Scatter Matters: Regularities and Implications for the Scatter of Healthcare Information on the Web

Suresh K. Bhavnani
Center for Computational Medicine and Bioinformatics, Medical School, University of Michigan, Ann Arbor, MI 48109-1092. E-mail: bhavnani@umich.edu

Frederick A. Peck
School of Education, University of Colorado, Boulder, CO 80309-0249. E-mail: fredpeck@mail.com

Despite the development of huge healthcare Web sites and powerful search engines, many searchers end their searches prematurely with incomplete information. Recent studies suggest that users often retrieve incomplete information because of the complex scatter of relevant facts about a topic across Web pages. However, little is understood about regularities underlying such information scatter. To probe regularities within the scatter of facts across Web pages, this article presents the results of two analyses: (a) a *cluster analysis* of Web pages that reveals the existence of three page clusters that vary in information density and (b) a *content analysis* that suggests the role each of the above-mentioned page clusters play in providing comprehensive information. These results provide implications for the design of Web sites, search tools, and training to help users find comprehensive information about a topic and for a hypothesis describing the underlying mechanisms causing the scatter. We conclude by briefly discussing how the analysis of information scatter, at the granularity of facts, complements existing theories of information-seeking behavior.

Introduction

The Web has spawned the development of extensive Web sites in domains such as healthcare and e-commerce. For example, the National Cancer Institute's Web site contains thousands of pages with information about more than a hundred different cancers. Given such large collections of information, one might conclude that it is easy to obtain comprehensive information[1] about a topic like cancer by visiting one such Web site. However, although users of search engines and domain portals can easily find information for questions that have *specific* answers (e.g., "What is a melanoma?"; Eysenbach & Kohler, 2002), they have difficulty in finding answers for questions requiring *comprehensive* information composed of many facts (e.g., "What are the risk and prevention factors for melanoma?"; Bhavnani et al., 2003; Bhavnani et al., 2005b).

Finding comprehensive information about a healthcare topic is critical because an increasing number of people use information from the Web for a wide range of tasks leading to real-world actions and outcomes. For example, an estimated half of all American adults have searched online for healthcare information to become informed, to prepare for appointments and surgery, and to share information (Fox & Fallows, 2003). Furthermore, healthcare professionals have often emphasized the need for patients to get a comprehensive understanding of their disease (from a consumer's perspective) to improve their judgment in making healthcare decisions and to encourage higher treatment compliance (e.g., Sturdee, 2000; McGlynn et al., 2003).

Why is it difficult to find comprehensive information about a topic? One clue to this difficulty is provided by expert healthcare searchers who know which *combination* of sites to visit in which *order* (Bhavnani, 2001; Bhavnani et al., 2003). A recent study suggests that such expert behavior emerges because of the complex scatter of information across relevant Web sites with a distribution that is skewed towards few facts: A large number of sources have very few facts, while a few sources have many (but not all) facts about a topic (Bhavnani, 2005b). Even highly reputed healthcare sites do not contain comprehensive information about a topic because they often cater to different populations, and respond differently to rapidly changing healthcare facts. This complex scatter of information presents a difficult situation for searchers because they have to visit a combination of sites

Received June 8, 2009; revised July 28, 2009; accepted July 28, 2009

[1] Comprehensive information about a search topic contains all facts (e.g., claims and recommendations) considered important, for that topic, by experts in the field (Bhavnani et al., 2005b).

© 2010 ASIS&T • Published online 26 January 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21217

to find comprehensive information for different topics. As most novice searchers have difficulty acquiring strategies to deal with this complexity, they often end their searches with incomplete information (Bhavnani et al., 2006). Furthermore, because little is known about the regularities underlying information scatter, few solutions to this problem have been proposed.

Because the retrieval of incomplete information in domains like healthcare can lead to negative consequences, this article probes regularities within the scatter of facts across Web pages and Web sites and implications of these regularities to search and design. We begin by discussing existing research that motivates the need to understand regularities in the way facts are scattered across Web pages and Web sites. Next, we briefly describe the data collection method and results of our prior work in analyzing the distribution of facts across high-quality Web pages and Web sites. We then describe the results from two new analyses on the same data: (a) a *cluster analysis* of healthcare pages that suggests the existence of three page clusters with different densities of information and (b) a *content analysis* of the above-mentioned page clusters that reveals the role that each plays in providing comprehensive information. The results provide insights for the design of Web sites, search tools, and training to help users deal with information scatter with the goal of finding comprehensive information and to testable hypotheses about the underlying mechanisms causing the scatter.

Motivation to Probe Regularities in the Scatter of Facts Across Web Sites

Studies on novice healthcare searchers have shown several behaviors that distinguish them from expert healthcare searchers. These behaviors include relying on using general-purpose search engines to find relevant pages (Eysenbach & Kohler, 2002), searching without a search plan resulting in most sites found accidentally (Fox & Fallows, 2003), and often terminating searches with incomplete information (Bhavnani, 2001; Bhavnani et al., 2003). On the other hand, expert searchers (e.g., healthcare reference librarians), tend to have a definite search plan, and know which sites to visit in which sequence. For example, in an earlier study (Bhavnani, 2001) an expert healthcare searcher with the search task of finding flu-shot information followed a three-step search procedure: (a) access a reliable general-purpose healthcare portal to identify sources for flu-shot information, (b) access high-quality sources of information provided by the portal to retrieve general flu-shot information, and (c) access a specific pharmaceutical Web site that sells a flu vaccine to verify the information. Such repeated sequences relevant for specific topics in a domain, and referred to as *search procedures*, enabled expert healthcare searchers to find comprehensive information quickly and effectively. In contrast, novices were unable to infer such knowledge by just using Google (Bhavnani, 2001).

Why do experts visit many different sites to find healthcare information? Our prior research suggests that this difficulty arises because information on the Web, even for narrow well-defined topics, tends to be scattered across different Web sites. For example, in a recent study (Bhavnani, 2005b), we found that although physicians independently agreed that patients need to know 14 facts about melanoma risk and prevention (e.g., having fair skin increases your risk of getting melanoma), none of the top-ten Web sites with melanoma information provided all those facts. Furthermore, more than 75% of the pages had less than half of the total facts. Although the pages were retrieved from the top-ten Web sites with high-quality melanoma information, why were there so many pages from high-quality Web sites with so few facts? As described below, existing research provides few clues to answer this question because few studies have analyzed regularities within the distribution of facts across Web pages and Web sites.

Several studies have analyzed the distribution of content across information sources at different levels of granularity, which include the distribution of articles across journals (Bradford, 1948), the distribution of words within a book (Zipf, 1949), the distribution of articles across online databases (Tenopir, 1982; Lancaster & Lee, 1985; Hood & Wilson, 2001), the distribution of images across databases (Bhavnani, 2005a), and the distribution of facts about a topic across Web pages and Web sites (Bhavnani, 2005b; Over, 1998; Halteran & Teufel, 2003). In each case, the studies analyzed the relationship of one variable (e.g., number of relevant articles) against another variable (e.g., number of journals) through a distribution analysis. Each of the resulting distributions was highly skewed in its upper tail, with a slow descent of the lower tail. This consistent result has led researchers to believe that the phenomenon of skewed distributions is a stable property of how information tends to exist across information sources (Bates, 2002), a phenomenon commonly referred to as *information scatter*.

Some of the above-mentioned researchers have explored how additional variables relate to the two distribution variables with the goal of deepening their understanding of the skewed distributions. For example, Bradford suggested that each topic had a core set of journals that had approximately one third of all the relevant articles and the rest of the articles being scattered across a decreasing level of less-core journals. Hood and Wilson (2001) speculated that the more interdisciplinary a topic, the more scattered its articles tend to be across databases. In our earlier study (Bhavnani, 2005b), we speculated that the amount of detail (e.g., a single sentence versus a paragraph) about a fact could explain why there were so many pages with few facts about a common healthcare topic across high-quality healthcare sites.

Other researchers have speculated about the underlying mechanisms that cause the observed skewed distributions. For example, Zipf (1949) speculated why in any stream of speech (such as in a book or a conversation), there tends to be a high frequency of a few words and a low frequency of many other words, leading to the well-known Zipf distribution.

DOI: 10.1002/asi

He claimed that this highly skewed distribution was the result of a *vocabulary balance* between two opposing forces: (a) the *force of unification* that motivates a speaker to use a small vocabulary of general-purpose words to describe a large range of concepts and (b) the *force of diversification* that motivates a listener to require a large vocabulary of specialized words to describe the same concepts.

Recent studies have revealed other important phenomena about Web content. For example, Bar-Ilan and Peritz (1999) described how Web pages 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, Snyder, Rosenbaum, Martinson, & Callahan, 1998, Bar-Ilan, 1998, 2000a, 2000b). For example, Cronin et al. identified 11 different source types of pages (homepage, conference page, etc.) 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. Finally, numerous studies of online content in different domains, such as consumer health and science, have analyzed the accuracy and completeness of specific Web sites (Allen, Burke, Welch, & Rieseberg, 1999; Beredjiklian, Steinberg, & Bernstein, 2000; Biermann, Golladay, Greenfield, & Baker, 1999; Davison, 1997; Griffiths & Christensen, 2000; Impicciatore, Pandolfini, Casella, & Bonati, 1997; Jiang, 2000; McClung, Murray, & Heitlinger, 1998; Soot, Moneta, & Edwards, 1999; Bichakjian et al., 2002; see Eysenbach, Powell, Kuss, & Sa, 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. showed the presence of misleading, inaccurate, and un-referenced information in online science publications.

