THE RETRIEVAL OF HIGHLY SCATTERED FACTS AND ARCHITECTURAL IMAGES: STRATEGIES FOR SEARCH AND DESIGN

SURESH K. BHAVNANI
School of Information, University of Michigan

A**bstract.** The development of huge sources of information in online domains like healthcare, e-commerce, and design, coupled with powerful search engines, suggests that finding comprehensive information about a topic is straightforward. However, recent studies show that while novices can easily find information for questions that have specific answers (e.g. What is a melanoma?), they have difficulty in finding answers for questions requiring a comprehensive understanding of a topic (e.g. What are the risk and prevention factors for melanoma?). This article argues that an important explanation for this difficulty is the phenomenon of *information scatter*: as the number of information sources about a specific topic increases, the information across the sources begins to follow a Zipf-like distribution, where a few sources have a large amount of information, and many sources have very little information.

To illustrate the phenomenon of information scatter, this article presents examples from an ongoing study of how facts related to common healthcare topics are distributed across high-quality sources. These results are compared to results from a small study to explore how images of buildings designed by a well-known architect are distributed across high-quality image sources. The results from both studies suggest that the distributions of facts and images across relevant sources are Zipf-like, and pinpoint the kind of search knowledge needed to address such scatter. These results suggest the need for the development of systems and training that are "distribution conscious", to assist users in finding comprehensive information about topics across information domains.

1. Introduction

To address the widespread ineffective use of complex authoring applications such as computer-aided-drafting (CAD), my research collaboration with Drs. Ulrich Flemming and Bonnie John focused on

identifying and teaching general and efficient strategies (Flemming et al., 1997, Bhavnani et al., 1999, Bhavnani et al., 2001.) This strategy-based approach has led to the development and testing of courses in three universities, including an undergraduate course[1] at the University of Michigan.

Recent research has revealed that the phenomenon of ineffective use is not unique to authoring applications. Despite the development of huge online resources in domains such as healthcare, and architectural design, coupled with powerful search engines, the retrieval of comprehensive information about a topic remains a challenge. For example, a recent study showed that while users of search engines and domain portals can easily find information for questions (e.g. "What is a melanoma?") that have *specific* answers (Bhavnani et al., 2003), they are far less effective when finding information for questions that require a *comprehensive* understanding of a topic (e.g. "What are the risk and prevention factors for melanoma?"). Given the rapid rise in the number of users who depend on the Web for their information needs in domains ranging from healthcare to architectural design, the retrieval of incomplete information can have a large impact on users' judgment in making important decisions.

Why do novice users have difficulty in finding comprehensive information? This article argues that an important explanation for this difficulty is the phenomenon of *information scatter*: as the number of information sources about a specific topic increases, the information across the sources begins to follow a Zipf-like distribution (Zipf, 1949), where a few sources have a large amount of information, and many sources have very little information. Such scatter of information requires strategic knowledge of which sources to visit in which order. This knowledge is neither easily inferred by using current search engines, nor from domain portals. The distribution of how information at the level of granularity important to users (e.g. facts about a disease, and images about buildings) therefore needs close attention.

However, while there have been several attempts to understand how research articles are distributed across journals and databases (e.g. Hood & Wilson, 2001), little is known about how facts and images are distributed across pages and websites, and the possible reasons for those distributions. This article presents the results from an extensive ongoing study (Bhavnani, 2003) of how *facts* (e.g. fair skin increases your risk of getting melanoma) related to common healthcare *topics* (e.g. melanoma risk and prevention) are distributed across high-quality sources. These results are compared to results from a small study to analyze how *images* of buildings designed by a well-known architect are distributed across high-quality image sources.

---

[1] http://www.si.umich.edu/strategycourse/

The results from both studies suggest that facts and images across relevant sources follow Zipf-like distributions, and help to pinpoint the kind of knowledge needed to address such scatter. As such knowledge is not easily inferred from current general-purpose search engines or domain portals, the results suggest the need for a "distribution-conscious" approach to the development of search systems, webpages and sites, and training, with the goal of assisting more users find comprehensive information in vast and unfamiliar online domains.

2. The difficulty of finding comprehensive information

Several studies have shown that novice searchers of healthcare information have difficulty in finding comprehensive information. For example, in the healthcare domain, studies have shown that novice searchers begin their search by typing a few terms in search engines like Google (Eysenbach and Kohler, 2002, Fox and Fallows, 2003), access the resulting hits in the order presented (Bhavnani, 2001), do not check the reliability of their sources (Eysenbach and Kohler, 2002), and end their searches prematurely without accessing sources that in combination provide comprehensive information (Bhavnani, 2001).

