

Interactive Internet search: Keyword, directory and query reformulation mechanisms compared

Peter Bruza (bruza@dstc.edu.au)

Robert McArthur (mcarthur@dstc.edu.au)

Distributed Systems Technology Centre, University of Queensland 4072, Australia

Simon Dennis (s.dennis@humanfactors.uq.edu.au)

Human Factors Research Centre, University of Queensland 4072, Australia

Abstract

This article compares search effectiveness when using query-based Internet search (via the Google search engine), directory-based search (via Yahoo) and phrase-based query reformulation assisted search (via the Hyperindex browser) by means of a controlled, user-based experimental study. The focus was to evaluate aspects of the search process. Cognitive load was measured using a secondary digit-monitoring task to quantify the effort of the user in various search states; independent relevance judgements were employed to gauge the quality of the documents accessed during the search process. Time was monitored in various search states. Results indicated the directory-based search does not offer increased relevance over the query-based search (with or without query formulation assistance), and also takes longer. Query reformulation does significantly improve the relevance of the documents through which the user must trawl versus standard query-based internet search. However, the improvement in document relevance comes at the cost of increased search time and increased cognitive load.

Keywords: navigation versus ad hoc search, monitoring user behaviour to improve search, field/empirical studies of the information seeking process, testing methodology.

Introduction

The experiment reported in this article compares the search effectiveness of:

1. standard internet query search, as supported by the Google search engine (<http://www.google.com/>)
2. directory browsing, as supported by Yahoo (<http://www.yahoo.com/>)
3. phrase-based query reformulation as supported by the Hyperindex Browser (HiB; <http://www.purl.org/dstc/hib>)

One motivation of this study was to compare the effectiveness of straight keyword search against browsing a directory. Our assumption in this regard was directory-based search would lead to higher relevance of documents as the underlying collection has been vetted by editors. Also, by drilling down the directory to a leaf node containing a relevant document, we expected that other documents associated with the leaf node would have a high likelihood of relevance as well. Another motivation was to investigate whether query reformulation assistance does lead to more effective queries as compared to unassisted (i.e., straight) keyword search. Our underlying assumption here was query reformulation assistance would lead to more expressive queries (i.e., longer), which would translate into better document rankings.

Google was chosen as the mechanism to support standard internet query search because it seems to be one of the more effective newer generation search engines. In addition, its interface is less cluttered than the more established search engines.

Yahoo was chosen as it is the most established, and probably most comprehensive internet directory.

Short queries on the WWW are a well known phenomenon. For this reason a number of query formulation aids have appeared in conjunction with web-based search engines. The Hyperindex Browser (HiB) is a web-based tool used in conjunction with a search engine producing reformulations of a query in the form of linguistically well-formed phrases [3]. The user enters an initial query that is passed onto the associated search engine. The resultant document summaries are not presented directly to the user but are analyzed using a shallow natural language parsing technique. Phrases are derived from the documents in the result set and displayed to the user as candidate refinement possibilities for the initial query. For example, a query "surfing" would produce refinements like "Australian surfing", "tips on surfing", "surfing and surfboards" (see figure 1). The user can then select any of these refinements to become the new query, and

timing out on the network, but because subjects were given five minutes per query).

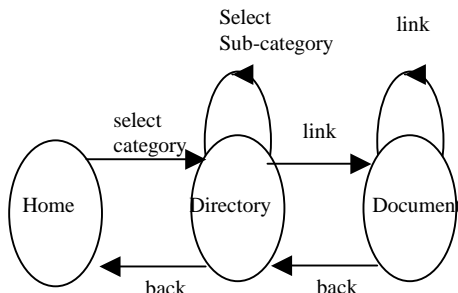


Figure 2: Search state transition diagram for Google.

