ELSEVIER

Automation in Construction 14 (2005) 724–735

AUTOMATION IN CONSTRUCTION

www.elsevier.com/locate/autcon

The retrieval of highly scattered facts and architectural images: Strategies for search and design

Suresh K. Bhavnani

School of Information, University of Michigan, United States

Abstract

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

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 [1], 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.

E-mail address: bhavnani@umich.edu.

0926-5805/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.autcon.2004.12.007

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 [2], 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 [3], 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 [4] 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.

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 highlight the contrast between the search behaviors of expert and novice searchers. For example, in the healthcare domain, novices tend to search by typing a few terms in search engines like Google [5,6], access the resulting hits in the order presented [7], and do not check the reliability of their sources [5]. Furthermore, they tend to end their searches prematurely without accessing sources that in combination provide comprehensive information [7].

In contrast, search experts follow effective search procedures to find comprehensive information [7–11]. For example, in a recent study [7] 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, and therefore retrieved incomplete and inaccurate information.

Why do experts visit different sites to find information, and why is it difficult for novices to do the same? Our research team hypothesized that experts visited many different sites because the facts related to the information topic they were searching were scattered across the Web. However, we found no studies that had analyzed how facts, and images related to a topic were distributed across websites, and the possible reasons for those distributions.

The distribution of content across sources has been analyzed at different granularities. Before the advent of the Web, Bradford [11] demonstrated that articles the distribution of articles across journals was highly skewed, and Zipf [2] showed a similar distribution of words across books. Recent studies of Web content have revealed the dynamic nature of the Web [12,13], and other studies have constructed typologies of the context in which query terms occur on webpages [14–17]. Furthermore, numerous studies of online content in different domains such as consumer health, and science, have shown that online information is often incomplete or inaccurate [18–27] (see [28] for a review). In contrast to the above studies on web content, a number of studies have focused on how webpages are linked together [29–32].

The above studies have therefore begun to shed light on the complex and dynamic nature of the Web. However, we have not found any studies that analyze how facts and images related to a topic are distributed across relevant webpages and websites.[1] As discussed above, we believe that such studies can help explain the complex behavior of search experts, and why it is so hard for novice searchers to find comprehensive information on the Web. In this paper, we present two such studies. We first summarize the results of our on-going study on how facts related to a healthcare topic are distributed across high-quality sites. To explore whether the results from the above study generalize across domains, we present a new study that examines the distribution of architectural images across high-quality image resources on the Web.

The above two studies are not designed to reflect how users search the Web for healthcare information or images. Instead, the studies are designed to analyze how information is scattered across pages and sites, and to pinpoint the knowledge required to deal with such scatter. The goal is to use this understanding to suggest novel approaches that assist users in finding comprehensive information.

3. Distribution of online healthcare information

A recent study [4] explored how facts related to common healthcare topics were distributed across high-quality healthcare sources on the Web. The study consisted of two inter-rater experiments whose data collection and results are presented here to enable a comparison with a study on images described later. 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 Fig. 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) *specific* 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. Fig. 2A and B shows an example of each type of page. In contrast, there was

[1] The distribution of content across sources has been studied in limited corpuses for purposes other than understanding the nature of the distributions and their causes. For example, *aspects* of a topic have been used to evaluate systems in the Interactive Track of the Text Retrieval Conference [33], and Halteren and Teufel [34] use the presence or absence of *factoids* to evaluate automatic summarization.

[2] These topics were selected from an earlier study [36] 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).

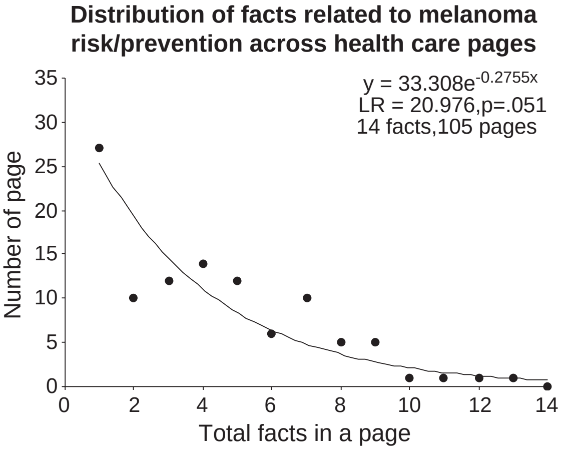


Fig. 1. The distribution of risk/prevention facts across relevant pages in high-quality sites is highly skewed (best-fitted by a discrete exponential curve, Likelihood Ratio=20.967, p=0.051 where significant fit is >0.05), with no page containing all the facts.

a smaller proportion of *general* pages about the topic that contain many, but not all, facts in medium amounts of detail. Fig. 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 get 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 prematurely with incomplete information [35].

