Computer Science Department of The University of Auckland
CITR at Tamaki Campus (http://www.tcs.auckland.ac.nz)

CITR-TR-24 June 1998

Intelligent Interpretation of World Wide Web Queries

Jacky Baltes [1] and Bryce Ewing [1]

Abstract

The problem of finding information on the Internet or on intranets is becoming increasingly more difficult, because of its explosive growth and lack of structure. This paper describes IWWWQ, a system that improves the ranking of a search engine's results by incorporating other features of a page, including the page type, content of a page, page formatting, and the relationship between pages. An empirical evaluation shows that incorporating these features improves the ranking of the results returned by a popular search engine. Furthermore, the evaluation showed that a combination of features is necessary to obtain this improvement. Finally, it shows that the algorithms used are relative insensitive to specic parameter settings.

[1] CITR, Tamaki Campus, University Of Auckland, Auckland, New Zealand

Intelligent Interpretation of World Wide Web Queries

Jacky Baltes   Bryce Ewing
Computer Science Department
Tamaki Campus
University of Auckland
Postal Bag 92019
Phone: +64 (9) 373-7599 Ext. 8744
Fax: +64 (9) 308-2377
Email: j.baltes@auckland.ac.nz   bryce@123.co.nz

Keywords:information retrieval, learning, applications

Abstract

The problem of finding information on the Internet or on intranets is becoming increasingly more difficult, because of its explosive growth and lack of structure. This paper describes IWWWQ, a system that improves the ranking of a search engine's results by incorporating other features of a page, including the page type, content of a page, page formatting, and the relationship between pages. An empirical evaluation shows that incorporating these features improves the ranking of the results returned by a popular search engine. Furthermore, the evaluation showed that a combination of features is necessary to obtain this improvement. Finally, it shows that the algorithms used are relative insensitive to specific parameter settings.

1 Introduction

The explosive growth of the Internet has lead to an enormous amount of information that is now available online. Ideally, this information is only a few keystrokes away. The reality, however, is that the distributed, dynamic, disorganized structure of the Internet makes finding information a difficult and tedious task.

The Internet is a collection of hypertext documents called pages. These pages are connected via hyperlinks (so called *links*). A page is *correct* for a query if it contains information that is relevant to the query. A *hit* is a page that is classified by a search engine as a being relevant to a given query. Note that not all hits are necessarily correct.

There are three major approaches to help a user find information: manual, automatic, and enhanced methods.

1.1 Manual approaches

Manual or user created collections, such as indices (Yahoo! [Yah97]) or special topic pages (e.g., the Mining Co. [Min97]) are pages that are maintained by users or administrators, who are responsible for these pages. Often the authors are experts in their respective fields.

Manual methods are labor intensive, but return correct information. However, since the process of finding information is tedious, manual indices are often incomplete, that is, pages that are relevant to a query are not listed. Also, the highly dynamic nature of the Internet means that often pages returned by a query are out of date, e.g., when a page moved to a different location.

1.2 Automatic approaches (Search engines)

Most search engines on the Internet use fully automated methods. They use variations and enhancements of well known information retrieval techniques. The simplest search methods use boolean queries. A query consists of keywords that can be connected by boolean operators such as `and`, `or`, and `not`. Fuzzy boolean queries are extensions of boolean queries that rank the hits based on some metric, e.g., the number of matched terms. To make sure that all relevant information is returned, automatic query expansion adds terms that often occur together to the query (e.g., adding the search term `war` when the user entered `battle`). This means that it increases the number of returned pages.

Most popular search engines fall into this category, including Lycos [Lyc97], Altavista [Dig97], Excite [Exc97].

Current state of the art in search engines (e.g., Intelligent Concept Extraction from Excite [Exc96]) use a combination of statistical, machine learning, and database techniques to return as many documents as possible and rank them. Most search engines still use a form of keyword lookup. Furthermore, the search techniques are general ones, specific methods that are only appropriate for certain queries are not considered.