The above novice behavior is often in stark contrast to how expert searchers find information. Expert searchers know which sources to visit in which sequence (Kirk, 1974, Florence and Marchionini, 1995, Bhavnani, 2001). For example, in a recent study (Bhavnani, 2001) an expert searcher of healthcare information looking for flu-shot information had a three-step search procedure: (1) Access a reliable healthcare portal to identify sources for flu-shot information. (2) Access a high-quality source of information to retrieve general flu-shot information. (3) Verify that information by visiting a pharmaceutical company that sells flu vaccine. Such search procedures enabled experts to find comprehensive information quickly and effectively, compared to novices who were unable to infer such procedures by just using Google.

What motivates an expert to visit different sites to find information, and why is it difficult for novices to do the same? Our research team hypothesized that the reason why experts had to visit many different sites was because the facts related to the information topic they were searching were scattered across the Web. However, as described below, we found no studies that had analyzed how facts and images related to a topic were distributed across websites.

Pre-web studies of content distribution include the classic works of Bradford who demonstrated the highly skewed distribution of articles about a topic across journals (1948), Hood and Wilson who analyzed the skewed

distribution of articles across databases (2001), and Zipf who described the highly skewed distribution of different words across a book (1949). Recent studies of Web content have focused on the dynamic nature of online information. For example, Bar-Ilan and Peritz (1999) described how webpages retrieved through search engines for the topic "informetric" disappeared, reappeared, or changed over the study period of several months, and Wormell (2000) studied how information about the topic "modern welfare state" spread and evolved through different forms of publication. Other studies of online content have focused on constructing typologies of the context in which query terms occur (Cronin et al. 1998, Bar-Ilan, 1998, 2000a, 2000b). For example, Cronin et al. (1998) identified 11 different source types (homepage, conference page etc.) of pages retrieved from search engines that contained content about highly cited researchers; Bar-Ilan (1998) identified a range of different types of pages in which information about "Erdos" (a well-known mathematician) occurred.

Numerous studies of online content in different domains such as consumer health, and science, have analyzed the accuracy and completeness of online information (Allen et al., 1999, Beredjiklian et al., 2000, Biermann et al., 1999, Davison, 1997, Griffiths and Christensen, 2000, Impicciatore et al., 1997, Jiang, 2000, McClung et al., 1998, Soot et al., 1999, Bichakjian et al., 2002; see Eysenbach et al., 2002 for a review). For example, Bichakjian et al. (2002) found that even the top healthcare sites had incomplete information about melanoma, and Allen et al. (1999) showed the presence of misleading, inaccurate, and un-referenced information in online science publications.

While the above Web content studies and related studies on Web links (e.g. Barabasi and Albert, 1999, Thelwall, 2001, Vaughan and Thelwall, 2003, Klienberg and Lawrence, 2001) have begun to reveal the dynamic and complex nature of the Web, to the best of our knowledge none have attempted to analyze how facts and images related to a topic are distributed across relevant webpages and websites. The following two studies therefore fill an important gap in our understanding of how facts and images related to a topic are distributed across relevant websites. These studies are not designed to reflect how users search the Web for healthcare information. Instead, the studies are designed to analyze the current distribution of facts about a healthcare topic, and images about a design topic across high-quality sites. The goal is to understand how information is scattered across pages and sites, and to pinpoint the knowledge required to deal with such scatter. This understanding could suggest novel approaches that assist users in finding comprehensive information.

3. Distribution of online healthcare information

A recent and ongoing study (Bhavnani, 2003) suggests why finding comprehensive information is difficult. The study consisted of two inter-rater experiments. In the first experiment, two skin cancer physicians identified facts (e.g. High UV exposure increases your risk of getting melanoma) that were necessary for a patient's comprehensive understanding of five melanoma topics (risk/prevention, self-examination, doctor's examination, diagnostic tests, and disease stage[2]) at different levels of importance.

The second inter-rater experiment analyzed how the facts identified by the physicians were distributed across relevant pages from the top 10 sites on the Web with melanoma information. To identify the pages, three search experts iteratively constructed Google queries targeted to each fact and site, and collected the top 10 pages from each query. The process helped to identify 728 relevant pages across the five melanoma topics.

To measure how the facts were distributed across the retrieved pages, two judges were asked to independently rate the level of detail at which facts about a topic occurred in each relevant page using a 5-point scale: 0=not covered in page, 1=less than a paragraph, 2=equal to a paragraph, 3=more than a paragraph but less than a page, 4=entire page. Pages rated by judges as having zero facts (which were retrieved as they had at least one keyword in the query) were excluded resulting in 336 total pages. Both the above experiments had high inter-rater agreement.

The results showed that for each of the five topics, the distribution of facts across the relevant pages were skewed towards few facts, with no single page or single website that provided all the facts. For example, as shown in Figure 1, the distribution of melanoma risk/prevention facts was skewed (resembling a Zipf distribution) towards few facts, and no page had all the 14 facts identified by the physicians. The distribution was similarly skewed when only facts rated by doctors as being "very important" and "extremely important" were included in the analysis.

