Information Clues in Topical Web Searches. Is the Web Message Getting Through?

Mary Ann Epp

Abstract: When creators produce Web pages, they aim to attract searchers to their site. In many cases. the searchers are given the URL site's address. It is, therefore, not difficult to find the page provided the searcher has typed the URL address correctly and the site still exists at the time of searching. But what happens if the searcher does not know the address and is searching a topic or subject? Surfing for information on a topic can be frustrating and time consuming for the searcher. In a study on Web searches, this researcher encountered and recorded many difficulties in identifying information clues to Web sites from the document representation in the hit list. T'he study was based on research for completion of the Masters thesis in Library and Information Studies at the University of British Columbia in 1997. Implications for Web page design are discussed.

Résumé: Une recherche dans Internet pour de l'information sur un sujet peut devenir pour une expérience frustrante qui prend du temps En étudiant des outils de recherche, le chercheur a retrouvé et noté de nombreuses difficultés dans l'identification des indices informattonnels dans la représentation des documents des pages Web Le but de cette étude était d'identifier. de décrue et d'analyser la représentation des documents dans les listes d'occurrence relevées dans les recherches effectuées par quatre outils de recherche: *AltaVista, InfoSeek Ultra.* Lycos et *Open Text Index* À ces outils de recherche on avait accordé la tâche de fane une recherche sur les trots sujets suivants. l'édition sur commande, l'éducation à distance et le graphisme en relief. Le chercheur a constaté que plusieurs des pages Web figurant dans le résumé des résultats de la recherche affichaient peu ou pas d'indices informattonnels sur l'auteur. l'organisation. le mode électronique ou le pays d'origine. Une discussion des implications de l'etude pour la conception de pages Web est fournie

The purpose of the study was to identify, describe and analyze the document representations of sites revealed in the hit list of the searches conducted by four search engines: *AltaVista, InfoSeek Ultra, Lycos,* and *Open Text Index.* Three subjects were selected to be searched through the search engines: custom publishing, distance education, and tactile graphies.

Document representations are surrogates for the documents themselves and are sometimes called information containers (Lancaster, 1993). They are the first sets of words the searcher sees after completing a search using one of the Web's search engines, or browsers. The title and summary mirror the first part of the Web page itself. The surrogates provide the abbreviated package of information from which a searcher selects the whole document to view.

The Problem

Because the information on the Internet is extensive, unorganized, and not presented in a standardized manner, searchers encounter considerable difficulty in identifying

and evaluating the resources on the Web. Publishing documents on the Web does not require adherence to any guidelines or standards for bibliographic presentation or content. The Web has no comprehensive classification scheme to group similar documents together, nor does it have a collocation system to differentiate the versions of a title or gather together the works by the same author. The absence of publishing standards dealing with presentation of content in bibliographic terms makes it difficult, at times, to identify the information clues such as the source, author, authority, type of publication, and provenance of a publication.

The research problem or questions for this study focused on the presence or absence of information clues in the document rcprcscntation of material retrieved by search engines on the Web. Do the various fields of information in the document representations reveal many clues? How can a Web page designer increase the access by including the clues in the design?

Assumptions

The researcher assumed that a large number of information clues would assist the searcher in selecting relevant documents in the set of hits retrieved by the search engines. Conversely, it was assumed that a small number of clues would make it more difficult for searchers to find relevant documents.

The content of the Web is an ever changing phenomenon. Web documents appear, disappear, and change. The functional capabilities of the search engines are constantly evolving as well. Search engines that are highly rated in one research article may not be the same ones selected in another article, or at another time.

Operational Definitions

Search engines are defined as computer program applications developed to assist the users to search the World Wide Web. The four search engines, *AltaVista Vista, InfoSeek Ultra, Lycos,* and *Open Text Index* were selected because they represented four of the most common search tools covering a large number of Web documents.

Document representation is defined as the display information shown by a Web search engine in the hit list, or the container of the information. which includes various information fields, such as the title, the URL (Uniform Resource Locator), or address of the Web site, and the summary. A sample of a document representation is shown in Figure 1.

Information clues are defined as identifiers that may include the author, the source, the organization, the topic or subject, the country of origin, and the electronic genre of information.

The *electronic genre* is defined as the format of the information. Electronic genres are listed as the 'resource type" in the Dublin Core Metadata Element Set (1997). For the purposes of this study, electronic gem-es were limited to print formats. The most common electronic genres found through this study were indexes or links, advertisements, research articles or reports, and directories.

