Modeling Result-List Searching in the World Wide Web: The Role of Relevance Topologies and Trust Bias

Maeve O'Brien (maeve.m.obrien@ucd.ie)
Mark T. Keane (mark.keane@ucd.ie)

Adaptive Information Cluster, University College Dublin,
School of Computer Science and Informatics,
University College Dublin, Ireland.

Abstract

There are important cognitive issues surrounding the searching of lists of results returned to search engine queries that could significantly impact system and interface design. In this paper, we focus on result-list search examining two key issues: the influence of relevance topology of the list on first-click behavior, and the question of whether trust-bias occurs in such search. On both issues we advance some empirical and modeling results. These results are discussed in terms of their practical implications for Web designers and practitioners generally.

Keywords: search behavior; information navigation; predictive user modeling; empirical tests.

Introduction

The World Wide Web (WWW; Berners-Lee, T. Cailliau R. Groff J. & Pollermann B.,1992) has presented people with a whole new medium in which to search for information and, arguably, has transformed list-searching from a rather arcane, laboratory phenomenon into a ubiquitous cognitive act.

Currently, there are two predominant modes for locating information on the World Wide Web (WWW): browsing and searching. Browsing is the process of viewing pages one at a time and navigating between them sequentially using hyperlinks. Searching refers to the entering of a search query (usually a list of one or more keywords) to a search engine and the subsequent scanning and selection of links from the results returned. Researchers have developed a number of models of Internet browsing (see Brumby & Howes, 2004; Cox & Young, 2004; Miller & Remington, 2005; Pirolli, 2005; Pirolli & Card, 1999; Pirolli & Fu, 2003). Result-list search has received less attention in the Web context, though for decades it has been a mainstay of memory and attention experiments (see Eysenck & Keane, 2000). There are significant issues surrounding how people search result lists within the Web context. In this paper, we focus on two key issues: the significance of the relevance topology of the list, and the issue of trust bias in search engines. On both issues we advance new empirical and modeling results.

It has repeatedly been shown that people tend to favor items presented at the top of lists, an effect that has been replicated in the Web context (Joachims, Granka, Pang, Hembrooke, Gay, 2005; Keane, O'Brien & Smyth, in press). Keane et al. (in press), for example, showed, using a simulated Google (Brin & Page, 1998) interface, that when result-lists are systematically reversed in response to user queries, people tend to choose less-relevant results at the beginning of the list over highly-relevant results lower down the list.

This bias raises concerns surrounding the search-engines power to route traffic: highly-ranked pages typically benefit from a greater volume of traffic, and this heightened exposure obviously increases the volume of incoming links these top pages receive over time, which in turn increases their ranking prominence and the volume of traffic they receive etc. resulting in a rich-get-richer scenario (Baeza-Yates, Saint-Jean, Castillo, 2002; Cho & Adams, 2003; Cho, Roy, 2004).

A key issue surrounding this bias effect is whether it is specifically due to some level of trust in the particular search engine. For example, as people come to trust the fact that Google tend to deliver relevant links in the first three results, people may stop closely assessing results and just lazily click on these first links (c.f., Joachims et al., 2005). An obvious way to check this is to see whether the effects found by Keane et al. (in press) for the simulated Google interface, are also found when the same materials are presented as simple text lists.

In their study Keane et al (in press) found that whilst people generally tended to click on top results, there were instances where highly-relevant results lower down the list were clicked. This effect may be due to the different relevance distributions, or topologies, of result-lists. That is, a highly-relevant result proceeding many irrelevant results may stand a greater chance of being chosen over the same highly-relevant result proceeded by other relatively relevant results. In menu searching, Brumby & Howes (2004) have already shown that the extent to which people search through a list interacts with such relevance topologies. But, it is not clear whether such effects extend to search-engine result lists.

In this paper, we present an experiment and a model that investigate these two issues. In the experiment we systematically manipulated the relevance topologies of the presented lists, presenting them in either a Google interface or a text-list interface.

The Study

People were presented with a result-list searching task on one of two interfaces: a Google-simulated one or a text-based one (see figures 1&2). The results used to make up the result-lists were classified as being of either high, medium or low relevance. For each list materials sets were created where one highly relevant result was presented surrounded by either low- or medium-relevant results. The position of this highly-relevant result in the list was also varied (top or bottom). So, the design was 2 Interface (Google v Text) x 2 Relevance (Low-Relevant surrounding items v Medium-Relevant surrounding items) x 2 Position (high-relevant item at top of list versus high-relevant item at bottom of list). The dependent measure was based on the first-click made (i.e., the item first clicked on in the presented list).