To understand the underlying causes for the skewed distribution, we conducted a detailed analysis of the content within the pages. The analysis suggested that the skewed distributions are caused by a large proportion of (1) *specialized* pages about the topic that contain a few facts in a lot of detail, and (2) *sparse* pages about related topics that contain few facts in little detail. Figure 2A and 2B shows an example of each type of page. In contrast, there was a smaller proportion of *general* pages about the topic that

[2] These topics were selected from an earlier study (Bhavnani et al., 2002) on the real-world questions that were sent to an ask-a-doc site (which provides answers to healthcare questions from real physicians, and make the anonymized question-answer pairs publicly available).

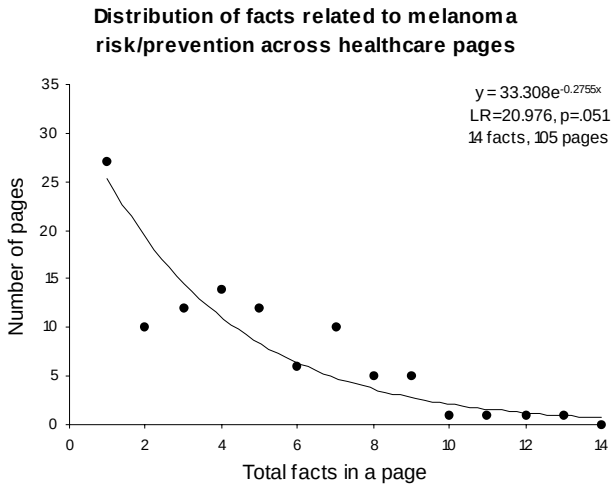

Figure 1. *The distribution of risk/prevention facts across relevant pages in high-quality sites is highly skewed (best-fitted by a discrete exponential curve, Likelihood Ratio=20.967, p=.051 where significant fit is >.05), with no page containing all the facts.*

contain many, but not all, facts in medium amounts of detail. Figure 2C shows an example of a general page.

The above study sheds light on the complex environment in which searching for comprehensive information often occurs. Searchers must visit a combination of pages and websites to find all the facts about a topic. Furthermore, because there are many more pages that contain few facts, there is a high probability that users will find such pages and end their searches early. As neither search engines nor domain portals address this problem, users have difficulty knowing when they have found all the relevant information, and often terminate their searches early with incomplete information (Bhavnani, 2003).

The analysis also provides a possible explanation for the behavior of search experts like healthcare librarians. We believe that such search experts visit a combination of select sources in a specific order when searching for comprehensive information because they have acquired an inherent understanding of the complexities in the distribution of healthcare information across sources.

While there is a skewed distribution of facts, little is known about the distribution of images across image sources on the Web. After all images are more discrete entities, easy to create, and easy to store, which leads to extensive databases such as Greatbuldings.com. Is the distribution of images

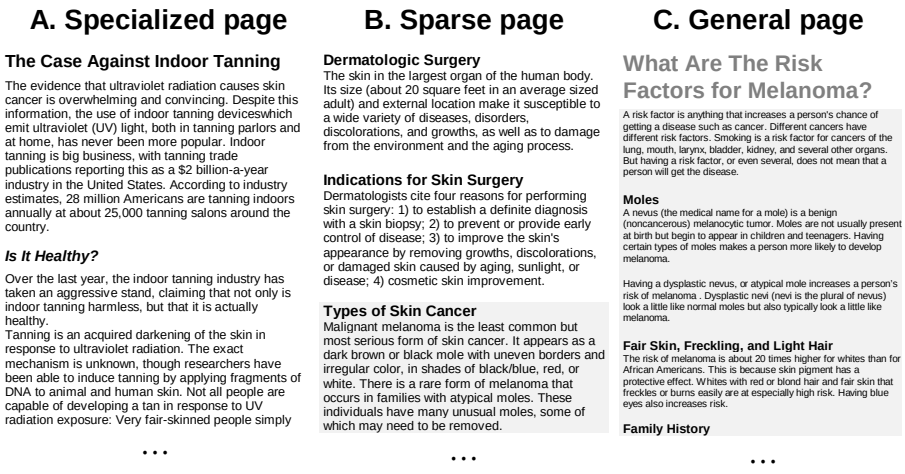

Figure 2. *Examples of three different webpage profiles. Specialized pages (A) have only one fact covered across the entire page, sparse pages (B) have only one fact covered in less than one paragraph, and general pages (C) have many facts in one or two paragraphs each.*

across high-quality image sources any different from the distribution of facts across high-quality healthcare sources?

