Abstract

Many experiments have shown that relevance feedback, whether in the form of interactive query expansion or automatic query expansion can enhance the relevance of documents retrieved by search engines. However, in real life scenarios users are reluctant to use suggested query expansion. A current look at seven different search engines shows that most offer term refinement; however, none are interactive, and none are placed in a way that draws attention to them.

1. Introduction

Extensive research has shown that relevance feedback in query modification helps to improve the quality of search results. However, most of this research was conducted in research settings, as opposed to real-life settings. Several researchers have looked at user's use of relevance feedback in real-life sessions (using the Excite and AltaVista search engines) and concluded that relevance feedback, while just as effective in these real-life scenarios, is rarely employed by users. The researchers concluded that search engines should modify their interfaces to enhance their relevance feedback features and encourage users to take advantage of them.

I researched seven internet search engines (About.com, Ask.com, AltaVista, Excite, Google, Lycos and Yahoo!) to determine their use and placement of relevance feedback features. I used two randomly selected queries, "Penguin" and "Inflatable Penguin" to compare search engine results as well as suggested term refinements. While all of the searches for "Penguin" returned results for Penguin Publishing, nearly all of the refinement suggestions concerned options for the animal. My search for "Inflatable Penguin" offered greatly varied results among the search engines for both retrieved web sites and query refinement options.

Although each website features some degree of relevance feedback—all of it automatic as opposed to interactive—there is little attention brought to the features. Using phrases such as "More Like This," "Related Searches," "Refine Your Results" and "Similar Pages," the search engines seem to offer these features to those who wish to use them, without really advocating their use. This may be due to the already cluttered nature

of the screen as well as users' reluctance to explore unfamiliar options because of time constraints and fear of encountering unwanted ads.

2. Background

Relevance is an essential aspect to information retrieval. In their examination of relevance, Schamber et al. (1990) begin their paper with the statement, "Since information science first began to coalesce into a distinct discipline in the forties and fifties, relevance has been identified as its fundamental and central concept" (755). Similarly, in their overview of the previous ten years of information retrieval, Robertson and Hancock-Beaulieu (1992) note that the 'relevance revolution' has had a major impact in thinking on the subject. They state, "There has been increasing acceptance that stated requests are not the same as information needs, and that consequently relevance should be judged in relation to the needs rather than stated requests" (458). What makes relevance especially difficult to assess is that a user's needs can change not only for every search but sometimes within a search itself. Schamber et al. conclude that the relevance judgments "refuse to 'hold still' for observation: the same item of information means different things to different individuals at the same time and different things to the same individual at different times…. Relevance, then, is a dynamic concept that depends on users' individual judgments to the quality of the relationship between information and information need at a certain point in time" (771).

2.1 Relevance Feedback

With this in mind, relevance feedback was introduced in the mid 1960s as an automated process for reformulating queries. Croft and Harper (1979) define relevance feedback as "the process of obtaining relevance information and using it in a further search" (339). Essentially, important terms that appear in retrieved documents can be added to the original query to enhance the search and push it towards a more refined result. As Magennis and van Rijsbergen (1997) note, "Queries can often be improved by adding extra terms that appear in relevant documents but which were not included in the original query" (324).

Researchers have experimented with various algorithms and vector space analyses to determine the most successful formula for calculating relevance feedback. While the details of their experiments vary depending upon the methods tested, one result remains consistent: an improvement over baseline searching is seen if relevance feedback is included.

For instance, Sparck Jones (1979) in an experiment designed to compare how much relevance information is needed to achieve performance improvement over "none at all" concluded that "very considerable improvements in performance can be obtained with relevance weights, even when these depend on very few relevant documents" (324). Indeed, the figures showed that "the average number of non-relevant documents retrieved is reduced with relevance weights. In real terms, the loss of some relevant documents is perhaps more than balanced by the huge reduction in non-relevant" (334).

Similarly, Salton and McGill (1983), in describing the SMART system note that, "it has been shown experimentally that the relevance feedback process can account for improvements in retrieval effectiveness of up to 50 percent in precision for high-recall (broad) searches, and of approximately 20 percent in precision for low-recall searches" (392).