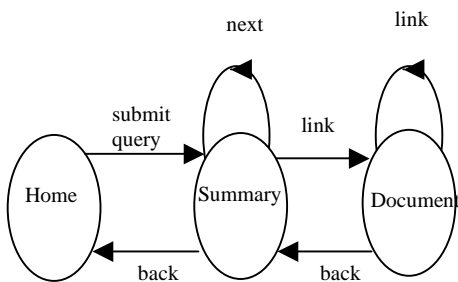


Figure 3: Search state transition diagram for Yahoo.

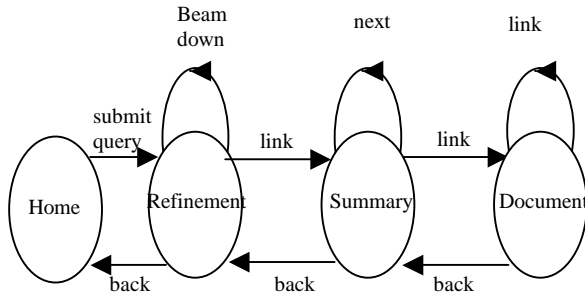


Figure 4: Search state transition diagram for HiB

Performance Criteria for Interactive Search

There are a number of dependent variables that can be used as measures of search effectiveness, for example, the traditional information retrieval measures such as recall and precision. However, these measures have a couple of important limitations that should be noted. Firstly, the document collection is the WWW where recall is impossible to measure. Moreover, most searches on the WWW are not concerned with finding

all of the relevant material. For this reason, we have chosen to measure the quality of the search results by having independent raters judge, on a seven point scale, the quality of the documents which were perused by subjects during their search. The mean of the ratings for all of the documents that the subject peruses during a search is termed “relevance rating”.

Secondly, when discussing interactive search it is important to record how long it takes a searcher to find the material for which they are looking. In the experiment subjects bookmarked what they considered to be relevant pages as they conducted the search, and we used the time to the first bookmark as an indicator of how long the engine was taking to present relevant pages.

Thirdly, and perhaps most importantly, it is crucial to measure the demands placed upon the user while interacting with the search mechanism – measures that capture the users’ experience during the search process. The first measure that we chose to collect was the amount of time that a user spends in each state of the search process – composing queries, assessing refinements, reading through document summaries and evaluating the documents themselves (see previous section).

In addition, however, we can look to the human factors literature for methods that more directly reflect the amount of cognitive load experienced by the user. We chose to employ a dual task methodology. As such measures are not common within the information retrieval literature we briefly discuss the measurement of cognitive load in the following section.

Cognitive Load and Dual Task Methodology

The dual task technique has a long history in the human factors literature (see [10] for a summary). The idea is to have subjects do two tasks simultaneously so that any excess cognitive resources that the subject does not use on the primary task (e.g., internet searching) are employed by the secondary task. The amount of effort they employ on the primary task is assumed to be inversely proportional to their performance on the secondary task.

A large variety of secondary tasks have been employed, from finger tapping [6], to random number generation [2]. After a number of trials, we chose a digit-monitoring task in which subjects listened to a stream of digits (from one to five) and responded when a digit was repeated. This task seemed to impose a considerable but not overwhelming load on the subjects and provided two simple measures of secondary task performance: reaction time (time to respond to a when a dual digit is heard) and the miss rate (how many times a dual digit is not responded to).

When employing dual task methodology, it is important to keep in mind its limitations. First, current theories of human resources contend that there are multiple pools and that tasks that draw on different pools will not necessarily impact upon each other. In this case, the secondary task performance may not be sensitive to differences in primary task demands.

Secondly, the secondary task should not precipitate a change in the strategy that subjects use on the primary task. For instance, if it were the case that the digit-monitoring task caused subjects to process only a couple of refinements, where they would have processed many more had they not been required to perform the secondary task, then we are no longer measuring the task of interest. Previous work has shown that subjects were less loaded when perusing query refinements than when perusing the document summaries generated by the Excite search engine [4].

Finally, the interpretation of high cognitive load can be problematic. In general, high load leads to increased fatigue, poor learning and an increased probability that users will fail to attend to relevant detail (i.e. a loss of situational awareness). However, if load is high because the user is engaged in task relevant activities, such as deciding between relevant refinements, then high load may result in superior performance as measured by relevance or time to first bookmark. As a consequence, it is important to view cognitive load in the light of relevance and search time.

