Augmenting and Limiting Search Queries

Elaine G. Toms, Luanne Freund, and Cara Li
Faculty of Information Studies
University of Toronto
Toronto, Ontario M4W 3V8 Canada
{toms, freund}@fis.utoronto.ca; cara.li@utoronto..ca

Introduction

Web queries tend to be significantly shorter and less complex than queries used in earlier types of information systems (Jansen & Pooch, 2000; Lawrence & Giles, 1999; Spink et al, 2001). Yet, there is general belief that enriched queries and query reformulation will lead to improved results (Belkin et al, 2001). In our research we are examining the sorts of tools that could assist with the creation of enriched queries and in turn improve the search process and the user's search experience.

In the work reported here we assessed the use of two types of tools: one to assist the user in targeting and, thus, restricting the query, and a second one to assist in augmenting the query. We speculated that certain types of tools are more useful for certain types of information tasks. In particular we targeted the standard informational request in which a suitable response could be culled from many different Web pages, and secondly, the 'know-item' task, in which a specific Website exists to solve the problem. We anticipated that the tool to enable query augmentation would be more useful for informational tasks, as it would encourage amplification of the query to contain many more words and phrases that represent the information task. On the other hand, 'know-item' searches suffer when the exact title or some exact content is not known in advance and the limiter would enable more focussed searches.

Our hypotheses were:
1) Restricting a search to selected Internet domains would improve results for "know-item" information problems
2) Amplifying the query were additional keywords would improve results for information oriented information problems.
We tested these hypotheses in a within-subjects experiment using a novel information retrieval experimentation system prototype, called WiIRE which was run on the Web and required no researcher-participant communication.

Methods

Participants

Twenty-four participants (15 females and 9 males) took part in the study. About half were under 28, the majority of which were young – between 18 and 22; all but one was under 43. Most had either an undergraduate degree or were currently enrolled in one. All have used the web for two to four years. Forty-six per cent use a computer for 11 or more hours a week and use the web for an almost equal additional amount of time. About half use a computer daily for home, work or school purposes.

About 75% use search engines almost daily and, in general, claim to find what they are looking for. Daily, about 90% use search engines for work/ school related project and about 67% for entertainment. The search for information in other areas was very different: very few search for shopping (13%), travel (8%), health (21%), or government information (17%) on a daily basis, and about 25% never search in for those types of information. Because the data source used for this study was the United States government websites in the *.gov* domain, and the participants were Canadian, we also inquired about their previous search experience in this domain and their familiarity with United States government structure and agencies. About 25% searched monthly in the *.gov* domain, but no one searched more frequently than that. Approximately 70% have never (knowingly) searched in the *.gov* domain. Nearly two-thirds confessed to being unfamiliar with the US government structure and agencies. Thus our participant group was unfamiliar with the US government and infrequent searchers of its web pages, but they were well acquainted with searching the Web in general.

Search Engine
The Panoptic search engine(see http://www.panoptic.com) developed by Australia's National Research Agency (CSIRO) and the Australian National University was used in the study. Panoptic is a probabilistic search engine which accepts natural language query input as well as advanced query operators. The search engine was operating on a static collection of United States government web pages crawled in January 2002. The document collection contains 18 gigabytes of data from 1 million web pages. This particular set-up was created in an attempt to do a controlled experiment in a Web-like laboratory.

Tools
To test our concept of enhancing the query creation process, we developed two types of interface tools tailored to the data source used in this study.

a) ***Limiters: Agency Locator and Acronym Identifier***
Both serve to limit the search. We know from experience with government information that finding the right department and sub-department is often key to the right information. The intent of these tools was to enable the searcher to restrict the search to a particular agency.

! *Agency Locator* contains a select group of government agencies, departments and offices, selected from the Louisiana State University library's hierarchical directory of Federal government agencies (www.lib.lsu.edu/gov/tree). This list was presented in a collapsible tree format. Selecting one or more branches on the tree restricted the search to the web domain of that particular agency or group of agencies. The list was arranged by broad subject areas.

! *Acronym Identifier* contains a select list of common acronyms used in/by government agencies. Because the URLs for *.gov* sites contain agency, program, office, etc. acronyms, the contents of the URL provide valuable clues about an item on the results list. The acronym list could be used both to restrict the search to that particular group(s) and also to help interpret the URL, providing value-added information to the user. In in-house tests on the web in general, we have found that users interpret the content of URLs and use that information in making choices from a results list. In this case, we wanted to simplify that process. The tool worked as an alphabetical look-up list.