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 Greatbuildings.com. Is the distribution of images across high-quality image sources any different from the distribution of facts across high-quality healthcare sources?

4. Distribution of online building images

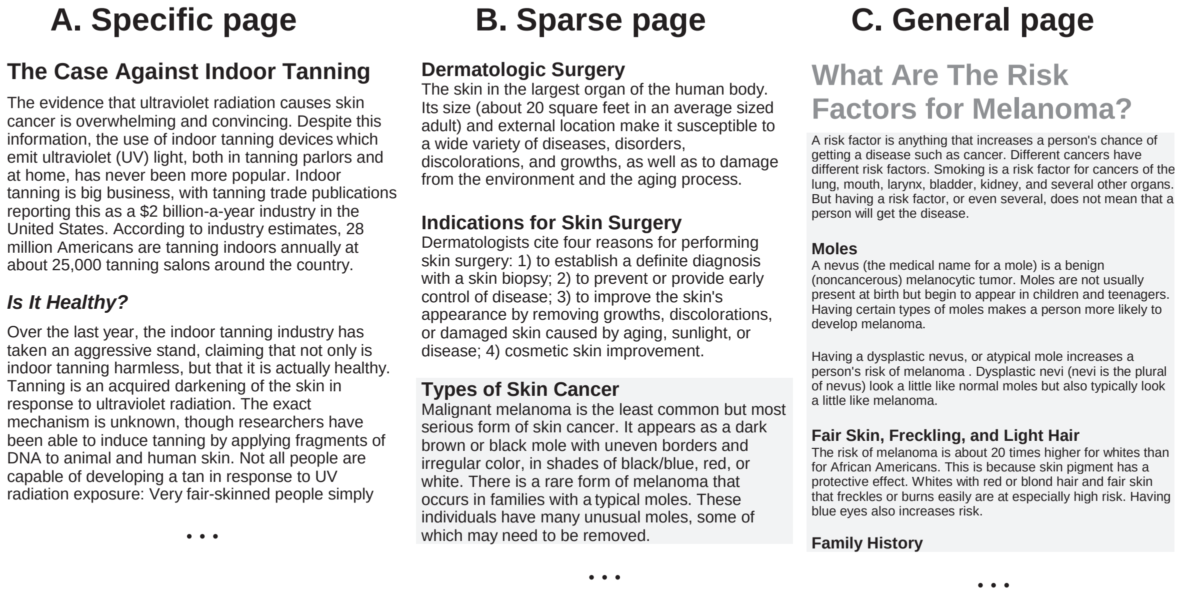
We conducted a small study to understand how architectural building images are distributed across high-quality image resources on the Web. Similar to the study described in Section 3, our study of images is not intended to reflect how users search the Web. Rather, the study is designed to analyze the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

4.1. Method

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 A 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

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



A. Specific page

The Case Against Indoor Tanning

The evidence that ultraviolet radiation causes skin cancer is overwhelming and convincing. Despite this information, the use of indoor tanning devices which emit ultraviolet (UV) light, both in tanning parlors and at home, has never been more popular. Indoor tanning is big business, with tanning trade publications reporting this as a $2 billion-a-year industry in the United States. According to industry estimates, 28 million Americans are tanning indoors annually at about 25,000 tanning salons around the country.

Is It Healthy?

Over the last year, the indoor tanning industry has taken an aggressive stand, claiming that not only is indoor tanning harmless, but that it is actually healthy. Tanning is an acquired darkening of the skin in response to ultraviolet radiation. The exact mechanism is unknown, though researchers have been able to induce tanning by applying fragments of DNA to animal and human skin. Not all people are capable of developing a tan in response to UV radiation exposure: Very fair-skinned people simply

. . .

B. Sparse page

Dermatologic Surgery

The skin in the largest organ of the human body. Its size (about 20 square feet in an average sized adult) and external location make it susceptible to a wide variety of diseases, disorders, discolorations, and growths, as well as to damage from the environment and the aging process.