4. Distribution of online building images

We conducted a small study to understand how architectural building images are distributed across high-quality image resources on the Web. Similar to the study described in Section 3, our study of images is not intended to reflect how users search the Web. Rather, the study is designed to analyze the current distribution of images across high-quality image sources, given an a priori list of images that are important for a specific search task.

Because we found no databases of user queries related to architectural image retrievals, we asked an architectural reference librarian at the University of Michigan to identify a typical search topic to find images. Drawing from her experience in helping architectural students and professors find online information, she recommended the following search topic: “Architectural images of houses built by Frank Lloyd Wright in Wisconsin.”

To find images related to the above search topic, we conducted the study in two parts. The first part of the study identified FLW houses in Wisconsin.

We first retrieved all the houses built by FLW in Wisconsin from the Frank Lloyd Wright foundation[3] website. Appendix 1 shows a comprehensive list of 35 FLW Wisconsin houses that were retrieved. However, the librarian informed us that not all houses in this list were of equal architectural importance and therefore many may not be relevant for a realistic search task. We therefore identified a subset of houses from the above list by selecting only those that had been studied by researchers, and recorded in a typical encyclopedic volume of FLW houses. This was done by searching for research papers on FLW houses in the AVERY database (using the query "Frank Lloyd Wright and Houses and Wisconsin"), and from the set of volumes entitled "Frank Lloyd Wright: Selected Houses" (Pfeiffer, 1989). A union of the houses from both sources yielded 12 houses (shown bold in Appendix 1). We refer to this set of houses as *architecturally important* houses built by FLW in Wisconsin.

The second part of the study analyzed how images of the above houses were distributed across high-quality image sources. To identify the high-quality image sources of architectural images relevant to our search task, we accessed three sources: (1) links to image databases identified by the architectural reference librarian and available on the University of Michigan website, (2) links to sources provided by the Public Broadcasting System (PBS) site dedicated to the Ken Burns documentary on FLW, and (3) links provided by the FLW foundation. We refer to this set of websites as *high-quality image sources*.

The first column of the table in Appendix 2 shows the resulting set of 20 high-quality image sources[4] categorized into four site genres based on their content description: (1) General image databases for art and architecture (e.g. Great Buildings). (2) University architectural slide collections (e.g. SPIRO Architecture Slide Library from University of California Berkeley). (3) FLW-specific sites (e.g. FLW foundation). (4) Architecture societies, journals, and museum exhibits (e.g. Society of Architectural Historians).

Next, we recorded the occurrence of images of the architecturally important FLW buildings, within the high-quality image sources. A search expert attempted to use three methods to search for the images within each site: (1) site-specific search using Google (e.g. "Bogk site:www.greatbuildings.com".) (b) site-provided search box, and (c) navigation links provided by the site[5].

---

[3] http://www.franklloydwright.org/index.cfm?section=research&action=thework

[4] We excluded all sources that specialized in specific types of images (e.g. images of Islamic architecture), and portals with only links to other sites.

[5] Not all methods worked for all sites because only some had search engines, and were accessible through Google. One site was under construction and was therefore dropped from the analysis.

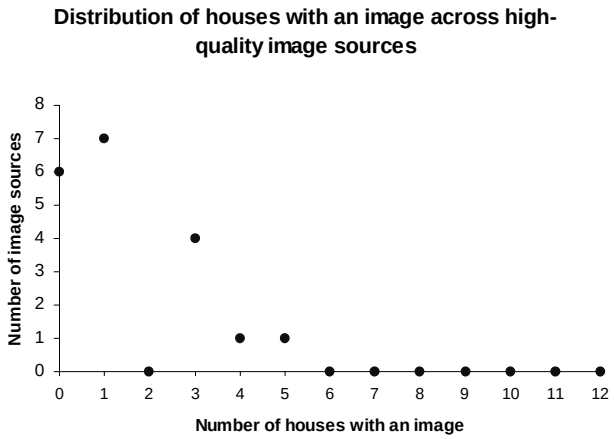

Figure 2. *The distribution of architecturally important houses designed by FLW in Wisconsin with at least one image across high-quality image sources is skewed towards fewer houses.*

To understand the distribution of images, we plotted the number of image sources that contained an ascending number of buildings with images. As shown in Figure 2, the distribution is skewed to the left where there are many image sources that contain a few images, and very few sources (toward the right tail) that contain many but not all the images. Furthermore, no image source had more than five images. Although the plot shows the general shape of the distribution, the sample size is too small to fit a curve (e.g. power vs. discrete exponential) to the data.

An analysis of how specific house images occurred in specific databases revealed that four houses were not present in any image source. A subsequent search in Google found two of the four missing houses on personal websites. The remaining two houses therefore could not be found at all (See Appendix 2 for the details). The scatter of images was therefore a very real phenomenon even within high-quality image sources on the Web.

