Independent Study Report
University of Pittsburgh
School of Information Sciences
135 North Bellefield Avenue, Pittsburgh PA 15260, USA
Fall 2004

An Implementation and Analysis on the Effectiveness of Social Search

Jae-wook Ahn

jahn@mail.sis.pitt.edu

1. Introduction

The explosive growth of the Web and its information contents addresses the need for the design of effective tools which can help people to find out proper information they need in an efficient way. Various Web search services have been developed and used so far but their quality of services in terms of user needs has been far from perfection and the problems of these services have been continuously pointed out.

Most of these Web search services are based on the traditional approach of information retrieval, which assumes that the query space and the document space are identical. However, in a real situation, especially in the new Web environment, it is not quite true. Web search service users are formulating very ambiguous and short queries unlike experimental settings where a lot of queries were formulated and refined by domain experts and the length of them was long enough to express users' information needs. Most users are not familiar with expressing their needs in exact query terms which appears in the document space and the number of terms used for their queries are just two or three in average (Lawrence, and Giles, 1998). This situation brings the mismatch between the query space and the document space (Freyne J. and Smyth B, 2004).

Also, most of the Web search engines adopt a "one size fits all" approach. Different users get the same set of search results if they use the same query with other users when they use search engines based on this approach. They are not personalized and are context-independent, so they can't serve different users' different needs.

Conventional rank information alone cannot provide users with enough information to find out relevant documents they need. Therefore, various new ideas have been developed to overcome the limitations of the traditional approach. For example, Google is exploiting link-connectivity information in addition to the traditional term based retrieval model. It applies higher weights to the documents which has more in-link counts and let them appear closer to the beginning of the retrieved result list (Brin, and Page, 1998). This approach can be understood as for finding out popular ones in terms of links among documents.

Social search is another attempt to improve the traditional approach. Like link-connectivity based approach, it makes use of new features to promote the effectiveness of search results. A group of different users who share the same interest can use similar query terms for the same task and their search experiments can be shared. Based on this observation, social search approach exploits past search histories. When a user enters a query, the social search system looks up the search history of the group where the user belongs to and can provide better search results by re-ranking with the clues extracted from the search history or by providing the user with more evidences in addition to the baseline term matching retrieval systems.

We have designed and implemented a social search system by extending a social adaptive navigation system KnowledgeSea. KnowledgeSea let users refer to other users' activities such as navigating and annotating on the documents to help them to find out proper materials. Users can easily find out which document was viewed a lot of times and which document was positively or negatively annotated by other group users. The new social search capability shares this characteristic of KnowledgeSea and let users retrieve search results by directly entering queries while maintaining the social navigation support features. In order to test users' need for this system and to find out if their behavior is different with that of conventional search systems, we conducted user surveys and transaction log analysis.

2. Backgrounds

2.1 Social navigation

Social navigation (Dieberger, Dourish, Höök, Resnick, and Wexelblat, 2000) research tries to explore methods for organizing users' explicit and implicit feedback in order to support information navigation in a meaningful way (Brusilovsky, Chavan, and Farzan, 2004). This approach includes two features. The first feature is to support a known social phenomenon, which means that people tend to follow the "footprints" of other people. The second important feature is self-organization, which allows social navigation systems to function with little or without manual endeavors of human administrators or experts. The well known exemplar systems based on this approach are Web auctions or Weblogs.

Jon Dron and others (2001) also introduced CoFind (Collaborative Filter in N Dimensions), which structures and selects learning resources for teachers. This system was inspired by the concept of Stigmergy. Stigmergy is a word coined by Grasse and it refers to systems employed by termites when building mounds (Heylighen, 1999). When termites build mounds or ants form trails, they can produce mounds and trails by following their colleagues' traces in a collaborative way. These outputs can become stronger as time passes and more group members participate. They also can dissipate if a specific cause runs out and the members' participation decreases, in such a way that when food runs out, the trail to the location of the food dissipates as time passes and ants follow less after they found out no more food from there.

