

Evaluation of a Tool for Visualization of Information Retrieval Results

Aravindan Veerasamy
veerasam@himalaya.cc.gatech.edu
College of Computing
Georgia Tech
Atlanta, GA, USA

Nicholas J. Belkin
belkin@scils.rutgers.edu
School of Communication, Information & Library
Studies
Rutgers University
New Brunswick, NJ, USA

Abstract

We report on the design and evaluation of a visualization tool for Information Retrieval (IR) systems that aims to help the end user in the following respects:

- As an indicator of document relevance, the tool graphically provides specific query related information about individual documents
- As a diagnosis tool, it graphically provides aggregate information about the query results that could help in identifying how the different query terms influence the retrieval and ranking of documents.

Two different experiments using TREC-4 data were conducted to evaluate the effectiveness of this tool. Results, while mixed, indicate that visualization of this sort may provide useful support for judging the relevance of documents, in particular by enabling users to make more accurate decisions about which documents to inspect in detail. Problems in evaluation of such tools in interactive environments are discussed.

1 Introduction

The disadvantages of Boolean IR systems are well known. Best-match (i.e. ranked output) systems address several of these problems by allowing users to submit unstructured queries, and by ranking the retrieved documents in (presumed) order of relevance. However, such systems also introduce new problems, or exacerbate problems that are not so severe in Boolean systems.

For instance, understanding why a document (or set of documents) was retrieved is relatively straightforward in exact-match systems, since all members of the set are required to contain exactly the query specification. Furthermore, the ordering within the set of retrieved documents is typically based on relatively well-understood formal characteristics of the documents, such as date of publication or alphabetical order by title or author. In best-match systems, on the other hand, neither the matching rule nor the ranking rule is easily understandable. The former is usually based on characteristics and algorithms which don't have simple relationships with the unstructured query; the latter is intended to reflect complex conceptual relationships between the query and the individual documents, and between the documents themselves.

Furthermore, query reformulation may be more difficult in Boolean than in best-match systems. Obtaining a manageable output set size in Boolean systems (the most typical re-formulation task) may be less demanding than attempting to rearrange the list of retrieved documents in a best-match system by manipulating an unstructured query. This is especially difficult when the rules for ordering and matching are not well understood.

Current best-match IR systems take relatively little account of these issues. In response to a user's query, most systems display surrogates (title, source, author ...) of the top 'n' retrieved documents, in a list, with some number(s) indicating the rank, or reason for being in that rank, i.e. a retrieval status value (RSV). Some systems display by default more information about the first retrieved document; most require the user to request such information (e.g. the full text of the document) explicitly. The only explanation of *why* the documents are ranked the way they are is typically the RSV, about which there is no further information than the number itself. More explanation may not be necessary in situations where the top retrieved documents are all clearly relevant. But when the user needs to modify the query in order to get better results, understanding the causal relationship between query and document ranking becomes very important. Having an accurate idea of *why* a list of documents was retrieved, of *how* they were ranked, and of *what* is sub-optimal about the ranking could be useful in effective query reformulation.

Knowledge about the relationships between query and ranking of retrieved documents is not in itself sufficient for effective query reformulation. It is also necessary that the user be able to manipulate the query effectively after the problem has been identified. For instance, knowing that an important query concept is missing in most of the retrieved documents is not sufficient for effective query reformulation. One must then be able to find the right words (or other techniques) for increasing the importance of the concept in the query. Without the ability to take corrective action once the problem is diagnosed, the diagnostic information is of little value.

A possible means for addressing problems of this sort is to display to the user something about the documents which relates them directly to characteristics of the query, and which relates them to one-another. Highlighting query terms in the text display of retrieved documents attempts to accomplish the former, and the indication of RSV is an attempt to accomplish the latter. However, neither of these techniques appears to give sufficient information to guide effective query reformulation. Graphical displays of the characteristics of retrieved documents (*visualizations*) which are relevant to their retrieval and ranking is one obvious approach to this problem.