Automatic search engines use web crawlers, i.e., programs that automatically scan the Internet and collect information about the different pages. This means that the information in automatic search engines is more complete and more up to date than that of manual methods. For example, Lycos traverses the whole Internet, that is, it visits every page on the Internet, in about one month [Sea97].

These automated search engines provide more complete and more up to date results than the user generated ones. The problem is that even specific queries often return hundreds of documents, most of which are not relevant to the intended search target. To overcome this problem, a number of systems have been developed that use more sophisticated methods.

1.3 Enhanced search engines

There are two major problems with automatic search engines: the number of hits returned, and the incorrect ranking of these pages. For example, Altavista, a popular automatic search engine, returns 1438 documents for the following query `"John Anderson Homepage Manitoba"`[1]. This query is a very specific query to locate a friend of mine. More general queries may return hundreds of thousands of hits. Furthermore, John's homepage is ranked 31 in the list, which means that the user will only find it after visiting 30 irrelevant pages.

Therefore, a number of systems try to enhance automatic search engines by exploiting additional features of the Internet.

There has been a lot of interest in the design of so-called "knowbots," autonomous software agents that search the Internet for specific information. For example, Doorenbos et al describe a system to help a user do a comparison shopping on the Internet [DEW97]. However, these systems are specific to a given task.

WebWatcher uses the links between pages to suggest other interesting pages to a user [AFJM95].

Malmberg and Zhang describe a system that improves search engine results by preferring pages with bi-directional links between them [MZ97]. The idea is that two pages have similar topics if they both contain a link to the other page (or a link to the "neighborhood" of the other page).

[1] In this paper, queries use a conjunction of search terms unless otherwise specified

1.4 IWWWQ

IWWWQ is based on the assumption that current search engine technology returns all correct pages and also a large set of incorrect ones. The problem is that the correct pages are not always at the top of the list, but are hard to find among the other hits.

Therefore, IWWWQ uses the hits of an automatic search engine, but tries to improve the ranking of the returned pages by using additional features of the Internet.

Section 2 introduces the motivation behind IWWWQ and describes its design. In section 3, we will present the results of an evaluation of IWWWQ. Section 4 concludes and gives directions for future research.

2 Design of IWWWQ

This section describes the design of IWWWQ, a system that uses a combination of features to determine how relevant a given page is to a query. Instead of scanning the text of a page for keywords, IWWWQ's ranking of a page is based on the page type, the content, the formatting, and the relationship to other pages. The following subsections will describe each of these features in detail.

2.1 Page Type

Although the Internet is uncentralized and uncontrolled by nature, a number of different page types have emerged. *Homepages* are pages designed by people to provide personal information about themselves, such as contact information, resume, and links to topics of interest to the author. The links on these pages may point to pages that have different topics. For example, a user may be interested in C programming and New Zealand and his homepage may contain links to pages about each topic.

Link pages, sometimes referred to as jumpstations, are manually created index pages, that attempt to organize information about a specific topic. These pages contain little information, but contain lots of links to pages with similar topics.

Information pages are the major source of information on the Internet. They contain actual information about specific topics, for example a Java tutorial, the current exchange rates, or local news.

A *graphics page* contains mostly graphics, for example live video feeds from certain locations. They contain little accessible information, since graphics are difficult to interpret automatically, IWWWQ classifies these pages separately. Also, some pages are really interfaces to applications (e.g., online games or front end to a database), so called applet pages. Since these pages do not provide any information to aid in classifying them, they are ignored by IWWWQ.

The rapidly changing structure of the Internet means that often pages are moved to other locations or removed completely. A user can waste a lot of time trying to access these pages. IWWWQ automatically moves these pages (so called *invalid pages*) to the bottom of the list. The pages are not completely removed since a page may only be temporarily inaccessible.

IWWWQ currently supports the classification of information, link, graphics, and applet pages. Pages are classified into a category by computing (a) the number of words, (b) of links, (c) of graphics, and (d) of Java applets on a page. An information page is categorized by a minimum number of words (500) and a maximum links to words ratio (10%). A link page contains a minimum number of links (30) and a minimum links to word ratio (10%). Pages that do not contain the minimum amount of words or links are categorized into Java or graphics pages depending on whether they predominantly contain graphics or Java applets.