Experiment

Subjects

Fifty four subjects were recruited from the undergraduate psychology pool at the University of Queensland and received credit for their participation. A pretest questionnaire was administered to gather demographic, computer usage and computer attitude information: 21 of the subject were male, 31 female and three did not respond. The mean age was 20.02 years with a range of 17-37 years. 47 of the subjects owned their own computer. Subjects were asked to rate on a five point scale how long they had used computers, how often they use computers and how often they use the internet. In addition, they were asked 13 questions on a seven point scale. The first 10 of these were combined to provide a measure of attitude towards computers. The last three were combined to provide a measure of computer anxiety. On all demographic and computer related measures one way analysis of variance was conducted to ensure that there were no significant differences between the subjects assigned to each of the search engines. The experiment was conducted between August and November 1999.

Design and Materials

A single factor design was used. Search engine was manipulated between subjects and could be Yahoo, Google or the Hyper Index Browser (HiB).

Eighteen queries were carefully generated to be a broad brushstroke of interesting internet queries (see table 1). These were divided into three sets of six queries and each subject saw one of the sets.

Table 1: Queries used in Experiment

1.1	Find pages listing jokes referring to Monica Lewinsky.
1.2	You are planning to move to Florida. Find pages listing jobs in the Florida area.
1.3	Find pages containing women's wave surfing competition results over the last two years.
1.4	Find pages about dyslexia.
1.5	Find pages that discuss clothing sweatshops.
1.6	Find pages that describe current or planned explorations or scientific investigations of Antarctica.
2.1	You own a personal computer that runs Windows '95. Find pages describing software that will test if it is Y2K compliant.
2.2	Find pages from which you can buy a pair of running shoes (online or at an address provided by the page).
2.3	Find pages that inform you which drugs are used to treat depression.
2.4	Find pages that discuss the disposal of long-lived radioactive wastes.
2.5	Find pages that discuss in vitro fertilization.
2.6	Are there any reliable or consistent predictors of mutual fund performance?
3.1	Find recipes for different varieties of carrot cake.
3.2	Find prices of Toshiba notebook computers.
3.3	You want to go skiing in Europe. Find pages describing a package holiday.
3.4	Find pages that discuss the concerns of the United States government regarding the export of encryption technology.
3.5	What makes Deep Blue capable of beating a human chess player?
3.6	Find pages that provide information regarding traveling in India.

Procedure

Before completing the experiment subjects were given two questionnaires to complete. The first collected demographic, computer attitude and computer anxiety information.

In the second, subjects answered a series of domain knowledge questions about the queries for which they were going to be searching. For instance, if they were going to be searching for pages of women's wave surfing (q1.3) then they were asked for the names of women surfers that they knew. Their responses were rated as either high or low domain knowledge by the experimenter. Our intention in asking these questions was to attempt to factor out variations between subjects in domain knowledge and to see if these variations might favour one engine over another. In particular, we hypothesized that the engines that provide some structure to the information domain such as Yahoo and the HiB might help subjects with little domain knowledge. Unfortunately, analysis of the results indicated no differences most probably as a consequence of the domain questions being an insufficiently precise measure and so we will not report any of these results.

During the experiment, subjects conducted searches using the PRISM browser developed at the Key Center for Human Factors and Applied Cognitive Psychology. The browser allows an experimental procedure to be administered and records the URLs that subjects visit, the amount of time they spend in each, whether they bookmark a page and the degree of cognitive load they are under as they conduct the search (as measured using a dual digit monitoring task). The software also classifies pages into different types using pattern matching on the URL of each page. In this experiment, pages were classified into Yahoo, Google or HiB home pages, Yahoo directory pages, Google Summary pages, HiB refinement pages, HiB summary pages and document pages (c.f. state transition diagrams).