A further problem in IR systems in general has to do with the multi-stage nature of presentation of results. The initially-presented surrogates are meant to provide a concise picture of what a document is about. Based on these surrogates, the user

may request more detailed information about particular documents which look promising (or for which the surrogate information is equivocal). In some cases, this might be the “full” bibliographic information about the item, in others an abstract, and in many systems now, the full text of the item. Thus, as the user progresses through the stages of display, that which is displayed is more complete and informative, allowing increasingly accurate relevance judgments. But, the information in the later stages of display is also more time-consuming to peruse. Therefore, it is important for the searcher to be reasonably certain that it is worthwhile doing this inspection. The information displayed in the earlier stages thus serves as a filter which supports the user in deciding which documents do not need further inspection (either because they are obviously good or obviously bad), and which documents do justify the further effort.

Thus, displaying a great deal of information at the surrogate stage of display could be a useful device for judging the relevance or usefulness of the document. The advantage is that when the user requests the second-stage display, it is more likely that that document will be relevant to the user than if there were less information in the first stage. The disadvantages to this strategy are that since there is more text to display in the first stage, fewer items can be presented, and more time must be spent in perusing the first-stage display. Thus, the total number of documents seen by the user may well be fewer, although the quality of the decision-making may be higher.

If, on the other hand, one chooses to display less information at the initial surrogate stage, then the decision about whether to look at the more complete display is less secure. Hence, the proportion of second-stage documents which turn out to be relevant is likely to be low. The advantage of seeing more documents, more quickly, in the first stage is thus offset by the additional time that is spent perusing non-relevant documents in the second stage.

A possible means to addressing this problem is to display information about the document in the first stage in some form that does not require as much perusal time and screen space as text. Graphical displays of the characteristics of documents which are significant in supporting the decision to peruse or not (*visualizations*), could support set-at-a-time perusal of documents, rather than document-at-a-time perusal of text displays.

It will not escape the reader that the suggested solutions to the two classes of problems that we have raised here are rather similar, and could, indeed, be instantiated by the same sort of display. We present a visualization tool which is intended to address these problems in IR systems, and a preliminary evaluation of this tool. The remainder of this paper is organized as follows. We first present a description of the visualization tool, and a rationale for the features of this tool with respect to the problems in IR interaction that we have discussed. We then discuss some related work in IR visualization that addresses this type of problem, and draw some comparisons between that work and ours. We follow with a description of the experiments we conducted to evaluate the visualization tool, and the results of those experiments. We conclude with some comments on the implications of our results, on future work, and on the implications of our evaluation experience for evaluation of interactive IR in general.

2 Visualization tool

2.1 Description

The visualization tool is an adjunct to a basic interface for IR. This interface is structured as indicated in Figure 1, with a query window, a display of titles retrieved, and the full text of a document. This serves as the baseline interface interaction which is compared to the visualization tool. A screen snapshot of the visualization tool is shown in Figure 2. The visualization corresponds to the query “how has affirmative-action affected the construction-industry construction projects and public works”.

The visualization consists of a series of vertical columns of bars. There is one column of bars for each document. The left-most vertical column of bars corresponds to the document ranked 1 and the rightmost vertical column corresponds to the document ranked 150. In each vertical column there are multiple bars -- one each for each query word. The height of the bar at the intersection of query word row and a document column corresponds to the weight of that query word in that document. Moving the mouse cursor over the vertical columns highlights the column directly beneath the mouse cursor and simultaneously highlights the title corresponding to that document in a title display window.

Apart from the query words typed in by the user, the visualization also shows the distribution information for words added by the system due to relevance feedback. Thus, all the words internally used by the system in computing the query results are shown in the visualization. This window is scrollable, in case the number of words in the query exceeds the vertical space. The words in the visualization are also stopped and stemmed. The basic interface, and the visualization tool, utilize the INQUERY retrieval engine, version 2.1p3 [Callan, Croft, Harding, 1992]. We use all of the default features of that system, including their relevance feedback, stemming and stoplist algorithms, but do not use any of the structured query facilities.