The above results therefore present a complex situation for users searching for a comprehensive list of images on the Web. To find images of ten of the twelve buildings, a user must know to visit a minimum of four sources: University of Michigan, A Digital Archive of America, Wright in Wisconsin, Broadacre All-Wright, and personal websites identified through the Google image database.

Our ongoing analysis of how images occur within webpages has revealed different page profiles. Some pages provide images of many houses on a single page, while other pages provide several images of a single house, and

yet other pages present images of houses focused towards tourism, rather than focusing on building design. Such page profiles appear very similar to the general, specific, and sparse pages that we observed in the melanoma study presented earlier. Future research will probe deeper into the validity of these preliminary observations.

5. Discussion

The results from the two studies presented in this article have important similarities and differences. In both studies we began with either a list of facts or images identified from non-web sources. The list of healthcare facts was identified from two skin cancer physicians, and the list of architecturally important FLW houses in Wisconsin was identified from paper publications. Next, we identified high-quality Web sources for each list. Finally, we conducted a rigorous search to find the facts and images in the respective high-quality sources, and plotted the distributions of facts and images across the high-quality sources.

As discussed earlier, neither study was designed to simulate how a user would search, but rather how comprehensive information for a specific task was distributed across high-quality sources. We hypothesized that the distribution would enable us to pinpoint the difficulties that users would have if they desired comprehensive information for the kind of tasks that we analyzed.

The analysis showed that the distribution of facts across high-quality pages, and the distribution of images across high-quality sites were skewed (following a Zipf-like distribution) towards few facts and few images respectively. Furthermore, no page or site had all of the facts or all of the images. Finally, the facts and images appear to be configured in general, specific, and sparse page profiles. An important difference in the two studies was that while each healthcare fact was present in at least one webpage, two houses had no images in any of the high-quality image sources.

The above results pinpoint the kind of knowledge that users must have when searching for comprehensive information about healthcare. When searching for facts about a topic, users must know that some pages have breadth information spanning many facts with medium levels of detail (general pages), while others have few facts in a high amount of detail (specific pages). In addition, users also need to know that they have to visit more than one general page to get all the relevant facts. For example, a user must visit at least two sites to obtain breadth information of all the facts about risk/prevention (e.g. Cancer.org and Harvard.edu,), and at least four sites to obtain depth information about each fact (e.g. AAD.org, Skincarephysicians.com, Cancer.org, Skincancer.org).

A similar situation appears to occur when finding images about a topic. As discussed, to find the 10 images of houses on the Web, users have to visit at least four sites from different genres. (Although we have not yet completed a more detailed analysis, the above result does not take into account the quality of the image, which might require users to visit even more sites.) As two images (to the best of our knowledge) are not present anywhere on the Web, users must know when to abandon searching for them. These results begin to reveal the complexity of the knowledge that a user needs to know when searching for comprehensive information about a topic. Because conventional search tools like Google and MEDLINEplus do not provide this kind of information about relevant pages, the lack of such knowledge often leads users to end their searches early, leading to the retrieval of incomplete information (Bhavnani et al., 2003).

One might argue that content providers must strive harder to make sure that the information they provide on relevant pages is complete. However, we have come to believe that such an argument does not acknowledge the nature of information, especially as provided on the Web. Information on the Web (even in the best sites) is created by different authors, with different intentions (Eysenbach et al., 2002), and targeted to different audiences resulting in high variability along many dimensions. While there might be pages that comprehensively cover topics that have a small number of facts or images, we believe that facts related to a vast number of topics will often have a scattered and complex distribution. This is the nature of most information on the Web. Hence we must understand it, and design for it.

6. Implications for developers, authors, and trainers

The analysis of the distributions helped to identify the knowledge required by users who wish to get a comprehensive understanding of a topic. As discussed below, this understanding has direct implications for search engine developers, page authors and designers, and trainers.

As discussed in the introduction, the current paradigm for search engines, is to “get you to the right site” (Thottam, 2001, p. 33). However, our studies (Bhavnani, 2001, 2002) have shown that such an approach is more appropriate for questions that have specific answers (e.g. What is a melanoma?) rather than for a comprehensive understanding of a topic. For search tasks that require a comprehensive understanding of a topic, our results suggest that search engine developers need to explore approaches that deal with how information is scattered across webpages and sites, often leading to a situation where there is no single page or site that contains all the information.