Although the above-mentioned studies have focused on the analysis of *content*, there has also been interest in the analysis of *links* between Web pages, which include the testing of computational models to explain regularities in how Web pages are linked. For example, Barabasi and Albert (1999) showed that there was a large number of Web pages that have a few incoming links and a few number of Web pages that had many incoming links. They and others have tested a variety of models for this and other link-related phenomena (e.g., Kleinberg & Lawrence, 2001). For example, the *preferential attachment* model (Barabasi & Albert, 1999) suggests that the probability of a page getting a link is based on the number of links it already has, whereas the *growth* model (Huberman & Adamic, 1999) suggests that the probability of a page getting a link is based on the size of the Web site to which the page belongs.

Although the above studies reveal the complex and dynamic nature of content and links on the Web, little is understood about the regularities underlying the scatter of facts across pages. The analyses described in this article build on our previous reported work (Bhavnani, 2005b) on the scatter of facts across Web pages, with the goal of understanding regularities within the scatter, and the implications of those regularities on approaches to help users find more comprehensive information.

Distribution Analysis: Data Collection and Prior Results

In an earlier study (Bhavnani, 2005b), we explored the question of why finding comprehensive information is difficult. The study comprised of two inter-rater experiments, whose data collection is briefly described here because of its relevance to the analyses in the rest of the article.

In the first experiment, two skin cancer physicians who had extensive experience in researching the information needs of patients, identified facts[2] (e.g., high UV exposure increases your risk of getting melanoma) that were necessary for a patient's comprehensive understanding of the following five melanoma topics:

- Self-examination in the diagnosis of melanoma (henceforth abbreviated to *self-examination*)
- Doctor's examination in the diagnosis of melanoma (henceforth abbreviated to *doctor's examination*)
- Diagnostic tests used in the diagnosis of melanoma (henceforth abbreviated to *diagnostic tests*)
- Disease stages used in the diagnosis of melanoma (henceforth abbreviated to *disease stage*)
- Descriptive information related to melanoma risk and prevention (henceforth abbreviated to *risk/prevention*).

The above five topics were selected from the most common question categories about melanoma retrieved from an ask-a-doc site containing real-world questions (Bhavnani et al., 2002). The two physicians rated the facts on a 5-point *fact-importance scale*: 1 (*not important to know*; and will be dropped from the study), 2 (*slightly important to know*), 3 (*important to know*), 4 (*very important to know*), 5 (*extremely important to know*).

In the second inter-rater experiment, we analyzed how the facts (identified by the physicians) were distributed across relevant pages from the top-ten[3] Web sites with melanoma information. To identify the Web pages, three search experts iteratively constructed 590 Google queries targeted to each fact and site and collected the top-ten hits from each query.

[2] A fact is defined as a statement about a topic, agreed upon by experts in the field (Bhavnani, 2005). For example, facts can be claims (e.g., having fair skin increases your risk of getting melanoma) or recommendations (e.g., confirm your self-diagnosis by consulting a local health care provider).

[3] The top ten websites related to skin cancer were defined as the union of all the sites pointed to by the melanoma page in MedlinePlus, and the top 5 most comprehensive sites identified in a recent study of online melanoma information (Bichakjian et al., 2002). The melanoma page in MedlinePlus was not included because unlike the rest of the pages in our study, it had no content but contained only links to other pages.

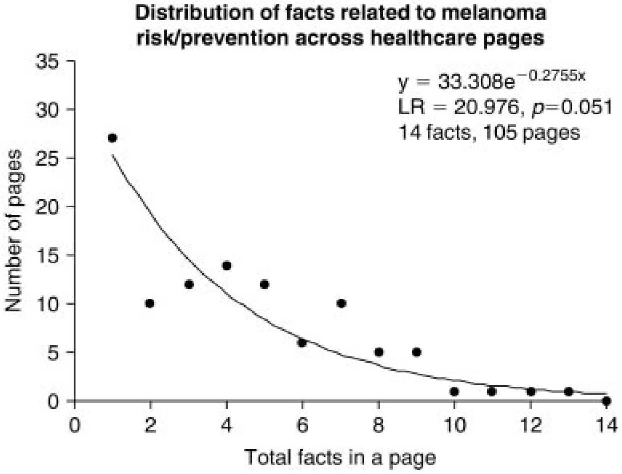


FIG. 1. The distribution of risk/prevention facts across relevant pages in high-quality sites is skewed toward few facts (best fitted by a discrete exponential curve, likelihood ratio $=20.967$, $p=0.051$ where significant fit is >0.05), with no page containing all the facts (Bhavnani, 2005b).

Similar to earlier studies (e.g., Hood & Wilson, 2001) we removed duplicate Web pages within each topic,[4] in addition to other pages such as news items, pages for health professionals, non-English pages, dictionary pages, personal homepages, and broken links. This process helped to identify 728 relevant and unique pages across the five melanoma topics.

Next, we measured how the facts were distributed across the retrieved pages. We asked two raters to independently rate the amount of information of each fact using a 5-point *fact-depth 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*). Note, that this scale enabled us to measure both the number of facts on a page and the amount of information about each fact in each page. Both the above-mentioned experiments had high inter-rater agreement (see Bhavnani, 2005b for details).

Pages rated by judges as having zero facts (but were retrieved as they had at least one keyword in the query) were excluded. This resulted in a total of 336 pages. The results showed that for each of the five topics, the distribution of facts across the relevant pages was skewed towards few facts, with no single page or single Web site that provided all the facts. For example, as shown in Figure 1, the distribution of melanoma risk/prevention facts was skewed towards few facts in its upper tail, and no page had all the 14 facts identified by the physicians. Furthermore, more than 75% of the pages had less than half the facts. The distribution was similarly skewed when facts rated by doctors only as being "very important" and "extremely important" were included in the analysis. The analysis also revealed that the minimum number of Web pages and Web sites that users had to visit to find all facts was dependent on the topic being searched.

These results are similar to the scatter results described by Hood and Wilson (2001), who reported that the number of databases necessary for comprehensive coverage of journal articles about a topic was dependent on the topic being searched. The results are also in agreement with numerous studies showing that many online healthcare sites provide inaccurate or incomplete information (see Eysenbach et al., 2002 for a review). Furthermore, the results also describe how the information is distributed *across* pages and sites.

These results reveal the complex information environment often encountered by healthcare searchers seeking comprehensive information. Because information is scattered across Web sites, searchers must visit a combination of pages and Web sites to find all the facts about a topic. Furthermore, this combination of sites is different across topics. Finally, as discussed earlier, because Web sites cater to different populations with different intentions (Eysenbach et al., 2002) and healthcare information rapidly changes, it is not possible to design Web sites to contain all the information about a topic. As neither search engines nor domain portals address the problem of information scatter, users have difficulty knowing when they have found all the relevant information and often prematurely end their searches with incomplete information (Bhavnani, 2001; Bhavnani et al., 2003; Bhavnani et al., 2005b).

The retrieval of incomplete information can affect real world decisions such as treatment compliance (Sturdee, 2000). The scatter of information therefore merits careful investigation. As we discussed earlier, although several studies have analyzed the scatter of articles of a topic across journals (Bradford, 1948) and across databases (Hood & Wilson, 2001), their results cannot explain why there were so many Web pages with few facts *even* in high-quality Web sites. One key difference between articles and facts is that although articles can occur within journals and databases in only one level of detail (i.e., the entire article), facts can occur within pages in different amounts of detail (e.g., in a line, in a paragraph, in many paragraphs, or in an entire page). In the next section, we explore whether the amount of detail about a fact can help to reveal regularities within the scatter of information.

Cluster Analysis: Exploring Regularities in the Scatter of Facts Across Web Pages

The main goal of this study was to analyze whether there were regularities in the skewed distribution of facts across Web pages. More specifically, we wished to answer the question: Why were there so many pages with few facts compared with the few pages with many facts? An informal analysis of the pages suggested that in addition to the number of facts (fact-breadth), the pages also differed in the *amount* of information about each fact (fact-depth), with both dimensions

[4] Pages that were relevant to more than one topic were included in the analysis as separate pages. Although these pages were identical, they had to be rated for a different set of facts for each of the topics for which they were relevant.

potentially interacting to define pages of different fact density. To explore this informal observation, we performed a cluster analysis on the 336 melanoma pages (identified in our previous study).

Method

Cluster analysis (Aldenderfer & Blashfield, 1984) is typically used as an exploratory data analysis tool to sort different objects into groups (called clusters) such that the similarity (based on the input variables) between two objects within a group is greater than its similarity to objects in other groups. The most common technique used to generate clusters is called K-means (Hartigan, 1975). Although other more complex clustering methods exist, such as hierarchical and fuzzy clusters, we chose to begin our exploratory analysis with an approach to see if we could get a simple explanation underlying the information scatter. However, although this technique can be used to divide data into clusters, it is not designed to generate the optimal number of clusters that best characterize the data. Various techniques have been employed to determine the optimal number of clusters. For example, K-means is often run for many different cluster numbers, and the resulting diagrams are manually inspected to determine which of them generate the most homogenous and meaningful clusters. Recently, there has been active research (e.g., Figueiredo & Jain, 2002) in trying to automatically determine the optimal number of clusters.

For our analysis, we used a two-step process: (a) automatically estimate the number of clusters and (b) automatically identify the boundaries of the clusters. The resulting clusters were then inspected to determine if they were meaningful.