2.2 Response to problems of IR interaction

In support of query reformulation, the visualization makes the connection between the query and retrieved documents explicit by graphically displaying the contribution of each query word to the retrieval of each document. The higher the contribution of a particular query word to the retrieval of a document, the taller the bar at the intersection of the corresponding query word and document. The absence of a bar at the intersection illustrates the absence of the term in the document. Absence of an important query concept in a number of retrieved documents points to a problem situation which the user needs to work on. The visualization makes relations between the documents themselves explicit, since the characteristics which have led to their rank (the number and contribution of matching terms) are explicitly displayed.

In support of informative first-stage display, the visualization provides a great deal of information useful for deciding whether to view the full text of a document in a highly condensed way, and allows many document surrogates to be displayed at one time. The presence or absence of specific significant words in any document can be quickly seen, and it is possible to identify sequences of documents which do, or do not have important contributions from specific query words.

Figure 1. Sample querying session. The window in the top-left corner is the query entry window. Immediately below that is another window where the titles of retrieved documents are displayed. To the bottom right is another window where the full texts of documents are displayed.

For the example search topic (“How has affirmative action affected the construction industry?”), there are two facets that are central: “affirmative action” and “construction industry”. From the visualization tool, we can immediately see that most of the documents are concerned with the “construction industry” and only a portion of the documents have the term

“affirmative action”. We can also see that the “affirmative action” concept is spread sparsely throughout the top 70 documents. The graphical format of presentation has some important advantages in that it is more condensed than an equivalent text display.

Figure 2. Visualization of results. The highlighted vertical column corresponds to document ranked 14. The title of document ranked 14 document will also be highlighted in the title display window. Clicking the highlighted vertical column brings up the full text of that document

From the visualization, one gets an immediate idea of how the different query words influence the document ranking (as given by the height of the bars). One can see that the concept “affirmative action” is not well represented in the retrieved documents. This suggests that synonyms and words related to that concept must be added to the query to reinforce that query concept in subsequent search iterations. From the visualization tool, one can infer that the system interprets “public” and “project” as two separate words and that the contribution of those two words to the retrieval of almost all documents is uniformly low (as given by the height of the bars). One can probably improve the situation by making “public projects” a phrase, thereby retrieving documents that have these two words in close proximity. Gaining such overall information about the query results by reading the document text is at best cumbersome if at all possible.

3 Related visualization work

A number of visualization schemes for information retrieval systems have been proposed. The Perspective Wall [Card, Robertson & Mackinlay, 1991] is a visualization scheme which supports browsing of documents. While such a system can not handle qualitative document classifications such as library subject catalogs, it is very useful for visualizing documents based on data which are linear in nature (like date of publication). A nice way of integrating different visualization schemes for efficient navigation through the hypermedia space has been proposed by [Mukherjea (1995)]. These schemes are primarily useful for navigational tasks. Other visualization schemes such as those of [Korfhage (1991)], [Spoerri (1994)], [Hemmje, Kunkel & Willett (1994)] have facilities for viewing a large document space. But visualizing the document space along more than 3 - 4 dimensions simultaneously becomes

very cumbersome using their systems. The visualization scheme in our tool can gracefully handle more query word dimensions. Many systems are tailored towards easy construction of queries [Spoerri (1994)] [Aboud, et al. (1994)] [Arents & Bogaerts (1993)] but do not pay much attention to the display of query results.

TileBars [Hearst (1995)] visually shows the query term distribution and overlap in retrieved documents. The term distribution in retrieved documents is shown right beside the title of the document. In a number of respects, the reasons and motivations for Hearst's work are similar to those of our visualization [Veerasamy, Navathe & Hudson (1995)] [Veerasamy & Navathe (1995)] [Veerasamy (1996)]. There are some important ways in which TileBars differs from the visualization that we propose.

- TileBars provides information on how different query facets overlap in different sections of a long document. Our visualization scheme does not provide information at that fine level of granularity.
- TileBars presents the document surrogates in a list, making it more difficult than in our tool to gain an overall picture of the query word distribution for a whole set of documents in one glance.
- TileBars seems best suited for long documents, while our visualization scheme does not seem to be constrained by length.