Finally, Salton and Buckley (1990), in a review of basic feedback procedures noted that, "Collections that perform relatively poorly in an initial retrieval operation can be improved more significantly in a feedback search than collections that produce satisfactory output in the initial search" (362). Their research includes many formulas to calculate average precision and improvement when compared to no feedback at all. In the end, they concluded, "In view of the simplicity of the query modification operation, the relevance feedback process should be incorporated into operational text retrieval environments" (363).

2.2 User Preferences

One reason why using relevance feedback enhances a search is the observation that readers are unsure of the best way to formulate an initial query in a search engine. Beaulieu et al. (1997), noted, "It seems that most users are not aware of formulating their query in any particular way or able to articulate why they have typed in particular

terms…. The majority of users tended to start with a simple query and then react to what the system did" (45).

Historically, the most common comparison in experiments was the effectiveness of different search models. Different models are frequently compared to Boolean models to see which system is more effective at retrieving documents. A quick look at the research indicates that much of this research was done within the first two decades of the 'discovery' of information retrieval.

However, as Saracevic et al. noted in 1988, although systems and models were compared, little was known of the preferences of the users themselves. They write, "Users and their questions are fundamental to all kinds of information systems, and human decisions and human-system interactions are by far the most important variables in processes dealing with searching for and retrieval of information. Nevertheless, it is nothing but short of amazing how relatively little knowledge and understanding in a scientific sense we have about these factors" (175).

Since then, research has been conducted into the amount of interactivity users prefer, and whether or not such interactivity produces better results. Two methods in particular have been tested: interactive query expansion and automatic query expansion. In the interactive model, users have control over which terms they wish to add to their queries. While a computer assists in the retrieval of terms, it is the user who either chooses the terms or assigns weights to them. In the automatic model, the computer does the work behind the scenes producing what is typically referred to as 'magical' results.

Harman (1992) conducted an interactive experiment to find the usefulness of multiple iterations of feedback. In her experiment, "the user sees ten documents, selects the relevant ones, the system automatically reweighs the terms and adds 20 new terms, and then ten more documents are shown to the user" (8). In theory, this iteration process could be continued indefinitely, or at least until the same ten documents were narrowed down and repeated. The results indicated that multiple iterations of feedback are successful, and that "users should be encouraged to continue feedback until no more relevant documents are found, [and] to look through more documents, at least another screenful, even if no relevant documents are found, unless they have a clear idea of a better query" (9).

2.3 Automatic v Interactive Text Retrieval

It has generally been assumed that users prefer automatic text retrieval to the interactive models. Marchionini (1992) includes these assumptions about end users: "They want answers rather than pointers; they want document delivery rather than information retrieval… [and they] want to achieve their goals with a minimum of cognitive load and a maximum of enjoyment" (156). Indeed, most of the search models, including Boolean, Vector Processing and Probabilistic accord more power to the computer than the user.

However, the cognitive model is more interactive, assigning decision making to the user. Borlund (2000), in her study of interactive retrieval systems explains, "The main purpose of this type of evaluation is to determine how well the user, the retrieval mechanism and the database interact in extracting information, under real-life operational conditions. In this approach the relevance judgments have to be made by the original user in relation to his or her personal information need which can always change over session time" (74). One experiment in particular concluded that users prefer to have a degree of interactivity with the system when deciding on the most relevant results. Koenemann and Belkin (1996) conducted an experiment in which they tested relevance feedback through a transparent and an opaque system. Their aim was to determine "how a relevance feedback component impacts the information seeking behavior and effectiveness of novice searches in an interactive environment" (3). There were three interfaces in total: opaque, transparent and penetrable. With the opaque interface, relevance feedback was treated as a 'magical' tool. Searchers would mark which documents they felt were relative and then be given new results. The transparent interface was essentially the same, except that the terms which the computer added to the search were included with the results. The final interface was penetrable, it allowed the users to see the terms that the computer suggested, and to choose the ones that they wanted before the search was concluded.