Figure 1. Plain text interface

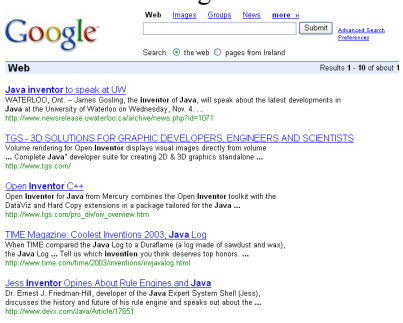

Figure 2. Google Interface

Method

Participants. Forty students from University College Dublin were paid to participate in the study.

Materials. Participants were required to answer sixteen Computer Science related questions in the experiment. For each question, the most commonly generated query (known from previous work by Keane et al., in press) was used to select a set of candidate links from the Google API. The criteria listed in Table 1 were used to classify results into three distinct relevance sets: high, moderate and low. The distinctiveness of these relevance sets to one another was verified in a rating study in which 10 raters were presented with a sample of 48 high, moderate and low results and asked to rate their "relevance to the questions posed" on a scale involving three options: "probably lead to target answer", "possible but unlikely to lead to target answer", and "unlikely to lead to target answer". An ANOVA analysis of these rankings revealed that the three groups were reliably different to one another, $F(2, 449) = 371.46$, $p < .05$. Post hoc Tukey tests showed that all pair-wise comparisons between the three groups were reliably different to one another, though the high-relevant items were markedly different to the others (see Figure 3).

Table 1: Criteria for Classifying Links

High-relevant The top Google result where: • the answer to question was in title/blurb accompanying the link url (e.g. "Java inventor James Gosling.."). • there was an exact match between the query terms and words contained in title/blurb (i.e., if the query terms were "Java inventor" then the text in the accompanying blurb/title should be "Java inventor" rather than "inventor of Java")
Moderate-relevant The top 9 Google results where: • the answer to the question was not contained in title/blurb, or link text. • all query terms were contained in title or blurb, but not as exactly matching phrases (i.e. "inventor of …the Java" as opposed to "Java inventor" to the query "Java inventor").
Low-relevant links • title/blurb contains some but not all query terms (i.e. broadly related to topic of e.g. either "Java" or "inventors") • the result is ordered >100 in Google's result-lists to the query • the answer to the question not contained within title/blurb, or actual link page.

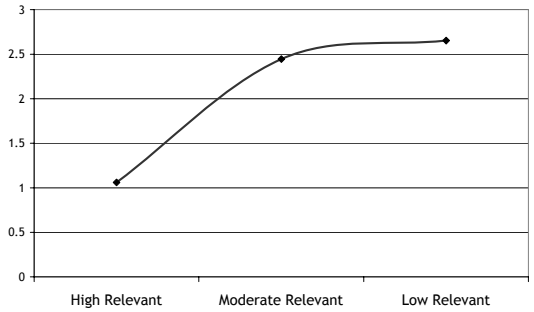

Figure 3. Average rating from 1 'probably lead to target answer' to 3 'unlikely to lead to target answer' of the high, moderate and low relevance results

Four distinct groups of result lists were created from these classified links to realize the material-conditions in the experiment: the top-moderate, top-low, bottom-moderate, bottom-low groups. The top-moderate materials were result-lists in which a high-relevant result is in first position followed by moderate-relevant results. The top-low materials were result-lists in which a high-relevant result is in first position followed by low-relevant results. The bottom-moderate materials had the high-relevant result in last position in the list preceded by moderate-relevant results. The bottom-low materials had the high-relevant

result in last position in the list preceded by low-relevant results.

Procedure: Participants were presented with the 16 randomly ordered questions on Computer Science (e.g., "Who invented Java?"). Next to each question was a link they clicked to reveal a result list. They were asked to find the answer to each question by clicking on the links within the presented result-list. Each result consisted of a clickable title, some snippets of the page's content (with highlighted matching content words), a web-address link, and other document components such as the document size, type, date and so on. The result-lists were presented as either a Google results page, or a minimally-formatted, text list of results (see figures 1&2). Each participant answered four questions using the top-moderate material set, four the top-low material set, four the bottom-moderate material set, and 4 bottom-low material set.