4 Experiments to test the effectiveness of visualization

4.1 General conditions

We discuss two experiments for testing the effectiveness, usability and acceptability of the visualization tool by comparing searching with an interface using the visualization, versus searching with the same interface, but without the visualization tool. The underlying retrieval engine used in these experiments was INQUERY version 2.1p3, from the University of Massachusetts, Amherst, generously made available to us by Prof. Bruce Croft [Callan, Croft & Harding (1992)]. We developed the graphical user interface using Tcl/Tk on top of INQUERY.

The experiments were conducted as part of the TREC-4 interactive track (Harman, 1996). Thus, the task for the searchers in the experiment was the TREC-4 interactive track task:

Find as many documents as you can which address the given information problem, but without too much rubbish. You should complete the task in about 30 minutes or less.

The "information problems" were chosen from the 25 adhoc topics used for the TREC-4 interactive track, and the database was the TREC Disks 1 and 2 database of the full texts of about 550,000 documents.

Both experiments were designed to test the usefulness of the visualization tool for addressing the two problems that we have discussed and that motivated the design of the tool:

- efficiency and effectiveness in discovering relevant documents; and,
- effectiveness in supporting query reformulation.

In order to test the former, we predict that searchers using the visualization tool will make better decisions about which documents to look at (or not look at) than those without visualization. We operationalize this difference with the following dependent variables:

- the number of documents saved per search (*s-p-s*). Since search times are more-or-less constant (about 30 minutes) across searchers, this measure reflects efficiency in being able to see more documents.
- the proportion of documents whose full text was viewed that were judged relevant by TREC evaluators (interactive trec precision or *i-t-p*). This measure indicates the quality of the documents which were chosen for viewing.
- the proportion of documents whose full text was viewed that were saved by the searcher (interactive user precision or *i-u-p*). This measure also indicates quality of documents which were chosen for viewing, but is indicative of the relationship of the display to the searcher's own concept of relevance to the problem, rather than being dependent upon the external relevance judgments.

To test the latter, we use:

- precision of the search, measured in the required manner for the TREC-4 interactive track; that is, as the proportion of documents saved by the searcher that were judged relevant by the external judges. This measure is indicative of the effectiveness of retrieval performance.

For all of these measures, higher numbers mean better performance.

The subjects for both experiments were undergraduate student volunteers who were registered in a one-credit hour course on library searching in the College of Computer Science at Georgia Tech. All subjects had prior computer experience, the majority with more than four years. All subjects were majoring in an engineering discipline, and had varying levels of experience with the Georgia Tech Electronic Library Catalog. They had no other IR experience than that offered by the class. Two different groups of subjects were used in the two different experiments.

All the subjects in both experiments followed the same general introductory and tutorial procedure. They were asked to fill out a background questionnaire about their computer and IR experience, major, and so on. They then had a hands-on tutorial (about 1 hour) on how to use the version of the system they would be using for the first experimental search. They were then asked to do a practice search on TREC topic 224 ("What can be done to lower blood pressure for people diagnosed with high blood pressure? Include benefits and side effects.") for 15 minutes. They then did the assigned searching tasks (details differ between the two experiments), during which they were instructed to "think aloud", which was recorded on audio tape. All the user interaction with the system was logged. After each search, they completed a search evaluation questionnaire. At the end of the session, a structured interview on their use of the system was administered. All subjects did three runs of the system: one practice run and two test runs.

4.2 Experiment 1

Thirty-six subjects were randomly divided into three groups of twelve each. Twenty-four of the 25 TREC-4 interactive track topics were randomly divided into twelve pairs. Each of the twelve pairs of search topics was randomly assigned to one of the searchers in each group, one to be searched in the "first" condition, the other to be searched in the "second" condition for the group of which the searcher was a member. The topic pairs were searched in the same order in all groups. Thus, the same twelve of the 24 topics were searched in the first condi-

tion for all three groups, and the other twelve were searched in the second condition for all of them.