Inputs for both the above-mentioned analyses were fact-breadth and fact-depth. Fact-breadth of a page was defined as the total number of facts for a topic that occurred on that page. Fact-depth of a page was defined as the maximum depth of any relevant physician-identified fact on that page. (We choose maximum depth of any fact on each page because many pages had a single fact that dominated most of the page. Average depth would, therefore, have masked this important distinction between pages.) Both the above-mentioned measures were determined from the *fact-depth* scale (described earlier), which was used by judges in our previous study (Bhavnani, 2005b) to determine the presence and amount of information about each fact on each page. Below, we describe the techniques that we used to implement the two-step cluster analysis procedure.

Estimate number of clusters. To estimate the number of clusters that best categorized the data, we used the *Minimum Message Length* (MML) criterion (Figueiredo & Jain, 2002). This method (see Appendix A for an explanation of MML) enabled us to estimate the optimal number of clusters based on fact-depth and fact-breadth for pages across all the topics and for each of the five topics. Because the number of facts across topics was not the same, fact-breadth (number of facts) was normalized to a percentage of total facts. This normalization enabled all the pages to be collapsed for the overall cluster analysis. Because MML requires interval-level inputs and our fact-depth scale was an ordinal variable, we converted each value in the fact-depth scale to its corresponding mean number of words. This mapping was done by randomly selecting 25% of the pages (evenly distributed across levels of max-detail and topic) and averaging the number of words that described a fact with the maximum detail on each page. The resulting mapping was as follows: max-detail = 1 mapped to 22.93 words, max-detail = 2 mapped to 66.07 words, max-detail = 3 mapped to 119.73 words, and max-detail = 4 mapped to 513.57 words. The MML algorithm was run for one to five clusters.

Estimate boundaries of clusters. To determine the cluster boundaries for all pages and for pages within each topic, we used the K-means algorithm (provided by SPSS, version 11.5). Inputs to K-means were the same two variables used for MML (fact-depth and fact-breadth), with number of clusters provided by MML.

The analysis was first done for the 336 Web pages across all the five topics. To understand variations between each of the topics, the analysis was repeated for pages within each of the five topics.

Results of Cluster Analysis

We first describe the cluster analysis results for the Web pages from all five topics. This is followed by the cluster analysis results for pages from each of the five topics.

Analysis of all Web pages across all topics. The lowest MML value represents the optimal number of clusters in the data. For our analysis, the lowest MML value was obtained for three clusters (see Appendix A for an explanation). This meant that three clusters best characterized the data for the 336 pages across the five topics. This result was, therefore, used to determine the subsequent cluster boundaries using K-means.

The K-means cluster analysis for all the 336 pages across five topics (with three clusters as input) produced the results shown in Figure 2. For clarity, the clusters are plotted with the Y-axis, showing the original fact-depth ordinal scale used by the raters. (See Figure 6 in Appendix B for the same clusters plotted with the Y-axis showing the mean number of words for each level of fact-depth that was used as input to MML and K-means.)

As shown, the analysis identified the boundaries of three page clusters. The lower right-hand cluster (denoted by "•") contains pages with many facts in low-to-medium detail (bounded by fact-breadth >40% and fact-depth = 1–3). These pages were labeled *general pages* for the topic. The top cluster (denoted by "▲") contains pages with a few facts, one of which is described in high detail (bounded by a maximum of 80% fact-breadth and fact-depth = 4). These pages were labeled *specific pages* for the topic. The lower left-hand cluster (denoted by "+") contains pages with

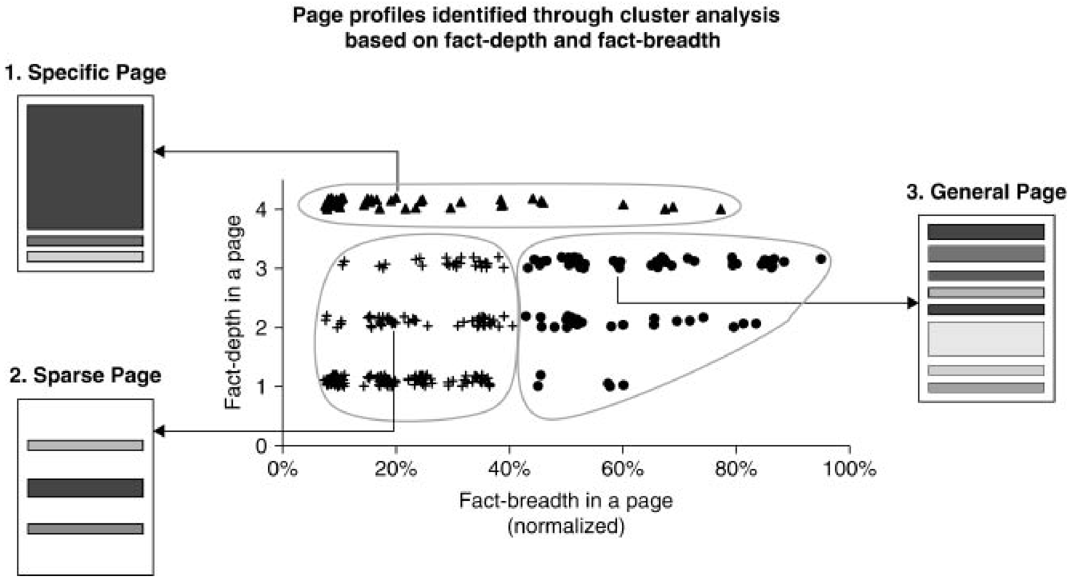


FIG. 2. Results of the cluster analysis that generated the boundaries of three page clusters: specific pages (▲), sparse pages (+), and general pages (●). The schematic drawings of each page profile represent pages closest to the centroid of each cluster, and the patterned rectangles represent different amounts of information about different facts. To enable comparison with other subsets of the data shown later, the fact-breadth on the x-axis has been normalized, and to enable comprehension of the clusters based on how the raters analyzed the Web pages, the y-axis shows the original ordinal scale used by the human raters to analyze the pages (Figure 6 in Appendix B shows the same figure with the y-axis representing the number of words used for the MML analysis that required interval data.) To create this figure, random jitter was added to each point to enable the separation of overlapping points.

TABLE 1. The total pages in each cluster, the cluster centroid, and the cluster boundary for the three clusters shown in Figure 2.

Cluster profile	General	Specific	Sparse
Total pages (number of pages, % of pages)	80, 23.8%	49, 14.6%	207, 61.6%
Centroid (mean fact-breadth, mean fact-depth)	60.1%, 2.6	19.3%, 4	19.9%, 1.5
Boundary (fact-breadth range, fact-depth range)	43%–93%, 1–3	7%–77%, 4	7%–38%, 1–3

Note. There are many more (76%) sparse and specific pages which have few facts, compared with general pages (24%), which have many facts. Because each topic had a different number of facts, fact-breadth has been normalized as a percentage.

relatively few facts in low-to-medium detail (bounded by fact-breadth <40% and fact-depth 1–3). These pages were labeled *sparse pages* for the topic. The cluster boundaries shown in Figure 2 are not jointly exhaustive because our corpus did not contain pages with every possible combination of fact-depth and fact-breadth. For example, because no page contained 100% of the facts, no cluster boundary includes pages with 100% of the facts.

The cluster diagram also shows that the distinction between sparse and general pages is less clear near the boundary of the two page types (at approximately 40% fact-breadth). This suggests that the there is a continuum in the number of facts, rather than a sharp boundary between the two page types. However, the archetypical pages at the centroid of each cluster have distinct information densities. As shown in the schematic drawings in Figure 2, there is a clear difference between an archetypical sparse page with few facts and low detail (fact-breadth = 20%, fact-depth = 2) and an archetypical general page with many facts in medium detail (fact-breadth = 60%, fact-depth = 3). (A subsequent analysis of content discussed in the Content Analysis: Understanding the Role of Page Cluster section provides further evidence about the distinction between sparse and general pages). Because each topic had a different number of facts, the schematic drawings assume a topic with 14 facts (the maximum number of facts in any topic).

The cluster analysis of all 336 pages therefore provided a more detailed view of the distribution[5] of facts across all the pages. As shown in Table 1, specific and sparse pages together dominated (76%) the overall percentage of pages (specific = 14.6%, sparse = 61.6%), both of which contained a relatively low mean number of facts (specific = 19.3%, sparse = 19.9%). Both these pages form the left part of the distribution causing the high skew towards few facts (similar to the distribution of only risk/prevention facts show in Figure 1). In comparison, there is a smaller percentage (23.8%) of general pages that contain a relatively higher mean

[5] Similar to Figure 1, the overall distribution for all 336 pages was also skewed and best fitted by a discrete exponential curve, $y = 142.736e^{-3.54x}$.

number of facts (60.1%). The distribution of facts is skewed towards few facts because there are many more specific and sparse pages compared to general pages.

Analysis of Web pages from each topic. As stated earlier, to understand the variation between topics, the above analysis was repeated for each topic. The optimal number of clusters across the five topics (as estimated by the MML) ranged between two and four: self-examination = 3, doctor's exam = 2, diagnostic tests = 4, disease stage = 4, and risk/prevention = 3. Therefore, we ran separate K-means analyses with two, three, and four clusters, for each of the five topics, and inspected the resulting 15 cluster diagrams.

For each topic, the different cluster inputs affected only the clusters below fact-depth = 4 (the sparse and general pages in Figure 2). With two clusters as input, K-means generated only one cluster below fact-depth = 4. With three clusters as input, the clusters were similar to the one shown in Figure 2. Finally, with four clusters as input, K-means generated three clusters below fact-depth = 4. An analysis of the 15 cluster diagrams suggested that three clusters provided the most meaningful groupings of pages for each of the five topics. A higher dimensionality clustering could reveal more interesting clusters in future research.