Indications for Skin Surgery

Dermatologists cite four reasons for performing skin surgery: 1) to establish a definite diagnosis with a skin biopsy; 2) to prevent or provide early control of disease; 3) to improve the skin's appearance by removing growths, discolorations, or damaged skin caused by aging, sunlight, or disease; 4) cosmetic skin improvement.

Types of Skin Cancer

Malignant melanoma is the least common but most serious form of skin cancer. It appears as a dark brown or black mole with uneven borders and irregular color, in shades of black/blue, red, or white. There is a rare form of melanoma that occurs in families with a typical moles. These individuals have many unusual moles, some of which may need to be removed.

. . .

C. General page

What Are The Risk Factors for Melanoma?

A risk factor is anything that increases a person's chance of getting a disease such as cancer. Different cancers have different risk factors. Smoking is a risk factor for cancers of the lung, mouth, larynx, bladder, kidney, and several other organs. But having a risk factor, or even several, does not mean that a person will get the disease.

Moles

A nevus (the medical name for a mole) is a benign (noncancerous) melanocytic tumor. Moles are not usually present at birth but begin to appear in children and teenagers. Having certain types of moles makes a person more likely to develop melanoma.

Having a dysplastic nevus, or atypical mole increases a person's risk of melanoma . Dysplastic nevi (nevi is the plural of nevus) look a little like normal moles but also typically look a little like melanoma.

Fair Skin, Freckling, and Light Hair

The risk of melanoma is about 20 times higher for whites than for African Americans. This is because skin pigment has a protective effect. Whites with red or blond hair and fair skin that freckles or burns easily are at especially high risk. Having blue eyes also increases risk.

Family History

. . .

Fig. 2. Examples of three different webpage profiles. Specific pages (A) are mostly devoted to one fact (although they may contain other facts as well), sparse pages (B) have very few facts covered in less than one paragraph, and general pages (C) have many facts in one or two paragraphs each.

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” [37]. A union of the houses from both sources yielded 12 houses (shown bold in Appendix A). 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 B 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]

4.2. Results

Most of the image sources in our study consisted of databases, which automatically constructed web-

[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.

pages containing only images of a single building. We therefore focused our analysis on how the images of specific buildings were distributed across the image sources as against webpages. This level of granularity enabled a more consistent analysis for all the sites.

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 Fig. 3, 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 B 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.

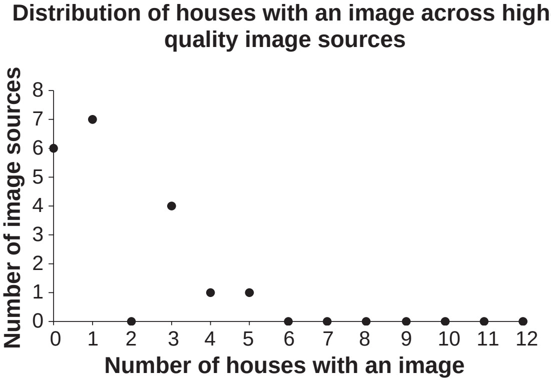


Fig. 3. The distribution of architecturally important houses designed by FLW in Wisconsin with at least one image across high-quality image sources is skewed towards fewer houses.

Our ongoing analysis of how images occur within webpages has revealed different site profiles. Some sites provide images of many houses (e.g. Digital Archives of America had four houses with an average of 10 images each). Other sites provide several images of a single house (e.g. Taliesin Preservation had one house with 15 images). Finally, yet other sites present images of houses focused on topics such as building conservancy rather than on building design (e.g. Frank Lloyd Wright Building Conservancy had one house with 2 images). Such site profiles appear very similar to the general, specific, and sparse page profiles 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 and site profiles respectively. 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), other pages have few facts in a high amount of detail (specific pages), while yet other pages have few facts in little detail (sparse 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 [1].

Given the above results, we reasoned that one way to address the scatter of information is to encourage content providers to make sure that the information they provide on relevant pages is complete. We believe that such an approach is not prudent, as it does not acknowledge the nature of information, especially as provided on the Web. This is because information on the Web (even in the best sites) is created by different authors, with different intentions [28], and targeted to different audiences resulting in high variability along many dimensions. Although 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 online information, and by understanding its nature, we can focus on new approaches to design for it.

6. Towards *distribution conscious* designs of search systems, websites, and training

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.