In the end, users of the penetrable system did best overall, performing 15% better as a group than subjects in the opaque and transparent models. However, the users seemed reluctant to fully use the system. Koenemann and Belkin note, "Subjects in the

penetrable condition marked a comparable number of documents as being relevant but were quite selective in using suggested feedback terms…. Subjects entered fewer terms manually. Indeed, subjects commented that they preferred the 'lazy' approach of term selection over term generation" (9-10).

Belkin et al. (1999) conducted a study that compared two methods of term suggestion for user-controlled query expansion. They were 'user control' over suggested terms, implemented as positive relevance feedback and 'magical term' suggestion, as a form of Local Context Analysis. They conclude, "term suggestion without user guidance/control is the better of the two methods tested, for this task, since it required less effort for the same level of performance" (1). In addition to the flippant yet instructional answer, "users want magic," they offer another possible reason for this: "user control itself was not enough to overcome the effect of the other factors which might affect preference, in particular that of effort" (7). They concluded that "magical term suggestion is likely to be a better mode of support for query modification than user-controlled term suggestion; in that control of term suggestion is less important to users of IR systems than is ease of use of a term suggestion feature" (8).

Similarly, Anick (2003) notes that while interactive relevance feedback is seen to be effective, "it has been difficult to implement in practice because of the reluctance of users to make the prerequisite document relevance judgments" (88). Given users' reluctance to interact, search engines in real situations are typically less interactive and allow the computer to do most of the work. Anick notes that systems often generate "search suggestions from the top ranked documents regardless of their actual relevance, using linguistic and other heuristics to select and order the terms back to the user" (88).

2.4 Real-Life Searches (The Internet)

Within the confines of an experiment, users prefer to have control over the proceedings; however, this does not necessarily translate into real life situations. Marchionini notes, "Humans prefer heuristic processes to algorithmic processes because they are more interesting and because they reduce complexity to simpler judgmental operations (Tversky & Kahneman, 1974)" (157). Research has shown that, despite the conclusions of Koenemann and Belkin, an automated system can be, if not more

effective, then just as effective as a manually created search. In their study of users of the ENQUIRE Okapi project, Beaulieu et al. concluded that there should be a combination of automatic and interactive resources: "The complexity of the retrieval task is such that the system takes the lead, but the user must be given sufficient support to recognize when it is opportune to intervene" (73).

Despite the evidence of increased success and better results through relevance feedback, in real situations users are typically satisfied with their initial results and are disinclined to further refine their search. In fact, Sparck Jones and Willet (1997) have noted that many of the original studies of relevance feedback were created in, "a wholly impersonal and abstract form via the exploitation of relevance judgments already available with the test collections. This has been justified as simulating a desirable real situation where the end user is not required to think about query reformulation, but just to press a yes/no button and lie back and enjoy letting the system do all of the real work" (171).

Spink et al. (2000), noting the experimental research cited earlier, were surprised to discover that in real life situations (a study of the Excite search engine), users behaved differently than in the experiments. They state, "We found it surprising that relevance feedback was so seldom utilized" (323). According to their research, "Query modification was not a typical occurrence. This finding is contrary to experiences in searching other IR systems [such as Spink and Saracevic's (1997) experiments with DIALOG], where modification of queries is more common" (320). In fact, they concluded that when compared with traditional IR studies, "relevance feedback on the Web is used half as much as in traditional IR searches. More complicated IR techniques, such as Boolean operators and term weighting, are used more frequently by Web users" (323).

The researchers acknowledge that users in traditional studies have a 'training' period to familiarize themselves with the system before the experiment begins. They concluded that some users presumably were unfamiliar with Excite, which affected the relevance feedback findings. They noted that, "approximately one in three of the possible relevance feedback queries were judged not to be relevance feedback queries, but instead a blank or null first query… From observational evidence, some novice users

'click' on the search button prior to entering terms in the search box, possibly thinking that the button takes them to a screen for searching" (324).

2.5 Perceived User Success

In a survey of online research from 1982-1987, Walker (1988) found that there is generally perceived success in initial searches primarily because users are not very demanding of the system. In the Okapi series, "perhaps 80% of subject search sessions are short and successful, the searcher being satisfied by one or two apparently relevant items in the first dozen or so" (433).