Figure 3A-E shows the results from the K-means cluster analysis with three clusters for each of the five topics, and Appendix C shows the boundaries, centroids, and percentage of pages for each cluster. In each case, the x-axis (fact-breadth) has been normalized for graphical comparison after the analysis. As shown, the analysis revealed clusters for each of the five topics that were mostly similar in their boundaries to those shown in Figure 2. The right-hand clusters (denoted by "•") in each diagram has fact-breadth > 40% and fact-depth = 1–3. Although the top clusters (denoted by "▲") in each diagram have a very wide range of fact-breadth, all of them are limited to fact-depth = 4. The lower left-hand clusters (denoted by "+") in each diagram is bounded by fact-breadth <= 40%, and fact-depth = 1–3. The above-mentioned boundaries matched the boundaries of the respective general, specific, and sparse clusters for all pages across the five topics shown in Figure 2.

An inspection of the specific clusters revealed the greatest variation in fact-breadth for three topics: doctor's examination was concentrated at 70% of fact-breadth, diagnostic tests was concentrated at 20% of fact-breadth, and disease stage ranged from 10%–80% of fact-breadth. In contrast, the specific cluster for self-examination and risk/prevention had almost the same range of fact-breadth (10%–60%.) We believe this variation in fact-breadth occurred because the former set of topics had fewer specific pages in addition to fewer total pages, compared with the latter set.

Overall, the above cluster analyses results suggest that there are potentially interesting regularities within the scatter of facts across Web pages. These regularities are in the form of three page clusters that vary in information density through a complex interaction between fact-breadth and fact-depth. These page clusters were labeled general pages (containing many facts with medium detail), specific pages (containing a range of facts with one having a lot of detail), and sparse pages (containing few facts in low-to-medium detail). Therefore, pages with few facts consisted mainly of pages from two different clusters (sparse and specific). Because these two page clusters dominate the overall number of pages, they make the distribution of facts across Web pages highly skewed towards few facts.

Although the cluster analysis suggested that the pages can be separated into three different page clusters (based on the number of facts and the amount of text about the facts), we needed to understand the role they played in providing healthcare information. For example, what role did the sparse pages (containing few facts in little detail) play in providing healthcare information? Were sparse pages poorly written, and, if so, why were there so many of them in the top-ten healthcare sites with melanoma information?

Content Analysis: Understanding the Role of Page Clusters

An informal analysis suggested that in addition to differing in information density (variations in fact-depth and fact-breadth), the pages also differed in *topic scope* (defined as how broad or narrow was the content of the page). For example, some pages focused on cancer (a broad scope) with a brief mention of a melanoma fact, while other pages focused on a specific fact, such as the importance of UV protection in the prevention of melanoma (a narrow scope). Could topic scope reveal the role of the page clusters? To answer this question, we analyzed the relationship between page clusters (general, specific, and sparse pages identified in the cluster analysis) and topic scope.

Method

We used the page title to analyze the topic scope of the 336 pages from the five melanoma topics. We analyzed page titles instead of the entire page content because (a) we wished to avoid possible circularities in the event that the two dependent variables (page type, and topic scope) that we were attempting to correlate were actually not independent to each other, and (b) Web page authors of high-quality sites design titles that reflect the content of their pages. Furthermore, a check of all the pages revealed that each of the 336 pages had meaningful page titles (e.g., "Skin Cancer" and "Moles"). To demonstrate the spread of title scope within a topic, Appendix D shows all 105 page titles on the topic of melanoma risk/prevention. To understand variations between each of the topics, the content analysis was repeated with pages for each of the five topics.

Topic scope was operationalized in terms of categories based on a hierarchical skin cancer taxonomy, developed by skin cancer physicians using real world skin cancer questions (Bhavnani, 2003) and similar to the taxonomy developed by Pratt, Hearst, and Fagan (1999). The titles of each page were transcribed into a separate document

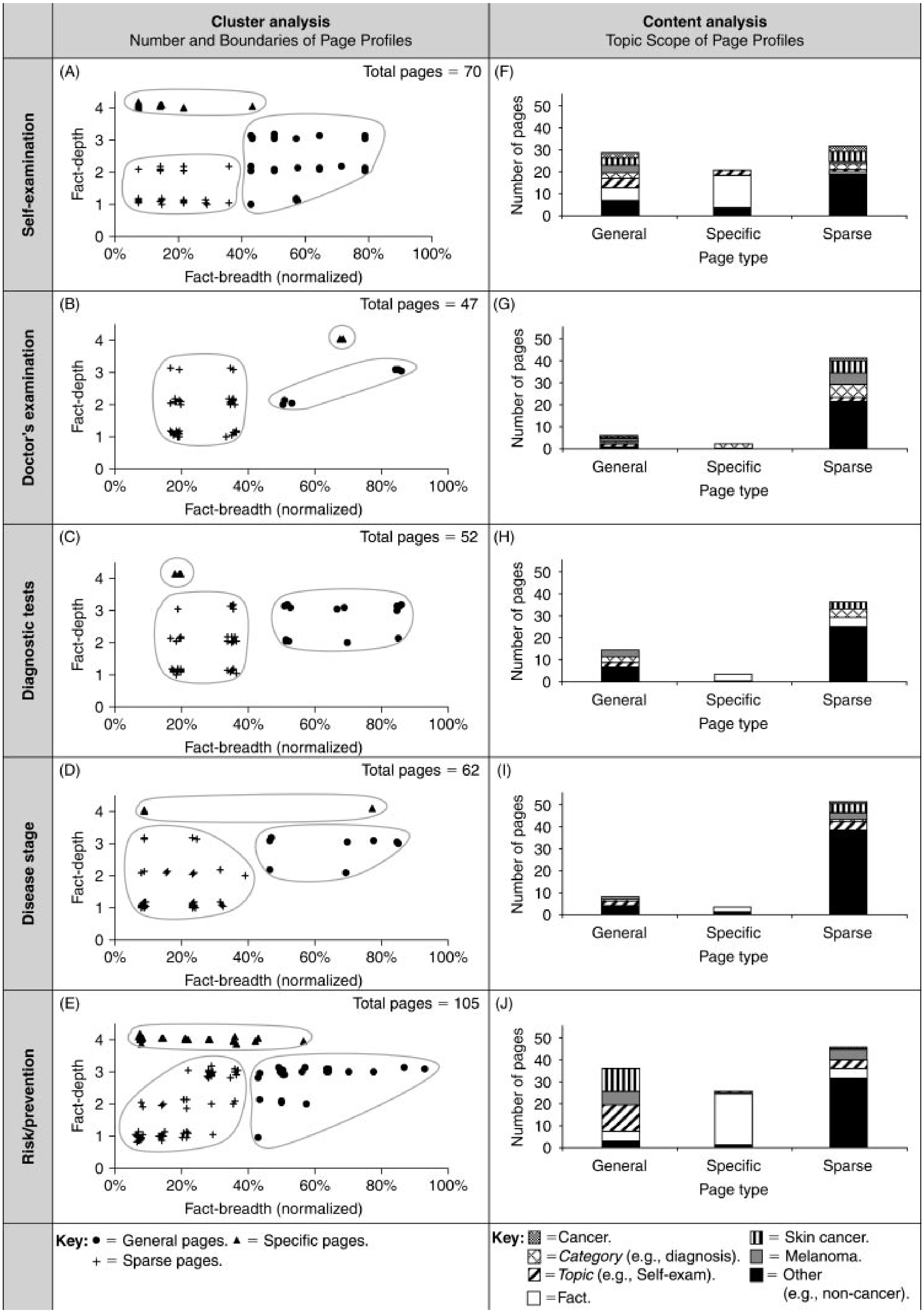


FIG. 3. Cluster analysis (A-E) for the five melanoma topics showing the number and boundaries of the three page types. To enable comparison across the five melanoma topics, the fact-breadth on the x-axis has been normalized, and to enable comprehension of the clusters based on how the raters analyzed the Web pages, the y-axis shows the original ordinal scale used by the human raters to analyze the pages. Content analysis (F-J) of the same five topics showing the relationship of the three page types to topic scope.

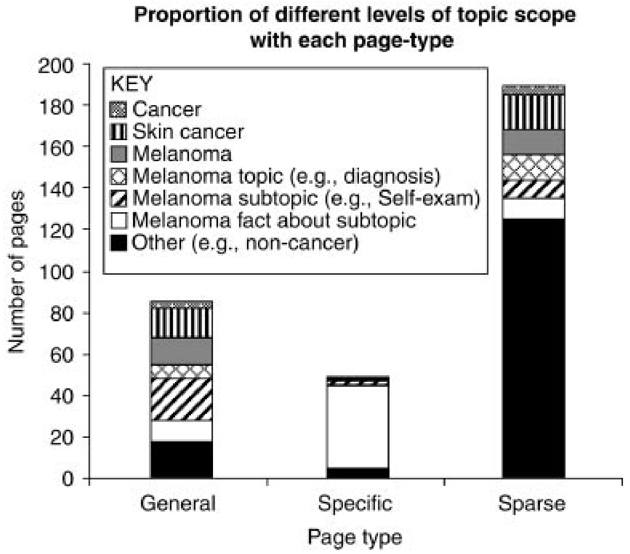


FIG. 4. A content analysis for all pages revealed that the pages in different clusters were written at different levels of topic scope. General pages were written across all levels of topic scope, specific pages were written at mostly the "fact" level, and sparse pages were written mostly for categories outside the taxonomy.