6.1. Design of search systems

The current paradigm for search engines is to “get you to the right site” [38, p. 33]. However, as discussed above, for comprehensive questions there is often no “right site”. Instead, information is scattered across many sites, and users must therefore visit multiple sites in order to obtain comprehensive information. Thus, our results suggest that search engine developers need to explore approaches that deal explicitly with this information scatter.

A few studies have explored the above idea from a domain-independent approach [39], and our own work from a domain-dependent approach [1,35]. 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 by guiding users to follow a general-to-specific search procedure. When a user selects a topic, the Strategy Hub responds by first suggesting a combination of general pages that together provide an overview of all the relevant facts, followed by specialized pages that focus on specific facts. A pilot study [1], and a more recent experiment [35] have revealed that Strategy Hub users are much more effective in retrieving comprehensive information compared to traditional search approaches. Furthermore, we are exploring approaches to automate the development the Strategy Hub through of the use of content analysis tools [40].

"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 [1] 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.

6.2. Design of websites

As discussed earlier, pages and websites tend to have different densities of information. However, our studies [35] have shown that while expert searchers find comprehensive information by first visiting general pages (which have broad overviews of the topic) before visiting specific pages (which have details of a few facts), novices tend not to follow such an approach. Our current studies are showing that this might be because websites do not make salient the different page profiles and do not link the pages in a way that encourages users to navigate from general, to more specific pages. This suggests that page authors and website designers should make more explicit which pages provide general overviews, and which pages provide detailed descriptions of specific facts. This can be accomplished through the use of metadata, through better design of menus and links, or by adding appropriate text in the page itself. Such designs should encourage more users to easily navigate between the different page profiles to find comprehensive information.

6.3. Design of instruction

Our results also suggest ways in which search instruction can be improved. Specifically, students should be taught that facts and images tend to be scattered across pages and websites in different amounts of detail, and that such a distribution makes searching for comprehensive information different from searching for a specific fact or image. Furthermore, students should be taught procedures for visiting relevant sources in a particular order. For example, one technique is to read several general pages that provide an overview of the topic, followed by more specific pages (see [41] for other search techniques to deal with information scatter). This is particularly important because most users select sources provided by Google in the order that they are presented [7], which typically does not follow a general to specific ordering. Finally, when alternate search engines that take into account information scatter become available, then trainers should encourage students to use 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 to identify the knowledge required to deal with it.

Results from the study on the distribution of healthcare information, and a study on the distribution of building images, suggest that the distributions of facts across pages, and images across relevant high-quality sources are skewed towards pages having few facts, and sites having few images respectively, with no page or site having all the facts or images for any topic. Furthermore, our analysis revealed that sources contain different densities (general, specific, sparse) of relevant information, and that the skewed distributions occur because there are many more specialized and sparse sources compared to general sources. Future research should explore how this explanation holds across even more widely differing online domains.

The two studies also reveal the difficulties that users face when trying to find comprehensive information in an unfamiliar domain. Because information is scattered across many heterogeneous sources, users must acquire the knowledge to structure their search in order to obtain comprehensive information. However, such knowledge is not currently provided by general purpose search engines or domain portals. Because this lack of knowledge can lead to the retrieval of incomplete information, the results of our study suggest a "distribution-conscious" approach to the development of search systems, webpages and sites, and training.

The results also suggest that search experts visit a combination of select sources in a specific order when searching for comprehensive information because they have acquired an understanding of the complexities in the distribution of healthcare information across sources. Future research that probes these complexities should suggest important approaches that can 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 thank M. Bates for guiding me towards the formal study of information distributions, and appreciate the diligent statistical analysis done for this project by F. Peck and R. Hu. I also thank A. Abernathy, 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, and to the anonymous reviewers for their suggestions.

Appendix A

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 B

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

High-quality sources for building images	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
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	

References

[1] S.K. Bhavnani, C.K. Bichakjian, T.M. Johnson, R.J. Little, F.A. Peck, J.L. Schwartz, V.J. Strecher, Strategy hubs: next-generation domain portals with search procedures, Proceedings of CHI'03, 2003, pp. 393–400.

[2] G.K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge, MA, 1949.

[3] W. Hood, C. Wilson, 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) (2001) 1242–1254.

[4] S.K. Bhavnani, Why is it Difficult to Find Comprehensive Information? Implications of Information Scatter for Search and Design. Journal of the American Society for Information Science and Technology (in press).