In each case, we used Panoptic's domain specific search feature to limit by specified domain(s).

b) Augmenter: Keyword Finder
The *Keyword Finder* was devised to suggest keywords, as well as give the user some notion of the kind of words that exist in the index and how they are represented in the database. We created this tool from Panoptic's 10-million word keyword index. Part of our problem was reducing Panoptic's index to something that was humanly digestible. Using a step by step process that used a combination of simple heuristics and WordNet, we reduced the list to around 20,000 words which were presented in an alphabetic list that also contained the keyword frequency information.

Interface
The interface to the Panoptic system was embedded in a testing environment (see last section in Methods). The interface was built using a three-frame design. The top frame was used only for experimental purposes. The bottom two frames contained and controlled access to the Panoptic Search Engine. The query input box was positioned on the left side while the query limit/augment tools were contained in the right frame. The tool in place depended on the experimental condition. When search results were returned, the search page on the left was refreshed to now contain both the query input box as well as results. Twenty hits were displayed per page. When a URL was selected from the results, a new 'website' window open to the right containing the website. Thus the website could be viewed in context with the results page, eliminating the need for constant backward movements. Each successively selected URL was presented in the website window.

Queries were processed with stop words filtered. Conventionally Panoptic presents results in tiers: the first tier contains all of the query words while documents which partially match are presented in subsequent tiers. If queries are very long, the keywords are processed in order of decreasing rarity and the most common words are ignored.

One of the following two options was present at the interface:
a) In the Limit condition, all selections from the Acronym Identifier and Agency Locator inserted the domain name for a domain restricted search to the Limit box in the Query frame. Subsequent additions were aggregated in this box. When the search was executed, the Limit box contents were appended to the query string and served to restrict the search to URLs containing the domain string.
b) In the Augment condition, all selections from the Keyword Locator were added to the query box. The resulting query string contained both manually entered terms and terms added through use of the Keyword tool. In this condition there was no Limit capability.

Under both conditions, we used a "bookbag" metaphor as a means for participants to indicate which pages from the results list were useful for the given task. Participants were asked to mark the checkbox beside relevant results and click on the "Add to bookbag" button. The URLs, the title of the webpage, and rank on the hitlist of each selection was stored in a database together with details of the participant and task.

Variables
The independent variables which were assessed using a series of metrics were:
1. Interface:
    a). Limit: the Acronymn Identifier and Agency Locator tools
    b). Augment: the Keyword Finder tool
2. Task Objective:
    a). Document: objective of the search is a 'known-item' – a Web page
    b). Information: objective of the search is a set of pages on which the solution will be found.

Information Problems
The eight information problems used in the study were devised by the Interactive Track, 11th Text Retrieval Conference specifically for use with the *.gov* data source (see Table 1). Four information problems were randomly assigned to each of the 24 participants so that, overall, twelve people addressed each problem. Problems were characterized according to type: a) whether they required Information or Documents to respond to the problem as defined above. Those of type Information needed certain facts or details, while those defined as type Document required a specific web page or pages to respond to the information problem.

#	Information Problems	Type
1	You are travelling from the Netherlands, and want to bring some typical food products as gifts for your friends in the United States. What are three kinds of food products from the Netherlands that you are not allowed to bring into the US?	*Information*
2	You are concerned with privacy issues related to electronic information and would like to know what laws have been passed by the US Congress regarding these issues. Identify three such laws.	*Document*
3	A friend has a private well which is the family's only source of drinking water. Locate a US publication, which contains guidelines for the maintenance of safe water standards for private well use.	*Document*
4	You are not sure about the safety of genetically engineered foods, and would like to find more information and research on this topic. Name four potential types of safety problems that have been raised.	*Information*
5	You are interested in learning more about what measures the US government has taken since 2001 to prevent Mad-Cow Disease. Identify three such measures.	*Information*
6	Name/find three research programs/projects that investigate the treatment/causes of	*Information*

#	Information Problems	Type
	dwarfism.	
7	You are planning a cycling expedition along the Silk Road in Central Asia. Find a website that is a good source information about health precautions should you take.	*Document*
8	You are planning to travel to the northeast territories of India and wonder if there are any problems/restrictions for tourists. Find a website that is a good source of information about such problems/restrictions.	*Document*

Table 1. Information Problems Used in the Study

Experimentation Environment – *WiIRE*

To manage this experiment we created a novel testing environment called WiIRE, the Web Interactive Information Retrieval Experimentation System. All experimental tasks and participant communications were contained within a web-based system. All data, from agreement-to-participate (the traditional consent form) to questionnaire, were captured into a database, or in server logs. For more details about this system, see Toms, Freund and Li (2002).