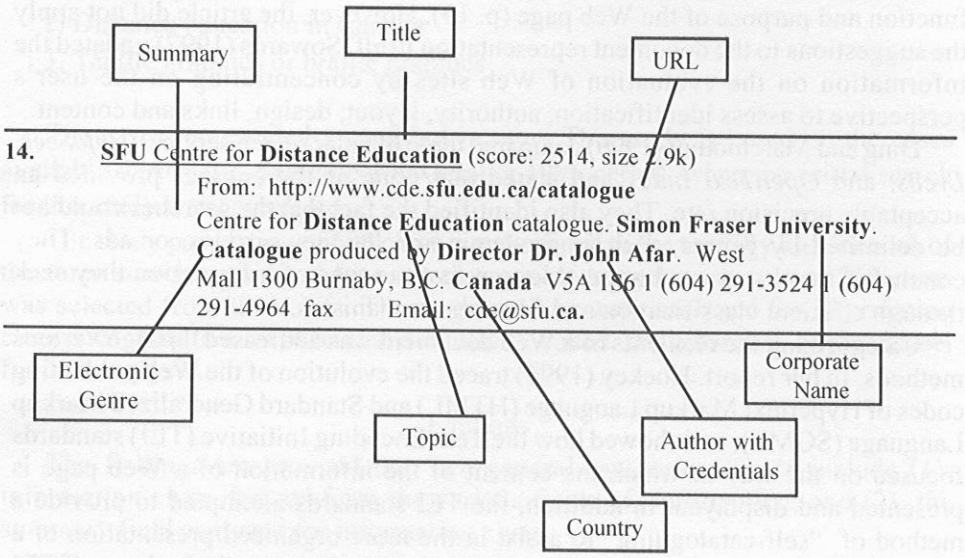

Figure 1: Sample of a document representation

Literature Search

While there was a substantial body of literature on the design and creation of Web pages, the literature tended to concentrate on the aesthetic aspects and the role of the Webmaster rather than on the content (or the representation of content in bibliographical terms) or access. The literature search for this study focused on the content representation and access issues related to that content.

In their study on information-seeking behaviors, Blaise Cronin and Carol Hert (1995) described the scholarly foraging of the Net and showed how the plethora, plurality, and potential instability of resources on Web were a problem for effective searching.

Although they covered a limited set of information clues, Marsha Tate and Jan Alexander (1996) provided an evaluation checklist for various categories of Web pages: advocacy, business/marketing, informational, news, and personal. Their criteria included accuracy, authority, objectivity, currency, and coverage. They provided useful examples of poor Web pages and good Web pages based on these criteria.

According to a survey on Web home page designs for education libraries, Mark Stover and Steven Zink (1996) found very few pages with good designs. They identified some of the sketchy literature pertaining to Web page design principles but said that guidelines for representing the content remained fragmented. They said that many of the pages did not attempt to include fundamental principles of information organization (p. 15). They analyzed the role of the webweaver to plan, analyze, design, and implement the layout, language, techniques for visual communication and targeting the potential audience. Stover *et al* identified the elements of home pages as the descriptive title, a header that names the organization, the purpose, link choices, and contact information. They noted the importance of

the date, the authorship, and formats based on the Yale guide which focuses on the function and purpose of the Web page (p. 19). However, the article did not apply the suggestions to the document representation itself. Sowards (1997) updated the information on the evaluation of Web sites by concentrating on the user's perspective to assess identification, authority, layout, design, links and content.

Ding and Marchionini (1996) compared three popular search services: *InfoSeek,* Lycos, and *OpenText Index* and stated that none of the engines provided an acceptable precision rate. They also identified the fact that the searches could not be delimited by genres, such as academic articles, newsgroups, or ads. They concluded that users need to be able to select the type of information they seek through efficient classification and filtering mechanisms.

Categorizing the elements of a Web document was addressed through various methods. In her report, Hockey (1996) traced the evolution of the Web publishing codes of Hypertext Markup Language (HTML) and Standard Generalized Markup Language (SGML), and showed how the Text Encoding Initiative (TEI) standards focused on the way in which the content of the information of a Web page is presented and displayed. In addition, the TEI standards attempted to provide a method of "self-cataloguing" to assist in the more organized presentation of a document and its retrieval. However, Hockey pointed out the limitations off TEI in its categorization and classification of the document as a whole. In the TEI system, genre headings referred only to literary genres, such as poetry or drama, and did not to the wide range of electronic genres on the Web, such as directories, journals, conference proceedings, listservs, and the like. She stated that the description of a document using TEI standards fell short of the kind of bibliographic description associated with Library cataloguing, such as MARC codes and AACR2 rules. These library codes identify fields of information and formats of works to characterize the document and to enable the document to be indexed appropriately.