Going into further detail of the Okapi system, Beaulieu et al. note that user satisfaction is simply not reliable because "users declare satisfaction even when systems perform poorly" (66). However, it is user satisfaction or dissatisfaction that directly affects whether the user will refine their search. Beaulieu et al. note, "Since end-users generally seem to be satisfied with so little, it may be that those who bother to or need to reformulate will always be in a minority" (67).

Walker concludes, "If these 'successful' sessions are the important ones, and if users' expectations do not increase, it is probably not worth trying to do much more in the way of improving online catalogues" (433).

Although they disagree with his suggestion that online catalogues should not be improved, the observance that users do not use feedback on the web is supported by Spink et al. They researched the Excite search engine web site. At the time, Excite was employing a feature called 'More Like This.' With this feature, the user would click on a link to regenerate the result using the relevance of the link chosen. (This feature appears to no longer be in use at Excite.)

The research showed that few users used feedback or even reformulated their queries. Their initial discovery was that, "most users did not use many queries per search, with a mean of 2.8 queries per search. Most users searched with one query only and did not follow with successive queries" (318). They also noted that "the mean number of terms per query was 1.98 for the relevance feedback population and 2.2 for the larger population. Assuming that lengthier queries are a sign of a more sophisticated user,

it appears that the relevance feedback population does not differ considerably from the larger population of Excite, and possibly, Web users" (325).

A similar study was done by Anick of the AltaVista search engine. Citing Beaulieu et al. concerning user reluctance to use relevance feedback, he suggests that "the additional task of judging feedback terms is itself a difficult one which users will avoid" (90). More telling, however, is this result of tests conducted at AltaVista: "many users did not even notice feedback terms when embedded within an already textually full results page. Those that did notice them often interpreted them as directories or advertising" (90). Indeed, the AltaVista feedback option was so under-utilized that Anick explained that the "word-cluster based refinement tool offered on the AltaVista web site several years ago received scant user attention and was eventually scrapped" (88).

The biggest drawback with real life search engine research is that the information is gathered anonymously, therefore, there is no definitive or even measurable explanation for why the users did what they did with regard to query modification. At best, the researchers are able to collate inferences from typical users in real-life settings. As Robertson and Hancock-Beaulieu note, "transaction logs are very useful but they are limited because they, provide information only about what users did, not what they thought" (464).

2.6 Reasons and Suggestions

Throughout the research, several possible reasons were generated for the low use of relevance feedback on the web. One possible reason is its low success rate. Spink et al. state, "Although it is successful 63 percent of the time, this implies a 37 percent failure rate or at least a not totally successful rate of 37 percent…. It points to the need for an extremely high success rate before Web users consider it beneficial" (326-327).

In a previous study of Excite, Jansen, Spink and Saracevic (1999) suggested that those who used relevance feedback features were simply more determined to get the results they wanted: "As for user characteristics of the relevance feedback population, they do not appear to differ in terms of sophistication from the other Web users, but they exhibit more doggedness in attempting to locate relevance information. This could be for several reasons. One may suspect that the subjects they are searching for are more

intellectually demanding. A cursory analysis of the query subject matter and terms does not support this conclusion" (8).

Given the conclusions and possible reasons for why people do not use relevance feedback in their searches, it has lead to a fairly consistent question: should interfaces of search engines change to enhance their feedback capacities? Spink et al. conclude, "Our findings emphasize the need to approach the design of Web IR systems, search engines, and even Web site design in a significantly different way than the design of IR systems as practiced to date" (327). Jansen et al. suggest, "At the very least it points to the need to tailor the interface to support these patterns if the goal is to increase the use of relevance feedback" (9). Finally, Anick concludes his study with these questions, "Would showing more/fewer feedback terms help or hurt uptake? Would the 'new search' option be more widely (and correctly) used if it were displayed more prominently? Should the underlying feedback ranking formula be altered to display relatively more container phrases?" (95).