2.2 KnowledgeSea

KnowledgeSea is a Web-based social navigation support system. It organizes Web-based open and closed corpus C language teaching materials including tutorials and slides. Closed corpus based adaptive hypermedia systems such as InterBook (http://www.contrib.andrew.cmu.edu/~plb/InterBook.html) requires manual indexing of features like keywords by domain experts because closed corpus includes only limited set of documents. On the other hand, open corpus based systems require automatic indexing which is generally done in the field of information retrieval because this corpus incorporates almost unlimited set of documents (Brusilovsky, Chavan, and Farzan, 2004). The Web environment is a typical example of the open corpus.

Therefore, KnowledgeSea's Web-based mixed (open and closed) corpus environment introduced the need for the characteristic of self organization and it actively incorporated the adaptive social navigation features into the system.

In order to implement this mixed corpus based social navigation, KnowledgeSea makes use of the semantic map concept (Brusilovsky, and Rizzo, 2002). The map is two dimensional and it is divided into number of cells. It is organized by self organizing map (SOM) algorithm and semantically related keywords and documents were assigned for each cell and the location of cells on the map represents the semantic relatedness. Contents of neighboring cells are semantically related.

By clicking on one of these cells, users can start the navigation and can locate proper materials they need. Background colors of the cells indicate the popularity of the cells. As more users click and visit a cell, the background color of the cell gets darker. When they click a cell, they can see a list of documents and can choose a document from the list.

The same logic to represent cell popularities by color lightness is applied to the representation of documents. Each item of the list provides two types of information, traffic and annotations. "Human-like" icons and colors provide users with popularity information and "thumbs-up" or "thermometer" icon and colors provide users with annotation information. If a document is popular among the group where a user belongs to, the background color of the icon gets darker. The foreground color of the icon gets lighter if the user clicked the document fewer times than other group members. Just like popularity, darker background color of an annotation icon indicates there are a lot of annotations for the document. "Sticky-note," "thumbs-up," and "question-mark" icons indicate "General," "Praise," and "Question" annotations respectively. A red "thermometer" icon indicates the overall annotations are positive, and a blue icon indicates the overall annotations are negative. Therefore, users can navigate socially by referring to other users' behaviors and opinions by looking up these icons and colors provided by KnowledgeSea.

2.3 Social search

Social search or collaborative search is an approach to promote the effectiveness of web search by relying on past search histories (Freyne J. and Smyth B, 2004). They implemented and tested I-SPY which is based on the concept of social search. This system is based on the observation that for specialized topic searches, the number repetition of query terms is higher than that of general topic searches. Therefore, they stored query-document frequency matrix from past search histories of the community users and re-ranked search results by looking up these query-document frequencies. They reported improvement of search results by this approach.

This study tried to implement social search capability to the existing KnowledgeSea social navigation system. Along with the browsing mode provided by KnowledgeSea, we added a search interface and let users can directly search for documents they needed. Because it is based on KnowledgeSea and share the corpus and database with KnowledgeSea, the users could retrieve search results with social navigation information and make use of it.

3. System design and implementation

As described above, the search functionality was added to the main social navigation system of KnowledgeSea. KnowledgeSea was designed based on the concept of knowledge map, which was organized by SOM algorithm. Users can start navigation from this map, find out documents they need and follow another links which have other users' social navigation and annotation information as well as conventional navigation functionality of plain HTML pages.

With the additional search functionality, users are able to directly find out documents by entering query terms and continue their social navigation. A search query text field was located just above the knowledge map and users can search or use the map by their preferences.

3.1. Documents

The document collection which was indexed by this system is identical to that of KnowledgeSea. It is a collection of C language educational materials including tutorials and slides. KnowledgeSea has a database which contains the URL's of target documents and provides the document contents with social navigation information in a real time fashion. The search system fetches these URL's, downloads and indexes them to make them searchable.

