

The First Click is the Deepest: Assessing Information Scent Predictions for a Personalized Search Engine

Karen Church, Mark T. Keane & Barry Smyth

Adaptive Information Cluster, Department of Computer Science,
University College Dublin, Belfield, Dublin 4, Ireland
{karen.church, mark.keane, barry.smyth}@ucd.ie

Abstract. “First-click behavior” describes one of the most commonly occurring tasks on the Web, where a user submits a query to a search engine, examines a list of results and chooses a link to follow. Even though this task is carried out a billion times a day, our understanding of the factors influencing this behavior is poorly developed. In this paper, we empirically evaluate information scent predictions for first-click behavior in the use of a personalized search engine, called I-SPY. Our experiments show that the predictive accuracy of current information foraging approaches is not good. To conclude, we advance a framework designed to understand first-click behavior and guide future research.

1 Introduction

Almost every time someone opens a web browser they carry out a very simple behavioral sequence leading to their first click to some distant website. This sequence involves the user submitting a query to some search engine, scanning a list of returned results and choosing to click on a selected link. Though this, apparently simple, “first-click behavior” is incredibly commonplace, it is still not wholly clear how it should be modeled and what factors influence the behavior. The best predictive models we have of the behavior are information foraging and scent theories of web usage [5, 6, 7, 8, 25, 26]. Hence, in this paper, we report an empirical evaluation of information scent predictions based on an empirical study of web usage in a personalized search engine, called I-SPY [29, 30]. We find that these approaches do not make accurate predictions, prompting us to re-assess the cognitive basis of first-click behavior.

In the next section, we outline information scent approaches and the I-SPY system. Then, we sketch the empirical study of I-SPY. Next, we describe our empirical evaluation of information scent techniques and present the results found when these techniques are applied to the I-SPY data. In part response to the predictive failure found, we advance a general framework for assessing first-click behavior with a view to understanding what might be important in I-SPY. This same framework also helps us understand what it is that information scent approaches are trying to capture and how they may need to be modified to do better in the future.

2 Information Scent & I-SPY

Taxonomic studies have shown that users adopt several distinct types of behavior when using the Web, one of which is the specific goal-driven search for answers to specific questions [3, 23]. In the present paper, we are concerned with this type of directed search where a user has to answer a specific question by accessing a web page and does this by entering a query and selecting a link from a result list to meet that information need. The key contribution to be made by adaptive, personalized systems in this area is to increase the relevance ordering of result lists, so that the most relevant sites are in the initial positions of the result list. This research goal has become more acutely important with the emergence of the mobile Internet and the shrinking screen real estate available for presenting information.

Over the past few years, several tools and techniques have been developed for evaluating the usability of websites [2, 5, 6, 7, 8, 18, 26, 27]. In general, this work takes an information foraging approach, seeing human information seeking as being analogous to animals foraging for food [25]. This approach casts users as followers of information scents when web searching [2, 5, 6, 7, 8, 26]. The basic idea is that users will assess the distal content - namely, the page at the other end of the link - using proximal cues, the snippets of text or graphics that surround a link [5, 6, 7, 8]. By comparing these cues with their information goal, the user chooses the link that best meets their current goal, namely, the link with the highest information scent [2, 5, 6, 7, 8]. Two main flavors of information foraging have been advanced for usability assessment, the Cognitive Walkthrough for the Web [2] and the InfoScent Bloodhound Simulator [8]. To date, these approaches have been mainly used to provide (semi)-automated usability assessments of websites. However, they also make predictions about link-choices made by users in first-click behavior.

The Cognitive Walkthrough for the Web (CWW) is a theory-based inspection method used to evaluate how well a website supports users navigation and information search tasks [2]. CWW uses Latent Semantic Analysis (LSA) [19, 20] to calculate the information scent of a link given a specific user information need. In essence, it works by calculating the LSA-derived similarity between the users information goal (i.e., the query) and the text surrounding a given link. CWW has been shown to successfully predict uninformative/confusing links on analyzed websites [2].

The InfoScent Bloodhound Simulator [8] is an automated analysis system that examines the information cues on a website and produces a usability report. Bloodhound uses a predictive modeling algorithm called Web User Flow by Information Scent (WUFIS) that relies on information retrieval techniques (i.e., TF.IDF analyses) and spreading activation to simulate user actions based on their information needs [6, 8]. In this paper, for reasons of space, we concentrate on CWW rather than Bloodhound. It should be pointed out that CWW does better than Bloodhound.