Clifford Lynch (1997) listed the types of resources in the Web's "chaotic repository" of books, papers, raw scientific data, menus, meeting minutes, advertisements, video, audio recordings, and transcripts of interactive conversations. He said that because the Web lacked standard identifiers, robot programs could not extract the routine infonnation such as author, date, length and subject matter (known as metadata). Lynch suggested a solution that would attach the categories of metadata to a Web page for retrieval by a Web crawler (p. 5).

Metadata were defined by the Dublin Core Metadata and Warwick Frameworks and were simpler elements than those in traditional library cataloguing (Weibel, 1997). They included: title, author. subject or keywords, description, publisher, other contributors, date, resource type, format, resource identifier, source and language.

In a recent article, Lynch (1998) reviewed the role of identifiers to authenticate documents, reference other works, and activate search strategies. He identified International Standard Book Numbers (ISBNs), Unifonn Resource Locators (URLs), and Serial Item and Contributor Identifiers (SICIs) as examples of useful labels.

Research Design

The researcher selected three topics to search on the Web as case studies for this report. The intent was to select topics broad enough to retrieve a substantial but manageable number of hits for each search. The topics selected were the following keyword phrases:

1. Custom publishing or custom textbooks
2. Distance education in Canada
3. Tactile graphics or braille graphics

The cases represented a snapshot in time. This study is based on the cases available during the September to October 1996 period, when the initial search for this study was completed.

The study population of 600 cases was confined to the top fifty hits for each of the three search topics and four search engines. A random sample of 260 cases was selected from these hits. The random cases were analyzed for information clues according to a set of textual definitions, coded, and entered in a spreadsheet for data analysis.

Findings

The findings are grouped into two general categories. These include (1): information clues derived from the various fields of information; and (2) the summary field analyzed for information clues.

Information Clues

An aggregated count of the major information fields in the document representations was produced to determine the overall presence of information clues. The information clues included authors, titles, organizations. genres, language, country, and elements of print works such as chapters and journal volume numbers. The clues were analyzed in relation to the variables: search engines, search topics and rankings. Overall, 1,674 clues were discovered out of a maximum number of 3,120 possibilities, or fifty-four percent of the total potential cases.

Arranged by search topic, there was a wide disparity between the number of clues for each topic. “Custom publishing” showed substantially more clues than either “distance education” or “tactile graphics”, which had the least.

The fields of information were analyzed for the number of information clues in each field. The results are shown in Table 1. In a large number of cases, there was a limited number of information clues. For example, the title field displayed few clues to the name or author of a document. Similarly, the summary field did not often show identifiers for organizational or personal responsibility. The summary field was not even very helpful in denoting geographical clues to the country of origin. Of the major variables for which information could potentially be available in the title, URL, and summary document fields in this study, six percent exhibited no information clues.

Summary Field

Because the summary field mirrors the first few lines of the Web site and is the major paragraph of information presented in the hit list, it represents the best reflection of the Web site itself. The summary field provided the highest number of clues of all the possible fields for the identification of genre. Authorship is a useful clue to assist in selecting the documents a reader wishes to pursue. Although personal authorship was rarely revealed in the summary field, over half the cases in the summary field showed organizational responsibility for the document.

Table I: Cases with Information Fields that Reveal No Information Clues

INFORMATION FIELDS	NO CLUES*
Title Field. Name Clues	158
Title Field Clues Other Than Name	86
Title Field Genre	195
URL Field Organization	59
URL Field Language	0
URL Field: Country	145
URL Field Genre	205
Summary Field Organization	124
Summary Field Author	204
Summary Field Country	201
Summary Field Language	1
Summary Field Genre	68
TOTAL NO CLUES (out of a possible 3,120 clues)	1446
*n=260 for each field	

Organizations included companies, educational institutions, government agencies, and community groups. Companies represented a little more than double the number of clues on organizations. Educational institutions were revealed in almost a third of the cases.