The three groups were defined according to the combination of conditions or treatments. Group wo:w (for WithOut:With) did the initial tutorial, the practice search and the first search without the visualization tool. An intermediate tutorial after the first search introduced the visualization tool and the search for the second topic was done with the visualization tool. Group “w:w” (for With:With) used the visualization tool for all the searches and the introductory tutorial. Group “wo:wo” (for WithOut:WithOut) did all the searches and the introductory tutorial without the visualization tool. In both the w:w and wo:wo groups, an intermediate tutorial on the interface with which they were working was introduced between the two searches to match the intermediate tutorial of the wo:w group.

This “within subjects” design was used in order to control for user differences, and to account for any possible learning effects from search 1 to search 2. It was predicted that performance on the various measures would improve from first search to second search in the wo:w group, more than in either the wo:wo or w:w groups.

4.3 Experiment 2

In this experiment, 36 subjects were randomly divided into two groups, one with the visualization tool (“viz”), the other without (“noviz”). Three search topics were chosen for searching by all eighteen searchers in each of the two groups, always in the same order. The searchers in the two different groups followed the same pattern of participation as those in experiment 1, but without any intermediate tutorial, and with the practice search time extended to 30 minutes. We picked topic 242 (“How has affirmative action affected the construction industry?”) for the practice search. The first “experimental” search was on topic 236 (“Are current laws of the sea uniform? If not, what are some of the areas of

disagreement?”), and the second was topic 203 (“What is the economic impact of recycling tires?”).

This “between-subjects” design was used to control for the effects of search topic difference, and to have larger numbers of subjects in the two conditions. It was predicted that performance in the viz group would be better than performance in the noviz group for each topic.

5 Results

In this paper, we report only on results with respect to the performance measures we have defined. Results from the questionnaires with respect to use and usability of the two systems, and with respect to interaction measures and “thinking aloud” will be reported in subsequent publications.

5.1 Experiment 1

The results of experiment 1, displayed in Table 1, are something of a disappointment. There are no significant differences (using the Wilcoxon Matched-Pairs Signed-Ranks Test, one-tailed at $p \leq .05$) between any of our four measures between the without- and with-visualization treatments in the wo:w group. Furthermore, there are no significant differences between any of the matched without-visualization/visualization groups (i.e. between the second searches of wo:w and wo:wo groups, the first searches of wo:w and w:w groups, and both searches of the wo:wo and w:w groups). There is no consistent pattern on any of these measures from first to second search (i.e. there appears to be no learning effect, nor does it appear that one of the sets of twelve topics is in general more difficult than the other, nor is it the case that any of the groups does consistently better or worse for either search). These very mixed results lead us to think that our experimental design in this case suffers from two significant problems. The first is great inter-subject and inter-topic variability; the second is that we have too few subjects for each condition to adequately test significance of any differences that may exist.

Condition	Precision			s-p-s			i-t-p			i-u-p		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
wo:w.1	0.712	0.750	0.257	2.193	1.630	2.671	0.491	0.505	0.296	0.459	0.435	0.210
wo:w.2	0.531	0.635	0.402	2.686	1.915	2.848	0.419	0.405	0.270	0.454	0.445	0.204
wo:wo.1	0.566	0.530	0.286	4.039	1.250	7.289	0.385	0.415	0.187	0.429	0.420	0.163
wo:wo.2	0.513	0.670	0.342	3.500	1.500	5.437	0.344	0.315	0.302	0.379	0.425	0.201
w:w.1	0.586	0.515	0.284	2.319	2.170	1.301	0.402	0.285	0.251	0.443	0.450	0.164
w:w.2	0.496	0.415	0.356	2.480	1.560	3.096	0.433	0.385	0.299	0.407	0.390	0.241

Table 1. Summary results for experiment 1. w = with visualization, wo = without visualization. The order of w and wo in the condition column indicates the order of application of conditions for that group, the number following the two indicates for which of the two conditions, first or second, the value is given. s-p-s = documents saved per search; i-t-p = interactive trec precision; i-u-p = interactive user precision.