Procedure

Initially, each participant was given a prelude to the study followed by an introduction to the experimentation system, which started with the consent form. On agreement, this led to a demographics and search experience questionnaire and an overview of the study. The pattern beyond that point included an introduction to the new tool and practice time to use it. When participants indicated readiness to begin a task, they were assigned by the system a search topic. For each topic, participants completed a pre-task questionnaire, searched for information on that topic, and then completed set of post-task questions that contained both closed and open-ended questions to capture some of the data that would normally have been handled via interview. This process was repeated with the second interface and the last two search topics. At the end, participants were assigned a final summary questionnaire.

Data Analysis

Most of the data was captured to an MS-Access database or in server logs. Because participants came to our lab, we collected data using *WinWhatWhere*, a client-side transaction log file. The MS-Access database was converted to an SPSS data format. The URLs of all sites viewed was retrieved from the client-side transaction log files. The queries submitted by participants were retrieved from the server logs. In addition, the server logs and the client-side transaction logs were blended to retrieve time data and selected process data.

Results

The purpose of this study was to assess the effectiveness of two tools to aid query creation. To do so we examined several factors that affect the search process: 1) what the participants brought to the process – their knowledge and experiences that may affect outcomes; 2) effectiveness of the interface to supported the user in the query creation/reformulation processes and in the interpretation of results; 3) the performance of the system: how well it interpreted the query and provided a suitable set of results; 4) outcomes: how well the user and the system were able to accomplish the task, and how well the user perceived the process. The metrics were based on participant assessment, and on objective observation collected via logs, the outcome, and on independent assessment.

Pre-Task: Prior Knowledge and Experience of Participants

No participant had searched any of the information problems used in this study prior to this study. Participants indicated a lack of expertise in the topic area: 70% were unfamiliar with the topic and a further 21% were somewhat familiar, while 84% indicated having no or little expertise in the topic area at all. About 50% believed that there was information available for each information problem while 25% believe that was little information on the topic, and a similar proportion (50%) of participants believed that the information problems would be easy to solve. In summary, while participants had a lot of search experience, but no formal training (as described earlier under participants), and had no knowledge or expertise with the topics, they believed that there was information out there, but it might not be easy to find. Furthermore, there were no significant difference in perceptions according to task objective (F(1,4)=.934, p=.448) or to interface (F(1,4)=1.468, p=.219).

Analysis of the Queries

Participants created 395 queries, an average of 4.1 queries per task. The queries were evenly divided between the Limit and Augment conditions. But in the task objective condition, 54% were created for Information problems versus 46% for Documents. There was an interaction of task objective and interface. Fewer queries were created for Documents than for Information in the Keyword condition.

Each query was composed of, on average, 4.3 keywords, ranging from one keyword to fourteen. Of these Panoptic used 3.7 in the retrieval process. There was no interaction of task and interface (F(1,391)=.178, p=.673). There was also no main effect of task (F(1,391)=.95, p=.482), but there was one of interface (F(1,391)=6.935, p=.009). Participants in the Augment condition created significantly larger queries (4.6) than those in the Limit condition (4.0).

As previously described, each information problem was represented as a sentence or statement to participants. Typically one keyword per query did not come directly from the assigned information problem, but was devised by the participant. There was no interaction of task objective and interface concerning the number of words in those queries that did not also appear in the original information problem statement (F(1,391)=.144, p=.704), but there were main effects of both task and interface. Those in the Limit condition added significantly fewer words to queries than those in the Augment condition (F(1,391)=27.370, p<.0001). Similarly, those searching for Information added significantly more keywords than those searching for Documents (F(1,391)=15.260, p<.0001).