Most of the recognizable electronic genres in the summary field were ads, followed by electronic texts, articles, directories, indexes, listservs and discussion lists. As might be expected, each of the three topics had its predominant type of document or format. A commercial topic such as “custom publishing” listed a large number of ads. “Tactile graphics," an emerging research topic, showed the largest number of articles or research reports. “Distance education” sites were related more to the institution in which the mode of education was delivered and, understandably, had more indexes, directories and courses listed.

Overall, the topic of “custom publishing ” indicated the highest number of organizational clues while “tactile graphics” revealed the lowest number of organizational clues in the summary field. As might be expected. the number of organizational clues for business companies was far greater for the topic of “custom publishing” than for either “distance education” and “tactile graphics.” Educational institutions accounted for eighteen of the clues for distance education. The results are shown in Figure 2.

There were very few non-organizational clues identified in the summary field. Non-organizational clues included personal names, events, journals or other words not related to a corporate name. Seventy-eight percent of the cases showed no non-organizational identification. Very few cases revealed personal names with credentials. Therefore, there was little to be gleaned from the summary on the authorship of a document.

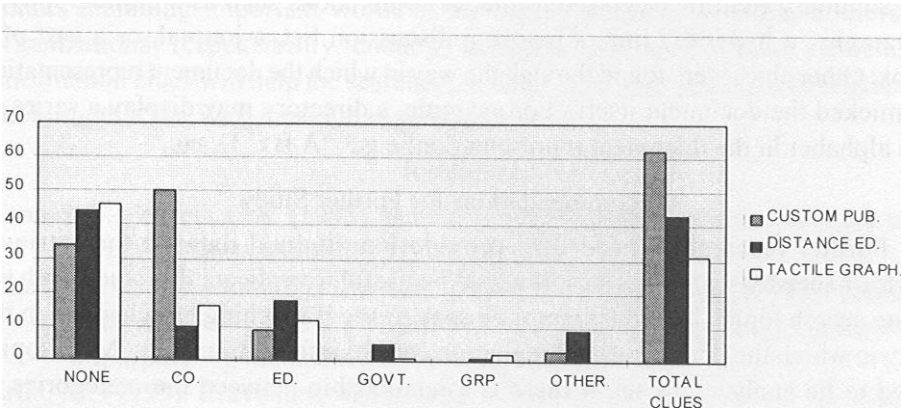

Figure 2: Search Topic and Organization in Summary Field

Discussion of Findings

A major outcome of the study was the relatively low number of information clues (just over 50%). The low incidence of information clues to organization, authorship, and even relevant topics places a large burden on searchers to weed out, filter and sort information before they link to the actual documents for further examination. When there are no information clues, the searcher has two choices: to ignore the entry and move onto the next one, or to link to the full document by clicking on the link to see if it reveals more information. This low level of clues has significance for training in Web searching and for training in Web design.

In this study, the title field was not very productive in revealing useful clues to the document. Personal names, organizational names, and other non-name identifiers were not very plentiful. In many cases, the title field did not deliver clues to the type of document either. The summary field was the most useful part of the document representation for identifying the electronic genre and the organizational name. However, personal authorship was rarely identified. Therefore, these fields of information need to be improved to reveal more useful information about the document itself. The Web designer should place important words at the beginning of the Web page to increase the number of information clues. The lack of authentication is problematic for the searcher because the searcher must take the time to click to the document itself to determine the authenticity of the Web site instead of determining the validity from the document representation. If the words are not in the document itself, they cannot be found through search engines.

The electronic genres were found most often in the summary field. However, it is necessary to know how information is structured to identify the type of clues available for electronic genres. Specific words in the document representation in

the summary field revealed the format of some books such as a bibliography, a biography, a hypertext link, a report, a discussion list, a journal, or a part of a book. Other clues were found through the way in which the document representation mimicked the document itself. For example, a directory may display a series of the alphabet in the document representation; e.g., "A B C D, etc."

Recommendations for Further Study

Further research is needed to provide longitudinal data on the retrieval performance by search engines. It would be useful to replicate this study with the same search topics for a different time span to see if anything has changed in the way in which the document representations reveal information clues. More topics need to be analyzed to see if there is a relationship between the categories of topics, whether they are commercial, academic, or a hybrid, and the number of information clues. The demographics of Web authors might also be analyzed to see what factors influenced their design of the Web sites in terms of information clues.