5.2 Experiment 2

The results of experiment 1 led to the design of experiment 2, whose results are displayed in Table 2. The rows in Table 2 are in the order that the topics were searched. In order to test

the significance of these results, it is necessary to compare them topic-by-topic, without cumulation, to maintain the assumption of independence, since each searcher did three searches (including the practice search, 242) in the same condition. To test for significance of results, we used the Mann-Whitney U test with $p \leq .05$, one-tailed. For precision,

there is no significant difference between noviz and viz for any of the three topics. For s-p-s, the trend is in favour of viz in all three cases, but significantly so at the chosen level only for topic 242 (although for topic 236 it only just misses). For i-t-p, again the trend is nominally in favor of viz, but is again significant only for topic 242. For i-u-p, the same trend holds, and again viz is significantly better than noviz only for topic 242.

For three of the four measures we can see that there are obvious topic differences which cannot be accounted for by a learning effect, since the direction is wrong. Two points are important to note here. First, it appears that topic 242 was “easier” than the other two topics, and that topic 242 benefited most from visualization. Second, it is clear that differences in topics are likely to affect results averaged over topics, unless there are also quite large numbers of searchers for each topic.

Condition/ Topic	Precision			s-p-s			i-t-p			i-u-p		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
noviz.242	0.912	1.000	0.140	1.084	1.000	0.706	0.371	0.365	0.132	0.344	0.325	0.132
viz.242	0.898	0.880	0.113	1.899	1.750	0.888	0.545	0.530	0.143	0.454	0.410	0.177
noviz.236	0.227	0.145	0.232	1.347	1.210	0.939	0.085	0.045	0.101	0.376	0.390	0.145
viz.236	0.215	0.140	0.244	2.249	1.670	1.701	0.105	0.080	0.080	0.472	0.440	0.174
noviz.203	0.424	0.470	0.178	1.326	1.085	0.956	0.217	0.215	0.096	0.347	0.330	0.161
viz.203	0.392	0.330	0.226	1.676	1.500	1.205	0.259	0.280	0.070	0.357	0.370	0.157

Table 2. Summary results for experiment 2. viz = with visualization, noviz = without visualization. The numbers following the condition designation indicate the topic searched. s-p-s = documents saved per search; i-t-p = interactive trec precision; i-u-p = interactive user precision.

Interpreting these results is somewhat difficult, although they are a bit more promising than those of experiment 1. It is of some interest that only topic 242 showed significant differences between the noviz and viz groups. This might be explained by that topic's being for some reason more suited to visualization than the other two. Although the numbers of relevant documents for the three queries are rather similar (242: 38; 236: 43; 203: 33), on the basis of median precision reported by all of the TREC-4 interactive track participants, topic 242 is “easier” than topics 203 and 236 (0.2368 vs 0.1515 vs 0.0465, respectively). This of course follows the pattern of precision results by the searchers in experiment 2, but it is not clear how this would explain the apparently beneficial effect of visualization for this topic. An alternative explanation might be that visualization of this sort is helpful for naive searchers, but loses its effect as they become more experienced with the IR system. On the basis of the data we have available, there is no way to decide between these alternatives.

In any event, it seems reasonable to accept, on the basis of the results of experiment 2, that there could indeed be some value to visualization of the sort we have tested here. However, this statement certainly must be very tentative, and subject to much more testing. The results of experiment 1 do not lead to any such conclusion. It must be said, however, that the very mixed nature of these results may well be an effect of the experimental design, and in particular of the inability to take proper account of what may be very large topic differences and searcher differences. Of course, another possible reason for the seeming lack of effect of visualization is the implementation that we chose. This issue needs further investigation.

6 Conclusions

The study reported here intended to demonstrate the potential of visualization to support particular kinds of interactions in IR, and to test one implementation of such visualization. Although the results of our experiments are mixed, it appears that some of them are positive enough to justify further such ex-

periments. But there are some other serious implications of our results.

We are not aware of other work reporting comparisons of visualization tools for IR with equivalent non-visualization interfaces. Our experience suggests that it is important to conduct more such studies, in particular to move beyond assuming the efficacy of visualization to demonstrating it in experimental environments. Our study also demonstrates the severe problems that arise in conducting interactive IR experiments. These include the problems of finding enough subjects to account for inter-subject differences, and of being able to account for inter-topic differences. Balancing these two demands is an exceedingly difficult problem, which is currently severely exercising the TREC-5 interactive track participants.