3.2. Stemming and stopwords

Terms collected from the documents are stemmed according to Porter's stemming algorithm (Porter, 1980). Very common or rare terms which are stored in a stopword list were excluded. However, due to the characteristics of the document collection, which is a C language tutorial pages, some stopwords such as “if”, “for”, and “while” should be stored in the index because users can use them as query terms and retrieve documents containing them. Therefore, a C keyword list was constructed and they were also included in the index. The identical process is applied to query terms when users enter queries.

3.3. Term weighting

The terms stored in the index of the search system were weighted by their importance for each document. The weighting scheme used is TF-IDF, which means TF (Term Frequency, frequencies of terms for each document) multiplied by IDF (Inverse Document Frequency, inverse of the number of documents where a term appears). TF means how many times a given term appeared in a document and indicates the importance of the term in the document. IDF means the degree of concentration of a given term in the document corpus. Therefore, if a term appears in a small number of documents with high frequencies within them, it is more highly weighted than other terms. For queries, the same weights 1 were used for every term.

3.4. Retrieval model

The vector space model (Salton, 1989) was used for representing documents and queries. Documents in the corpus and users’ queries were represented as vectors. Each element in document vectors represents a term and it has TF-IDF weight. If a term appears in a document, it has the weight of term frequency in the document multiplied by inverse document frequency in the corpus. Elements in query vectors also represent terms and they were represented as binary, that is, the weight of the term is 1 if it appears in the query or zero otherwise.

By comparing a query and document vectors, we can produce a list of documents which are similar to the user query. They are ordered by their similarity and we can therefore achieve a ranked set of retrieved documents. This process is done by calculating Cosine similarity coefficients among query and document vectors as the following equation. Cosine similarity coefficient was calculated with equation 1, where x and y represents query and document vectors.

$$Sim(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \qquad \textbf{(1)}$$

With these similarity values along with some metadata such as document collection, document titles, and social navigation information were provided. The documents with more than 0.01 similarity values were returned and 20 documents per page were displayed.

3.5. Implementation of social search

In addition to the conventional features of this search engine, social navigational features are supported. In response to users' queries, a set of documents sorted by their similarity with the query are retrieved and ranks, document titles, sources of the documents, and similarity scores are displayed for one record. Along with this conventional information, social navigation information is also displayed with proper icons and different foreground/background colors at the end of each record.

The search service shares traffic and annotation database with KnowledgeSea and it can retrieve social navigation information from the database and show it along with search results. Also, when a user clicks any record and views its contents, a document display window of KnowledgeSea is opened. With this window open, traffic information in database is automatically updated and users can make annotations just like when they annotate using KnowledgeSea.

This system supports two types of social navigation information.

1) Traffic of a document – how many times users selected and viewed the corresponding document
2) Annotation on a document – annotations made by users to the corresponding document

Traffic includes user traffic and group traffic. User traffic means the traffic of the current user who is using the system and group traffic means the traffic of other users of the group where the current user belongs to. Annotations include "Praise," "General," and "Question" types and it also represents whether they are positive or negative.