All of these suggestions imply changes, whether cosmetic or structural, to current search engines. But how practical is that? Gerry McGovern, (2002) a noted consultant of web content management states that redesigning your web page can not only be costly it can be counterproductive, especially if it alienates existing users. He notes that the basic look and feel of Yahoo! has remained unchanged for seven years, and concludes, "The irony is that the re-design, demanding much effort and expense, can do real damage. The people who use your website most will be among your most valuable customers. Unless the original design was deeply flawed, they will likely hate any changes you make. A new website design means they have to re-learn how to get to parts of the website they regularly visit" (web site).

3. Personal Research

On March 30, 2004, I did an informal survey of seven search engines: About.com, Ask.com, AltaVista, Excite, Google, Lycos and Yahoo!. I wanted to see if any of these engines used relevance feedback, either interactive or automatic, when they display their results. I arbitrarily chose the word 'penguin' for my search. I also used the more

specific search ‘inflatable penguin’ to see if a more specific query would yield different suggestions. I will compare the results given and the suggested search refinements.

I chose random word queries as opposed to an actual information need because I wanted to compare the options that were given—both in terms of web site and query refinement suggestions—rather than the ‘usefulness’ of the of the term suggestions. As the research indicated that the mean queries per search was between 1.98 and 2.8, and that most users chose to use only one term, I chose to use one term for my initial search, with a refinement of two words. As the results were significantly different between the two searches, I did not do any further refinements. In the appendices, I have included all of the suggested feedback options, as well as the Sponsored results and the Top Five Non-Sponsored results for comparisons sake.

I also investigated the sites to see how their search engines work and how they explained their technology to the public. I sent an email to each company asking for information, but none have returned my questions. Although each site has an ‘About Us’ section, there is very little about the mechanics of the search engines available on the sites. This is how the sites explain themselves to the public [note that all quotes are [sic] from the site, including grammar and repetitions:]

About.com: The About network consists of hundreds of Guide sites neatly organized into 23 channels. The sites cover more than 50,000 subjects with over 1 million links to the best resources on the Net and the fastest-growing archive of high quality original content. Topics range from pregnancy to cars, palm pilots to painting, weight loss to video game strategies. No one has greater depth and breadth than About.

AltaVista: During the spring of 1995, scientists at Digital Equipment Corporation’s Research lab in Palo Alto, CA, devised a way to store every word of every HTML page on the Internet in a fast, searchable index. This led to AltaVista’s development of the first searchable, full-text database on the World Wide Web. Most advanced Internet search features and capabilities: multimedia search, translation & language recognition, and specialty search

Ask.com: Imagine a search engine that could read your mind, one that understands what you are thinking when you type in a query. What you’ve imagined is Natural Language Processing (NLP). When we chat with friends, we speak in casual, conversational tones. It should be the same thing when you’re looking for information online. With NLP, Jeeves is able to

understand the context of what you are asking, and he can thus to offer you answers and search suggestions in the same human terms in which we all communicate. You can see this technology in action in our related search terms and editorially selected answers. It's natural because it's what comes to us innately.

Teoma, which means 'expert' in Gaelic, is unlike any other search engine out there. Now, we could throw a lot of fancy terms at you, like *refinement* and *relevance* and *advanced* algorithms. And all of these describe what makes Teoma so powerful. But, what's really important for you to know is that Teoma adds a new dimension to your search results-authority. Instead of ranking results based upon the sites with the most links leading to them, Teoma analyzes the Web as it naturally occurs - in its subject-specific communities - to determine which sites are most relevant. Teoma is unique from any other search technology because it analyzes the Web as it actually exists - in subject-specific communities. This process begins by creating a comprehensive and high-quality index. Web crawling is an essential tool for this approach, and it ensures that we have the most up-to-date search results.

Google: PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search. Of course, important pages mean nothing to you if they don't match your query. So, Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines all aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query.

Excite, **Lycos** and **Yahoo!** offer no relevant information about their searching techniques.

3.2 Search Results:

About.com

Offers no suggestions for search refinement; however, it does offer topics within the 'About Network' such as: birding.about.com

PENGUIN: Top Non-Sponsored Search Result:

Penguin Group--http://www.penguin.com

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Giant Emperor Inflatable--http://www.penguin-place.com

AltaVista

Offers an option called 'More pages from this site' which links to other pages within the web site. Also offers suggestions under the heading 'Refine Your Search.'