[5] G. Eysenbach, C. Köhler, 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 (2002) 573–577.

[6] S. Fox, F. Fallows, 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).

[7] S.K. Bhavnani, Important cognitive components of domain-specific search knowledge, Proceedings of TREC'01, 2001, pp. 571–578.

[8] T. Kirk, Problems in library instruction in four-year colleges, in: John Lubans Jr. (Ed.), Educating the library user, R.R. Bowker, New York, 1974, pp. 83–103.

[9] S.K. Bhavnani, Domain-specific search strategies for the effective retrieval of healthcare and shopping information, Proceedings of CHI'02, 2002, pp. 610–611.

[10] V. Florance, G. Marchionini, Information Processing in the Context of Medical Care. SIGIR '95, 1995, pp. 158–163.

[11] S.C. Bradford, Documentation, Crosby Lockwood, London, 1948.

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

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

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

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

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

[17] J. Bar-Ilan, Results of an extensive search for S and T indicators on the Web—a content analysis, Scientometrics 49 (2) (2000) 257–277.

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

[19] P.K. Beredjiklian, D.J. Bozentka, D.R. Steinberg, J. Bernstein, Evaluating the source and content of orthopedic information on the Internet: the case of carpal tunnel syndrome, Journal of Bone and Joint Surgery. American 82 (2000) 1540–1543.

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

[21] K. Davison, The quality of dietary information on the World Wide Web, Clinical Performance and Quality Health Care 5 (1997) 64–66.

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

[23] P. Impicciatore, C. Pandolfini, N. Casella, M. Bonati, 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 (1997) 1875–1879.

[24] Y.L. Jiang, Quality evaluation of orthodontic information on the World Wide Web, American Journal of Orthodontics and Dentofacial Orthopedics 118 (2000) 4–9.

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

[26] L.C. Soot, G.L. Moneta, J.M. Edwards, Vascular surgery and the Internet: a poor source of patient-oriented information, Journal of Vascular Surgery 30 (1999) 84–91.

[27] C. Bichakjian, J. Schwartz, T. Wang, J. Hall, T. Johnson, S. Biermann, Melanoma information on the Internet: often incomplete—a public health opportunity? Journal of Clinical Oncology 20 (1) (2002) 134–141.

[28] G. Eysenbach, J. Powell, O. Kuss, E.-R. Sa, 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) (2002) 2691–2700.

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

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

[31] L. Vaughan, M. Thelwall, 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) (2003) 29–38.

[32] J. Kleinberg, S. Lawrence, The structure of the Web, Science 294 (2001) 1849–1850.

[33] P. Over, TREC-6 Interactive track report, Proceedings of TREC'98, 1998, pp. 73–82.

[34] H. Halteren, S. Teufel, Examining the consensus between human summaries: initial experiments with factoid analysis,

HLT-NAACL 2003 Workshop: Text Summarization (DUC03), 2003, pp. 57–64.

[35] S.K. Bhavnani, C.K. Bichakjian, T.M. Johnson, R.J. Little, F.A. Peck, J.L. Schwartz, V.J. Strecher, Strategy Hubs: domain Portals to Help Find Comprehensive Information. Journal of the American Society for Information Science and Technology (in press).

[36] S.K. Bhavnani, C.K. Bichakjian, J.L. Schwartz, V.J. Strecher, R.L. Dunn, T.M. Johnson, X. Lu, Getting patients to the right healthcare sources: from real-world questions to strategy hubs, Proceedings of AMIA'02, 2002, pp. 51–55.

[37] B.B. Pfeiffer, Frank Lloyd Wright: selected houses, A.D.A. Edita, Tokyo, 1989.

[38] J. Thottam, Search smarter, On magazine, 2001 (November), pp. 33–37.

[39] J. Carbonell, J. Goldstein, The use of MMR, Diversity-based reranking for reordering documents and producing summaries, Proceedings of SIGIR'98, 1998, pp. 335–336.

[40] F.A. Peck, S.K. Bhavnani, M.H. Blackmon, D.R. Radev, Exploring the use of natural language systems for fact identification: towards the automatic construction of healthcare portals. Proceedings of ASIST'04 (in press).

[41] M.J. Bates, Speculations on browsing, directed searching, and linking in relation to the Bradford Distribution, Proceedings of CoLIS4, 2002, pp. 137–150.