At the start of the search phase, subjects were given written instructions on how to use the engine to which they had been assigned including the restrictions that we placed on each of the engines. In the case, of Yahoo they were told to use only the directory mechanism and not the query box.

They were also given instructions on completing the dual task. During each search a random series of digits between one and five were played into their headphones. So that they would not become accustomed to the rate of the digits and hence switch attention to the dual task in a rhythmic fashion rather than maintaining attention on the dual task, the digits were timed to have a mean inter-digit interval of 5 seconds with a uniform random variation around this mean of 1.5 seconds. Subjects were required to hit the escape key when a digit was repeated. To further ensure that subjects would have to maintain attention

on the dual task and that data was collected uniformly across all phases of a search, a double digit was forced every five iterations if one did not occur by chance. In pilot testing these values seemed to provide a balance between the collection of enough data to monitor cognitive load while ensuring that the subject continued to see the Internet search as their primary task.

After reading through the instructions, subjects conducted seven searches. The first was a practice trial to allow them to become familiar with the dual task procedure and none of the data from this trial is reported. Each question appeared on the screen. When they had read and understood the question they clicked on a continue button which took them to the home page of the engine to which they had been assigned. They then had five minutes to find as many pages relating to their query as they could. Any that they felt were relevant they were asked to bookmark. The question remained in a text box at the top of the browser throughout the experiment. Subjects completed the entire experiment within one hour.

Results

Relevance: Relevance judgements are made in the context of the quality of the documents that the subject is seeing and subjects will adjust their criteria appropriately (raising it if they are seeing many relevant documents and lowering it if they are seeing many irrelevant documents). Because search engine is a between subjects factor in this experiment, no direct comparisons of user generated relevance judgements (of which bookmarks are one type) can be made. For this reason, independent raters were employed and we have not analysed the number of pages bookmarked across engines.

Independent relevance judgements were compiled for 504 documents perused for Yahoo subjects, 794 for Google and 648 for HiB. (In reality, the figures for documents accessed were higher, but due to the dynamic nature of the WWW it was not possible to assess relevance for some URLs). Figure 6 shows the mean relevance ratings from the independent raters of the documents that were retrieved by subjects for each of the queries and as a grand mean. These results include all documents that the subjects saw, not only those that they bookmarked. Note that these analyses were conducted with the unique documents as the random factor. As a consequence, because different queries generate different numbers of documents they are weighted somewhat differentially in the analysis of the grand mean.

One way analysis of variance was applied to the grand mean and to each of the queries and the * or ** postceding each title (in figure 6) indicate the results that were significant at the 0.05 and 0.01 levels respectively. Table 2 shows the results of posthoc analyses of the differences between engines for those engines with a significant one way ANOVA.

Table 2: Means and Significance levels of relevance differences for the grand means and significant queries.

	Yah	Goo	HiB	HiB v Goo	Yah v Goo	HiB v Yah
Grand Mean	4.3	4.1	4.5	.002	.146	.174
Florida Jobs	4.7	3.8	5.1	.010	.114	.411
Sweat shops	3.6	5.7	4.4	.026	.001	.150
Run Shoes	5.4	3.9	4.5	.184	.001	.044
Depress Drugs	3.5	4.1	5.0	.027	.236	.004
Dispos Nuke Waste	3.5	4.2	5.2	.031	.236	.005
Toshiba Nbooks	4.9	3.7	3.2	.351	.020	.005

Time to First Bookmark: There were significant differences between the amounts of time it took for subjects to register their first bookmark as a function of engine, $F(2,51) = 23.59$, $p < 0.001$. Yahoo took the longest with a mean of 137 seconds. Next was the HiB with a mean to first bookmark of 117 seconds and the fastest engine was Google with a mean time of 75 seconds.

Time in State: Figure 5 shows the time spent in each state as a function of engine. Firstly, note that the amount of time spent in the home state differed by only two seconds across engines. Secondly, there was only a mean difference of six seconds between the time spent in Google summaries versus the time spent in the HiB summaries. The extra time spent in the HiB refinement state was being taken primarily from the time spent in documents of Google. Finally, note that the subjects spent the least time in documents when using the HiB, followed by Yahoo and then Google.