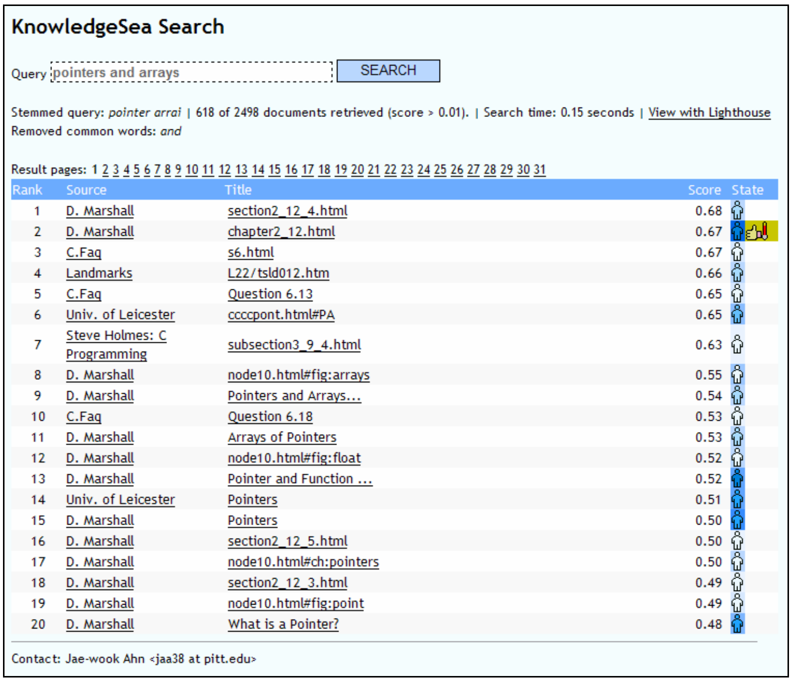

Figure 1. KnowledgeSea search interface

Traffic part of social navigation makes use of "human-like" icons. The blue background colors of the icon represent the group traffic. As group members select and view the corresponding document, the group traffic increases and the blue background gets darker. The foreground colors of the icon means user traffic. If the user has viewed the document more than the average user of the group, the icon is darker than the background. If the user has viewed the document less than the average, the icon is lighter than the background. Figure 2 shows two different records with same similarity scores. Even though their similarity with a given query is identical, the traffic information is different. We can easily see that group users have visited the second record more times than the first record by its darker background color. We can also see that the current user visited these records similar times to group users because the foreground and background colors of these records are identical.

Figure 2. Traffic information

The number of annotations is represented as the darkness of background colors. As users annotate a document more, the yellow background color of the annotation part gets darker. For three different types of "General," "Praise," and "Question" annotations, "sticky-note" "thumbs-up," and "question-mark" icons were used respectively. In order to represent whether the overall annotations for a document are positive or negative, "thermometer" icons were used. For positive annotations, red colored "thermometer" icons were used and for negative annotations, blue colored "thermometer" icons were used. From the example record in Figure 3, we can find out that it has a lot of annotations (darker background), "Praise" annotations ("thumbs-up" icon), and the annotations are positive (red "thermometer icon").

Figure 3. Annotation information

4. Research design

4.1. Research questions and hypotheses

This study is based on the following questions.

1) Do users agree with the need for the search functionality of social navigation?
2) Do they consider the social navigation information provided by group users more important than document ranks when they select a document among the search results?

The first question is about whether the real users will need the new search capability with social navigation support along with the baseline KnowledgeSea system. The other one is related to the situation when the users retrieved documents using the social search. The search result provides users with two types of different information at the same time, conventional similarity ranks and social navigation information. Therefore, users should select documents on the basis of these information and we are interested in the type of information users depend on more. Based on these questions, we have established two hypotheses.

1) Users will need the social search capability and will use it meaningful times.
2) Users will actively select documents with higher social navigation scores. Especially, they will select lower ranked documents (appeared lower in the retrieved list) if the documents have high group traffic and/or positive annotations.

4.2. Data collection

To answer these questions, user survey and usage log analysis of the system were conducted. For the survey, following questions were asked to the students of INFSCI 0012 Introduction to Programming class at the School of Information Sciences, University of Pittsburgh.

1) The availability of search interface in KnowledgeSea was important.
2) Unlike traditional search engines that return the list of results ordered by relevance, KnowledgeSea search also shows you using standard color metaphors how many visits you and your group made to the pages found. This feature was useful in deciding which pages in the list of search results to visit.