We were interested in applying these approaches to a developed adaptive system, the personalized search engine I-SPY [29, 30]. I-SPY is an example of an adaptive information retrieval system. See, Micarelli & Sciarrone [22] and Pierrakos et al, [24] for other related work on adaptive information filtering and Web personalization.

I-SPY implements an adaptive collaborative search technique that enables it to selectively re-rank search results according to the learned preferences of a community of

users. Effectively I-SPY actively promotes results that have been previously favored by community members during related searches so that the most relevant results are top of the result list [29]. I-SPY monitors user selections or *hits* for a query and builds a model of query-page relevance based on the probability that a given page will be selected by the user when returned as a result to a specific query [1, 13, 29, 30].

I-SPY has previously been shown to be capable of generating superior result rankings, based on its collaborative model of page relevance [1, 13]. For instance, we know from the results of a live-user trial, designed to compare the search effectiveness of I-SPY users against a control group, that I-SPY's promoted results are likely to be valuable to searchers. This study provided us with access to comprehensive search logs reflecting detailed search behavior information including the queries submitted by control and test groups, the results returned and promoted, the results selected, and their positions when selected. If the information scent approaches accurately capture user behavior we should find that the results chosen by people are indeed those with the highest information scent.

3 Evaluating I-SPY

In this section we describe key aspects of the I-SPY evaluation. Although we do not evaluate the I-SPY system in this paper, we have included relevant details regarding the I-SPY evaluation to illustrate the environment and conditions in which we are attempting to evaluate information scent predictions for first-click behavior. For further details regarding the I-SPY evaluation see, [1,14, 15, 29].

The I-SPY evaluation took place over two separate sessions and involved asking two separate groups of 45 and 47 Computer Science students, to answer a series of 25 questions on topics in Computer Science and Artificial Intelligence. In the first session, the I-SPY collaborative search function was disabled so the results presented to participants were drawn from a Meta search engine (using Google, AllTheWeb, Wisenut and HotBot). The students taking part in the first session served as a control group against which to judge the students taking part in the second session, where I-SPY's collaborative search function was enabled. When each person used I-SPY, they entered one or more query terms and were presented with a list of up to 20 results; consisting of a result number/rank, a title and a description/summary (or blurb).

A substantial amount of web search behavior data was generated from the I-SPY experiment. A total of 811 distinct queries were logged with 10,445 unique pages being returned in result lists. From these lists, a total of 427 unique pages were clicked on by users. All of this information was collected and archived for analysis in assessing the information foraging approaches to which we now turn.

4 Does CWW's Information Scent Predict First-Click Behavior

Cognitive Walkthrough for the Web [2] is a usability inspection technique for assessing information search tasks that uses Latent Semantic Analysis (LSA) [19, 20]. CWW calculates the information scent of a given piece of link text by finding its similarity (as computed by LSA) to the query terms used (or more commonly an elaborated description of the query). LSA captures patterns of co-occurrence between

words based on a text corpus analysis, in its most commonly used form, of general reading up to first-year college level in the US. LSA has been shown to predict some aspects of language comprehension and priming in various cognitive tasks [10, 11, 17, 19, 21]. In our analysis of the I-SPY data we looked at the information scent values computed from LSA for long/short queries compared against each and every piece of link text returned in the results lists (see Church, Keane & Smyth [9], for details).

4.1 Method

Data & Analysis Procedure. The analysis was performed on the data common to both I-SPY sessions. Limiting our dataset to just queries and urls common to both sessions of the I-SPY experiment provided us with valuable before and after ranking information as well as key web search behavior details. This common dataset consisted of 132 valid distinct queries and 2,571 results/urls.

Queries and results were sub-classified. For queries, we had short queries (the actual terms input by users) and long queries (an elaboration of the original question posed to the user). For results, we distinguished between (i) the title of the link alone, (ii) the blurb text alone (i.e., the summary text given back by the search engine) and (iii) both title and blurb. All six possible pairings for every unique query and result (2,571 distinct test items) were submitted to the LSA website [20]. In all, it took roughly 6 days of computing time to gather these scores. The short query - title case revealed the best results of all six possible pairings. One of the problems with LSA is that it does not contain many of the specialist terms used in Computer Science (we would assume that this is a general problem that would attach to any specialist domain). Hence, we filtered out those queries-result pairs that had large numbers of unknown terms to LSA (reduced set to 97 queries).