Dual Task: Dual task performance can be measured either in terms of the number of times the subjects fail to respond when a digit was repeated (the miss rate) or in terms of the amount of time it takes for them to respond when a digit is repeated. Previous work has found the miss rate to be a useful measure [4]. In this experiment, however, miss rate was not significant, $F(2,51)=0.30$. However, reaction time did show a significant difference between the Google summaries state ($m=1546ms$), and the HiB refinement state ($m=1815ms$) $F(1,34)=4.21$, $p = 0.05$, but not between the Google summary state and the HiB summary state ($m=1752$), $F(1, 34)=2.14$ or the Google summary state and the Yahoo directory state ($m=1766$), $F(1,18)=2.86$.

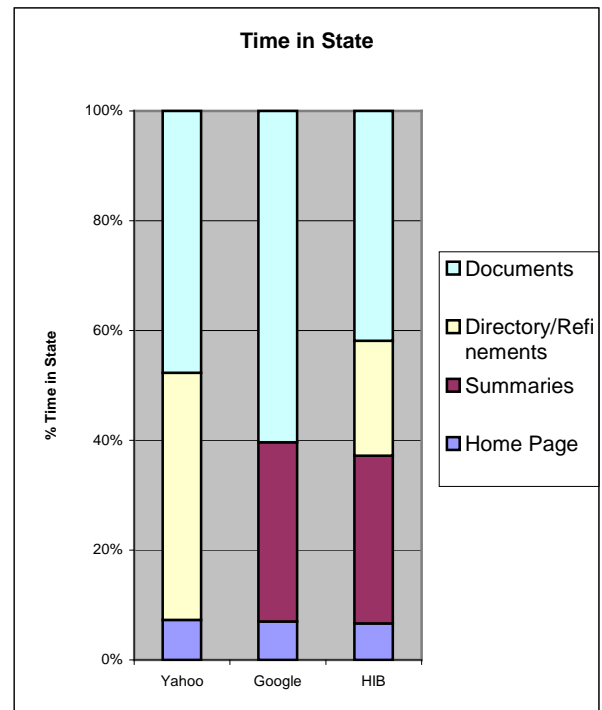


Figure 5: Time in State as a function of Engine.

Discussion and Conclusions

The time to first bookmark results are not surprising. The HiB involves one or more refinement states before the user beams down to the document summaries. A subject must traverse through the Yahoo directory from a root classification, to a leaf classification, before document summaries are seen. Proceeding through the refinement, or directory, states take time - time which the Google subjects will not experience, since a query directly results in a ranking of document summaries. This also explains why Google subjects perused more documents than either HiB or Yahoo subjects.

In terms of relevance, Table 2 shows that Google was superior to both HiB and Yahoo on the Sweatshops query. "Sweatshop" is an unambiguous term with fairly high discriminating power. Such terms feed well into statistical ranking techniques and this may explain Google's superiority. Contrast this with the "depression drugs" and "radioactive waste" queries. These queries involve more general terms which open up the problem of vocabulary mis-match. For example, in the query dealing with the disposal of radioactive waste, the HiB readily produces a refinement "radioactive waste management", which is a useful phrase for retrieving relevant documents. None of the subject's queries formulated for the Google search engine employed the term "management". Further support comes from the number of unique terms submitted to the retrieval engine through either Google or the HiB. For Google there were 166 while for the HiB there were 198. Important to note here is that the HiB query term vocabulary is in part derived from the titles of retrieved documents. In this sense, our study contributes to the evidence that phrases generated from titles or abstracts assist in the discovery of useful query terms [9].

More generally, this experiment shows that query reformulation support via the interactive use of phrases (HiB) does lead to higher relevance of documents perused than unassisted search (Google). The improved relevance does not stem from HiB refinements (average 2.95 terms) being longer than Google queries (average 2.97 terms). An interesting point to note is that the HiB subjects spent the least time perusing the documents, but those they did peruse tended to be those with higher relevance.

