DRTC Workshop on
Semantic Web
8th – 10th December, 2003
DRTC, Bangalore

Paper: B

The User and the Digital Library

Carolyn Watters
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada

Abstract

Presentation discusses the designing of Digital library as par the user expectations. Presentation discusses issues like Query formation, Document Matching, Ranking of search results and presentability.

In this presentation we consider the impact on the user of the design and implementation of Digital Libraries, with an emphasis on current research results. While the concept of a Digital Library can be traced to Vannevar Bush's *memex* (1) and earlier, implementations became feasible with access to computers in the mid-1960's and we read of Licklider's (2) *library of the future*, Nelson's (3) *hyperspace* and Lancaster's *paperless library*. There is little doubt that the modern era of the Digital Library arrived with widespread access to information provided by networks and the Internet and recognized in 1996 by the United States' National Science Foundation funding for Digital Libraries (4). The term Digital Library now carries a variety of meanings to a variety of constituents from electronic access to the catalogs of a conventional library, electronic access to databases of electronic documents, remote access to digital objects, Internet access to the World Wide Web, and now to support the semantic Web. The term Digital Library is also used to refer to collections where the user group may be very specific, i.e., ACM DL for the computing community or HyNIC for the hypertext community, or where the collection may be defined for the individual user. Furthermore a Digital Library is no longer restricted to notion of a collection of controlled and archived material but includes dynamic collections of objects that be versioned and recombined in new ways so that that the roles of users and authors begin to merge.

Although it is not necessary to remind this audience of readers of the impact of S.R.Ranganathan (5) on libraries, it is relevant to consider his advice in the light of the theme of users in an expanded notion of library.

1. Books are for use.
2. Every book its reader.
3. Every reader its book.
4. Save the time of the reader.
5. Library is a living organism.

The digital library is a reaffirmation of these ideals. In this presentation, however, we concentrate on how the retrieval system can be tailored to support the information needs of individual users. In this context, we need to consider three factors: articulation of query, task motivation, and loss of intermediaries. First, the query that a user presents to

the system is not itself a definition of the information need but a representation of what the user's current understanding of what query might be successful in locating that information. Second, the success of queries is related to the task for which the information is intended. A given query can generate results that would satisfy some users and not others in unpredictable ways. If it is difficult for users to articulated exact queries, it is equally difficult for a system to interpret these queries within the context of an unstated task. Finally, satisfying users of Digital Libraries, who are engaged largely in self-directed electronic access, is made more difficult by the nearly universal substitution of the human intermediary with search mechanisms.

What then can retrieval systems do to be more responsive to users' stated and unstated needs? By and large, digital libraries, with including the Web, are based on indices representing the source and search engines that manipulate those indices. Within that framework there are four points in the retrieval process where interventions can be made to affect the results of a search for the individual user: query formulation, selection of documents that match the query, ranking of these documents, and presentation with possible feedback from the user. Assuming an appropriate archival system is in place, we now consider how the retrieval system can modify these factors to improve user service and satisfaction with Digital Libraries, especially over the web.

Improving the Query

Research (6) has shown that longer queries produce better search results for users and that the quality of the terms chosen for use in a query affects the search results. Retrieval systems can then improve the user experience by modifying the user query to improve the choice of terms and/or to increase the number of terms used in the query. Croft's clarity measure (7) is an excellent example of comparing the characteristics of the terms the user has selected against the characteristics of the vocabulary of the document set to predict how successful the query will be for the user. Extending the query can be done by the use of user profiles of expressed user interests, direct relevancy feedback such as Rocchio feedback, or subject thesauri for adding related terms. These techniques, however, have difficulty in coping with users who have a wide range of changing interests, multitask

during information sessions or depend on users taking the time to provide relevancy feedback between iterations.

Improving the Match

Typically retrieval systems select documents based on a match of the user's query terms and the vocabulary characteristics of the documents in the collection. Machine learning, co citation, and user profiles are often used with varying degrees of success to improve the matching algorithm. The genre of documents is a potentially powerful attribute that could be used more aggressively to match not only the query terms but the task requirements of the user. Most importantly, the role of metadata that captures the semantics of the document as well as the structure or syntactic characteristics of the document holds great promise in improving the quality of matching performance.