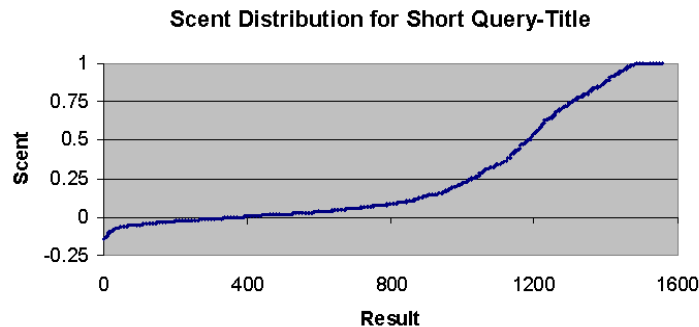


Fig. 1. Plot of the Scint Distribution for the Short Query-Title Pair in which the Results are Reordered by Scint

4.2 Results & Discussion

Overall, CWW does not do a good job of predicting the links clicked on by users in the I-SPY study. The correlations found are weak and, more damagingly, appear to mainly relay simple string-matching between the query and result.

Properties of the CWW Scent Values. Most of the scent values generated by LSA were low: the minimum score was -0.14 , the maximum was 1 ($M = 0.26$, $SD = 0.35$). Figure 1, above, shows the distribution of scent values for the short query-title pairing. In general, LSA generates high scent values when the text it receives has a high percentage of word-to-word matches and very few word-to-word mismatches (see our later analysis using string matching).

Does CWW's Information Scent Predict Link Choice ? The crucial question for CWW is whether its scent values predict the pages chosen by people in the study. To carry out this evaluation we extracted a subset of I-SPY data that included only the common hit data (i.e., links that were chosen by users, all of whom entered an identical query with the same question in mind). This set consisted of 110 distinct queries and 1,218 chosen links. If CWW predicts people's link choice then the hit score (i.e., the number of people choosing a given url) should correlate with its scent value. Unfortunately, this correlation is low. Table 1 shows the correlations between the short query and the three possible versions of the result (as title alone, blurb alone and title and blurb together). The correlations for long queries were worse, all ≤ 0.04 .

Table 1. Correlations between the Information Scent of the Short Query - Result Pair and the Hit Score

Result Type	Correlation	Classification
Short Query - Title	0.10	Very Weak
Short Query - Blurb	0.11	Very Weak
Short Query - Title + Blurb	0.13	Very Weak

Another perspective on this data can be gleaned from Figure 2 below, which shows scent values plotted by their hit scores for the best pairing (i.e., the short query – title + blurb). The most obvious conclusion from the graph is that many high-scented links have low hit scores and some low-scented links have high hit scores.

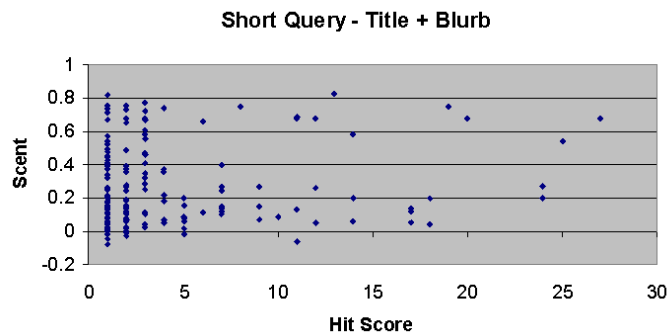


Fig. 2. Plot of the Scent Value by Hit Score for the Most Highly Correlated Query-Result Pair

One of the problems with this analysis is that some of the queries contain terms that are not in LSA's corpus¹. A fairer test would be to perform the same analysis for query-result pairs in which all the terms were known by LSA. However, even after removing all unsupported query-result pairs, the correlations did not improve appreciably, the highest being still low ($r = 0.15$).

Finally, to give CWW the best possible chance, we picked the most highly correlated cases (from the query-result matrix) to see what the best possible correlation could be. Figure 3 shows the result of this analysis. Note that each of the points here represent the scent value x hit score of a query-result pairing where that pairing could be any of the 6 possibilities (e.g., short query-title, long query-blurb, etc). In this best case, the correlation is better and moderate ($r = 0.5$). However, this good news must be tempered by the size and selective nature of the dataset.