IWWWQ prefers information pages over link pages. Java and graphics pages are assigned a lesser ranking.

2.2 Content

The pages of the Internet are made up of a wealth of information. IWWWQ collects all features that are related to the data on a page in its content group. The content group includes:

- amount of data – the number of words per page

- creator of the page: the software used to create a page. For example, a page that was created with LaTeX2HTML is likely to contain more information about a topic (HTML version of an academic paper) than a page created with Microsoft's Homepage tool.

- date modified – the last time the page was modified. Pages that have been modified recently are likely to contain more up to date and therefore more useful information than pages that have not been maintained recently.

- web counters – a counter of the number of times the page is accessed. Pages that have been accessed often by other users are likely to contain useful information.

IWWWQ uses the amount of data in ranking pages. The idea is to penalize pages that contain too little or too much data. IWWWQ assigns a linear penalty function to pages with more than or less than the optimum number of words (500), thus preferring pages with about one page of text on them.

2.3 Formatting

Another powerful source of information is the formatting of a page. Pages on the WWW use HTML, a structured language derived from SGML [Int86]. Therefore, the formatting of a page is readily available. For example, the title, headings for the different sections of a page, and address (URL) of a page can be of value when trying to determine the relevance of a hit.

Pages with the search term in the title, headings or URL are more likely to be relevant to the query. Therefore finding how many search terms are in each of these areas can improve the ranking of the query results. IWWWQ computes the ratio of search terms in the title, the headings, and the URL to the total number of words in these three categories. Pages with a high ratio are preferred over ones with a low ratio.

2.4 Relationship between Pages

Apart from the pages themselves, the structure of the Internet and the relationship between pages can help in determining the relevance of a page to a query. Two features can be used in determining this relationship:

- hierarchy of pages in a site – pages about a given topic that are located at the same site are generally organized in a tree structure.
- links between pages – the links between pages can be used to determine pages that have similar content.

Firstly, using the address of a page, it is easy to determine pages that exist on the same site. Assuming that sites are well organized in themselves, then it is advantageous to go to the root of the subtree about a given topic. For example, given that a query returns two pages `P1` and `P2`, with URLs `www.site.com/area1` and `www.site.com/area1/subtopic1` respectively, then IWWWQ will prefer `P1` over `P2`. The goal of IWWWQ is to find the highest page in the hierarchy that is still relevant to the topic (still included in the query results).

Secondly, the WWW can be thought of as a directed graph:

$$G_{WWW} = \{< P, Q > | P \text{ has hyperlink to } Q, P, Q \in H\}$$

with the nodes (P, Q) coming from the set of HTML-pages (H) and the edges being the links between the pages. Nodes that have edges between them generally have some relationship (even if it is just that the creator of the page is interested in the other page). If there are a lot of links into a certain node then this page will be of some interest to a large group of people, and it can be thought of as more important than a page with few edges into it [JMFA95]. Since pages may cover more than one topic or may be of general interest, looking at the whole Internet may be misleading. For example, lots of pages contain direct links to popular sites such as Excite, or Microsoft's homepage. However, for most queries, the Microsoft site is of little interest.

Instead, IWWWQ only takes the subgraph created by the result pages into consideration. The link to hits rating is the ratio of the number of links to this page from other pages in the hit list over the maximum number of links to any of the pages in the hit list. An example is shown in figure 1. In (I), page P4 is the page with the most incoming links (4) and has a link to hit ratio of 1. P3 is the next one with three incoming links and a link to hit ratio of 0.75. In (II), taking a subset of these pages (the search results are shaded in grey), P4 is no longer present as it is not relevant to this query, P3 is no longer as important, and P6 is considered more important.

2.5 Combining Results

Once the ratings for the individual features (page type, content, formatting, and relationship) have been computed, they must be combined to compute a summary score for a given page. Currently, IWWWQ calculates a linear weighted sum of the individual features as shown in table 1.