All search results clicked on (url, position etc), the timing of each transaction, question number, condition and student id were recorded. Participants were asked to complete a form detailing their answers to the questions and experiment sessions took between 1-1.5 hours.

Results & Discussion

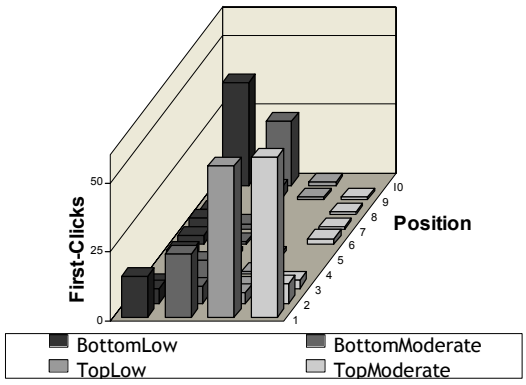

Figure 4. First-clicks in the Google conditions.

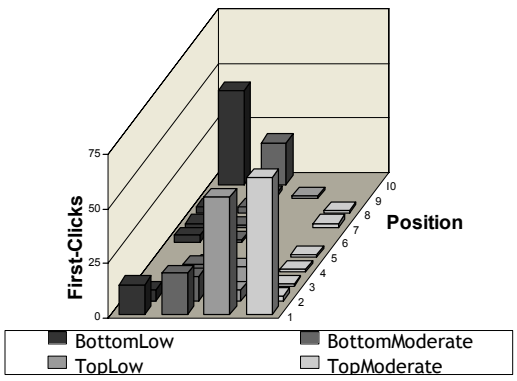

Figure 5. First-clicks in the Plain conditions.

The data was analyzed as a 2 x 2 x 2 ANOVA with Interface (Google vs. Text) as a between-subject factor and Relevance (moderate vs. low) and Position of the high-relevant, target answer (top vs. bottom) as within-subject factors. The dependent measure was the average position (1-10) of the first-clicks in the result list. First-click behavior was focused on as this presents a clean scenario for examining these factors (e.g., repeated query refinement in a progressive search would be a lot more complex).

The ANOVA analysis revealed a reliable main effect of Position, $F(1, 38) = 172.43$, $p < .05$ and Relevance, $F(1, 38) = 26.13$, $p < .05$. No reliable effect of Interface was found. There were no reliable interactions except for one between Position and Relevance, $F(1, 38) = 59.09$, $p < .05$. Exactly, what is going on in this experiment is best understood by breaking out the results in each Interface presentation. Figure 4 shows the results for the Google interface. In it we see that first-clicks in the top-moderate and top-low conditions are very similar, with most clicks occurring on the high-relevant result presented in first position. Clicks on other results further down the lists are rare.

The first-click behavior in the bottom-moderate and bottom-low conditions is different in two respects. First, we have a bi-modal type of response where some of the time people click on the first-placed result (even though it is not the target, high-relevant result), and more people pick on the high-relevant result placed last in the list. Second, this tendency to choose the high-relevant result at the bottom of the list over the less-relevant result in the first position, is much more pronounced when low-relevant results are present, than when moderate-relevant results are present (presumably, giving us the interaction between Position and Relevance).

Figure 5 for the Text interface basically shows the same pattern of results though there are some minor differences in the actual values found. Overall, the failure to find a main effect of Interface is readily observed in the similarity between Figures 4 and 5.

Three conclusions can be made from the evidence. First, position definitely matters: people are biased toward results presented towards the top of the lists. Overall, when a high-relevant link is presented first on the page it is clicked on 83% of the time, conversely when this same link appears as the tenth link it is clicked on only 43% of the time. Second, the relevance topology of the list matters: a highly-relevant result surrounded by low-relevant results will be picked out more readily than the same result surrounded by medium-relevant items. Overall, ignoring position, when surrounded by moderate-relevant results the high-relevant target is clicked on 56% of the time, but this rises to 69% in the context of low-relevant results. Furthermore, this effect can partially outweigh the effects of choosing first-placed items. Third, the bias found does not appear to be a search-engine specific "trust bias", as Joachims et al (2005) have suggested, but just something happens in lists of things, any lists.

