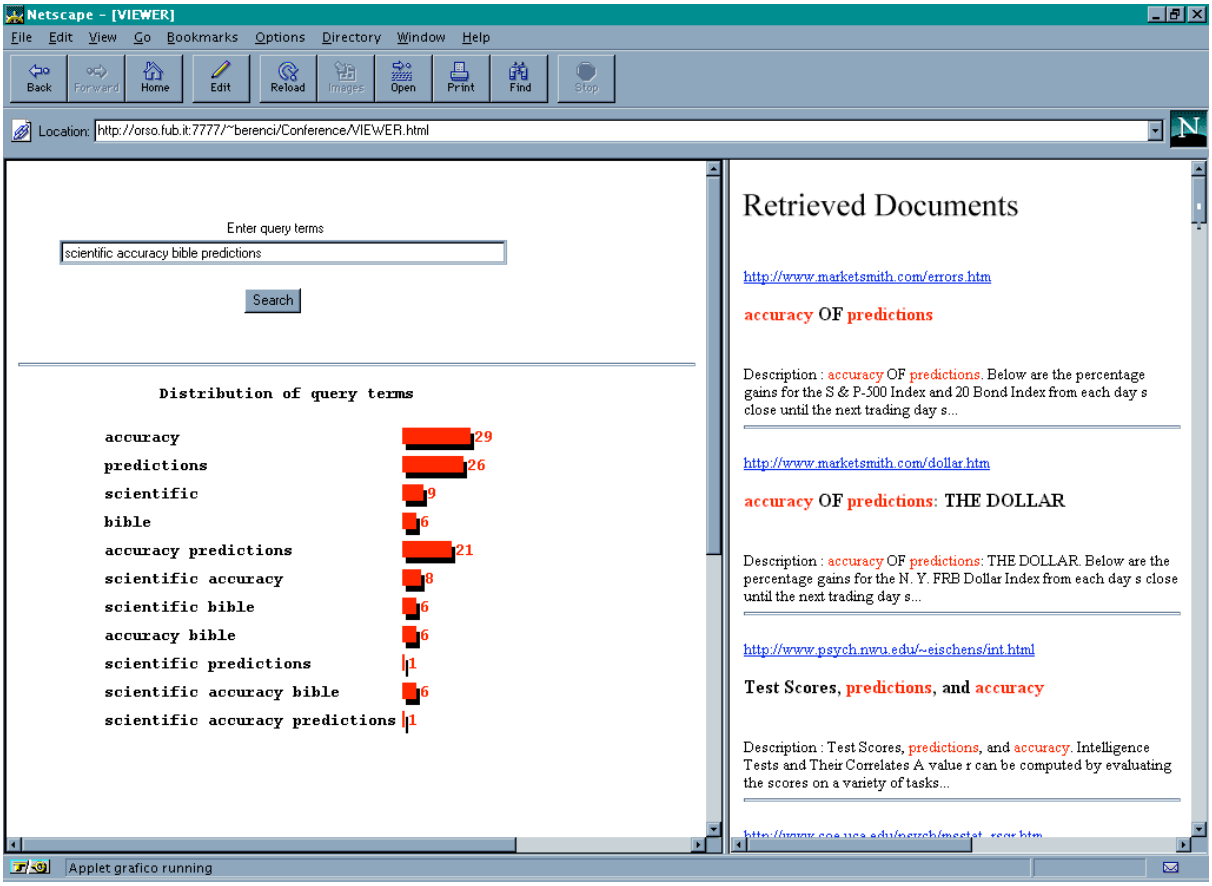

Figure 1: VIEWER visualization of Web retrieval results for the topic: "scientific accuracy of Bible predictions".

4. Experiment 1: Comparing VIEWER and AltaVista on single-query searches

4.1 Goal

The goal of the experiment was to evaluate the effectiveness of VIEWER in reordering the set of documents retrieved by a ranked output retrieval system in response to a query. We compared the performance of VIEWER, used as an interactive ranking system, with that of AltaVista, which produced the ranked document list given as input to VIEWER itself.

4.2 Subjects

We tested ten subjects in this experiment. The subjects were recruited in our institute; they had a computer science background and good knowledge of English. Each subject was provided with a short tutorial session on a training database to ensure that he or she could easily manipulate the interface used in the experiment.

4.3 Database and queries

We did not perform subject searching over the Web because it would be difficult to assess a system's retrieval effectiveness and do comparative studies in this unrestricted domain without biasing the results. Rather, we used a standard large IR corpus, containing a set of predefined topics and their associated relevance judgements. We experimented over the TREC4 test topics (201-250) and test collection, consisting of disks 2 and 3 (approximately 2 Gigabytes of text) of the TREC/Tipster collection. We chose the TREC4 topics because they are short (a one sentence topic description) and hence may better reflect an interactive situation. Each topic was manually transformed into a three to four term query by selecting terms from the topic. The set of queries generated this way had an average length of 3.9 terms; their complete description is available from: *http://www.fub.it/viewer/queries.txt*.

4.4 Implementation of ranking systems

AltaVista can be used for global Web search as well as for indexing and searching site specific information. We connected an AltaVista server to the test collection and executed the queries, determined as explained in Section 4.3, against the corresponding AltaVista database. The result was the ranked list used as a baseline in the comparison with VIEWER.

The implementation of VIEWER as a ranking system was not straightforward, since VIEWER cannot produce a document ranking by itself. We designed an interactive procedure that worked as follows. Each subject was shown the topic, the query extracted from it, and VIEWER's visualization of the distribution of query terms in the full-text documents retrieved by AltaVista in response to the query. Then the subject was asked to choose a sequence of views (without seeing the document summaries) by repeatedly selecting one of the views offered by VIEWER until all views had been selected. After each view selection, the number corresponding to its rank in the ordering chosen by the user was displayed on the screen.