Improving the Ranking

The order of retrieved documents from large digital collections is often as important as the selection of documents by the matching algorithm because few users will examine more than the few dozen top documents. Furthermore, one can imagine that different users with different contexts and different purposes will benefit from individualized document orderings. We will examine three examples, geospatial, recommender, and stereotypes. Geospatial ranking (8) provides access to documents satisfying a location-based search in order of their physical distances from some point. For example, a query on *tourist sites in India* could be ordered by their distance either from Delhi or from Bangalore, giving different orderings for each. Recommender systems, on the other hand, rank documents on the similarity of the document to a user profile based on past preferences. The third example is the user stereotypes of users to order the results. For example, in an experiment with medical news (9) the items where matched to the intended audience, i.e., physicians, care givers, and the general public. The results for individual queries could then be reordered to better suit the state of knowledge of the reader.

Improving the Presentation

Typically, retrieved results are presented as an ordered list of titles and possibly brief annotations. This style of presentation assumes that the most relevant items will be at or near the top of the list and that the user will consider the retrieved items in sequential order and stop when satisfied. Other tasks, however, might be better served by some higher level clustering of the results. For example, giving the user an overview of the types of genre found in the document set (academic report, home page, business white paper) or overviews based on semantic or classification schemes. Furthermore, research has shown that the actual layout of results has an impact on user satisfaction. For example, users much prefer a broadsheet layout over the familiar linear layout for selection of news stories, images, and audio or video choices.

In summary, retrieval engines for digital libraries must be aggressive in tailoring query modification, matching, ranking, and presentation to individual user characteristics and individual user tasks. This is clearly, not easily done for real users on real data. User profiles, for example, are problematic because user interests shift over time, users typically multitask their information requests and relevancy judgment is largely subjective. Furthermore, users vary widely in their ability to articulate clear information requests and digital libraries largely lack the reference librarian who bridges this gap in traditional libraries. While automatic detection of user task and interest is critical to the success of any user oriented digital library, past experience indicates that users are not willing to invest much personal time in providing feedback in terms of relevancy judgments on search results or in filling out and updating information questionnaires. Implicit measures are available, however, to provide information on use patterns. For example, users tend to save, print, or bookmark relevant information more often than nonrelevant information. It is also clear that little can be done to predict strictly ludic requests and so care is needed to identify when it is helpful to intervene and when it is better to let the user take the lead.

Users are remarkable for their diversity of interests and goals when accessing information through the web. Before we can design better systems we must appreciate the goals of the individual user for an individual search. The impact of Digital Libraries, both personal

and communal, can only become more dramatic as we continue to see improvements in networking and technology. Improvements in effectiveness and user engagement, however, will require major improvements in how the system responds to the user in satisfying his or her information goals.

References

1. Bush, V. 1945 As We May Think. *Atlantic Monthly*. 176:101-108.
2. Licklider, J.C.R. 1965. *Libraries of the Future*. MIT Press, Cambridge, Mass.
3. Nelson, T. 1980. *Literary Machines*. Mindfull Press.
4. Harper, S. What is a Digital Library? 1996. *KOLISS'96, International Conference on Digital Libraries and Information Services for the 21st Century.* Seoul, Korea.
5. Ranganathan, S.R.: *Prolegomena to Library Classification*, Bombay: Asia Publishing House, 1967.
6. Belkin,N., Kelly,D., Kim,G.,Kim,J-Y.,Lee,H-J.,Muresan,G.,Tang,M-C,Yuan, X-J. and Cool,C. 2003. Query length in interactive information retrieval.*Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. Toronto, Canada. 205-12
7. Croft, W.B, S. Cronen-Townsend, and V.Lavrenko. 2001. Relevance Feedback and Personalization: A Languge Modeling Perspective. *Proc. of the Joint Delos-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*. 49-54.
8. Watters, C. and G.Amoudi. 2002. GeoSearcher: Location Based Ranking of Search Engine Results. *Journal of the American Society for Information Science and Technology.* 53(y)
9. Watters,C., W.Zheng, and E.Milios. 2002. Filtering for Medical News Items. *Proc. of the American Society for Information and Technology Conference.* Nov. 15-19, Pittsburgh.
10. Ozmutlu,S, H.C. Ozmutlu, Spink, A. 2003. Multitasking Web Search and Implications for Design. Proc. of Annual Conf of ASIS&T. Long Beach. 416-21.
11. Shepherd, M., C.Watters, A.Marath. 2002. Adaptive Filtering for Electronic News. *Proc. of HICSS'35*. Jan 7-10.
12. Shepherd, M. C.Watters, and R.Kaushik. 2001.Lessons from reading E-news for Browsing the Web: The Role of Genre and Task. *Proc. of Annual Conference of ASIS&T.*
13. M.Shepherd, J.Duffy, C.Watters, and N.Gugle. 2001. Evaluation of User Profiles for Reading News. *Journal of the American Society for Information Science*. 52(2) 149-160.