First, the ratings for individual features are computed. Then a score for each feature group is computed by averaging the individual feature scores. This will result in a rating between 0 and 1. The ratings for each group of features are then weighted by a weight term and added to give the final score of a page. Unless otherwise specified, the tests run in the next section used a weighting of 1 for each feature group, which means that all features were considered to be equally important.

3 Evaluation

This section discusses the results of an empirical evaluation of IWWWQ. The evaluation attempts to establish three facts: (a) that IWWWQ improves the ranking of the results that are returned by an automatic search

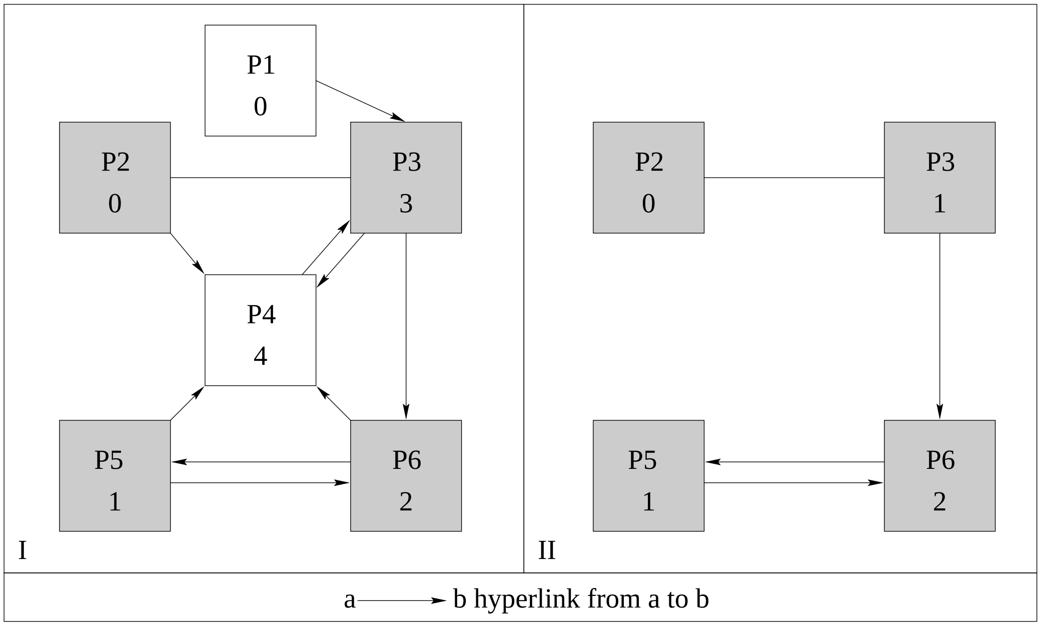

Figure 1: Example of link to hit ratio calculations

Weights	Group Rating	Feature Rating	Feature
W_f	R_f	R_t R_h R_{URL}	title rating heading rating URL rating
W_{ad}	R_{ad}	R_{ad}	amount of data rating
W_{pt}	R_{pt}	R_{pt}	page type rating
W_r	R_r	R_{cr} R_{lh}	common root rating links to hits rating

Table 1: Ratings of features and associated weights

Number	Query Terms	Abbrev.
1	genetic algorithms	ga
2	java programming resources	jpr
3	microsoft project price	mpp
4	new zealand auckland computer sale	nzacs
5	new zealand job employment	nzje
6	new zealand tourism	nzt

Table 2: Queries used in evaluation

engine, (b) that the individual features are useful in different queries and that a combination of these features is necessary, and (c) that the selection of the weights for the individual feature groups is robust, that is IWWWQ performs well over a large range of possible weight settings.

3.1 Performance

The evaluation compared the first 40 hits returned by Excite and reordered those queries with IWWWQ. The six different queries that have been used in this evaluation are shown in table 2. The selected queries range from general (`ga` and `jpr`) to very specific (`mpp` and `nzacs`).