As an illustration, Figure 2 shows an example screen produced during the search of the documents relevant to the topic about "illegal disposal of medical waste", which was translated into a query with the four corresponding terms. It should be noted that this query may present several partial match problems, because some terms may match out of context of their intended meaning and produce much broader concepts than the topic concept (e.g., "illegal waste disposal", "waste disposal", "illegal disposal", "illegal waste"). In terms of Bodoff and Kambil (1998)'s taxonomy, this problem is essentially one of reduced precision from decreased specificity of indexing (i.e., from the inclusion of the topic's less central concepts). Figure 2 shows a complete ordering of the views associated with the topic, as actually determined by a user in the experiment. The numbers in Figure 2 should not be confused with the number of documents contained in each view, which was not shown to the users in this experiment. The documents were

thus ranked according to the order chosen by the user to select the views.[3] Documents contained in more views were ranked based on the earliest view in which they occurred. In this way we obtained a partly-ordered retrieval output; we further ranked the (equally-ranked) documents within each view by using the ranking produced by AltaVista for those documents. As a result of this process, the final ranking built by the user corresponds to a particular sorting of the documents contained in the output returned by AltaVista. Each subject took about one and a half hour to execute the 50 queries.

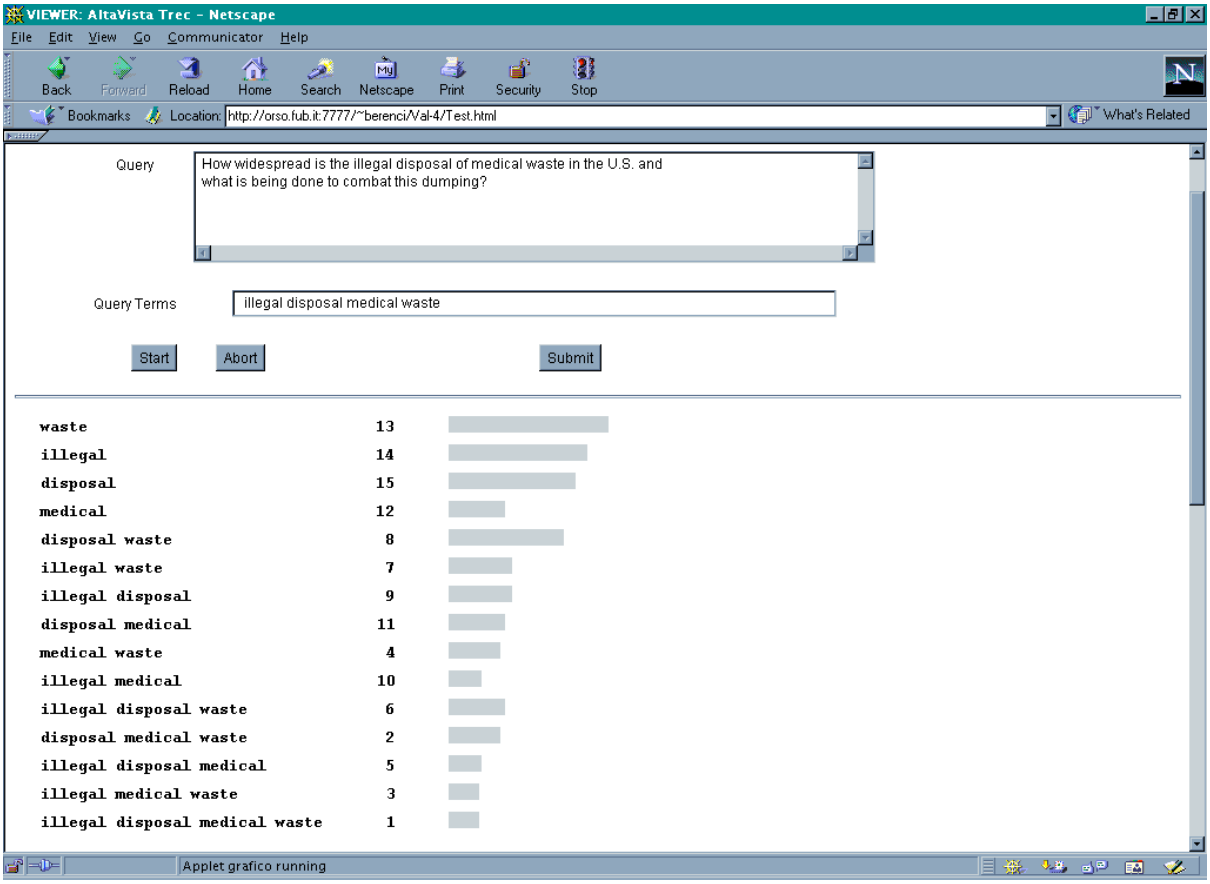

Figure 2. User-ordering of the views associated with the TREC-4 query: "illegal disposal medical waste".

4.5 Results

The results are displayed in Figure 3. The precision-recall curve was normalized considering, for each query, only the relevant documents that contained at least one query term; i.e, those that were actually retrieved and ranked by the two methods. The figure reports interpolated precision at eleven recall levels, averaged over the 50 queries;

[3] It should be noted that the order of views chosen by the user was very different from the one that would have been produced by an automatic system based on coordination level.

the results of VIEWER were averaged over the ten subjects. The performance improvement was therefore apparently consistent at all recall points, and the differences were statistically significant (with a combined *p* value of 5.35E-05). These results confirm and extend earlier findings obtained on two small test collections (Berenci *et al.*, 1998), and offer strong evidence that VIEWER can be effectively used by a user to reorder the documents returned by Web search engines, at least for short queries. This is a useful starting point to evaluate the utility of VIEWER, but its scope is limited by the fact that what we measured is a theoretical, rather abstract, aspect of performance. In practice, the operational conditions are very different, because users do not examine the whole set of documents retrieved in response to a query and because they usually question the system with several queries. This issue is addressed in the next section.