(to avoid bias from the page content) in a random order. A rater categorized each of the 336 page titles into one of the following categories, representing different degrees of *topic scope* going from broad to narrow: A = cancer; B = skin cancer; C = melanoma; D = melanoma topic (e.g., diagnosis); E = melanoma subtopic (e.g., self-examination); F = melanoma fact about subtopic. If a page did not fit into any of the categories, then the rater was asked to use the category G = other categories outside the above-mentioned categories.[6] The reliability of the rater was assessed by requesting a second rater to perform the same categorization on a random selection of 84 titles (25% of all page titles).

Results of Content Analysis

We first describe the content analysis results for all the Web pages collapsed over the five topics. This is followed by the analysis results for pages from each of the five topics.

Analysis of Web pages from all topics. The two raters had substantial[7] agreement (Kappa = 0.68) in their categorization of the page titles. Figure 4 shows the proportion of page titles rated at the different degrees of topic scopes in each page cluster across the five topics. As shown, the three page clusters had different topic-scope profiles. General pages were categorized at many different levels across the taxonomy. Specific pages were most often categorized at the level of a fact (e.g., a fact about melanoma self-examination). Finally, sparse pages were most often categorized as "other" (categories outside the topic scope).

A closer look at the overall content of the pages provided a deeper understanding of the above-mentioned results. General pages (which have many facts in medium amount of detail) play the role of providing an overview of facts about a topic. Figure 5A shows a general page close to the centroid of the general cluster shown in Figure 3E. As shown, the page provides many facts about melanoma risk/prevention, each discussed in a medium amount of detail. We also noticed that although the general pages were similar in information density, they varied in the way that the information was provided. Some general pages presented the facts either in the form of answers to frequently asked questions (FAQs), in short bulleted lists of facts, or with short paragraphs and bolded headings as shown in Figure 5A. However, a few general pages were long online brochures or booklets in PDF format, which discussed overview of facts related to many topics within melanoma (and, therefore, were common across many topics). These variations in page genres (Crowston & Williams, 1997) explain why the general pages had titles that were categorized in a wide range of topic scope.

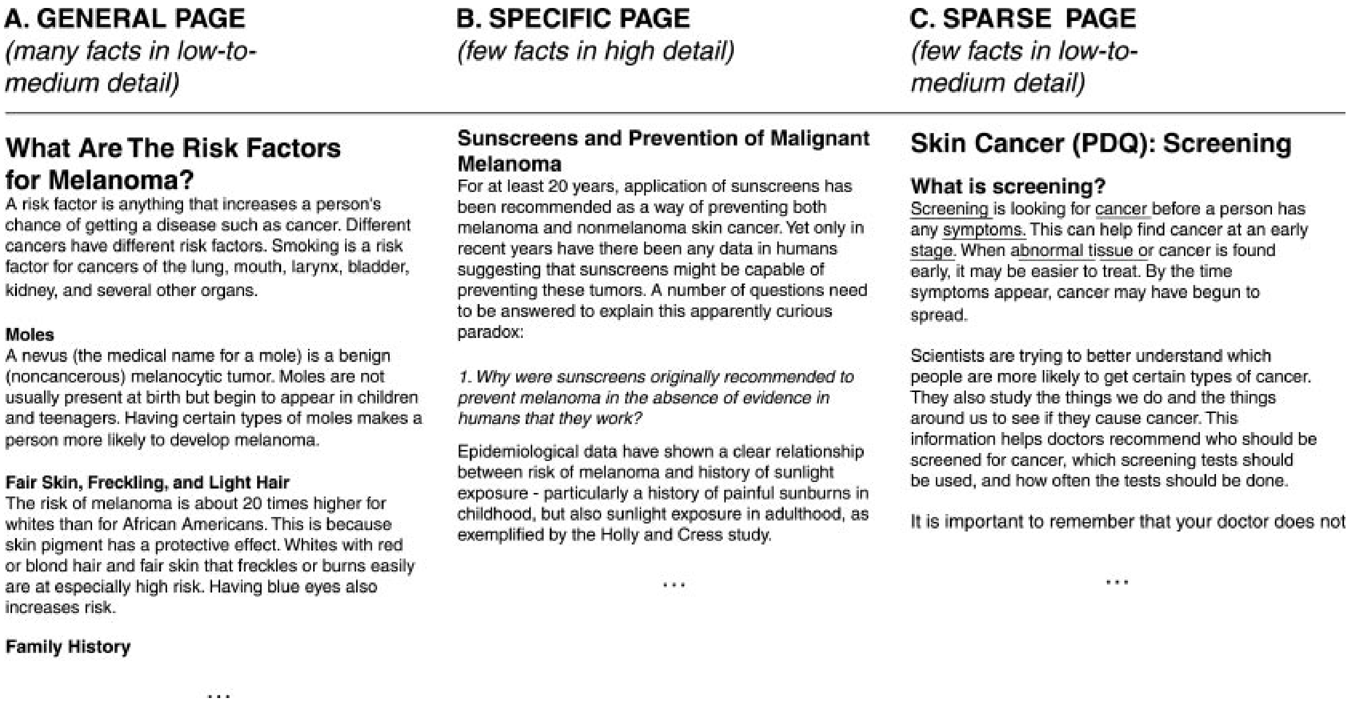
Specific pages (which have one fact in a lot of detail) play the role of elaborating a fact in the context of other facts. Figure 5B shows a specific page at the centroid of the specific cluster in Figure 3E. As shown, the page elaborates details about sunscreen, an important melanoma risk/prevention fact. Finally, sparse pages (which have few facts in little detail) play the role of connecting a relevant fact to other topics. Figure 5C shows a sparse page at the centroid of the sparse cluster in Figure 3E. This page is about skin cancer screening (a topic that is not directly related to melanoma risk/prevention), but it also provides a brief mention of two facts about melanoma risk/prevention.

The content analysis therefore revealed that the three page clusters, besides differing in information density (based on the cluster analysis), also differed in topic scope, which in turn suggested their possible roles. Furthermore, the content analysis suggested that sparse pages are not "poor" or "irrelevant" pages. Rather, sparse pages appear to play the role of briefly mentioning a melanoma fact, which could enable readers to understand the relationship between the main topic focus of the page and a melanoma fact. Finally, the content analysis also provided more evidence that overall, sparse, and general pages do have different content profiles and, therefore, confirms the results suggested by the cluster analyses discussed earlier.

Analysis of Web pages from each topic. To understand variations between topics, the above analysis was repeated for each of the five topics. Figure 3F-J shows the content analysis

[6] For example, if the topic being rated was melanoma risk/prevention, and the sub-topic was descriptive information about risk/prevention, then the other category provided to the raters consisted of the following: Statistical information about risk/prevention, melanoma terminology, melanoma diagnosis, melanoma treatment, melanoma prognosis, non-melanoma skin cancer, non skin cancer, and non-cancer.

[7] Kappa values that exceed 0.6 are considered substantial agreement (Landis and Koch, 1977).



A. GENERAL PAGE
(many facts in low-to-medium detail)

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.

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.

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

...

B. SPECIFIC PAGE
(few facts in high detail)

Sunscreens and Prevention of Malignant Melanoma

For at least 20 years, application of sunscreens has been recommended as a way of preventing both melanoma and nonmelanoma skin cancer. Yet only in recent years have there been any data in humans suggesting that sunscreens might be capable of preventing these tumors. A number of questions need to be answered to explain this apparently curious paradox:

1. Why were sunscreens originally recommended to prevent melanoma in the absence of evidence in humans that they work?

Epidemiological data have shown a clear relationship between risk of melanoma and history of sunlight exposure - particularly a history of painful sunburns in childhood, but also sunlight exposure in adulthood, as exemplified by the Holly and Cress study.

...

C. SPARSE PAGE
(few facts in low-to-medium detail)

Skin Cancer (PDQ): Screening

What is screening?

Screening is looking for cancer before a person has any symptoms. This can help find cancer at an early stage. When abnormal tissue or cancer is found early, it may be easier to treat. By the time symptoms appear, cancer may have begun to spread.

Scientists are trying to better understand which people are more likely to get certain types of cancer. They also study the things we do and the things around us to see if they cause cancer. This information helps doctors recommend who should be screened for cancer, which screening tests should be used, and how often the tests should be done.

It is important to remember that your doctor does not

...

FIG. 5. Example of three pages close to the respective centroid of each cluster shown in Figure 3E. (A) General pages have many facts in low-to-medium detail. The page shown discusses eight risk factors for melanoma in up to two paragraphs each. (B) Specific pages are mostly devoted to one fact (although they may contain other facts). The page shown is mostly devoted to a single prevention fact for melanoma (wearing sunscreen), and it mentions two other relevant risk/prevention facts in very low detail. (C) Sparse pages have few facts covered mostly in less than one paragraph. The page shown is mostly about skin cancer screening, although it briefly mentions three facts relating to melanoma risk/prevention. (Graphics and menus in these pages have been removed for clarity.)

results for each topic. As shown, similar to the results for all pages collapsed across topics, the general pages in each topic were categorized across many different levels in the taxonomy, and sparse pages were most often categorized as “other.” In contrast, there was some variation in the categorization of specific pages. For three topics (self-examination, diagnostic tests, and risk/prevention), specific pages were most often categorized at the “fact” level, which matches the overall categorization. However, the categorization of specific pages in two topics did not follow this pattern. In doctor’s exam (Figure 3G), all the specific pages were categorized at the “category” level, whereas in disease stage (Figure 3I), the specific pages were split nearly evenly between “fact” (2 pages) and “other” (1 page). Similar to the cluster analysis results, we believe these deviations from the overall pattern arose because of the small number of specific pages in each topic (2 and 3 total pages, respectively).