PENGUIN: (Top Non-Sponsored Search Result:
Penguin Computing-- www.penguincomputing.com

Refine your search suggestions: Initially focused on Linux: 'Linux Penguin' and 'Linux Penguin Logo' but then moved towards nature 'New Zealand,' 'Antarctica' and 'French.'

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Bullet Holed Messenger: Inflatable Penguin--korgy.kokoyashi.net/arukaibu/001663.html

Initially offered similar results: "Linux Penguin" and New Zealand" but also contained "Suse Linux," "Baseball Cap" and the inexplicable "Geeko."

Ask.com

Offers feedback suggestions under the heading 'Related Searches.'

PENGUIN: Top Non-Sponsored Search Result:
Penguin UK--www.penguin.co.uk

Offers eleven suggestions for 'penguin.' Primarily, the suggestions concerned the animal, focusing on penguin books, penguin habitats or Emperor penguins.

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Inflatable Linux Penguin--www.suse.com

Related Searches focused primarily on toys: 'penguin toys,' 'inflatable shark,' 'blow up animals,' but also diverged into these seemingly unrelated topics: 'Elvis wigs,' 'furry handcuffs,' and 'Hen Night L Plates.'

Excite

Offers the option to 'Refine Your Results.' Excite has the most sophisticated refinement options with collapsible menus and the number of suggestions for a topic in parenthesis.

PENGUIN: Top Non-Sponsored Search Result:
Indexable Books and Authors by Title and Name--www.penguin.co.uk

Results are given with a heading and then subheadings:
Books (12) [New Zealand (2) Gifts, Posters (2) Free Stuff (2)
Other Topics (6)]
Linux (10) [Computing (2) Mascot (2) Other Topics (6)]
Emperor, Adelie (9) [Animal (4) Birds (2) Antarctic Penguins (2)]
Collection, Gifts (6)

Animal (8)
CCM Swiveling (5) [Motorola v60 (2) Motorola V120 (2)]
Cards (6) [Spaced Arcade (3)]

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Airblown Inflatable Whale--www.creatableinflatables.com
Toys, Doll (10) [Angles, Dinosaur (3) Ebay (3) Will liven (2)
Other Topics (2)]
Airblown (10) [Gemmy (5) Gift, Christmas (2) Other Topics (3)]
Linux (10) [Chair (4) Mascot, Portable (3) Other topics (3)]
Signs, YARD (6) [Your milestone (4) B Class (2)]
Inch, Animals (4)
Emperor, Tall (4)

Google

Offers no alternative suggestions, but does contain a 'Similar pages' feature next to each result. When clicked, it automatically regenerates a new list of pages.

PENGUIN: Top Non-Sponsored Search Result:
Penguin Group--www.penguin.com

INFLATABLE PENGUIN: Top Non-Sponsored Search Result
Linux Journal Store-- store.linuxjournal.com

Lycos

Offers a 'Narrow Your Search' feature. For both searches, Lycos offered the same refinement suggestions.

PENGUIN: Top Non-Sponsored Search Result:
Penguin Group--www.penguin.com

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Giant Emperor Inflatable--www.penguin-place.com
Suggestions: Penguin Books, Poke the Penguin, Penguin Info, Penguin Putnam, Penguin Publishing

Yahoo!

Offers an option called 'More pages from this site' which links to other pages within the web site. Also offers 'Related' terms, starting with a small number (one line of results), which can be expanded by 'More' to produce two lines and 'Show All' which concluded with 100 results for 'penguin.' No related terms were given for 'inflatable penguin'

PENGUIN: Top Non-Sponsored Search Result:
Penguin UK--www.penguin.co.uk

INFLATABLE PENGUIN: Top Non-Sponsored Search Result:
Bullet Holed Messenger: Inflatable Penguin--korgy.kokoyashi.net/arukaibu/001663.html
No alternatives for 'inflatable penguin' but some of the 100 options for 'penguin' included: 'penguin game, penguin books, penguin facts, Linux penguin, Pittsburgh penguin, penguin jokes, penguin mints, penguin gore,

> macaroni penguin, tux penguin, extreme penguin, jackass penguin and penguin munsingwear.