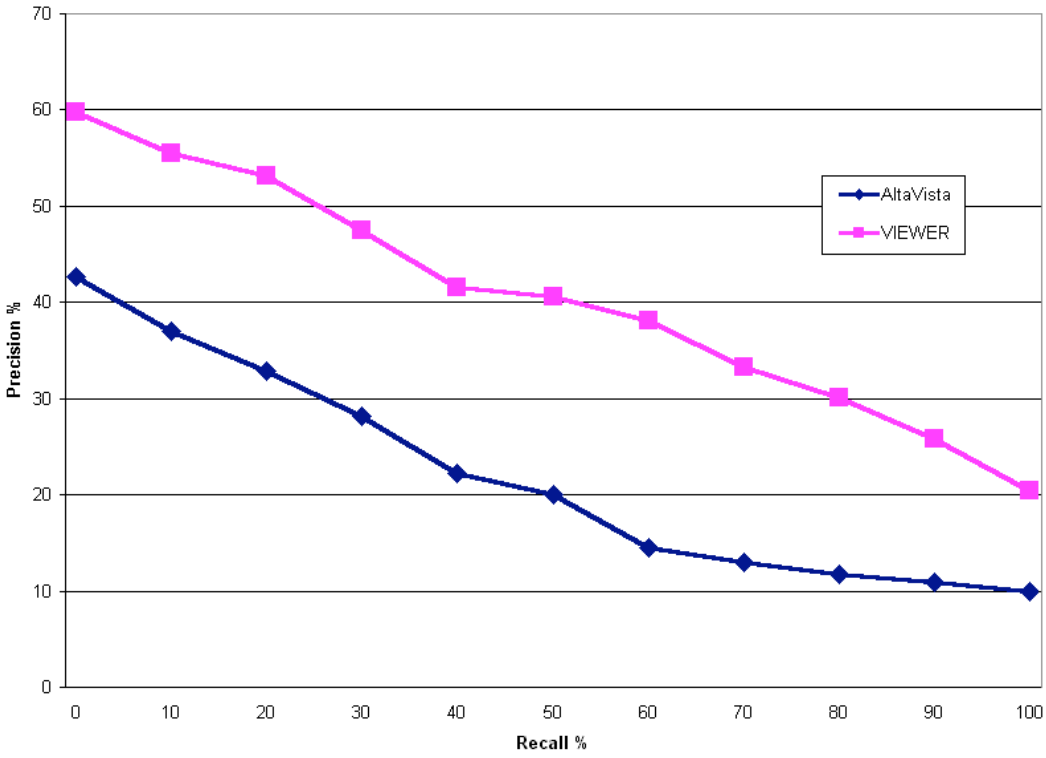

Figure 3: Comparison of AltaVista and VIEWER (used as an interactive ranking system) on short queries

5. Experiment 2: Comparing VIEWER and AltaVista on multiple-query searches

5.1 Goal

The goal of the experiment was to evaluate the effectiveness of VIEWER in contrast with AltaVista in a realistic search situation, involving free inspection and selection of results by users and possibility of query reformulation. In particular, we wanted to test

the hypothesis that VIEWER allows the user to focus on the relevant document summaries obtained in response to a query, without selecting the irrelevant ones, and that VIEWER helps user reformulate a query.

5.2 Subjects

We tested twenty subjects in the experiment. The subjects were undergraduate students at the University of Rome with the following main characteristics (ascertained through a pre-test questionnaire): basic computer experience, some familiarity with on-line document searches, and a good knowledge of English. Sixty dollars was paid to each subject for his participation.

5.3 Databases and topics

We used the same test collection as in the first experiment (TREC4) because its short topics favour the use of short queries and because we can consider our results to scale, given the size and the characteristics of the collection. In addition, using the same collection allows cross-experiment comparison. For this experiment we used only six topics, randomly selected from the 50 TREC4 test topics:

- Topic 202: Status of nuclear proliferation treaties: violations and monitoring.
- Topic 204: Where are the nuclear plants in the U.S. and what has been their rate of production?
- Topic 215: Why is the infant mortality rate in the United States higher that it is in most other industrialized nations?
- Topic 221: Steps taken by church, governments, community, civic organizations to halt carnage among youths engaged in drug or gang warfare.
- Topic 222: Is there data available to suggest that capital punishment is a deterrent to crime?
- Topic 240: What controls, agreement, technological advances or equipment are now in use or planned to assist in combating terrorism?

5.4 Implementation of retrieval systems

We minimized as much as possible the effect that having different interfaces has on performance. The interfaces to the two retrieval systems were implemented as Jave applets; they ran on the same machines (PC's) and used many identical interaction devices such as the topic window, the query formulation facility, and the document summary display-and-evaluation facility. In Figure 4 we show an example screen of the interface used to test VIEWER. The topic being searched is displayed in the left upper window, the current user query is shown right below, and the right window on the screen shows the summaries associated with a user-selected view (i.e., "nuclear

proliferation treaties"). The interface used to test AltaVista was like Figure 4, except that the left lower region of the screen, containing the VIEWER's display stage, was empty.

5.5 Experiment design

In our experimental design the independent variable is the mechanism for selecting the document summaries, which may be based either on VIEWER or on AltaVista. The test comprises two tasks that a group of subjects will have to perform: to retrieve summaries that are relevant to the given set of topics using either VIEWER or AltaVista. The dependent variable is the performance of a group of subjects in these tasks, whose variation may be attributed to the change in the level of the independent variable provided that other biasing factors are kept under control (Preece *et al.*, 1994). Our experimental setting attempted to ensure this condition.

To assign tasks to subjects we used an independent subject design (Robinson, 1983). The 20 subjects were randomly split into two groups with 10 subjects, and each group was assigned to one experimental condition only. The instruction for the task were given in a manner similar to interactive track specification at TREC: "find as many good document summaries as you can for a topic, in around 20 minutes, without collecting too much rubbish". The subjects were asked to label as relevant/nonrelevant each read summary; they did not have access to full-text documents, because this would have distracted them from the main focus of the experiment, i.e., the selection of relevant summaries.