Overall, the analysis for the individual topics showed that the different page clusters differed in topic scope similar to the overall pattern when all the pages were collapsed across the five topics. The general pages were mostly categorized by the raters in a wide range of topic scope, the specific pages were mostly categorized at the “fact” level, and the sparse pages were mostly categorized as “other.” The analysis, therefore, suggest the different roles played by each of the profiles in providing information about a topic: The general pages typically provide an overview of facts, the specific pages mostly provide detail elaboration of a fact, and the sparse pages typically provide a connection between a relevant fact to topics that are not directly relevant to the topic being searched.

Discussion of Results From Cluster and Content Analyses

Similar to earlier studies on information scatter (Hood & Wilson 2001; Bradford 1948), we introduced additional variables to explore the regularities within the scatter. In the cluster analysis, we analyzed how the amount of information about facts (fact-depth) interacted with the number of facts (fact-breadth) in a page to define different page clusters. The analysis suggested the existence of three page clusters with different densities of facts through a complex interaction of fact-depth and fact-breadth. Furthermore, the analysis showed that specific and sparse clusters (both of which had relatively few facts) dominated the overall number of pages, and, therefore, together make the distribution of facts across Web pages highly skewed towards few facts.

However, although the cluster analysis suggested that the pages could be divided into three clusters, we needed to understand what role each type of page played in providing healthcare information. We therefore performed a content analysis with a focus on topic scope, which revealed the role each page cluster played in providing information about a topic. The analysis suggested that general pages provide overviews of topics with content that varied in topic scope, the specific pages provide elaborations of particular facts, and the sparse pages connect the topic being searched to other topics. The large number of pages with few facts, therefore, does not appear to be a random occurrence or the result of pages that have poor content. Instead, the large number of pages with few facts appears to be the result of Web page authors creating pages with two different densities, each playing different roles in providing comprehensive information.

Fact-depth, fact-breadth, and topic scope are by no means the only variables that could be explored. We selected these variables because they appeared to be important in understanding the regularities in the scatter of facts. However, we have noticed other dimensions (such as document length and overlap of pages between topics; Bhavnani & Adamic, 2007; Adamic, Bhavnani, & Xiaolin, 2007) that also appear important and should be explored in future research. Similar to explorations by others such as Hood and Wilson (2001), the analyses that we have conducted are essentially correlations between variables, which is only the first step in trying to understand a phenomenon. Future studies with a wider range of variables should reveal which of them are more significant in explaining regularities within information scatter.

Similar to other information scatter studies, the limitation of this study is the number of topics that could be reasonably analyzed. This limitation is directly related to the difficulty of first identifying a comprehensive list of facts agreed upon by experts, and then accurately identifying those natural language statements in different pages. Although we have attempted to do the latter identification process through computational means (Peck, Bhavnani, Blackmon, & Radev, 2004), the process is only reasonably accurate compared with human raters. We therefore chose to do this matching process manually to understand as accurately as possible the scatter phenomenon. This manual process was time-consuming, resulting in the limited number of topics that we could analyze. Future research should explore computational methods that are more accurate to enable us to rapidly understand regularities underlying scatter in other healthcare topics and how that scatter changes over time.

Implications for the Design of Web Sites, Search Systems, and Training

Although information scatter is an important phenomenon to study in its own right, it hardly matters if most users have no difficulty in finding comprehensive information. However, several studies have shown that despite the use of powerful search engines and access to extensive sites, many users find it difficult to know when they have found all the relevant information and, therefore, when to end a search. In a pre-Web study, Blair and Maron (1985) showed that even expert searchers tended to stop searching before they found comprehensive information. More recently, our analysis has shown that users looking for healthcare information through Google and MedlinePlus (an extensive healthcare portal developed by the National Library of Medicine, and endorsed by experts) typically end their searches prematurely with incomplete information (Bhavnani et al., 2003; Bhavnani et al., 2005b).

The analysis of information scatter suggests that users looking for comprehensive information could benefit by a *general-specific-sparse* search strategy, in which they first read a few general pages (to get an overview of all the facts), followed by specific pages (to get detailed information about specific facts), followed by sparse pages (to understand how the topic being searched might be connected to related topics). Because a similar approach has been recommended by search experts (Kirk, 1974; Bhavnani et al, 2002; Bhavnani et al, 2003; Bhavnani et al., 2005b), we describe, below, how this search strategy has implications for the design of Web sites, search systems, and training.

Implications for the Design of Web Sites

Do Web sites organize their pages such that the general, specific, and sparse pages are directly linked? Such a question is important because traversability (the existence of a path) is a prerequisite to navigability (ability to find a path), which is important for users browsing Web sites to find healthcare information. To address this question, we are currently analyzing how general, specific, and sparse pages about a topic are linked within the top-ten Web sites with melanoma information. Our preliminary analysis has shown that Web sites do not provide links such that they guide users to visit general, specific, and sparse pages.

This suggests that similar to the *Strategy Hub* (Bhavnani et al., 2006), which provides *search procedures* (sequences of links annotated by their purpose) to guide users *across* sites, Web site designers could also provide explicit search procedures that guide users to general, specific, and sparse pages per topic *within* a site. Such approaches could help users quickly get a comprehensive understanding of a topic in vast unfamiliar domains such as healthcare.

Implications for the Design of Search Systems

Although the organization *within* a site is important, the organization of search results *across* sites is also vital to enable users find comprehensive information. As discussed earlier, neither current search engines nor domain portals guide users to deal with information scatter. There have been a few research prototypes that attempt to address information scatter. For example, we manually constructed a portal called the *Strategy Hub for Healthcare* (Bhavnani et al., 2006), which provides search procedures that guide users to visit a combination of general, specific, and sparse pages from reputed healthcare sites. A pilot study (Bhavnani et al, 2003) and a more extensive controlled experiment (Bhavnani et al., 2006) suggested that the search procedures provided in the Strategy Hub enable novice searchers to be more effective in retrieving comprehensive information about a topic when compared with similar users of Google and MedlinePlus. However, the current design of the Strategy Hub requires the manual coding of expert knowledge and, therefore, achieves greater accuracy at the expense of scalability.

In contrast to the above-mentioned approach, researchers have explored more automated and domain-independent approaches. For example, researchers in the Novelty Track (Soboro & Harman, 2003) within the Text Retrieval Conference (TREC) have focused on how to identify a small set of pages that together cover a topic completely with little overlap. Carbonell and Goldstein (1998) have also developed a domain-independent approach called the maximal marginal relevance (MMR) that re-ranks search results from a search engine so that each subsequent hit maximizes novel content. Although MMR reduces the possibility of consecutive hits having high overlapping content, it does not distinguish between novel information of the same fact versus a new fact. Thus, this approach addresses scalability at the expense of organization, i.e., it does not ensure that breadth information is provided before the depth information about specific facts.

The regularities in page profiles that we have observed suggested to us the following *Information Density* algorithm. This algorithm assumes that physicians will pool their knowledge to create a database of facts that they believe patients must know for a comprehensive understanding of specific healthcare topics. When a user selects a topic, such as melanoma risk/prevention, the algorithm will: (a) extract the corresponding list of facts for that topic from the database; (b) retrieve relevant pages for that topic using Google; (c) use content analysis tools such as latent semantic analysis (LSA; Dumais, Furnas, Landauer, & Deerwester, 1998), which can be used to dynamically determine the fact-depth and fact-breadth of each retrieved page; (d) use these calculated values to identify general, specialized, and sparse pages based on boundaries from the cluster analyses; and (e) present pages to the user in a suggested order (such as from general to specific). Our initial studies have revealed that LSA was reasonably accurate compared to a human judge in determining fact-depth and fact-breadth (Peck, et al., 2004), and we are exploring more sophisticated natural language analyses to improve the results. Future research should reveal whether the integrated tool discussed above enables users to retrieve more comprehensive healthcare information about a topic.

Implications for the Design of Training

Although much Web research in the information field has focused on how users search and retrieve information, very little research has focused on how the underlying structure of information can be used to teach users how to search. For example, Bates (2002) has suggested that users should use query searching when information is scattered and use browsing when information is more densely located. The analysis of regularities underlying information scatter at the granularity of facts provides an opportunity to explore how the structure of information can be exploited by searchers. For example, as stated earlier, the regularities in information density discussed in this article suggest that to find comprehensive information about a topic, users should follow a general-specific-sparse search strategy, in which they first read a few general pages (to get an overview of all the facts), followed by specific pages (to get detailed information about particular facts), and then followed by sparse pages (to understand how the topic being searched might be connected to related topics). Such a strategy reduces the probability of missing important information about a search topic, and it is an approach that is independent of features in search tools. Future courses can focus on teaching the declarative knowledge (e.g., the existence of general, specific, and sparse pages, and how to recognize them based on their content profiles) and the procedural knowledge (e.g., the order in which to visit the pages) to execute such a strategy. Regularities in Web *content*, such as described above, and in Web *structure*, such as patterns in link density (Barabasi & Albert, 1999; Huberman & Adamic, 1999; Klienberg & Lawrence, 2001), could together help in the systematic identification and training of search strategies based on regularities underlying information on the Web.