To complicate this nice picture somewhat, we should say that there was slight evidence of learning in the experiment: the overall percentage of people that first-clicked on the tenth link when that link appeared as the last result was 43%, but if only the final quarter of trials are considered this percentage increases to 58%. This learning effect does not vitiate the results reported however, as it is an effect that should work against the conclusions made. To put it another way, if we only used the first 75% of trials the effects reported would be more pronounced.

The Model

In this section we advance a preliminary model of first-click behavior drawing on the evidence presented in the study, and the existing literature on information navigation and search.

Information Navigation

In attempting to understand why a user clicks on one link over another, research into web navigation has focused on the interaction between people's assessments of link descriptors and the navigation strategy adopted. Miller (2005) has classified these navigation strategies into those that assess links from either a 'threshold' or a 'comparison' approach. From the threshold perspective, a link is selected if its relevance is above an established threshold. Otherwise, users proceed to the next link for assessment (e.g. Miller, 2005). From the comparison perspective, a user may first assess several links and then select the link with the highest relevance (e.g. Blackmon, Polson, Kitajima & Lewis, 2002; Blackmon, Kitajima & Polson, 2003, 2005; Chi, Rosien, Supattanasiri, Williams, Royer, Chow, Robles, Dalal, Chen & Cousins, 2003; Pirolli & Card, 1999; Pirolli & Fu, 2003; Pirolli, 2005).

Models that approach information navigation from a comparison perspective are generally built upon the principles of Information Foraging Theory (Pirolli & Card, 1999). Within this theory the efficiency of information seeking as an exploratory, goal-directed activity is improved if the information system gives the users some 'scent,' or indication of the utility of taking a particular information path. Scent-based assessments inform decisions about which information items to pursue so as to maximize the 'information diet' of the forager. Generally it is assumed that users assess all of the scents (links) in a choice set prior to selection (e.g. Blackmon, Polson, Kitajima & Lewis, 2002; Blackmon, Kitajima & Polson (2003, 2005), Chi, et al, 2003, Pirolli & Card, 1999; Pirolli, 2005, Pirolli & Fu, 2003).

This theory obviously makes it difficult to explain the results observed in the present study, and using a similar experimental methodology to that presented here Brumby and Howes (2004) have recently presented eye-tracking evidence that challenges this account. Based on this evidence Brumby and Howes (2004) and Cox and Young (2004) have proposed models of information navigation based on Young's Model of Menu Exploration (Young, 1998). Within these models links are assessed as long as the expected information gain of making another link (re)assessment exceeds the cost of the assessment. The cost of the assessment is calculated by a simple utility function which is dependent (due to a normalization assumption) on the assessment of other links. Importantly, these models assume that only a single link in the menu choice set leads to the required information. Taken in this context, the utility of assessing a link can be evaluated by the expected information gain in reducing the degree of uncertainty as to which of the items in the menu choice set actually leads to the required information.

When people consider search-engine results they do not assume that only a single result in the choice set leads to the required information. Rather, they discriminate between links on their likelihood to reach the information most efficiently. Thus, whilst these models do appear to work rather well at predicting Internet navigation, the normalization assumption makes it hard for models such as Brumby and Howes (2004) and Cox and Young's (2004) to be adapted to search behavior.

The threshold approach to link appraisal, on the other hand, appears to be a logical starting point to modeling information search. However, Miller and Remington's (2005) threshold model is neutral with respect to the actual order in which links are evaluated. Here we have shown the position of links in a result-list plays a crucial role in the links people select.

Joachims et al. (2005) have found that users tend to evaluate results sequentially from top to bottom. Joachims et al (2005) used eye-tracking to investigate how users interact with Google results pages. They noted that fixation time is roughly equal for results 1 and 2 though users tend to click substantially more often on result 1. After the second result, fixation time drops off sharply. There is a further drop both in the fixation time and number of clicks after result 5, which they attribute to the fact that users need to scroll to view these results. The model described below takes into account this evidence, and adapts principles from threshold-based models of information navigation to suit information search.

Model Input

The mean and standard deviations from each of the group ratings in the materials section were used to automatically generate random samples of 1000 records across each of the four within-subject conditions. In this way, the model initialised with relevance estimates similar to those provided by human raters.