4. Conclusion

The most frequently returned result for 'penguin' was Penguin Publishing, either their U.S., their U.K. or their computing divisions. In fact, the appendix shows that Penguin Publishing is in the Top 5 of each list. Despite this, the most commonly suggested alternatives were related to the animal penguin. In fact, terms like 'Emperor penguin," "New Zealand," and "Antarctica" were in nearly all of the recommended searches.

These results indicate that the search engines realize that even the most 'likely' result may not be even remotely close to the intended result. Further, even though the suggested results were off the topic of the main retrieval results, the suggestions all seemed to be linked to a similar theme. In other words, not only are we given the suggestion to search further for the animal penguin, but we are given an enormous variety of refinements to add to the search to specify which aspect of the animal we want to research. Some examples are: 'penguin pictures,' 'penguin gifts' or 'penguin habitats.'

When I altered my search with the refinement 'inflatable,' it changed the results quite dramatically. The seven searches produced five different and unrelated results. Most of the results pointed to places where one could purchase inflatable penguins, either from the Linux store or from an 'inflatables' store, but others seemed to point to web discussions about inflatable penguins. The 'Bullet Holed Messenger' result includes this line, "Nothing says 'I was drunk in Australia' like a 4 foot inflatable penguin!"

Further, while some of the suggested refinements continued to focus on the animal 'penguin' ('Emperor penguin,' 'New Zealand'), typically the suggestions focused on toys, novelties and apparently unrelated links: 'airblown,' 'Patrick Rogan,' 'potatobiker,' and 'geeko.'

In accordance with the evidence that relevance feedback works for refinement of searches, all of these sites include some degree of suggested feedback. Even About.com, which offers no links to external sites, does contain refinements within their own network. However, much like Anick's conclusions with AltaVista, I felt that many of the

suggested terms were either not prominently displayed, or seemed like ads, rather than refinements. The high placement of sponsored results—which appear before the "Non-Sponsored" Results on every page that offers them—and advertisements—typically banner ads at the top of the page—may also distract from the user's desire to click on unknown features.

A personal example is the 'I'm Feeling Lucky' feature of Google. The feature allows you to skip the step of looking at search results and takes you immediately to the highest ranked result. I had never used this feature, primarily because until I learned otherwise, I was sure it was some kind of contest like the pop-up ads that redirect you to an unwanted site.

From my own searching experience, I know that I will typically not browse the search engine's own page because I have gone to the engine to find something, not to see what features they offer. Also, like Walker suggests, I am usually satisfied with the results on the first retrieval page. If I am not, I have always manually updated the search. Now that I am more aware of these search refinements I will be more inclined to use them.

Overall, I agree with the conclusions of Jansen, Spink and Saracevic that users of relevance feedback options tend to be more dogged and willing to spend more time on the search engines themselves. However, I disagree with their suggestion of radically altering interfaces to encourage the use of relevance feedback.

Perhaps the removal of, or at the very least, the more discrete placement of sponsored and advertised sites would encourage users to explore more. However, this is extremely unlikely given not only the prevalence of ads, but on one site (Lycos) an even more obtrusive pop-up ad. Most sites do offer a 'What is this?' descriptor of their refinement suggestions, but once again, they are not prominently displayed which necessitates even more curiosity and exploration when time is at a premium.

At this point, it seems that relevance feedback on search engines will be used only by those who are aware of the features and have had success with them in the past, or those dogged users who are willing to investigate alternatives presented. Given McGovern's recommendations against redesigning websites, the only way that search engines can draw more attention to their refinement features is to make them bigger or

more obvious. Yet on an already cluttered web page, that may be more of a hindrance than a help. Perhaps a non-profit search engine that is genuinely interested in getting the best results for its users would focus on drawing attention to its refinement features. However, as long as web search engines are profit minded, and as long as searchers seem satisfied with their initial results, there is no real incentive for change.