Yahoo performed best with the shopping-related queries "Toshiba Notebooks" and "Running Shoes". This is due to the fact that such queries have been optimised by the Yahoo directory editors. If these queries are omitted from the analysis, then HiB would be superior to Yahoo as well as Google with respect to average relevance. Interestingly, this study did not reveal that directory-based search improved relevance over standard query-based internet search.

These experiments demonstrate that cognitive load complements the use of relevance judgements for evaluating interactive search processes. Relevance by itself, or even with the inclusion of a simplistic measure of effort such as the time to first bookmark, would have missed how much effort a user must employ when using these systems. Even though the HiB produced higher average relevance than Yahoo and Google, it came at a "cost" via the higher cognitive load when perusing HiB refinements.

It has been reported that users tend not to reformulate [5,8]. HiB subjects spent about 20% of their time in the refinement state. We conclude that users will reformulate when there is satisfactory support for this, but note that they may only take advantage of the support in a limited way by refining, on average, once after the initial query.

One can ponder to what extent the experimental methodology chosen influenced the results. In particular, users were given a fixed time (five minutes) per query (the motivation for this was to provide a uniform basis for analysis). Typical internet search does not follow this pattern. In future work, we are planning to allow the subjects to terminate the search themselves.

In conclusion, directory-based search using Yahoo does not seem to offer increased relevance over query-based search (with or without query formulation assistance) using Google, and also takes longer. Query reformulation using the HiB can significantly improve the relevance of the documents through which the user must trawl versus standard query-based internet search. However, the improvement in document relevance comes at the cost of increased search time and increased cognitive load when perusing query refinements.

Acknowledgements

The work reported in this paper has been funded in part by the Australian Research Council, Key Center for Human Factors and Applied Cognitive Psychology at the University of Queensland, and part by the Cooperative Research Centers Program through the Department of the Prime Minister and Cabinet of Australia. We gratefully acknowledge the research assistance of Rohan Clarke, Bernd Irmer, Israel Keys, Naomi Norman and Sonia Twigg.

References

- [1] Anick, P.G. and Tipirneni, S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. In Proceedings of the 22nd Annual International ACM SIGIR Conference (SIGIR'99) pp 153-159, 1999.
- [2] Baddeley, A. The capacity for generating information by randomization. Quarterly Journal of Psychology, 18, 119-130, 1966.
- [3] Bruza, P.D. and Dennis, S. Query re-formulation on the Internet: Empirical Data and the Hyperindex Search Engine. In Proceedings of the RIAO97 Conference - Computer-Assisted Information

Searching on Internet, Centre de Hautes Etudes Internationales d'Informatique Documentaires, pp 488-499, 1997.

- [4] Dennis, S., McArthur, R, and Bruza, P.D. Searching the World Wide Web Made Easy? The Cognitive Load imposed by Query Refinement Mechanisms. In Proceedings of the Third Australian Document Computing Symposium (ADCS'98), Department of Computer Science, University of Sydney, TR-518, pp 65-71, 1998.
- [5] Jansen, B., Spink, A., Bateman, J. and Saracevic, T. Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum*, 32(1): 5-17, 1998.

[6] Michon, J. A. Tapping regularity as a measure of perceptual load. *Ergonomics*, 9, pp 401-412, 1966.

[7] Moldovan, D.I. and Mihalcea, R. Using WordNet and Lexical Operators to Improve Internet Searches. *IEEE Internet Computing* 4(1): 34-43, 2000.

[8] Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(3), 1999

[9] Spinks, A. Term relevance feedback and Query Expansion: Relation to Design. In Proceedings of the 17th Annual International ACM SIGIR Conference (SIGIR'94) pp 81-90, 1994.

[10] Wickens, C. D. *Engineering Psychology and Human Performance*, Harper Collins: NY, 1992.