Firstly, we attempted an automatic evaluation using lists of pages created by experts in their fields and comparing the number of these pages in the first 10 results of Excite and IWWWQ. Unfortunately this method did not work since none of the pages in the lists were returned within the first 40 pages for our six sample queries. This result in itself was surprising to us and points to the need for better information retrieval methods on the Internet.

Therefore, we had a number of users use the system and rank the first 40 pages into four groups: very good, good, medium, and poor. Users were asked to classify the pages based on the following guidelines. Very good pages contain relevant information. Good pages contain links to relevant information pages. Medium pages are somewhat relevant to the topic, and poor pages are not relevant to the topic at all. The results of the individual users were averaged.

The evaluation compared the performance of IWWWQ to Excite. The "goodness" of a query result is calculated by the ordering and the quality of the returned pages. Very good, good, medium, and poor pages were assigned a quality rating of 3,2,1, and 0 respectively.

Of primary importance was the ranking of those pages. In other words, the very good and good pages should occur at the top of the list. To find

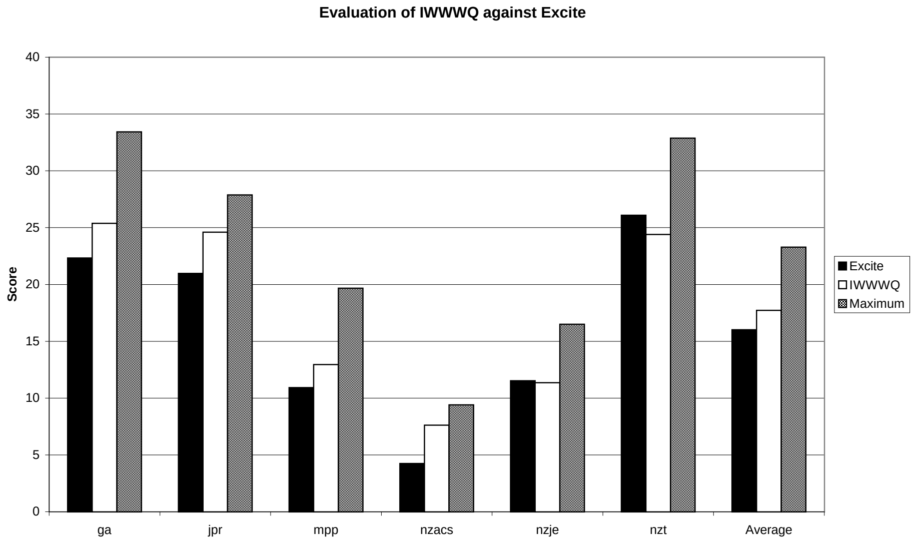

Figure 2: Results of IWWWQ compared to Excite and the maximum score

the score of a certain position in the list the quality score of the page (e.g., 3 for a very good page) is multiplied by the position of the page term (e.g., $i/40$ for page i out of a total of 40 pages). Given this position score, the score for a complete query result can be computed by summing the positions in the list giving the score for IWWWQ and Excite. This way, we can also compute a maximum score of a query, which is the score if all the pages are perfectly ordered (first all very good pages, then the good pages, and so on) which gives an indication of the quality of the search result.

Figure 2 shows a graph of the results of IWWWQ compared to the results of Excite and the maximum results for the perfect order of the search results. The average score shows that the use of IWWWQ has lead to an improvement in the ordering of the query results. Points to note are that with the exception of the `nzje` and `nzt` queries IWWWQ performed better than Excite. The increases do not seem to have any relationship with the overall maximum quality of the query, and there is still a lot of improvement required to achieve the maximum score.

3.2 Influence of the different features

To find the influence that each of the individual feature groups has over the final results the standard deviation of each group was calculated. These standard deviations are shown in the graph in figure 3. It can be seen that in half of the queries all of the features had some influence on the final weighting, the only feature that doesn't asper in all was the relationship group. This occurred since the WWW is such a large graph. There are few links between pages in a small set of pages (40 in our case), even if the pages are supposedly on the same topic. There are many reasons why a page may not have a link to a page, even if the two pages are closely related. For example, a company is not likely to maintain links to its competitors. The amount of data lead to the greatest standard deviation since there is such a large range of web page sizes.