There are a few attempts to explore the above idea from a domain-independent approach (Carbonell and Goldstein, 1998), and our own work from a domain-specific approach (Bhavnani et al., 2003). For example, we have built and tested a prototypical domain portal called the Strategy Hub for Healthcare that attempts to address the distribution of healthcare information across sources. When a user selects a topic such as descriptive information about melanoma risk/prevention, the Strategy Hub responds by suggesting a search procedure that first guides the user to visit a combination of general pages (that together provide an overview of all the relevant facts about melanoma risk/prevention), followed by specialized pages (that focus on specific facts about melanoma risk/prevention such as the danger of tanning booths). A pilot study (Bhavnani et al., 2003), and a more recent experiment have revealed that Strategy Hub users are much more effective in retrieving comprehensive information about melanoma compared to users of MEDLINEplus or Google.

"Distribution-conscious" systems such as the Strategy Hub for Healthcare therefore do not just provide a list of ranked hits, but rather an ordered set of hits that guide a user to webpages in a way that is conducive to acquiring a comprehensive understanding of the topic being searched. Providing an ordered set of hits in turn requires new approaches on how to design a search interface (Bhavnani et al., 2003) that provides a search plan, versus just providing a list of hits. The results from the analyses in this article should provide the incentive to explore the space of such "distribution-conscious" solutions more systematically.

Because pages tend to have different densities of information, page authors and website designers might consider making more explicit which pages are general overviews, and which pages are more detailed, through the use of metadata, through better design of menus, or by adding appropriate text in the page itself. Furthermore, new tools could be developed to find pages within large websites that help to distinguish between pages with different densities of facts.

The results of our analyses could also benefit trainers who teach how to perform effective searches. The analyses suggest that trainers should include in their instruction both, declarative knowledge (concepts), and procedural knowledge (how to perform a task) about distributions. Declarative instruction should include how facts and images tend to be scattered across pages and websites in different amounts of detail, and how such a distribution makes searching for comprehensive information different from searching for a specific fact or image.

Procedural instruction should go beyond teaching how to *find* relevant sources of information, but also how to *visit* relevant sources in a particular order. For example, such instruction could suggest that users should read several general pages that provide an overview of the topic, before reading

more specific pages. This is particularly important because most users select sources provided by Google in the order that they are presented (Bhavnani 2001), which typically does not follow a general to specific ordering. When alternate search engines that take into account information scatter become available, then trainers must steer students to those search engines for finding comprehensive information.

7. Summary and conclusions

Our research was motivated by three observations: (1) novice searchers have difficulty finding comprehensive information, (2) expert searchers know which combination of sources to visit in which specific order to obtain comprehensive information about a topic, and (3) while the above suggested that information was scattered, we found no studies that analyzed how information, such as facts and images, were distributed across relevant sources. We were therefore motivated to understand the scatter of information in different domains, and identify the knowledge required to deal with it.

Results from an ongoing study of the distribution of healthcare information, and a study on the distribution of building images suggest that: (1) the distributions of facts and images across relevant high-quality sources were skewed towards pages having few facts and few images respectively, (2) no page in any site had all the facts for any topic. Further analysis suggested the existence of general pages (that cover many facts in a medium amount of detail), specialized pages (that cover few facts in a high level of detail), and sparse pages that contain (few facts in very little detail). The skewed distributions appear to occur because there were many more specialized and sparse pages compared to general pages. Future research will explore how this explanation holds across domains.

The above results helped to pinpoint the kind of knowledge needed by a searcher to find comprehensive information when information is scattered across pages and websites. As such knowledge is not easily inferred from current general-purpose search engines or domain portals, the results provide direct implications for a “distribution-conscious” approach to the development of search systems, webpages and sites, and training.

The results also provide insights into the behavior of search experts like healthcare librarians. We believe that such search experts visit a combination of select sources in a specific order when searching for comprehensive information because they have acquired an inherent understanding of the complexities in the distribution of healthcare information across sources. These complexities need much more scrutiny than they have received in the past, and should be the focus of future

research. Such research should assist more users find comprehensive information in vast and rapidly growing online domains such as healthcare and architectural design.

Acknowledgements

This research was supported in part by the National Science Foundation, Award# EIA-9812607. The views and conclusions contained in this document should not be interpreted as representing the official policies, either expressed or implied, of NSF or the U. S. Government. I appreciate the diligent statistical analysis done for this project by F. Peck and R. Hu. I also thank A. Abernathy, M. Bates, D. Carter, C. Bichakjian, G. Furnas, T. Johnson, R. Little, R. Price, J. Nardine, F. Reif, V. Strecher, R. Thomas, and G. Vallabha, for their contributions to the data collection and analysis.

References

Allen, E. S., Burke, J. M., Welch, M. E., & Rieseberg, L. H. (1999). How reliable is science information on the Web? Nature, 402, 722.

Bar-Ilan, J. (1998). The mathematician, Paul Erdos (1913-1996) in the eyes of the Internet. Scientometrics, 43, 2, 257-267.

Bar-Ilan, J. (2000a). The Web as information source on informetrics? A content analysis. Journal of the American Society for Information Science, 51, 5, 432-443.