Implications for a Model of Information Scatter

The results from the current study motivated us to begin exploring the process through which the scatter of facts across Web pages occurs. As discussed earlier, the general and specific pages appear to trade off depth and breadth of facts. This trade-off in number and detail of facts suggests a process through which they are generated. Web page authors might be following a process of *accumulation* to progressively add facts in detail to a page until a length and detail *saturation threshold* is reached. At such a threshold (which is parameterized to model different domains), heeding design guidelines that advise authors to keep Web pages short and readable (e.g., Brinck, Gergle, & Wood, 2002), authors might be removing detail from these pages through the process of *abstraction*, in addition to creating new pages to elaborate particular facts in high detail through the process of *specialization*. We speculate that these two processes would lead to the creation of a large number of specific pages, while constraining the total number of general pages. When a fact such as *UV protection* becomes important, page authors might be compelled to add that fact into pages that are otherwise not relevant through the process of *permeation*. This process would lead to the creation of sparse pages. In time, the three processes—accumulation, specialization, and permeation—should lead to a large number of pages with few facts and few pages with many facts, resulting in the skewed distribution of facts across pages. Because this model turns on the concept of a page growing in information density until a saturation threshold is reached, we refer to the above processes collectively as the *saturation model of information scatter*.

The cluster and content analyses, therefore, suggested to us that the scatter of information is not completely random. Rather, it could be the result of a rational process through which the actions of many page authors collectively create the scatter of facts across the pages and sites that we have observed in the data. This conjecture of course needs to be rigorously tested through discussions and observations of real Web page authors, and through the development of a computational model that generates the different distributions. Such an approach would also allow comparisons between alternative models.

Summary and Conclusions

Our overall research to understand how facts are scattered across pages, and the regularities within such scatter, was motivated by the following two observations. (a) Despite the existence of huge Web sites and powerful search engines, novice searchers have difficulty finding comprehensive information about even common topics. (b) Expert searchers visit a combination of sources, often in recognizable sequences, to find comprehensive information. A prior study suggested that such expert behavior emerges because the distribution of facts related to common healthcare topics is skewed towards few facts: A large number of sources have very few facts, while a few sources have many (but not all) facts about a topic (Bhavnani, 2005b). However, neither our previous study nor other information scatter studies could explain why there were so many pages with few facts. We, therefore, conducted a cluster analysis (to understand the density of facts in each page) and a content analysis (to understand the role each type of page played in providing healthcare information) with the goal of probing regularities within the distributions.

A cluster analysis of 336 Web pages relevant to the five melanoma topics suggested regularities within the scatter in form of three page clusters, each with a different information density: General pages contained many facts with a medium amount of detail, specific pages contained few facts, one of which had a high amount of detail, and sparse pages contained few facts with little detail. The large number of pages with few facts in the distribution comprised a disproportionately large number of specific and sparse pages. A content analysis probed the role of the three page clusters in providing information. This analysis suggested that general pages provided overviews of facts, specific pages provided elaborations of particular facts, and sparse pages connected relevant facts to other topics that were distantly related. Each of the three page profiles, therefore, play different but important roles in providing comprehensive information about a topic.

The two analyses together helped us to understand the number, boundary, and role of the different page profiles, each of which helped to deepen our understanding of the regularities underlying the high scatter of facts across pages. Furthermore, the results suggest that the large number of pages with few facts is not a random occurrence or the result of pages that have poor content. Instead, these pages appear to be the result of rational decisions made by Web page authors in creating pages that play distinct roles in providing comprehensive information. These insights led to implications for the design of Web sites and search systems, training to help users find comprehensive information, and a framework by which to organize and direct future scatter studies.

Although prior research on information scatter was the primary inspiration for the current work, the analysis of scatter at the granularity of facts enables a more direct connection to existing information-seeking models and theories. Both the *berrypicking model* (Bates, 1989) and the *Information Foraging Theory* (Pirolli & Card, 1999) deal, in part, with searchers moving from one source to another during the search process. The berrypicking model describes *how* users search through the process of collecting "bits and pieces" of information from many different sources. The Information Foraging Theory predicts *when* a user will abandon searching in one information patch and move to another. The distribution of facts provides an empirical foundation to explain that these behaviors are possibly a necessary adaptation to the underlying structure of information. Furthermore, we believe that expert searchers (such as the healthcare librarian described in the Motivation to Probe Regularities in the Scatter of Facts Across Web Sites section) tend to begin searching by visiting sources that provide an overview of a topic (general sources) before delving deeper into specific details (specific sources), because they (a) have developed an inherent understanding that information occurs in different densities within different sources, and (b) use the general-specific-sparse approach as a *sensemaking*[8] strategy (Russell et al., 1993) to help them to quickly understand how information is organized in an unfamiliar domain.

Although much is known about the scatter of information at different granularities, the main contributions of this article are as follows: (a) to bring attention to the regularities in how facts about common topics are scattered across relevant pages and (b) to explore how that understanding can benefit the design of Web sites, search systems, training, and future research. This understanding should lead to new approaches that enable more users to retrieve comprehensive information when searching in vast and unfamiliar domains like healthcare.

Acknowledgments

This research was supported, in part, by a developmental grant from the University of Michigan's Center of Excellence in Cancer Communications Research (funded by NIH), and the NIH grant # UL1RR024986. The authors thank M. Cervone, G. Furnas, A. Ganesan, T. Johnson, R. Little,

[8] Sensemaking is used here to mean representations to organize needed information in ways that are conducive to performing real-world tasks as proposed by Russell et al. (1993).

S. Mathan, J. Olson, F. Reif, A. Rao., P. Saxman, V. Strecher, B. Suhm, R. Thomas, and G. Vallabha for their contributions.

References

Adamic, L.A., Bhavnani, S.K., & Xiaolin, S. (2007). Scatter networks: A new approach for analyzing information scatter on the Web. New Journal of Physics (Special Issue on Complex Systems), 9, 231.

Aldenderfer, M.S., & Blashfield, R.K. (1984). Cluster analysis. Beverly Hills, CA: Sage Publications.

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., & 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.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(5), 407–424.

Bates, M.J. (2002). Speculations on browsing, directed searching, and linking in relation to the Bradford Distribution. In H. Bruce, R. Fidel, R. Ingwersen, & P. Vakkari (Eds.), Emerging frameworks and methods. Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (pp. 137–150). Greenwood Village, CO: Libraries Unlimited.

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. Journal of Bone and Joint Surgery, 82, 1540–1543.

Bhavnani, S.K. (2001). Important cognitive components of domain-specific search knowledge. In E. M. Voorhees & D. K. Harman (Eds.), NIST special publication 500-250. The Tenth Text Retrieval Conference (pp. 571–578). Washington, DC: NIST.

Bhavnani, S.K. (2005a). The retrieval of highly scattered facts and architectural images: Strategies for search and design. Automation in Construction, 14(6), 687–776.

Bhavnani, S.K. (2005b). Why is it difficult to find comprehensive information? Journal of the American Society for Information Science and Technology, 56(9), 989–1003.

Bhavnani, S.K., & Adamic, L.A. (2007, February). Making sense of information scatter on the Web. Paper presented at the Human-Computer Interaction Consortium Workshop, Fraser, CO.

Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., et al. (2003). Strategy hubs: Next-generation domain portals with search procedures. In G. Cockton & P. Korhonen (Eds.). Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (pp. 393–400). New York: ACM Press.

Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., et al. (2006). Strategy hubs: Domain portals to help find comprehensive information. Journal of the American Society of Information Science and Technology, 57(1), 4–24.

Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., et al. (2002). Getting patients to the right healthcare sources: From real-world questions to Strategy Hubs. In Proceedings of the AMIA 2002 Annual Symposium (AMIA'02), San Antonio, TX (pp. 51–55).

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.

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness. Communications of the ACM, 28, 289–299.

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

Brinck, T., Gergle, D., & Wood, S. (2002). Designing Websites that work: Usability for the Web. San Francisco: Morgan Kaufmann.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, Diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 335–336). New York: ACM Press.

Chen, Y., & Leimkuhler, F.F. (1986). A relationship between Lotka's Law, Bradford's Law and Zipf's Law. Journal of the American Society for Information Science, 37(5), 307–314.

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.

Crowston, K., & Williams, M. (1997). Reproduced and emergent genres of communication on the World-Wide Web. In Proceedings of the Hawaii International Conference on System Sciences, Maui HI.

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

Dumais, S.T., Furnas, G.W., Landauer, T.K., & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In E. Soloway, D. Frye, & S.B. Sheppard (Eds.). Proceedings of the ACM SIGCHI Proceedings of the Conference on Human Factors in Computing Systems (pp. 281–285). New York: ACM Press.

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.

Figueiredo, M.A.T., & Jain, A.K. (2002). Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3), 381–396.

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. Retrieved July 3, 2003, from http://www.pewinternet.org/reports/toc.asp?report=95

Griffiths, K.M., & Christensen, H. (2000). Quality of web based information on treatment of depression: Cross sectional survey. British Medical Journal, 321, 1511–1515.

Hartigan, J.A. (1975). Clustering algorithms. New York: Wiley.

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.

Huberman, B.A., & Adamic, L. (1999). Growth dynamics of the World-Wide Web. Nature, 401, 131.

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. British Medical Journal, 314, 1875–1879.

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

Kirk, T. (1974). Problems in library instruction in four-year colleges. In J. Lubans (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.

Lancaster, F.W., & Lee, J.-L. (1985). Bibliometric techniques applied to issue management: A case study. Journal of the American Society for Information Science, 36(6), 389–397.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174

Larson, K., & Czerwinski, M. (1998). Web page design: Implications of memory, structure and scent for information retrieval In C.M. Karat, A. Lund, J. Coutaz, & J. Karat (Eds.). Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (pp. 25–32). New York: ACM Press.

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

McGlynn, E., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., et al. (2003). The quality of health care delivered to adults in the United States. New England Journal of Medicine, 348(26), 2635–2645.