The formatting had surprisingly little influence on the results. However, it can be seen that the importance of formatting is related to the overall quality of a query as shown in figure 2. For the queries with higher maximum scores (`ga`, `jpr`, `nzt`), the formatting is more important. This suggests that the formatting would have more of an effect on the final results if the results from Excite were of better quality.

3.3 Robustness of the weight assignment

To show that the performance of IWWWQ is not critically dependent on selecting the "right" parameter settings and feature weights, we tested the robustness of IWWWQ by varying the weights of individual features over a range from 0 to 32. All feature weights were initially assigned a value of 1, which means that they were considered to be equally important. This was also the parameter setting that was used in the evaluation described in the previous subsection. The results of this experiment are shown in figure 4. The Y-axis shows the percentage improvement of IWWWQ over Excite and the X-axis shows the selected feature weights.

On average, IWWWQ performed better than Excite. The percentage improvement ranged from about 6 to 14 percent. Although the weights for the individual features varied over a large range, the percentage improvement is relatively stable.

4 Conclusion

This paper shows that a variety of different features of a page can be used to improve the ranking of search engines.

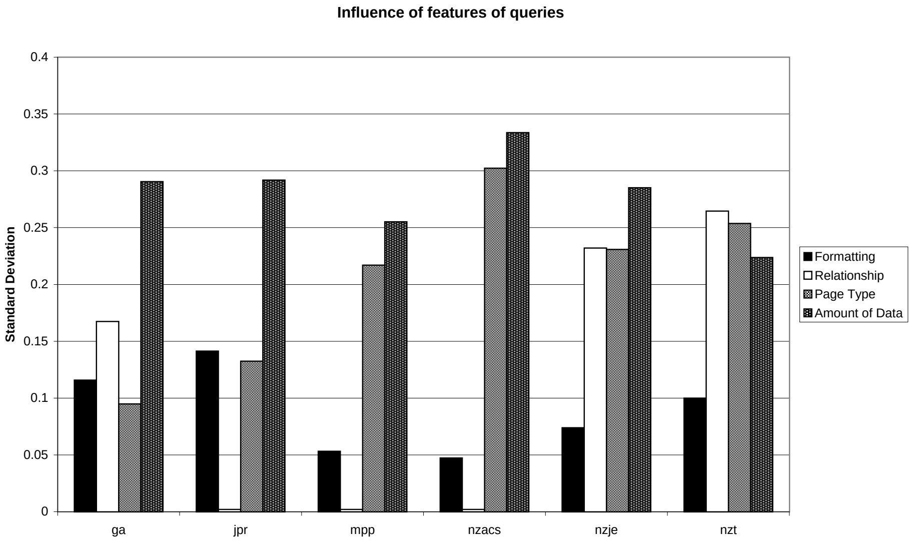

Figure 3: Influence of each feature set on the results

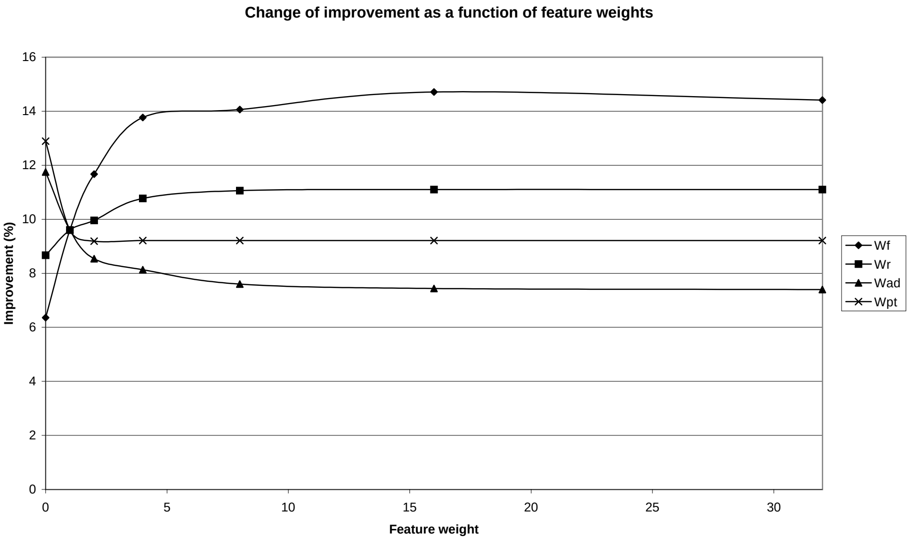

Figure 4: Robustness of IWWWQ weightings

However, more improvements are necessary to design a truly practical automatic search method. One interesting aspect of this research is that the amount of information that should be returned is dependent on the query itself. An often used metaphor used when discussing automatic searching is that of a library. What is a poor librarian to do when a patron comes up and simply says `"Travel"`. Current search engine technology tries to return all books with the word travel in them. However, an *intelligent* librarian would realize that the query is very vague. In this case, he could show a listing of the catalog with books that have `"Travel"` in the title. On the other hand, if a customer asks for `"Price of a Roadster XLE car in Auckland"`, the intelligent librarian should realize that this specific query is best satisfied by specific information (e.g., a list of car dealers and prices). The examples in the evaluation section have shown that a static set of feature weights (e.g., link to information ratio) is not sufficient, and we are currently investigating methods for adapting the weights based on the specificity of the query. The specificity of the query is determined by examining the number and type of hits.