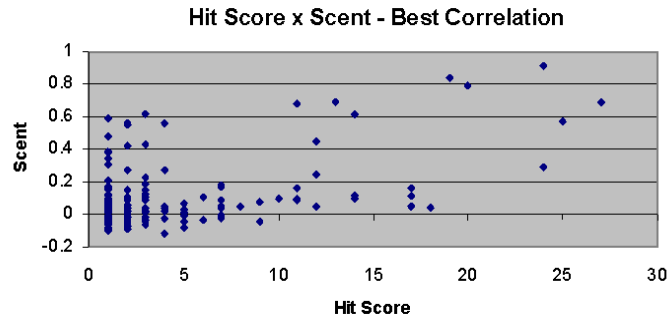


Fig. 3. Plot of the Scent Value by Hit Score for the Best-fit Correlations

Overall, we have to conclude that CWW can only generate reasonable predictions for a limited subset of queries (i.e., those for which it has a vocabulary) and even then it is not clear which one of 6 possible flavors of query-result pairing will best predict hits. This conclusion of limited appeal is further overshadowed by the possibility that CWW mainly appears to succeed by term matching.

Does CWW Succeed by Term Matching ? On the face of it, CWW's highest scent values seem to rely on string matching. To test this hypothesis, we applied a term matching model to the original dataset, using Tversky's Contrast Model of Similarity [32]. Tversky's model sees similarity as being based on the common and distinctive features of two items. The similarity between two objects a and b can thus be defined as,

$$S(a, b) = \theta f(A \cap B) - \beta f(A - B) - \alpha f(B - A) \quad (1)$$

where A represents the set of terms associated with a , B represents the set of terms associated with b , $A \cap B$ is the set of terms common to a and b , $A - B$ represents the set of features distinct in a , $B - A$ represents the set of features distinct in b . In this

¹ In these cases, LSA returns a value of N/A for the terms not in its corpus.

formula, f is usually a count of the features and θ , β and α are usually 1 or 0 (creating a family of models where one or another of the components can be cancelled out).

We then looked at the correlations between the scores produced by variants of the contrast model and CWW's scent values for the same items. We found that a simple term-matching model (i.e., the contrast model using only the common features component) was moderately-highly correlated with the CWW scent values ($r = 0.6$). Given that term matching is simpler and does not encounter the same vocabulary problems as CWW, it would appear to offer a better basis for predicting first-click behavior.

5 Bloodhound's Predictions for First-Click Behavior

We have carried out a similar analysis of the same data using the techniques developed in the InfoScent Bloodhound Simulator [8]. For reasons of space we cannot report these results here, suffice to say that the correlations are considerably worse than those found in CWW (see Church, Keane & Smyth [9], for details).

6 A Framework for Assessing First Clicks

Given our findings it is hard to escape the conclusion that, from a predictive perspective, we know very little about the basis of first-click behavior. As such, it would appear to be a good idea to step back from the problem and consider the main cognitive components of the context in which this task is being carried out. Card et al [5] have previously advanced a problem space of user's browsing behavior. However, we feel that their analysis is at a too fine-grained level to help us in this case and, maybe, really just provides us with a language with which to describe user behavior rather than a theory of the main parameters that impact that behavior. Therefore, in this section, we attempt to outline a general framework for understanding first-click behavior.

Broadly speaking, we can distinguish between the parts of the first click task that are represented in the user's head and those that are represented textually in the computer. The relevant data on the computer side are easily characterized and inspected. They include: the explicit question posed (if elaborated), the specific query terms used, the result lists returned, the ordering of those results, the distal pages to which these links refer and so on. On the human side, the relevant components are less easily characterized and not easily inspected. They include: the users' mental representation of the question, query and results, the user's background knowledge about the domain of the question, the user's knowledge of natural language, the user's knowledge of what ordered results entail, the user's similarity function for matching his/her information need to the presented result, the strategies the user normally employs when searching result lists, knowledge of previous searches and so on.