Bar-Ilan, J. (2000b). Results of an extensive search for S&T indicators on the Web - A content analysis. Scientometrics, 49, 2, 257-277.

Bar-Ilan, J., and Peritz, B.C. (1999). The life span of a specific topic on the Web; the case of 'Informetrics': a quantitative analysis. Scientometrics, 46, 3, 371-382.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509-512.

Beredjiklian P.K., Bozentka D.J., Steinberg D.R., & Bernstein J. (2000). Evaluating the source and content of orthopedic information on the Internet: the case of carpal tunnel syndrome. J Bone Joint Surg Am, 82, 1540-1543.

Bhavnani, S.K. (2001). Important cognitive components of domain-specific search knowledge. Proceedings of TREC'01, 571-578.

Bhavnani, S.K. (2002). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. Proceedings of CHI'02, 610-611.

Bhavnani, S.K. (2003). The distribution of online healthcare information: A case study on Melanoma. Proceedings of AMIA'03, 81-85

Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., & Lu, X. (2002). Getting patients to the right healthcare sources: From real-world questions to Strategy Hubs. Proceedings of AMIA'02, 51-55.

Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., Strecher, V.J. (2003). Strategy Hubs: Next-generation domain portals with search procedures. Proceedings of CHI'03, 393-400.

Bhavnani, S.K., John, B.E., and Flemming, U. (1999). The Strategic Use of CAD: An Empirically Inspired, Theory-Based Course. Proceedings of CHI'99, 42-49.

Bhavnani, S.K., Reif, F. and John, B.E. (2001) Beyond Command Knowledge: Identifying and Teaching Strategic Knowledge for Using Complex Computer Applications. Proceedings of CHI'01, 229-236.

Bichakjian, C., Schwartz, J., Wang, T., Hall J., Johnson, T., & Biermann, S. (2002). Melanoma information on the Internet: Often incomplete-a public health opportunity? Journal of Clinical Oncology, 20, 1, 134-141.

Biermann, J.S., Golladay, G.J., Greenfield, M.L., & Baker, L.H. (1999). Evaluation of cancer information on the Internet. Cancer, 86, 3, 381-390.

Bradford, S. C. (1948). Documentation. London: Crosby Lockwood.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, Diversity-based reranking for reordering documents and producing summaries. Proceedings of SIGIR'98, 335-336.

Cronin, B., Snyder, H., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. Journal of the American Society for Information Science and Technology, 49, 14, 1319-1328.

Davison K. (1997). The quality of dietary information on the World Wide Web. Clin Perform Qual Health Care, 5, 64-66.

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews, British Medical Journal, 324, 573-577.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E-R. (2002) Empirical studies assessing the quality of health information for consumers on the World Wide Web: A systematic review. Journal of the American Medical Association, 287, 20, 2691-2700.

Flemming, U., Bhavnani, S.K., and John, B.E. (1997). Mismatched Metaphor: User vs. System Model in Computer-Aided Drafting. Design Studies 18-, 349-368.

Florance, V. & Marchionini, G. (1995). Information Processing in the Context of Medical Care. SIGIR '95, 158-163.

Fox, S., & Fallows, F. (2003). Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access. Pew Internet and American live project: Online life report. Available: http://www.pewinternet.org/reports/toc.asp?Report=95 (July, 2003).

Griffiths, K.M., & Christensen, H. (2000). Quality of web based information on treatment of depression: cross sectional survey. BMJ, 321, 1511-1515.

Hood, W. & Wilson, C. (2001). The scatter of documents over databases in different subject domains: How many databases are needed? Journal of the American Society for Information Science, 52, 14, 1242-1254.

Impicciatore, P., Pandolfini, C., Casella, N., & Bonati, M. (1997). Reliability of health information for the public on the World Wide Web: Systematic survey of advice on managing fever in children at home. BMJ, 314, 1875-1879.

Jiang, Y.L. (2000). Quality evaluation of orthodontic information on the World Wide Web. Am J Orthod Dento-facial Orthop, 118, 4-9.

Kirk, T. (1974). Problems in library instruction in four-year colleges. In: Lubans, John, Jr. (ed.), Educating the library user, 83-103. New York: R. R. Bowker.

Kleinberg, J., & Lawrence, S. (2001). The structure of the Web. Science, 294, 1849-1850.

McClung, H.J., Murray, H.D., & Heitlinger, L.A. (1998). The Internet as a source for current patient information. Pediatrics, 101, 1-4.

Pfeiffer, B. B. (1989) Frank Lloyd Wright: Selected Houses. A.D.A. Edita, Tokyo.

Soot, L.C., Moneta, G.L., & Edwards, J.M. (1999). Vascular surgery and the Internet: a poor source of patient-oriented information. J Vasc Surg, 30, 84-91.