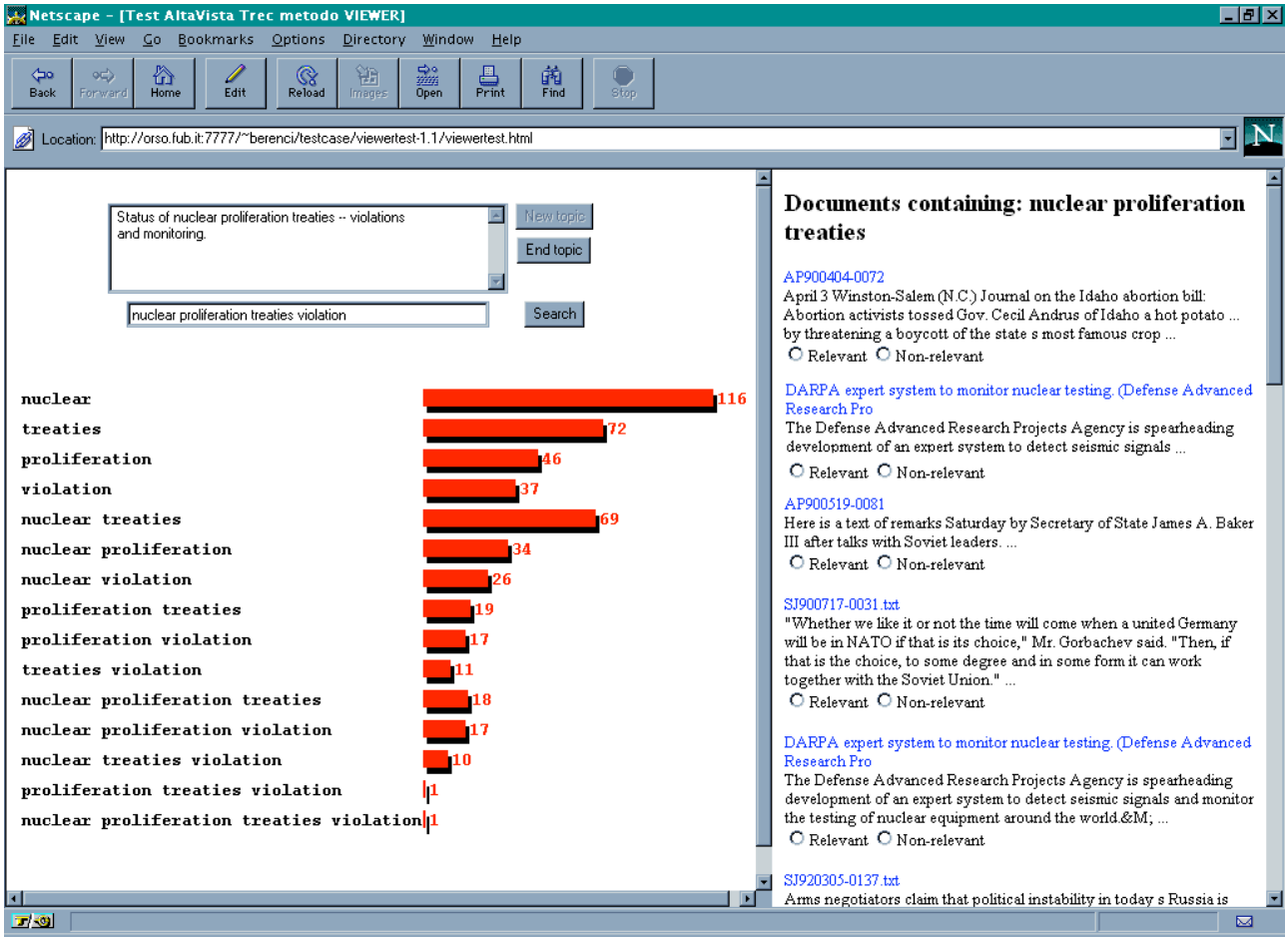

Figure 4. Example screen of the interface used to test VIEWER.

During each search, the subjects using AltaVista could formulate more than one query, seeing the summaries obtained in response to each query on a scrollable window. The subjects using VIEWER could also formulate more queries, but in order to see the summaries they had to select some view first. To be more precise, after submitting a query the subjects were not presented with a list of summaries, as with AltaVista, but with VIEWER's visualization stage relative to (at most) the first 150 summaries retrieved by AltaVista itself. At this point, they could select one or more views and read the summaries associated with each of them (see Figure 4).

5.6 Experiment operationalization

The experimental sessions took place in the "Human factors" laboratory at Fondazione Ugo Bordoni and lasted four days. We tested five subjects at a time, supervised by the same experimenter; each subject performed the task alone, in an acoustically-isolated room, with a video camera recording the session. The subjects were provided with a tutorial session of about an hour, including a search on a training topic. Great attention was paid to ensure that at the end of the training they could easily manipulate the

interface for specifying queries and seeing summaries, including the view mechanism. Then they did the six searches, with a five minutes break between one search and another. Twenty minutes were allocated for each search, although the subjects were allowed to give up at any time after 15 minutes. After each search, they completed a search evaluation questionnaire. At the end of the experiment, a more elaborate questionnaire on their use of the system was administered. The experimental sessions were fully digitized: the topic to be searched automatically appeared on the screen and all questionnaires were filled out by computer.

5.7 Performance measures

Since there are no established performance metrics that measure the effectiveness of interactive information retrieval systems, especially when visualization of retrieval results is involved, it is advisable to use different evaluation scenarios and measures. In particular, it seems useful to try to extend conventional measures of batch retrieval to the interactive context by taking also into account the dynamics of retrieval sessions (Carpineto and Romano, 1996; Buckley et al., 1999) and the user's opinions (Brajnik *et al.*, 1996). We focus on precision (i.e., the ratio of number of items retrieved and relevant to the number of items retrieved), because this is usually the primary concern for users engaged in on-line interactive searches. One definitional problem with batch measures, including precision, is that they are based on the notion of retrieved document, which is often difficult to define in an interactive setting. One approach (Veerasamy and Heikes, 1997) is to consider as retrieved documents only the documents that have been seen and judged to be relevant by the user, but this approach has the disadvantage that we might be measuring the users judgement more than the effect of the retrieval method. A more typical choice (e.g., Tague-Sutcliffe, 1992; Carpineto and Romano, 1996) is to rate a document as a retrieved document as soon as its description (the summary, in our case) is recovered, without considering the user's judgement. In this experiment we have taken the latter approach; i.e., all summaries labelled by a subject in an experimental session, whether subjectively labelled relevant or nonrelevant, were considered as retrieved by the system. Then, in order to consider the dynamics of the session, we measure how the precision of the interactive retrieval varies as a function of retrieved documents and time.