The key problem is that we do not have good techniques for acquiring and characterizing the knowledge that is brought to bear by users in choosing a link from a set of returned results. In theoretical terms, we need a well-developed cognitive model of this behavior. In practical terms, we need good proxies for this knowledge based on some analysis of the textual data we can explicitly enumerate in the task. In this sense a lot of the work to date can be characterized as proxies of varying goodness.

6.1 Some Proxies for User Knowledge

The usability methods employed here and many of the methods used to personalize and relevance-rank search engine outputs basically use some analysis technique that tries to approximate what people want using explicit data from the web context.

Link-Structure Analysis. Techniques that hinge on recommendation by analyzing link structure e.g., Google [4], essentially work on the assumption that the authority/relevance choices of a community of web-page builders, as indicated by their established links, will parallel the authority/relevance required by someone searching for a resource. The link structure created by the community is a proxy for the relevance ordering of the searching user.

Community-based Hits Analysis. Similarly, the techniques used by I-SPY, which hinge on analyzing the query-result choices of community users, also work on the assumption that what was good for others will be good for you. I-SPY's success is based on the closeness of this proxy to what the user is doing; using other people's choices to predict a new user's choice. In this respect, it is important to point out that I-SPY's builders assume that the community will be in some way representative of the user. This type of representative assumption is a familiar foundation for many approaches to lazy learning [31] and the degree to which it stands up in the context of I-SPY will depend largely on the focus of a particular community of searchers.

Corpus Analysis. CWW basically uses corpus analysis, based on LSA, to approximate a model of people's background knowledge for the words they use. On the face of it, this looks like a sound idea. But, other research has shown that LSA is not good at finding deep semantic similarity [12, 16]. This is exactly what our empirical analysis shows up. First, CWW fails because we fall off the edge of LSA's word knowledge (with specialist terms). Second, its generalization over word meanings is not powerful enough to be a good proxy to human knowledge.

Term-Frequency Analysis. Bloodhound makes heavy use of term frequency analysis in order to provide a proxy to people's knowledge of the domain. Our empirical studies show that varying the set of pages over which these values are computed (i.e., the domain) do not have a significant impact on the goodness of its predictions. It is hard to escape the conclusion that such term-frequency analyses are a poor proxy, on their own, for characterizing user knowledge.

7 Conclusions

Overall there are some positive and some negative conclusions to be made from the empirical analysis we have carried out here. Taking the bad news first, it is clear that current approaches using information scent do not do a good job of predicting the first clicks users make when presented with various results lists. In other words, we still need a good user model for this key behavior in web searching.

Happily, there are also a number of pieces of good news that we can take from this work. First, we outlined a methodology for the empirical evaluation of web search behavior. Second, we have shown that there are limitations to current information foraging theory that can be used productively to guide future theorizing. Third, with our presented framework, we have gained some perspective on the general nature of

first-click behavior. Fourth, we have seen that community-based hits analysis provides a reasonable proxy for first-click behavior, thus suggesting a fruitful direction for future work to characterize this behavior.

8 Acknowledgements

We thank the I-SPY research group at the University College Dublin, namely Jill Freyne, Evelyn Balfe, Peter Briggs and Maurice Coyle, who provided us with data from the I-SPY experiment. This work was funded by grants to the second and third authors from Science Foundation Ireland under Grant No.03/IN.3/I361.