Twenty per cent were simple keyword queries, while 52% contained a phrase. A further 28% contained a statement or question. About 34% of all queries created for Information problems were simple keyword queries compared to 4% for Document tasks. Seventy-seven per cent of all Document queries contained a phrase whereas only 33% of the Information queries had phrases. In the interface condition, between 50% to 60% were phrase queries, with the remainder about evenly split between keyword and sentence formatted queries. This pattern was retained regardless of interface used.

The content of the queries was further analyzed both from a linguistic perspective and for the use of syntax. Eighty per cent contained no advanced syntax, e.g., use of Boolean or “+” and “-.” The proportion was the same in both task and interface conditions.

Over the course of a search, 75% of the second and subsequent queries for the same task were modified in some way: 18% were expanded in scope, while 9% were reduced in scope and 30% were completely revised. The remainder were re-entered. Twenty-two per cent of those in the Augment condition were expanded compared to 13% in the Limit condition. In the task objective condition, 35% of those seeking Information were totally revised compared with 25% in the Document conditions.

In addition, we examined how the tools were used within the task type. There were no significant differences by either the number of times the Limit tool was used or by the number of agencies appended to the query by type of task objective. On average the Limit tool was used about twice per information problem for a total of two agency URLs appended per task. The Augment tool was used on average once per query; a new keyword was added to a query about once for every two information problems. Of these keywords about the same number were new to the information problem. That is, the words did not already appear in the information problem statement.

Search Engine

Aboutness is a measure of how well the page fits the task. This is not a user-centred relevance judgement, but an objective assessment of the search engine's performance. Each web page examined by the participants was assessed by an independent expert for each information problem. Aboutness is partially a user selection, but more so, evidence of the a search engine's ability to deliver what the user has requested. Users can only select from what the search engine delivers, but then must make judicious choices from those items. The rules are listed below:

Rating	**Meaning**	**% pages**
5	pages directly related to the topic and containing clear info on the topic	14%
4	pages that provide some information that is related, or leads directly to the answer	17%
3	pages that about the topic but may be broader or narrower that the topic	26%
2	tangentially related but not really in the topic area	14%
1	pages that are clearly not about the topic at all	29%

The percentage in the third column contains the proportion of URLs examined that were assessed to be at that rating level. The average aboutness for all pages examined by participants was 2.7; 57% of the web pages examined were related or somewhat related to the topic being searched. The differences within each condition was insignificant at each rating level. In the interface condition, the percentage of pages at each level was about evenly split. This was not the case in the task objective condition when the trend was to a larger proportion of the pages rated a "4" or "5" in the Information condition. Thus pages retrieved in the Information condition tended to have a larger number of pages rated as on or relevant to the topic than those in the Document condition.

About 30% of the first URLs examined had an aboutness rating of "1." All second and subsequent URLs selected had a positive change (or no change) in the aboutness rating. In general, 27% of the URLs represented a higher aboutness rating than the one examined prior to it. There appears to be no difference by task objective or interface.

Of the 837 items examined, 400 were added to the bookbag. Presence in the bookbag signifies that participants considered the item to be useful in responding to the information problem. On average, participants inserted, per task, about 4.5 items into the bookbag while examining about 6.2 URLs. Forty-seven per cent of those added to the bookbag were rated a 4 or 5 on the scale presented above; 10% were rated a "1", or clearly not about the topic at all. Thus about half of the items in the bookbag were related to the information task.

The items contained in the bookbag also captured the rank of the item on the results list. The range of items viewed varied from a rank of 2 to 100 with the average being 61 (S.D. = 27.5). Twenty items were presented on each results page which means that participants examined up to the first five pages of results. The proportion that appeared on each page are listed below:

Page 1 (rank 1-20):	7.5%
Page 2 (rank 21-40):	18%
Page 3 (rank 41-60):	26.3%
Page 4 (rank 61-80):	18.2%
Page 5 (rank 81-100):	29%

There was no interaction of task objective and interface (F(1,80)=.650, p=.423) and no main effect of interface (F(1,80)=2.052, p=.156), but there was a main effect of task type (F(1,80)= 4.078, p=.047). Those in the Information condition had an average lower rank than those in the Document condition. Thus the search engine appears to be outputting better pages for for those classified as informational than for .know-item searches.