These objective measures are complemented with the opinions of the users gathered from the questionnaires. In the questionnaires we focused on three main variables: user satifaction, utility of the system, and, for the subjects using VIEWER, the usage of views. For each variable we considered a number of aspects, e.g., for user satisfaction we measured interestingness, effort, and fun. For each aspect, the subjects were presented with a five-point rating scale, using both Likert scales and semantic differentials (Preece *et al.* 1994). For instance, user satifaction's aspects were measured through a semantic

differential with three pairs of bi-polar adjectives (boring-interesting, difficult-easy, and unpleasant-pleasant).

5.8 Experimental results

We present the results according to the type of variable measured and not to how it has been measured (i.e., objective versus subjective assessment). The section is split into two parts: results about the relative performance of the two systems, and results about the usage of views with VIEWER.

5.8.1 Performance comparison

Table 1 shows the average number of summaries per topic retrieved by the two systems (partitioning the summaries in relevant and nonrelevant) along with the average precision. Using AltaVista, the users retrieved many more summaries than with VIEWER, but the number of retrieved relevant summaries was very similar for the two systems, which means that the AltaVista users retrieved many more nonrelevant summaries. In fact, the precision of VIEWER was markedly better than AltaVista. The value of precision shown in Table 1 refers to a complete search, and thus ignores the dynamics of interaction. Figure 5 shows the precision of the two systems as a function of the number of retrieved summaries. The results were averaged over the topics and the subjects. The x axis in Figure 5 is restricted to 26 because this was the minimum number of summaries per topic retrieved by the subjects. The two curves show a similar behaviour: the precision initially increases until it reaches a peak after which it declines rather gracefully. This suggests that in an interactive retrieval setting the most relevant documents may not be the very first retrieved documents, as in automatic ranking systems, but those retrieved right after the first ones, probably as soon as the subjects tune in to the search domain and the search facilities. The results of Figure 5 show that the subjects using VIEWER obtained markedly better precision values than those using AltaVista throughout the session.

Table 1. Performance comparison at the end of search

	Retrieved Rel	Retrieved nonRel	Retrieved Total	Precision
AltaVista	7.1	53.9	61	0.116
VIEWER	6.4	24.75	31.2	0.206

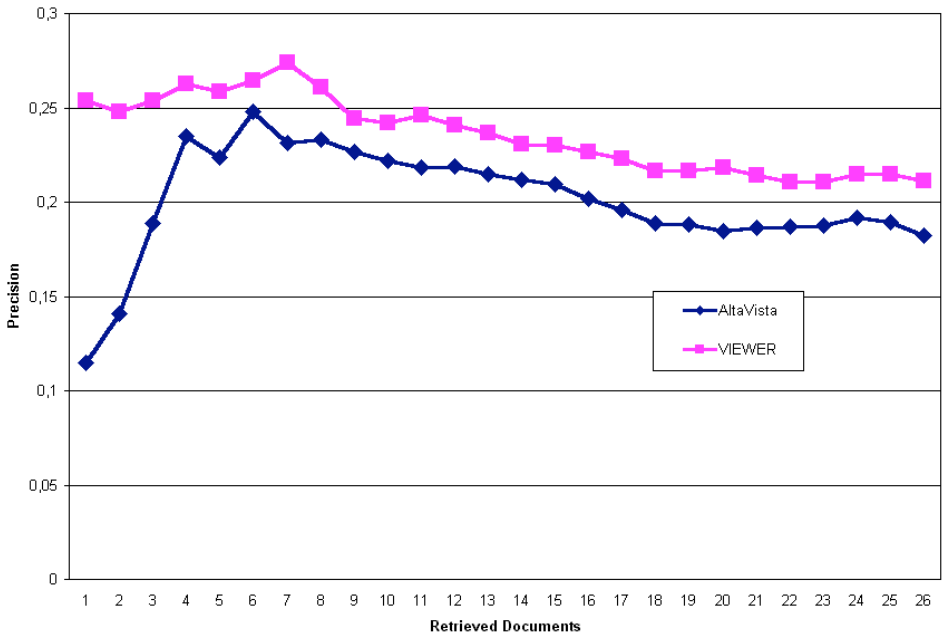

Figure 5. Interactive precision as a function of retrieved documents

Figure 5 tells us with which accuracy relevant summaries have been retrieved, but it does not say when they have been retrieved. This is an important piece of information because we might be more interested in a system that has a lower overall precision but is faster in retrieving a few relevant documents. Figure 6 shows how the precision varied as the search time increased. We restricted the search time to 15 minutes because this was the minimum search time actually taken by the subjects. Similar to Figure 5, Figure 6 shows that the precision of both systems quickly rises and then gracefully decreases as the session progresses. Figure 6 also shows that the precision of VIEWER was better than AltaVista at almost all time points. As in Figure 5, the differences appear to be relevant, although we must be cautios to generalize these results to different sets of topics due to the limited number of topics used in the experiment. Taken together, the results of Figure 5 and Figure 6 suggest that the retrieval of summaries occured rather uniformly throughout the session and across systems. That this was actually the case is shown in Figure 7.

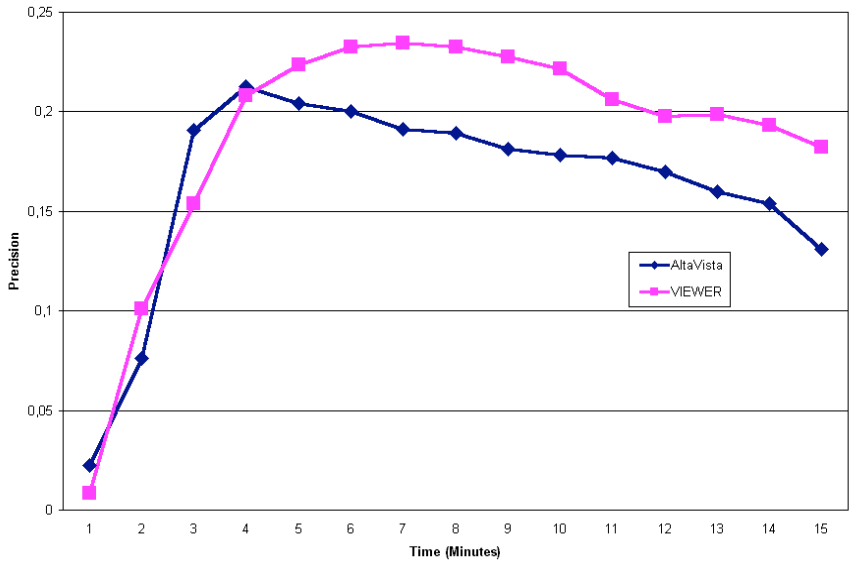

Figure 6. Interactive precision as a function of time.