Model Procedure

Based on Joachims' et al (2005) eye-tracking evidence the model begins by evaluating the relevance of the first and second results on the list. The model does not elaborate on this process, but rather takes these relevance estimates as input parameters (see Kaur & Hornof, 2005 for possible means of modeling relevance estimation). If the first result is more relevant than the second, and it is above a reasonable (static) relevance threshold, this result is selected

immediately without further result evaluation. In this way the model can account for the sharp drop off in fixation time after result two. If the first result does not match these criteria the model proceeds to evaluate the next result. As with the preceding result this result is compared to both to a static relevance threshold and the following result. Results are processed in this manner until the model reaches a suitable result.

To account for the observed drop off in fixation and clicks after result 5 the model treats having to scroll as a decision point whereby a minority of users (7%), rather than processing the remainder of the results opt to either trust that the top result is probably quite good, or click on the link that was perceived to be the best among the 5 processed so far (essentially lower their initial threshold). Joachims, et al. (2005) have noted that users tended to re-formulate their queries rather than scroll passed result 5. This was not an option in the present experiment, however, and so this possibility is not considered here.

Model Results

The model provides quite a good fit to the experimental results (see Figure 7). As in the experiment, position and result relevance exert a combined influence, when the highly relevant result is placed at the top of the list it is selected regardless of whether the remaining nine results in the list are moderately relevant or not at all relevant. When the highly relevant result is placed at the bottom of the list the model is less likely to first-click on the first result (21%, compared to 25% in the experiment), rather people are inclined to first-click on the tenth result. As in the experiment this tendency is stronger when the preceding nine results were moderately relevant to the question asked (27%, compared to 29% in the experiment) compared ones low in relevance (79%, compared to 57% in the experiment).

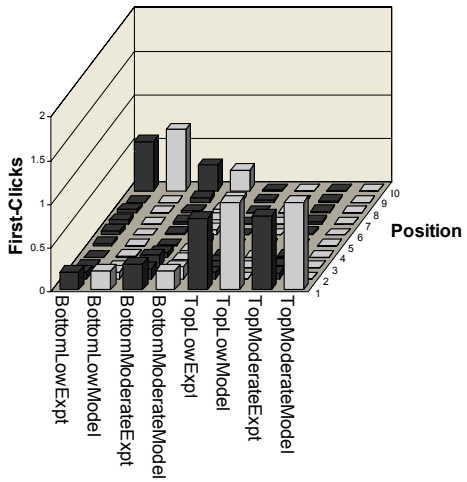

Figure 7. Comparison of experiment results (lighter grey) and model predictions (darker grey)

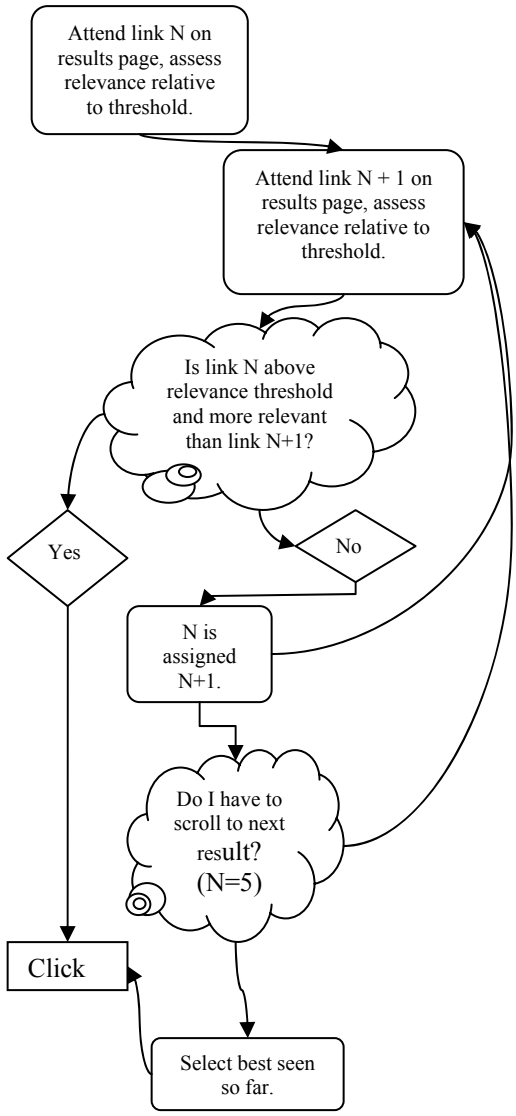

Figure 6. Model Procedure.