Thelwall, M. (2001). Extracting macroscopic information from web links. Journal of the American Society for Information Science and Technology, 52, 13, 1157-1168.

Thottam, J. (2001). Search smarter. On Magazine, November 2001, 33-37.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? Journal of the American Society for Information Science and Technology, 54, 1, 29-38.

Wormell, I. (2000). Critical aspects of the Danish welfare state - as revealed by issue tracking. Scientometrics, 48, 2, 237-250.

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley.

Appendix-1

Comprehensive list of 35 FLW houses built in Wisconsin retrieved from the Frank Lloyd Wright Foundation site (12 houses, shown in bold, were determined to be architecturally important).

1. Summer Cottage for Henry Wallis, Lake Delavan, WI
2. House for George W. Spencer, Lake Delavan, WI
3. House for Charles R. Ross, Lake Delavan, WI
4. House for Fred B. Jones, Lake Delavan, WI
5. House for A.P. Johnson, Lake Delavan, WI
6. House for Robert M. Lamp, Madison, WI
7. **House for Thomas P. Hardy, Racine, WI**
8. Tan-y-deri House for Andrew Porter, Spring Green, WI
9. House for Eugene A. Gilmore, Madison, WI
10. **House for Frederick C. Bogk, Milwaukee, WI**
11. Duplex Apartments for Arthur Munkwitz, Milwaukee, WI (demolished)
12. Duplex Apartments for Richards Company, Milwaukee, WI
13. Two Small Houses for Arthur L. Richard, Milwaukee, WI
14. **House for Herbert Jacobs, Madison, WI**
15. **Wingspread House for Herbert F. Johnson, Racine, WI**
16. House for Bernard Schwartz, Two Rivers, WI
17. **House for John C. Pew, Madison, WI**
18. **Solar Hemicycle House for Herbert Jacobs, Middleton, WI**
19. House for Richard Smith, Jefferson, WI
20. House for Patrick Kinney, Lancaster, WI
21. House for Willard Keland, Racine, WI
22. **House for Dr. Maurice Greenberg, Dousman, WI**
23. House for E. Clarke Arnold, Columbus, WI
24. **House for Albert Adelman, Fox Point, WI**
25. House for Eugene Van Tamelen, Madison, WI
26. **House for Arnold Jackson, Beaver Dam, WI**
27. House for Frank Iber, Stevens Point, WI
28. **House for Joseph Mollica, Bayside, WI**
29. House for Walter Rudin, Madison, WI
30. House for Duey Wright, Wausau, WI
31. **Cottage for Seth C. Peterson, Lake Delton, WI**
32. Lake Mendota Boathouse, Madison, WI (demolished)
33. **Taliesin I, II, and III (in different phases) Spring Green, WI**
34. House for Stephen M.B. Hunt, Oshkosh, WI
35. House for Charles Manson, Wausau, WI

Appendix-2

The occurrence of at least one image of architecturally important houses built by FLW in Wisconsin, within high-quality image sources.

		Architecturally important houses built by FLW in Wisconsin												
		1. Adelman	2. Bogk	3. Greenberg	4. Jackson	5. Jacobs I	6. Jacobs II	7. Mollica	8. Pew	9. Seth	10. Taliesin (res.)	11. Thomas Hardy	12. Wingspread	Total
High-quality sources for building images	***General architectural image databases***													
	1. A Digital Archive of American Architecture		●			●		●					●	4
	2. Archinform										●			1
	3. Cupola Buildings and Structures													0
	4. Great Buildings					●	●						●	3
	5. Library of Congress American Memory		●								●		●	3
	University slide collections													
	6. Digital Imaging Project (Mary Ann Sullivan Bluffton College)		●											1
	7. SPIRO (UC Berkeley, Architecture Slide Library)		●			●	●							3
	8. University of Michigan Library- Image Services		●			●	●				●		●	5
	FLW-specific sites													
	9. Broadacre All-Wright Site—Frank Lloyd Wright Guide											●		1
	10. Frank Lloyd Wright Building Conservancy										●			1
	11. Frank Lloyd Wright Foundation										●			1
	12. Spring Green, Wisconsin										●			1
	13. The Frank Lloyd Wright School of Architecture	*Site under construction*												
	14. Taliesin Preservation Commission										●			1
	15. Wright in Wisconsin									●	●		●	3
	Societies, Trusts, Museum exhibits													
	16. Architectural Record													0
	17. National Building Museum													0
	18. National Trust for Historic Preservation													0
	19. Society of Architectural Historians													0
	20. The Library of Congress: "Frank Lloyd Wright: Designs for an American Landscape 1922-1932"													0
	Total	**0**	**5**	**0**	**0**	**4**	**3**	**1**	**0**	**1**	**8**	**1**	**5**	