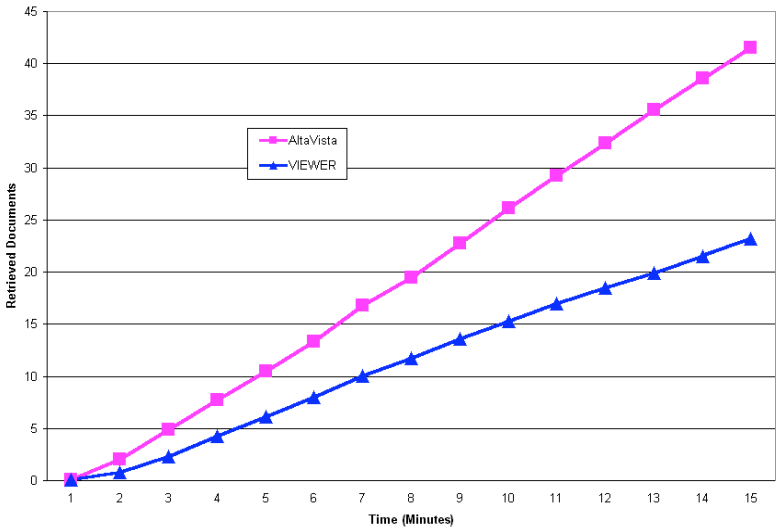

Figure 7. Number of retrieved documents (both relevant and nonrelevant) as a function of time.

We should emphasize that while in the above results about precision we used the "objective" TREC's assessors relevance judgements, we also computed the same curves using instead the subjective relevance judgements expressed by the subject during their search. The results, not shown due to space limitations, were very similar to those found with objective judgements. This represents additional important evidence in favour of our approach because in interactive information retrieval it is likely that user judgements, although less "objective", better reflect the context in which the search took place and its dynamic nature.

As mentioned in Section 5.7, we measured also the user satifaction and the utility of the system as perceived by the subjects involved in the experiment. For each of these two variables, we converted the rating scales into numerical values and computed the mean over the subjects and the aspects measured. The results are depicted in Figure 8,

where the scale ranges from 1 (least helpful) to 5 (most helpful). As shown in Figure 8, the subjective opinions of the users provide further evidence that VIEWER performed better than AltaVista, both with respect to utility and user satisfaction. For the latter variable, according to the user ratings VIEWER was slightly more difficult but much more interesting and pleasant than AltaVista.

Before concluding this section it is useful to look at how an automatic single-query retrieval system would fare on the same topics we used in the interactive multiple-query searches. It turns out that the retrieval effectiveness of an automatically-generated ranked document list would be much worse than that obtained with interactive retrieval. These are indeed difficult topics for an automatic system, because most relevant documents do not match the topic keywords. For instance, for topic 204, 221, and 240, AltaVista would retrieve no relevant document at all in the first 50 documents returned in response to a complete topic statement.

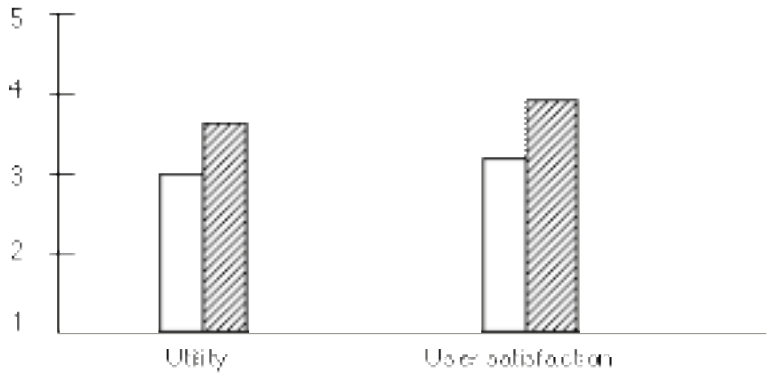

Figure 8. User ratings for Alta Vista (☐) and VIEWER (▨)

5.8.2 Views usage

On the whole, the VIEWER users submitted 500 queries and selected 693 views (about 1.4 views per query, or 11.5 per session). Table 2 shows for how many queries 0, 1, 2, or more than 2 views were selected. According to these results, most of the times the users selected a very limited number of views, thus reading only a small subset of the set of summaries returned in response to a query. In the limit, for 29% of the times, the subjects selected no views at all, which implies that they decided to formulate a new query by just looking at the view display, without reading any summary.

Table 2. Distribution of the number of views.

Number of views	Number of queries	Percentage
0	146	29%
1	222	44%
2	62	12%

>2	70	15%
Total	500	100%

The users preferred views with more terms, which were selected more often and were usually the users' first choice. Table 3 gives the number of times a view of a given length was selected as a first choice, as a second choice, and in total: users selected the three or four term views 45%+20%=65% of the times; the first selected view was usually one with 3 or 4 terms. We also found that the average length of the selected views usually decreased as more views were selected by the users. Table 4 shows that the first view selected was usually (72% of the times) the longest possible (i.e., the one with 4 terms, or the one with all the terms in the query if the query contained less than 4 terms). However, it is also important to note that for 100%–72%=28% of the times (see Table 4 again), the first view selected was not the longest possible: some other reasons (e.g., semantics of terms, size of associated set of documents) induced the users to select another view, usually one among those with the maximum length minus one.

Table 3. Distribution of the length of views.

View length	1st selected		2nd selected		Total	
1	8	2%	5	4%	31	5%
2	87	24%	46	35%	208	30%
3	140	40%	73	55%	314	45%
4	119	34%	8	6%	140	20%
	354		132		693	

Table 4. Distribution of the relative length of views (wrt query length).

Relative view length	Number of selections	Percentage
max	256	72%
max - 1	81	23%
max - 2	16	4.5%

The subjects were asked, through the questionnaires, to rate the importance of the criteria used to choose one view rather than another. The results, shown in Figure 9, suggest that the users made use of all major displayed clues about the relevance of retrieved summaries. This indication was also confirmed by other data gathered from the questionnaires, where the subjects positively rated their full understanding of the view mechanism.