We have kept the transaction records of how users behave when they browse, search, and select documents using this system. The transaction logger keeps track of users' navigational behavior. Therefore, we can find out whether users used browsing or searching mode to select an educational document from this data. Also, we can extract the similarity rank and the social navigation score of the documents when users selected and viewed them.

5. Analysis of results

5.1. User survey

9 students answered the survey questions. The results are shown in Figure 4. For the first question asking about the need for the search interface, about 88.9% of the students agreed the need for the search

capability for social navigation system. 11.1% of them were neutral, and no student expressed disagreement with this need. For the question asking the need for the social navigational functionality of retrieved results, 77.8% of the students agreed, 11.1% of them were neutral, and 11.1% disagreed.

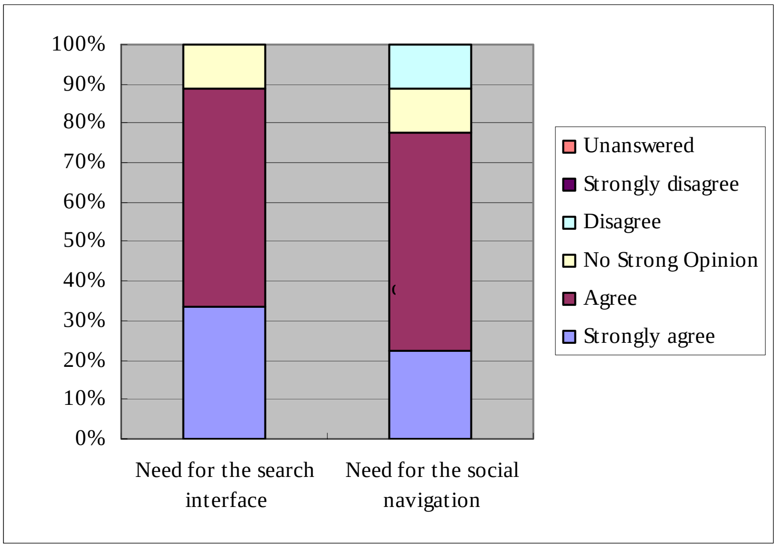

Figure 4. Students' attitude to the need for the search interface and the social navigation

5.2. Transaction log analysis

First, the transaction log for two months (from 10/19/2004 to 12/18/2004) was analyzed. This data contains the frequencies of each mode users had used before they finally located and opened a document. Users can choose a cell and browse using KnowledgeSea's baseline system, or directly search for relevant documents by entering queries to locate relevant materials they need. With this data, we can find out users' preference on each mode before they reached educational document. The result is summarized in Table 1. The most frequently used mode was browsing. Map mode and searching mode were used about 1.5 and 4 times less than the browsing mode respectively.

Map	Browsing	Searching	Total
299 (36.2%)	423 (51.1%)	105 (12.7%)	827

Table 1. Number of times used for each navigation mode

Part of the transaction log data contains more information about users' behaviors. For one month (from 11/16/2004 to 12/18/04), we have recorded data which can be used to analyze social search activities. The additional information includes document rank calculated by similarity, document ID, query string, number of traffics for the corresponding document made by the user himself and other users, the number of annotations, and annotation types. By analyzing this data, we can see if users preferred conventional rank information provided by the search engine or the popularity and annotations of other group users.

53 user activities were obtained within this time period. It represents the information users had perceived before they selected and opened a specific document. Table 2 shows the average rank and count of selected documents for two groups. One group is documents with social navigation information and the other is without such information. This result is corresponding to our expectation in a part and not in part. As we have expected, the users selected documents with social information slightly more times than documents without social information (29 to 24). The user survey and this analysis result support our first hypothesis that users will need the social search capability and will use it meaningful times.

We also have expected that the users would select documents with lower ranks (displayed in the lower part of the retrieved list) if they were popular and/or annotated. However, the overall average rank in Table 2 shows that users still preferred higher rank documents even though the documents were emphasized by social navigation information.