General Discussion

Search-engines are playing an increasingly important role in day to day life. Google alone is estimated to return results to an average of 200 million queries daily. Conceptually, Google models web surfers pursuing random walks over the entire WWW link structure. Surfer path information is viewed as an indicator of user interests, and this information is used to re-weight and re-rank the results of a text-based search (Brin & Page, 1998). Recent evidence suggests a need for a more informed approach than the random walk model implicit in Google (Baeza-Yates et al, 2002; Cho & Adams, 2003; Cho & Roy, 2004), and just as surfing information has improved text-based search results, the development of explicit cognitive models describing the strategies people employ in Internet search holds the potential to improve surf-based search results.

A detailed analysis of the knowledge-strategy issues involved in Internet search requires a combination of techniques: ecologically valid search-log analysis, knowledge abstraction from real users, usability studies, more lab-based studies such as the eye-tracking studies carried out by Brumby et al, 2004 and Joachims et al, 2005, and in this present paper, we have seen how a carefully-controlled lab study can be used to inform the development of a suitable cognitive model.

Acknowledgments

This research was supported by the Science Foundation Ireland under grant No. 03/IN.3/I361 to the second author.

References

Baeza-Yates, R. A., Saint-Jean, F., and Castillo, C. (2002). Web Structure, Dynamics and Page Quality. *Proceedings of the 9th international Symposium on String Processing and information Retrieval* (pp. 117-130). A. H. Laender and A. L. Oliveira, Eds. Lecture Notes In Computer Science, vol. 2476. Springer-Verlag, London.

Berners-Lee, T. Cailliau R. Groff J. & Pollermann B. (1992) World Wide Web: the information universe. *Electronic Networking: Research, Applications, and Policy.* 2(1).

Blackmon, M. H., Kitajima, M., and Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* CHI '05 (pp 31-40). ACM Press, New York, NY.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2003). Repairing usability problems identified by the cognitive walkthrough for the web. *Proceedings of the SIGCHI Conference on Human Factors in Computing CHI '03* (pp. 497-504). ACM Press, New York, NY.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves CHI '02 (pp. 463-470). ACM Press, New York, NY.

Brin, S. & Page L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the Seventh International World Wide Web Conference.*

Brumby, D. P. & Howes A. (2004) Good enough but I'll just check: web-page search as attentional refocusing. *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 46-51).

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., and Cousins, S. (2003). The bloodhound project: automating discovery of web usability issues using the InfoScentπ simulator. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '03 (pp. 505-512). ACM Press, New York, NY.

Cho, J., Roy, S., and Adams, R. E. (2005). Page quality: in search of an unbiased web ranking. Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data. SIGMOD '05. (pp. 551-562). ACM Press, New York, NY,

Cho, J. and Roy, S. (2004). Impact of search engines on page popularity. *Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04* (pp. 20-29). ACM Press, New York, NY.

Cox, A. L. & Young R. M. (2004) A rational model of the effect of information scent on the exploration of menus. *Proceedings of the Sixth International Conference on Cognitive Modeling.*

Eysenck, M. W. & Keane M. T. (2000) *Cognitive Psychology: A Student's Handbook*. Psychology Press UK.

Furnas, G. W. (1997). Effective view navigation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* (pp. 367-374) S. Pemberton, Ed. CHI '97. ACM Press, New York, NY.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. SIGIR '05 (pp. 154-161). ACM Press, New York, NY.

Kaur, I. and Hornof, A. J. 2005. A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. *CHI '05.* (pp. 51-60) ACM Press, New York, NY.

Keane, M. T. O'Brien M. & Smyth B. (in submission) Are people biased in their use of search-engines?

Miller, C. S. & Remington R. W. (2005) Modeling information navigation: implications for information architecture. *Human-Computer Interaction.* 19. 225–271.

Pirolli,P.(2005) Rational Analyses of Information Foraging on the Web. *Cognitive Science* 29. 343–373.

Pirolli, P. & Card S. K. (1999) Information Foraging. *Psychological Review.* 106, 643-675.

Pirolli, P., & Fu, W-T.F. (2003). SNIF-ACT: a model of information foraging on the world wide web. *Proceedings of the 9th International Conference on User Modeling*, Johnstown, PA.

Young, R. M. (1998) Rational analysis of exploratory choice. M.Oaksford, & N.Chater (Eds.), Rational models of cognition. Oxford University Press.