Outcome

Completeness of the Task

Completeness assessed the set of URLs examined by a participant for a task to answer the question: how completely could the information problem be resolved with the web pages collected in the bookbag? The following rules were used:

Rating	Rule
1	0% of the problem was answered/responded to
2	about ¼ of the problem was answered
3	about ½ of the problem was answered
4	about ¾ of the problem was answered
5	100% of the problem has been answered

Thirty-three per cent of the tasks were consider “complete” as verified by an external expert, while 42% were considered incomplete, that is the former were rated 5 on the scale while the latter were rated 1, and the other ratings were fairly evenly distributed between 7 and 9 per cents. Overall, about half a problem was complete. While there appears to be no differences by interface, more tasks in the Information condition (40%) were complete compared with those in the Document condition (25%) (Chi-Square= .044, df=4).

User perception: process

In general, about 52% of participants found the topics very easy to search, a proportion that is increased to 70% if the median “Somewhat” is also included. Approximately 42% claimed that they had more than enough time to do the search and a further 29% said that it was “Just right.” About 76% indicated that their previous knowledge did not help with the search (which was not surprising considering the initial indication of their knowledge and expertise in the topic areas). About 47% indicated that their search experience was satisfying, while a further 21% were “neutral.” For each of these measures there was no main effects or interaction of the factors.

User perception: outcomes

Participants were tri-modal about what was learned in course of doing the searches. About 1/3 learned very little and 1/3 learned a lot about the topic. In this case there was a borderline interaction of task objective and information (F(1,88)=3.723, p=.057). More participants in the Limit condition felt that they learned a lot while doing Document searches. The exact reverse was true for those in the Augment condition. About 51% were satisfied with their search results which is increased to 67% if the median “Somewhat” is included. Fifty-seven per cent indicated they would likely recommend their search results to a friend who was seeking the same information and a further 17% indicated they “maybe” would do so. About 61% were certain that they found answers to the search topics. For each of these measures there were no main effects of task objective and interface and no interaction.

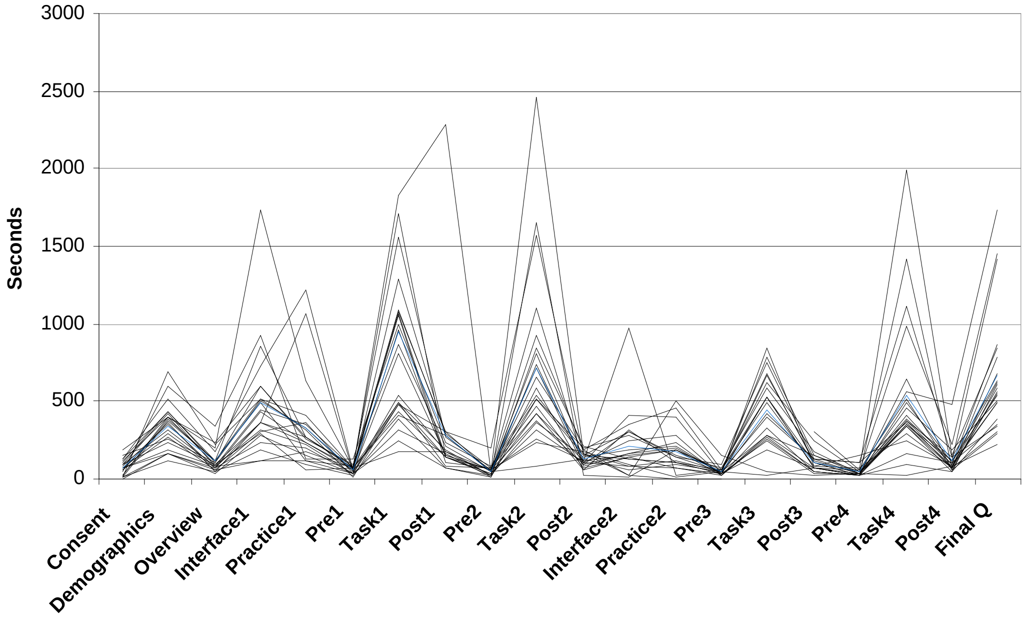
Analysis

This research assessed the effectiveness of two types of tools, one to restrict the search to Internet domain and one that enabled amplification. We anticipated that the Limit tool would be more useful for Document tasks, and the Augment tools for the Information task. Results are not clear cut and we cannot conclude that our hypotheses were supported. For many metrics there were no anticipated interaction effects from which we can conclude that one tool is more suited to a certain type of task. Surprising in the results was the high rank of relevant hits when compared with the typical Web search result in which users rarely go beyond the first couple of pages of hits. When combined with the aboutness and completeness rating, we are suspicious about both the ability of the search engine to find good pages. In addition, participants were unfamiliar with the topics to be searched and with US government information and their ratings and personal perception at the end illustrates those issues.