	With social navigation	**Without social navigation**
Average rank	6.48	8.54
Selection count	29	24

Table 2. Average rank of documents with and without social navigation information

Table 3 and 4 shows the number of documents viewed per query by popularity and annotations. Users can distinguish popular documents among group users with higher group traffic by looking at the background colors of the “human-like” icons and they can also distinguish positively annotated documents by looking at the colors of the “thermometer” icons. They selected and viewed about 1.3 times more documents when they retrieved results which include documents other group members had viewed before them. For positive social annotation, they selected and viewed 2.4 times more documents among the retrieved results than they retrieved results without positive annotations. To summarize, users tried more items among their retrieved set when they saw higher traffic items or positively annotated items.

	Average	**Total**	**Number of queries**
With group members’ traffic	2.69	35	13
Without group members’ traffic	2	18	9

Table 3. Number of documents viewed per each query (by traffic type)

	Average	Total	Number of queries
With positive annotations	4.5	18	18
Without positive annotations	1.94	35	4

Table 4. Number of documents viewed per each query (by annotation type)

The data above shows that the retrieved documents with social navigational information was popular among the users. However, in terms of the average rank of the selected documents, the average rank score of the documents with social navigation information was less than the others. This is not corresponding to our expectations because we had expected the users to choose lower rank documents if they are popular or with positive annotations even though their similarity score is lower.

Table 5 can explain why this happened. Not all of the documents stored in KnowledgeSea database appear on the map. Therefore, there can be some documents which can be found solely by searching, not by referring to the map or browsing. Among the total 53 document views, 31 were the documents which did not appear on the map. This means when users were returned search results, the items could be less affected by baseline social navigation capabilities of KnowledgeSea such as the map and browsing. In this sense, we could assume that popularity and annotation information was not accumulated enough in the earlier stage of our experiment and affected the average ranks.

Number of documents	Total number of views	Ratio
31	53	0.58

Table 5. Number of documents not appeared on the map

In order to test this assumption, we divided the transaction log data into 5 chunks by week and analyzed them. The results are provided in Table 6 to 9. We can observe an increase of the document views both with and without high traffic in Table 6. Table 7 shows the average rank of the document views with higher traffic increased and in the last week the average rank of views with high traffic got bigger than that without high traffic. This means users selected items which were displayed at the lower part of the retrieved results as time passed and group users' traffic information accumulated. Table 8 and 9 show the similar results with Table 6 and 7 in terms of positive annotations. As time passed, the number of positively annotated document views and the average rank increased even though they were not sharply distinguishable as the case of traffic.

Week	**1**	**2**	**3**	**4**	**5**	**Total**
Views with high traffic	0	0	0	5	5	10
Views without high traffic	4	0	10	23	6	43
Total views	4	0	10	28	11	53

Table 6. Weekly view counts by popularity

Week	1	2	3	4	5
Rank of views with high traffic	NA	NA	NA	6	6.4
Rank of views without social high traffic	6.5	NA	8.4	8.74	3.33

Table 7. Weekly average ranks by popularity

Week	1	2	3	4	5	Total
Views with positive annotations	0	0	2	4	0	6
Views without positive annotations	4	0	8	24	11	47
Total views	4	0	10	28	11	53

Table 8. Weekly visits by annotations

Week	1	2	3	4	5
Rank with positive annotations	NA	NA	10	7.25	NA
Rank without positive annotations	6.5	NA	8	8.42	4.73

Table 9. Weekly average ranks by annotations

The overall average rank of the documents with social navigation information viewed by users was not lower than the average rank of the documents without such information. At the earlier stage of our experiment, users were inactive in selecting documents with social navigation and they showed a tendency to select higher rank documents as they used conventional search engines. However, the analysis of transaction log shows that as time passes, users have selected actively those documents provided with social navigation information even though they were in lower rank positions. This result supports our second hypothesis.