Another evaluation problem raised by our study is how to measure the effectiveness of visualization tools. The problems with using precision as a measure for evaluating interactive IR are now well-known, especially if precision is decided according to relevance judgments from experts, rather than the searchers. It is also the case that for certain functions of visualization, precision is an inappropriate measure. But we do not have available a suite of accepted alternative measures for evaluating the effectiveness of systems with respect to these functions. So it was necessary for us to invent some new measures which appear appropriate to the IR tasks that we wished to support. Whether these were good choices also needs to be further investigated.

In conclusion, we find that this study has given some support for the general idea of visualization as a tool for enhancing user interaction with search results, and for the specific tool with which we implemented this idea. We also find that the level of support for these statements from this study is not high, and that it is necessary to conduct further studies, with better designs, before we can become confident in the value of visualization for these purposes, as opposed to other tools for interaction. Finally, we find that our study has shown, again, the necessity of developing better measures and methods for the evaluation of interactive IR systems, and the

necessity of rigorous comparative evaluation of visualization in IR.

Acknowledgments

Support from the ARPA contract No. F33615-93-1-1338 to the first author is appreciated. The work of the second author was in part supported by NIST Cooperative Agreement No. 70NANB5H0050.

References

[Aboud, et al., 1993] Aboud, M., Chrisment, C., Razouk, R. and Sedes, F. (1993) Querying a hypertext information retrieval system by the use of classification. *Information Processing Management*, v. 29, 387-396.

[Arents & Bogaerts, 1993] Arents, H.C. and Bogaerts, W.F.L. (1993) Concept-based retrieval of hypermedia information -- from term indexing to semantic hyperindexing. } *Information Processing Management*, v. 29, 387-396.

[Callan, et al., 1992] Callan, J., Croft, W.B. and Harding, S. (1992) The INQUERY retrieval system. In: *DEXA-3: Third International Conference on Database and Expert Systems Applications*

[Card, et al., 1991] Card, S., Robertson, G. and Mackinlay, J. (1991) The information visualizer, an information workspace}. In: *CHI '91: Proceedings of CHI 91 Human Factors in Computer Systems..* New York: ACM.

[Harman, 1996] Harman, D. (1996) *TREC-4, Proceedings of the fourth Text REtrieval Conference*. Washington, DC: GPO.

[Hearst, 1995] Hearst, M.A. (1995) TileBars: Visualization of Term Distribution Information in Full Text Information Access} In: *Proceedings of CHI 95*. New York, ACM.

[Hemmje, et al., 1994] Hemmje, M., Kunkel, C. & Willett, A. (1994) LyberWorld -- A visualization user interface supporting full text retrieval, In: *SIGIR '94, Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*. London: Springer Verlag, 249-259.

[Korfhage, 1991] Korfhage, R (1991) To see, or not to see -- Is that the query? In: *SIGIR '91: Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 134-141.

[Mukherjea, et al., 1995] Mukherjea, S., Foley, J. and Hudson, S. (1995) Visualizing Complex Hypermedia Networks through Multiple Hierarchical Views. In *CHI 95. Proceedings of the Conference on Human Factors in Computing Systems*. New York: ACM.

[Spoerri, 1994] Spoerri, A. (1994) InfoCrystal: A visual tool for information retrieval and management.. In: *CHI '94: Human Factors in Computing Systems Conference Companion*. New York: ACM, 11-12.

[Veerasamy, 1996] Veerasamy, A. (1996) Interactive TREC-4 at Georgia Tech. In: Harman, D., ed. *The Fourth Text REtrieval Conference (TREC-4)*. Washington, DC: GPO, in press.

[Veerasamy & Navathe, 1995] Veerasamy, A. & Navathe, S. (1995) Querying, Navigating and Visualizing a Digital Library Catalog. In: *Proceedings of the Second International Conference on the Theory and Practice of Digital Libraries*.

[Veerasamy, et al., 1995] Veerasamy, A., Navathe, S. and Hudson, S. (1995). Visual Interface for Textual Information Retrieval Systems. In: *Proceedings of the Third Conference on Visual Database Systems. IFIP 2.6*