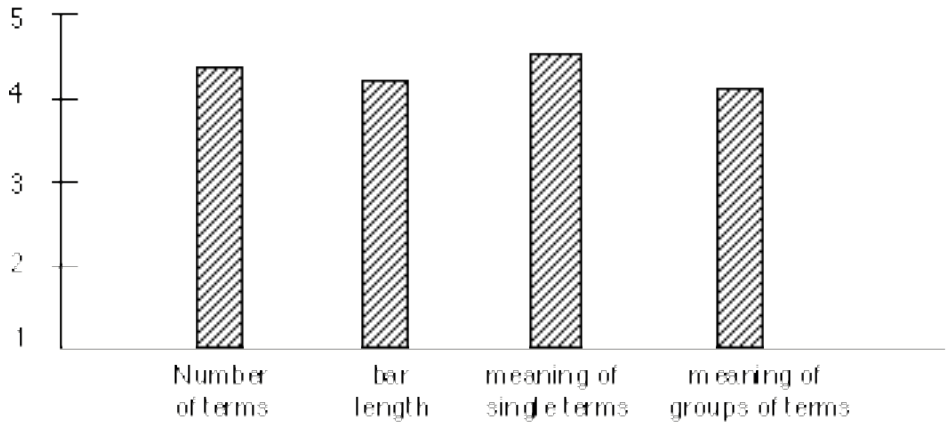

Figure 9. User ratings of criteria used for view selection.

One of the objectives of the experiment was to test the hypothesis that VIEWER supports query reformulation. The results confirm the validity of our hypothesis, as demonstrated by both the number of queries and the query length (number of terms in each query). The average number of queries per session was higher for VIEWER (8.3 versus 6.7), and Figure 10 shows that the average number of queries per minute was higher for VIEWER during almost the whole session.

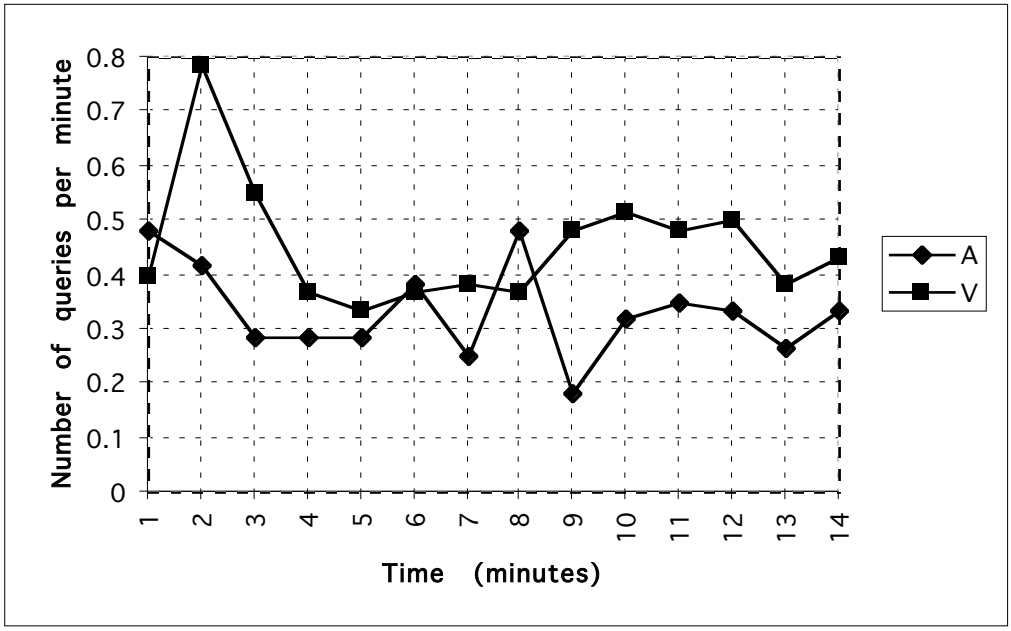

Figure 10. Average number of queries during a search session.

The average query length was also higher for VIEWER (3.8 versus 3.1). Figure 11 shows that average query length slightly increased as more queries were submitted by the users, for both AltaVista and VIEWER. VIEWER stimulated longer queries from the beginning of the session (the length of the first submitted query was 3.3 for

VIEWER and 2.85 for AltaVista), and this difference remained approximately constant for the rest of the session. As long queries are customarily difficult to formulate for real users, this can be taken as an indication that VIEWER supported the subjects in doing so.

The subjects opininons were consistent with these results. The users declared that they sometimes used views for adding, removing, or modifying terms in the query. The subjects also declared that: views were easy to learn (average score 5 on a 1–5 scale), they were useful for query reformulation (4.3), they improved search effectiveness (4.7), they allowed to spare time (4.6), and they did not hinder the search (4.9).

Taken together, the results of Figure 7 and Figure 10 suggest that when passing from AltaVista to VIEWER much of the user effort shifted from inspection to evaluation of results. Using AltaVista, the users formulated fewer queries and retrieved more results, which implies a prolonged direct inspection of the results obtained in response to a query. With VIEWER it is just the opposite: the users retrieved fewer results and formulated more queries. This suggests that they were engaged in an accurate view-based evaluation of the results, which decreased the amount of results actually inspected in response to a query and spurred the formulation of new queries. This observation is also confirmed by the results of Table 2, which show, as already remarked, that the subjects sometimes formulated a new query without retrieving any result of the current query.