6. Conclusions

In this study, we added a social search capability to a social adaptive navigation system KnowledgeSea and tested its usability. We implemented a service which shares traffic and annotation information with KnowledgeSea and let users make use of social navigation features within our social search system. We expect this new feature will improve the effectiveness of our system and the social search capability will overcome the limitations of the traditional Web search services.

By implementing this system, we tried to find out if users really needed the social search capability and if they would show behaviors which are different from when they use traditional search services. Users tend to select documents displayed in the upper part of the retrieved result set. However, with additional clues like group user traffic and positive annotations implemented in our system, we could expect a change in users' document selection behaviors.

Therefore, we established two hypotheses. First, users will need the social search capability and will use it meaningful times. Second, users will actively select documents with higher social navigation scores. To test these hypotheses, we conducted a survey and analyzed the transaction log. According to the survey, very high number of users agreed with the importance of search interface and the usefulness of the social navigation support for the search interface. We were also able to find out from the transaction log that the search interface was used in a significant number of times even though the existing map and browsing system of KnowledgeSea were used more often.

To analyze users' document selection behavior in terms of social navigation information, we observed selection counts and the ranks of the selections and tried to find out if users were willing to select and view documents with higher group user traffic and positive annotations. Our assumption was that users would actively select and view popular and positively annotated documents and they would select such documents even though the documents had lower rank score and appeared at the lower part of the retrieved result set. However, overall average rank of documents with social navigation information was higher than those with such information unlike our expectation. Therefore, we tried to see how users' behaviors change as time passes and found out users tended to select lower rank documents as their usage history accumulates.

From these results, we can conclude that users' inactive selection behaviors in the earlier stage of the experiment was caused from the Cold-Start-Problem, which happens at the earlier stage of social navigation systems when enough user history information is not collected. Therefore, we can expect users to exploit popular and positively rated items more and more actively with a system that supports social search and social navigation features as the accumulation of social navigation information increases.

Conventional rank information for information retrieval system is not enough for support users to select relevant documents. With the help of group users' tacit or explicit evaluation on that information, user can more effectively complete their task to find out documents they really need.

7. References

Brin S. and Page L. (1998) The Anatomy of A Large-Scale Web Search Engine. In *Proceedings of the Seventh International World-Wide Web Conference*.

Brusilovsky, P., Chavan, G., and Farzan, R. (2004) Social adaptive navigation support for open corpus electronic textbooks. In: P. De Bra (ed.) *Proceedings of Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004)*, Eindhoven, the Netherlands.

Brusilovsky, P., and Rizzo, R. (2002) Using maps and landmarks for navigation between closed and open corpus hyperspace in Web-based education. *The New Review of Hypermedia and Multimedia* 9, 59-82.

Dieberger, A., Dourish, P., Höök, K., Resnick, P., and Wexelblat, A. (2000) Social Navigation: Techniques for Building More Usable Systems. *Interactions* **7**(6), 36-45.

Dron, J., Boyne, C., and Mitchell, R.: Footpaths in the Stuff Swamp. In: Fowler, W. and Hasebrook, J. (eds.) (2001) *Proc. of WebNet'2001*, World Conference of the WWW and Internet, Orlando, FL, AACE, 323-328

Freyne J. and Smyth B. (2004) An Experiment in Social Search. In: P. De Bra (ed.) *Proceedings of Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004)*, Eindhoven, the Netherlands.

Heylighen F. (1999) Collective Intelligence and its Implementation on the Web: algorithm to develop a collective mental map. *Computational and Mathematical Theory of Organizations* 5(3), 253-280.

Lawrence S. and Giles C.L. (1998) Context and Page Analysis for Improved Web Search. *IEEE Internet Computing*, July-August, 38–46.

Porter, M.F. (1980) An algorithm for suffix stripping, *Program*, 14(3), 130-137.

Salton, G. (1989) Automatic Text Processing, Addison- Wesley Publishing Co., Reading, MA, 313-326.