References

Anick, P. (2003), 'Using terminological feedback for web search refinement—a log –based study,' SIGIR '03, Toronto, Canada.

Beaulieu, M., Do, T., Payne, A., Jones, S. (1997) 'The ENQUIRE Okapi Project' British Library Research and Innovation Report 17. 1997.

Belkin, N.J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., Lobash, L., Park, S.Y., Savage-Knepshiel, P., Sikora, C. (1999), 'Relevance feedback versus local context analysis as term suggestion devices: Rutgers' TREC-8 Interactive Track Experience,' The Eighth Text Retrieval Conference, TREC-8

Borlund, P., (2000), 'Experimental components for the evaluation of interactive information retrieval systems,' *Journal of Documentation*, vol. 56, no. 1 January 2000. 71-90.

Croft, W.B. and Harper, D. J. (1979), 'Using probabilistic models of document retrieval without relevance information,' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 339-344.

Harman, D. (1992), 'Relevance feedback revisited,' 15th Annual International SIGIR 1992.

Jansen, B. J., Spink, A. & Saracevic, T. (1999), 'The use of relevance feedback on the web: Implications for web IR system design,' 1999 World Conference on the WWW and Internet, Honolulu, Hawaii.

Jansen, B. J., Spink, A., and Saracevic, T. (2000), 'Real life, real users, and real needs: A study and analysis of user queries on the web,' *Information Processing and Management*. *36*(2), 207-227.

Koenemann, J., and Belkin, N. (1996), 'A case for interaction: A study of interactive information retrieval behavior and effectiveness,' CHI 96 Electronic Proceedings (Accessed from http://www.acm.org/sigchi/chi96/proceedings/papers/Koenemann/jk1_txt.htm on 4/2/04).

Magennis, M. and van Rijsbergen, C.J. (1997), 'The potential and actual effectiveness of interactive query expansion,' SIGIR 1997, Philadelphia.

Marchionini, G. (1992), 'Interfaces for end-user information seeking,' *Journal for the American Society for Information Science*, 43 (2): 156-163.

McGovern, G. (2002), 'Think twice before re-designing your website,' http://www.gerrymcgovern.com/nt/2002/nt_2002_01_14_redesign.htm (January 14, 2002) (Accessed April 2, 2004).

Robertson, S.E., and Hancock-Beaulieu, M.M., (1992), 'On the evaluation of IR Systems,' *Information Processing and Management*, 28(4), 1992, 457-466.

Salton, G., and Buckley, C. (1990), 'Improving retrieval performance by relevance feedback,' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 355-364.

Salton, G., and McGill, M.J. (1983), 'The SMART and SIRE experimental retrieval systems,' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 381-399.

Saracevic, T., Kantor, P, Chamis, A.Y., Trivison, D, (1988), 'A study of information seeking and retrieving,' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 175-190.

Schamber, L., Eisenberg, M.B., and Nilan, M.S., (1990), 'a re-examination of relevance: toward a dynamic, situational definition,' *Information Processing and Management*, 26(6), 1990, 755-775.

Sparck Jones, K (1979) 'Search term relevance weighting given little relevance information' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 329-338.

Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997).

Spink, A., Jansen, B, Ozmultu, H.C., (2000), 'Use of query reformulation and relevance feedback by Excite users,' *Internet Research: Electronic Networking Applications and Policy*. 10, 4, 317-328. 2000.

Spink, A. and Saracevic, T. (1997), "Interaction in information retrieval: selection and effectiveness of search terms," *Journal of the American Society for Information Science*, Vol. 48 No. 8, pp. 728-40.

Tversky, A., and Kahneman, D., (1974), 'Judgment under uncertainty: Heuristics and biases,' *Science*, New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131. Stable URL: http://links.jstor.org/sici?sici=0036-8075%2819740927%293%3A185%3A4157%3C1124%3AJUUHAB%3E2.0.CO%3B2-M (Accessed April 5, 2004).

Walker, S. (1989), 'The Okapi online catalogue research projects,' in Sparck Jones, K. and Willet, E. (Eds.), *Readings in Information Retrieval*. (Morgan Kaufmann, New York 1997) 424-435.