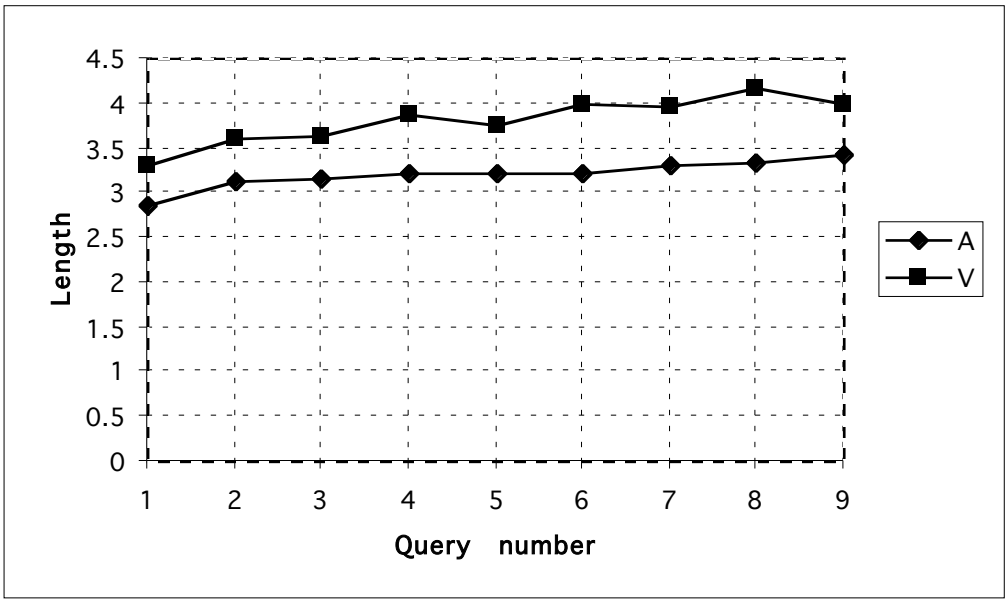

Figure 11. Distribution of query length in the search session.

6. Conclusions

Users engaged in information searches may be willing to use more clues about the relevance of retrieved documents. We showed the feasibility of using the view mechanism to give user more control over display and manipulation of retrieval results. In particular, from our experimental evaluation, two main conclusions can be drawn.

- The view mechanism allowed users to select relevant results with a higher precision, reducing the burden of collecting nonrelevant results.
- The view mechanism shifted the user effort from inspection to evaluation of retrieval results, increasing the number and the length of the submitted queries and increasing the user satisfaction.

One important design parameter of VIEWER is the amount of textual information extracted from the retrieval results and used to compute the views. If a centralized repository with all accessible documents is available, one can use a relatively high number of retrieved documents along with their full-text descriptions without sacrificing the response time of the interface. This was the case in our experiments. However, for ubiquitous searches on the Web, it may be necessary for efficiency reasons to use only a very limited number of the documents retrieved by the engines and to use the short document summaries, as provided by the engines, without downloading the full-text descriptions. The current version of VIEWER manages to keep the computational overhead small by using only a few tens of retrieved summaries. Of course, using only a small fraction of the amount of textual information theoretically available may affect the retrieval effectiveness of the view mechanism. While there are some recent results that suggest that this may not necessarily be the case (Zamir and Etzioni, 1998), an exploration of the main trade-offs involved here (e.g., efficiency versus effectiveness, centralized versus distributed) is an issue for future research.

Acknowledgments

We would like to thank Giovanni Romano for his help in preparing the first experiment and evaluating its results. We would also like to thank Giacinto Matarazzo for his help in designing and organizing the second experiment, and the whole "Human factors" group at Fondazione Ugo Bordoni for making their facilities available for the experiments. This work has been carried out within the framework of an agreement between Telecom Italia and the Fondazione Ugo Bordoni.

References

Berenci, E., Carpineto, C., and Giannini, V. (1998). Improving the effectiveness of Web search engines using selectable views of retrieval results. *Journal of Universal Computer Science* 4(9), 737-747.

Blair, D. and Maron, M. (1990). Full-text information retrieval: further analysis and clarification. *Information Processing & Management*, 26(3), 437-447.

Bodoff, D., and Kambil, A. (1998). Partial coordination. I. The best of pre-coordination and post-coordination. *Journal of the American Society for Information Sciences*, 49(14), 1254-1269.

Brajnik, G., Mizzaro, S., and Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: a case study on user support. *Proceedings of SIGIR'96*, Zurich, 128-136.

Buckley, C., Mitra, M., Walz, J., and Cardie, C. (1999). SMART high precision: TREC-7. *Proceedings of TREC-7*.

Carpineto, C., and Romano, G. (1996). Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45, 553-578.

Chalmers, M., and Chitson, P. (1992). "Bead: explorations in information visualization", *Proceedings of SIGIR'92*, Copenhagen, Denmark, 330-337.

Clarke, C., Cormack, G., and Tudhope, E. (1997) "Relevance ranking for one to three term queries". *Proceedings of RIAO'97: Computer-assisted information searching on the Internet*, Montreal, Canada, 388-400.

Cooper, J., and Byrd, R. (1997). "Lexical navigation: visually prompted query expansion and refinement". *Proceedings of the 2nd ACM Digital Library Conference*, 237-246.

Hearst, M. (1995). "TileBars: Visualization of term distribution information in full text information access". *Proceedings of CHI'95*, Denver, Colorado, USA, 59-66.

Hearst, M. (1996). Improving full-text precision on short queries using simple constraints. *Proceedings of the Symposium on Document Analysis and Information Retrieval.*

Hemmje, M., Kunkel, C., and Willet, A. (1994). "LyberWorld - A visualization user interface supporting full text retrieval". *Proceedings of SIGIR'94*, Dublin, Ireland, 249-259.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994). *Human-Computer Interaction.* Addison Wesley.

Robinson, C. (1983). *Experiment, design, and statistics in psychology, 2nd edn.*, Harmondsworth: Penguin.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison Wesley.

Spoerri, A. (1994). "InfoCrystal: Integrating exact and partial matching approaches through visualization". *Proceedings of RIAO'94: Intelligent Multimedia Information Retrieval Systems and Management*, New York, New York, USA, 687-696.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing & Management*, 28, 4, 467-490.

Tombros, A., and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proceedings of SIGIR'98*, Melbourne, 3-10.

Van Rijsbergen, C. (1979). *Information Retrieval (second edition),* Butterworths, London.

Veerasamy, A., and Heikes, R. (1997). Effectiveness of a graphical display of retrieval results. *Proceedings of SIGIR'97*, Philadelphia, USA, 236-245.

Zamir, O., and Etzioni, O. (1998). Web document clustering: a feasibility demonstration. *Proceedings of SIGIR'98*, Melbourne, 46-54.