Participants in the Augment condition created significantly longer queries than those in the Limit condition, suggesting that the Keyword tool, either directly or indirectly affected the query. In data assessed to date, the use of the tool does not suggest that it affected ability to complete. The Agency finder was used by the most participants (21) and is the only tool that received an above average rating on a scale of one to seven. The keyword tool was the least transparent to users, and most of the comments indicated that participants either did not understand how to use it: "I don't understand the use of the keyword finder when you can just type the word by typing it in the box. (P21)", or found it too unwieldy to use: "It was very time consuming to scroll through the list… (P11)". From the comments, it seems likely that most participants did not understand how to apply this data to query formulation. Participants found the Agency tool "a great help in narrowing down the search area (P10)". Participants noted that it suited the government domain, but that it might not be helpful "in all subjects of interest (P14)". The low level of use of the Acronym tool seems to be related to its more specific application. The general thread of the comments was that it looked potentially useful, but the need for it did not arise: "I just didn’t need to use it in this instance, however, I think I'd use it right away if I didn't know the agency (P3) ". Thus, despite their overall performance participants understood the intent of the tools, except perhaps the Keyword Finder. So why did those tools not help? Prior to the start of the study, we were able to identify by limiting by agency more useful items in selected tasks, yet the participants were seemingly unable to do the same. Likely the problem is in its implementation.

WiIRE – the Web Information Retrieval Experimentation System prototype

WiIRE enabled the processing of participants in three groups, rather than the usual one-on-one, two-hour sessions; this resulted in considerable efficiencies in data collection. Form-based data was collected in a database while process-based data such as time and queries were acquired from web server logs. The other data critical to this study that could not be collected in the server logs was the URLs visited. For this we relied on a client-side transaction logs, a tool that would not be available to us with remote participants. However, we did incorporate a ‘bookbag' feature at the hitlist level, which stored items and their rank in a ‘bookbag.' This proved to be too cumbersome, as participants had to go back to the hitlist to add items to the bookbag, a task we plan to add at the page level in the next version.



The figure above illustrates the process taken by a participant in WiIRE. Noteworthy about this figure is the amount of time spent doing each activity in the experiment from reading the consent form to doing the final questionnaire. The bolded heavy black line is the average time for each activity. The most interesting factoid is the amount of time taken to do each task which decreased significantly over the course of the study. (In the figure, tasks are in order of completion and not by topic number.) This raises significant questions about the length of time that should be allotted for human experiments and the number of tasks that should be assigned before the process becomes onerous.

Conclusions

In this study, a modified search interface containing two types of tools designed to help with query creation were tested using the Panoptic search system. Overall results were inconclusive, as few participants could respond completely to a search topic and the pages examined were mostly not about the topic assigned, at least as determined by an external assessor. Part of the problem is likely due to the specialized nature of the database – the US government web pages, and to the lack of knowledge and experience of the participants with this type of data. Although the participants rated the experience slightly above average in their personal assessments, one has to question the effectiveness of the tools as well as the search engine. In addition a new web-based experimentation system was tested. In the lab it enabled bulk processing of participants without degradation in data quality. Noteworthy about its use was the significant drop in time taken to do a search task as the experiment progressed.

References

Belkin, N. et al. (2001) Rutgers' TREC 2001 Interactive TREC experience In: Text Retrieval Conference 2001, NIST, DARPA, ARDA, 452-461.

Jansen, B. J. & Pooch, U. (2000) A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology 52 (3), 235-246.

Lawrence, S. & Giles, C.L. Searching the web: general and scientific information access. IEEE Communications Magazine (January1999), 116-122.

Spink, A. Wolfram, D., Jansen, B.J., & Saracevic, T. Searching the web: the public and their queries. Journal of the American Society for Information Science, 52, 3(2001), 226-234.

Toms, E.G., Freund, L., & Li, C. (2003). WiIRE: a Web Interactive Information Retrieval Experimentation system prototype. (Submitted)