Another direction for future research is to improve the format in which the search results are being displayed. One reason for a large number of unrelated hits is that a query may be ambiguous. For example, if a user searches for `"Java"`, there are at least three different concepts that the user may have been interested in: the programming language, the coffee, or the island. The representation of these results in a list is too inflexible. Current work addresses this problem by organizing hits in a tree structure. We intend to use clustering techniques from machine learning to separate the hits into different clusters corresponding to different concepts.

References

[AFJM95] Robert Armstrong, Dayne Freitag, Thorsten Joachims, and Tom Mitchell. Webwatcher: A learing apprentice for the world wide web. *ftp://ftp.cs.cmu.edu/afs/cs/project/teo-6/web-agent/www/webagent-plus.ps.Z*, 1995.

[DEW97] Robert B. Doorenbos, Oren Etzioni, and Daniel S. Weld. A scalable comparison shopping agent for the world-wide web. In *Proceedins Autonomous Agents 97*, 1997.

[Dig97] Digital Equipment Corporation. The altavista search engine. WWW Homepage, 1997. http://www.altavista.digital.com.

[Exc96] Excite Inc. Information retrieval technology and intelligent concept extraction searching. *http://www.excite.com/ice/tech.html*, 1996.

[Exc97] Excite Inc. The excite search engine. WWW Homepage, 1997. http://www.excite.com.

[Int86] International Organization for Standardization. Information processing - text and office systems - standard generalized markup language (sgml). Geneva/New York, 1986.

[JMFA95] Thorsten Joachims, Tom Mitchell, Dayne Freitag, and Robert Armstrong. Webwatcher: Machine learing and hypertext. *ftp://ftp.cs.cmu.edu/afs/cs/project/teo-6/web-agent/www/mltagung-e.ps.Z*, 1995.

[Lyc97] Lycos International. The lycoos search engine. WWW Homepage, 1997. http://www.lycoos.com.

[Min97] Mining Company. The mining company search engine. WWW Homepage, 1997. http://www.miningco.com.

[MZ97] Ake Malmberg and Tingting Zhang. Intelligent agents for information retrieval in internet. In M. H. Hamza, editor, *Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing*, pages 327–330, 1997.

[Sea97] Search Engine Watch. Search engine features, search engine comparison chart. *http://searchenginewatch.com/features.htm*, 1997.

[Yah97] Yahoo! Inc. The yahoo search engine. WWW Homepage, 1997. http://www.yahoo.com.