9 References

1. Balfe, E., Smyth, B.: Case-Based Collaborative Web Search. In: Proceedings of the European Conference on Case-Based Reasoning, ECCBR'04, Springer (2004) Madrid, Spain.
2. Blackmon, M.H., Polson, P.G., Kitajima, M., Lewis, C.: Cognitive Walkthrough for the Web. In: Proceedings of the CHI 2002, ACM Press (2002) 463-470.
3. Broder, A.: A Taxonomy of Web Search. SIGIR Forum 36(2) (2002).
4. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the 7th International World Wide Web Conference, (1998) 107-117 Brisbane, Australia.
5. Card, S.K., Pirolli, P., Van Der Wege, M., Morrison, J.B., Reeder, R.W., Schraedley, P.K., Boshart, J.: Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. In: Proceedings of CHI 2001, ACM Press (2001) 498-505 Seattle, WA.
6. Chi, E.H., Pirolli, P., Chen, K., Pitkow, J.: Using Information Scent to Model User Information Needs and Actions on the Web. In: Proceedings of CHI 2001, ACM Press (2001) 490-497 Seattle, WA.
7. Chi, E.H., Pirolli, P., Pitkow, J.: The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site. In: Proceedings of CHI 2000, ACM Press (2000) 161-168 The Hague, The Netherlands.
8. Chi, E.H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., Cousins, S.: The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent™ Simulator. In: Proceedings of CHI 2003, ACM Press (2003) Fort Lauderdale, FL.
9. Church, K., Keane, M.T., Smyth, B.: Evaluating Cognitive & User Models of "First-Click Behavior" in a Personalized Search Engine. User Modeling and User-Adapted Interaction, *Submitted*.
10. Connell, L., Keane, M.T.: PAM: A Cognitive Model of Plausibility. In: Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, Erlbaum (2003) Hillsdale, NJ.
11. Connell, L., Keane, M.T.: What Plausibly Affects Plausibility. Concept Coherence and Distributional Word Coherence as Factors Influencing Plausibility Judgments. Memory & Cognition (2004).
12. French, R., Labiouse, C.: Four Problems with Extracting Human Semantics from Large Text Corpora. In: Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Erlbaum (2002) Hillsdale, NJ.
13. Freyne, J., Smyth, B.: Collaborative Search: A Live User Trial. In: Proceedings of the 26th European Conference on IR Research, ECIR'04, (2004) Sunderland, UK.

14. Freyne, J., Smyth, B.: An Experiment in Social Search. In: Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH-04, (2004) Eindhoven, The Netherlands.
15. Freyne, J., Smyth, B., Coyle, M., Balfe, E., Briggs, P.: Further Experiments on Collaborative Ranking in Community-Based Web Search. *AI Review: An International Science and Engineering Journal*, *In Press*.
16. Glenberg, A.M., Robertson, D.A.: Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, 43 (2000) 379-401.
17. Kintsch, W.: Predication. *Cognitive Science*, 25 (2001) 173-202.
18. Kitajima, M., Blackmon, M.H., Polson, P.G.: A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis. In *People and Computers XIV*, Springer (2000) 357-373.
19. Landauer, T.K., Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104 (1997) 211-240.
20. LSA Website: <http://lsa.colorado.edu>
21. Lund, K., Burgess, C., Atchley, R.A.: Semantic and Associative Priming in High-Dimensional Semantic Space. In: Proceedings of the 17th Annual Conference of the Cognitive Science Society, Erlbaum (1995) 660-665 Hillsdale, NJ.
22. Micarelli, A., Sciarrone, F.: Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System. *User Modeling and User-Adapted Interaction*, 14(2-3) (2004) 159-200.
23. Morrison, J.B., Pirolli, P., Card, S.K.: A Taxonomic Analysis of What World Wide Web Activities Significantly Impact People's Decisions and Actions. In: Proceedings of CHI 2001, ACM Press (2001) 163-164 Seattle, WA.
24. Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web Usage Mining as a Tool for Personalization: A Survey. *User Modeling and User-Adapted Interaction*, 13(4) (2003) 311-372.
25. Pirolli, P., Card, S.K.: Information Foraging. *Psychological Review*, 106(4) (1999) 643-675.
26. Pirolli, P., Fu, W.-T.: SNIF-ACT: A Model of Information Foraging on the World Wide Web. In: Proceedings of the 9th International Conference on User Modeling, (2003) Johnstown, PA.
27. Polson, P.G., Lewis, C., Rieman, J., Wharton, C.: Cognitive Walkthroughs: A Method for Theory-Based Evaluation of User Interfaces. *International Journal of Man-Machine Studies*, 36 (1992) 741-773.
28. Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley (1989).
29. Smyth, B., Balfe, E., Briggs, P., Coyle, M. and Freyne, J.: Collaborative Web Search. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI-03, Morgan Kaufmann (2003) 1417-1419 Acapulco, Mexico.
30. Smyth, B., Freyne, J., Coyle, M., Briggs, P., Balfe, E.: I-SPY: Anonymous, Community-Based Personalization by Collaborative Web Search. In: Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer (2003) 367-380 Cambridge, UK.
31. Smyth, B., Keane, M.T.: Adaptation-Guided Retrieval: Questioning the Similarity Assumption in Reasoning. *Artificial Intelligence*, (102)2 (1998) 249-293.
32. Tversky, A.: Features of Similarity. *Psychological Review*, 84(4) (1997) 327-352.