Recommendations for Design Improvements FOR WEB authors

Authors who wish to have their sites discovered need to include more information clues by designing their pages for easier retrieval and more meaningful selection. A template should be created for Web authors to assist them in the self-identification of the information clues on organization, country, language, content, authorship and genre. It would be useful for the author of Web pages to provide a geographical description of the location of the information either from the point of view of the originator or the point of view of the topic and its relevance to a geographical place.

The size of the title field in Web displays is currently quite limited. A template could assist authors to maximize the limited space by including the required core elements of the topic, authorship, and genre characteristics.

Significant progress has been made to identify the elements which make up a core metadata set as a new bibliographic standard for improved access to Web pages. As Stuart Weibel (1996) recommends. the metadata convention should be embedded in the HTML coding.

Conclusion

Although there are many advocates for better organization of the Web to improve access, the changes will likely not be implemented in the current Web architecture for some time. As a result, searchers are still limited to the access provided by search engines that use robots to collect words from the documents for retrieval purposes. Robots, and search engines relying on robots. are not able to pick the clues and list them in the document representation unless the authors include these searchable keywords in relevant fields. Because the summary field is a literal transcription of the text on the Web site page. authors of Web pages should construct the information on the page for more effective retrieval. This

entails including important words at the top of the site: the topic, authorship, organizational responsibility. country, language, and genre. These elements of information clues will help the searcher (or surfer) find documents more efficiently, ensuring that the Web message is getting through.

References

Cronin, B.. & Hert, C.A. (1995). Scholarly foraging and network discovery tools *Journal of Documentation, 51(4), 388-403.*

Dublin Core Metadata Element Set: Reference Description. (1997). [online]. January 15. 1997 [cited 17 April 1997]. Available from: World Wide Web: <http://purl.org/metadata/dublin_core elements/title>.

Ertel, M. (1995). Brave new world: what a working librarian should know about living on the internet. *Searcher: The Magazine for Database Professionals 3, 28-30.*

Hockey, *S.* (1996). *Describing electronic texts: the text encoding initiative and SGML* [online]. [cited 30 July 1996]. Available from World Wide Web: <URL:http://www.loc.gov/catdir/semidigdocs/hockey.html>.

Lancaster, F.W., & Warner, A.J. (1993). *Information retrieval today,* rev. retitled and expanded ed. Arlington, VA: Information Resources Press.

Lynch, C. (1998). Identifiers and their role in networked information applications. *Feliciter 44(2), 3* l-35

_____ (1997). Searching the internet *Scientific American* [online]. March 1997 [cited 3 April 1997]. Available from World Wide Web: <URL:http:// www.sciam.com/0397issue/O397lynch.html>

Ormondroyd. M. E., & Cosgrave, T. (1997). *How to critically analyze information sources* [online]. [cited 9 January 1997]. Available from World Wide Web: <URL:http://www.library.cornell.edu/okuref/research/skill26.html>.

Stover, M. & Zink, S.D. (1996). World wide web home page design: patterns and anomalies of higher education library home pages. *RSR: Reference Services Review, 24(3). 7-20.*

Tate, M., & Alexander, J. (1996). Teaching critical evaluation skills for World Wide Web Resources. *Computers in Libraries, 16(10*), 49-55.

Tomaiuolo, N.G., & Packer, J.G. (1996). An analysis of internet search engines: assessment of over 200 search queries. *Computers in Libraries, I6(6), 58-62.*

Wei Ding, W., & Marchionini, G. (1996). A comparative study of web search service performance. In *ASIS '96: Proceedings of the 59th ASIS annual meeting, Baltimore, Maryland, October* 21-24, 1996. Medford, N.J.: Information Today.

Weibel, *S.* (1996). *A proposed convention for embedding rnetadata in HTML* [online]. June 2, 1996 [cited I7 April 1997]. Available from World Wide Web: <URL:http://www.ocIc.org/-weibel/html-meta.html>.

Westera, *G.* (1997). *Robot-driven search engine evaluation overview* [online]. October 1996 [cited 9 January 1997]. Available from World Wide Web:

<URL:http://www.curtin.edu.../personal/senginestudy/>.

Younger, J.A. (1997). Resources description in the digital age. Library *Trends, 45(3), 462-481.*

Zorn, M.J., Emanoil, M., & Marshall, L. (1996). Advanced web searching: tricks of the trade. *Online, 20(3), 14-* 16.

AUTHOR

Mary Anne Epp is Director of Contract Administration, Library Services at Langara College, 100 W. 49th Avenue, Vancouver, B.C. V5Y 2Z6. E-mail maepp@langara.bc.ca