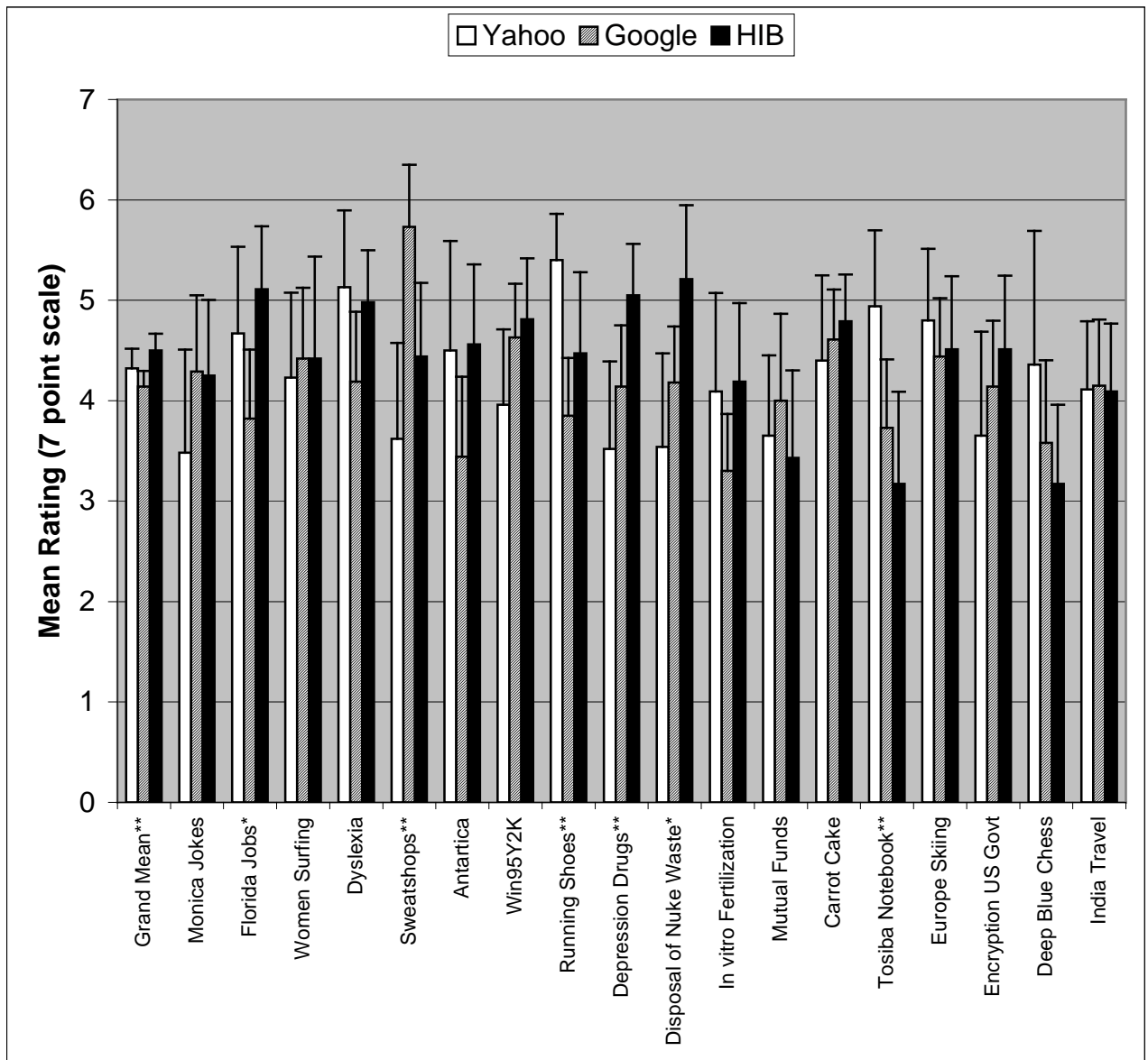


Figure 6: Relevance of the pages that subjects visited as a function of Query. Error bars represent the 95% confidence intervals. * = $p < 0.05$, ** = $p < 0.01$