Over, P. (1998). TREC-6 Interactive track report. In E.M. Voorhees & D.K. Harman (Eds.), NIST special publication 500-242. The Seventh Text Retrieval Conference. Washington, DC: NIST. Retrieved January 12, 2010, from http://maroo.cs.umass.edu/pub/web/getpdf.php?id=34

Peck, F.A., Bhavnani, S.K., Blackmon, M.H., & Radev, D.R. (2004). Exploring the use of natural language systems for fact identification: Towards the automatic construction of healthcare portals. In Managing and enhancing information: Cultures and conflicts. Proceedings of the 2004 annual meeting of the American Society for Information Science and Technology, Providence, RI.

Pirolli, P., & Card, S.K. (1999). Information Foraging. Psychological Review, 106, 643–675.

Pratt, W., Hearst, M., & Fagan, L. (1999). A knowledge-based approach to organizing retrieved documents. In Proceedings of the Sixteenth National Conference on Artificial Intelligence, Orlando, FL.

Russell, D.M., Stefik, M., Pirolli, P., & Card, S.K. (1993). The cost structure of sensemaking. In Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E., & White, T. (Eds.), (pp. 269–276). Proceedings of the ACM CHI 93 Human Factors in Computing Systems Conference. Amsterdam, The Netherlands:

Soboro, I., & Harman, D. (2003). Overview of the TREC 2003 novelty track. In E. M. Voorhees & L. P. Buckman (Eds.), NIST special publication 500–255. The 12th Text Retrieval Conference. Washington, DC: NIST. Retrieved January 12, 2010, from http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf

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

Sturdee, D.W. (2000). The importance of patient education in improving compliance. Climacteric, 10(2), 9–13.

Tenopir, C. (1982). Evaluation of database coverage: A comparison of two methodologies. Online Review, 6, 423–441.

Van Halteren, H., & Teufel, S. (2003). Examining the consensus between human summaries: Initial experiments with factoid analysis. In Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (pp. 57–64). Morristown, NJ: Association for Computational Linguistics.

Woodruff, A., Landay, J., & Stonebraker, M. (1998). Constant information density in zoomable interfaces. In T. Catarci, M. F. Costabile, G. Santucci, & L. Taranfino (Eds.). Proceedings of Advanced Visual Interfaces (pp. 57–65). New York: ACM Press.

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

Appendix A

The MML (Figueiredo & Jain, 2002) criterion finds an optimum number of clusters by balancing the cost of having multiple cluster centroids and the cost of the deviation of each data point from those centroids. When there are too many clusters, the centroid cost is high, but the deviation cost tends to be small. Conversely, when there are few clusters, the centroid cost is low but the cost of deviations from those centroids will tend to be large.

Because the input data were discrete, random jitter (drawn from a zero-mean Normal distribution) was added to each data point at different levels; this constituted the input data for a single "run," and there were 50 such runs. For each run, MML was used to calculate the optimal number of clusters. As shown in Table A1, for each level of jitter, we calculated how often a particular number of clusters was chosen as optimal. As shown, three or four clusters were optimal for a majority of runs. The difference between the three-cluster and four-cluster configuration was slight (essentially whether the jitter was able to sufficiently counteract the discreteness of the data).

TABLE A1. Number of clusters based on the amount of random jitter.

		Amount of Random Jitter									
		0.002	0.005	0.007	0.02	0.05	0.07	0.1	0.2	0.3	
No. of Clusters	1	0	0	0	0	0	0	0	0	40	40
	2	0	0	0	0	0	0	4	24	9	37
	3	0	1	2	7	40	43	45	14	0	152
	4	36	36	36	28	7	5	0	9	1	158
	5	14	10	11	11	2	2	1	1	0	52
	6	0	3	1	4	1	0	0	2	0	11
		50	50	50	50	50	50	50	50	50	

DOI: 10.1002/asi

Appendix B

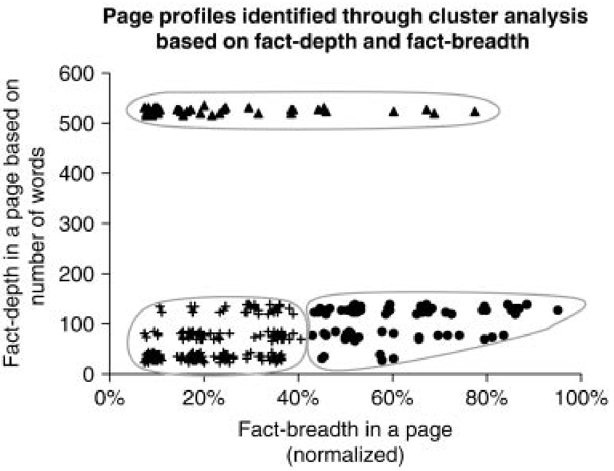


FIG. 6. Results of the cluster analysis that generated the boundaries of three page clusters: specific pages (▲), sparse pages (+), and general pages (●). The Y-axis is mapped to the mean number of words for each level of fact-depth. To create the above figure, random jitter was added to each point to enable the separation of overlapping points.

Appendix C

The boundary, centroid, and percentage of pages of the general, specific, and sparse page clusters for each of the five melanoma topics, as shown in Figure 3, and for all the topics collapsed, as shown in Figure 2. Fact-breadth is shown as a normalized percentage to enable a comparison across topics.

	Self-examination Total facts = 14	Doctor's examination Total facts = 6	Diagnostic tests Total facts = 6	Disease stage Total facts = 13	Risk/prevention Total facts = 14	All topics collapsed
General page cluster						
Boundary (fact-breadth range, fact-depth range)	43–79%, 1–3	50–83%, 2–3	50–83%, 2–3	46–85%, 2–3	43–93%, 1–3	43–93%, 1–3
Centroid (mean fact-breadth, mean fact-depth)	56.9%, 2.3	66.7%, 2.5	65.5%, 2.6	65.4%, 2.8	57.4%, 2.8	60.1%, 2.6
Pages (number of pages, % of total pages)	25, 35.7%	6, 12.8%	14, 26.9%	8, 12.9%	27, 25.7%	80, 23.8%
Specific page cluster						
Boundary (fact-breadth range, fact-depth range)	7–43%, 4	67%, 4	17%, 4	8–77%, 4	7–57%, 4	7–77%, 4
Centroid (mean fact-breadth, mean fact-depth)	11.1%, 4	66.7%, 4	16.7%, 4	30.8%, 4	20.5%, 4	19.3%, 4
Pages (number of pages, % of total pages)	18, 25.7%	2, 4.3%	3, 5.8%	3, 4.8%	23, 21.9%	49, 14.6%
Sparse page cluster						
Boundary (fact-breadth range, fact-depth range)	7–36%, 1–2	17–33%, 1–3	17–33%, 1–3	8–38%, 1–3	7–36%, 1–3	7–38%, 1–3
Centroid (mean fact-breadth, mean fact-depth)	16.9%, 1.3	23.9%, 1.6	24.3%, 1.6	15.2%, 1.3	19.9%, 1.8	19.9%, 1.5
Pages (number of pages, % of total pages)	27, 38.6%	39, 83.0%	35, 67.3%	51, 82.3%	55, 52.4%	207, 61.6%

Appendix D (Titles for 105 pages on the topic Melanoma Risk/Prevention)

1. Skin cancer
2. Malignant melanoma fact sheet
3. Malignant melanoma
4. Kids connection: The ABCDs of skin cancer
5. Ultraviolet index: What you need to know
6. Moles
7. Atypical nevus
8. Facts about sunscreens
9. Personal skin cancer risk profile
10. Dermatologic surgery
11. Who is most at risk for melanoma?
12. [MelanomaNet] > Prevention
13. MelanomaNet update: Recurrent and metastatic melanoma
14. Basic facts about melanoma
15. MelanomaNet update: Skin self-examination is a family affair
16. [MelanomaNet update] > Research update
17. MelanomaNet update: Precursor lesions and risk factors for melanoma
18. Risk factors update: Dysplastic nevi (atypical moles) as risk factors for melanoma
19. MelanomaNet update: Questions about benign pigmented lesions
20. [MelanomaNet] > Medical diagnosis: Dermatoscopy
21. Skin cancer
22. Aging skin update-September 2001: Protection against photoaging
23. Preventing actinic keratoses by protecting yourself against the sun
24. What are the risk factors for melanoma?
25. Can melanoma be prevented?
26. What is melanoma skin cancer?
27. What is melanoma?
28. Can melanoma skin cancer be prevented?
29. Can melanoma be found early?
30. Skin cancer
31. UV radiation and cancer
32. Skin cancer facts
33. Are some people more susceptible to sun damage?
34. [Cancer Prevention > Sun safety] > Sunlight and ultraviolet radiation
35. Melanoma skin cancer
36. Melanoma: treatment guidelines for patients
37. What are the risk factors for eye cancer?
38. What are the risk factors for vulvar cancer?
39. Vulvar cancer
40. Nonmelanoma skin cancer
41. Testicular cancer
42. [Cancer Prevention > Sun safety] > Take the sun quiz
43. [Your cancer risk > melanoma] > Preliminary questions
44. Skin cancer facts
45. Mohs micrographic surgery: a handbook for patients
46. Melanoma facts
47. [What is melanoma] > Risk factors
48. [Melanoma prevention] > Sun safety
49. [Melanoma prevention] > Am I at risk?
50. [Speaking with a doctor about melanoma] > Frequently asked questions
51. What causes melanoma?
52. Play it safe with the sun: a guide to protecting yourself from the sun and skin cancer
53. If you are concerned about melanoma…. Information about diagnosis and treatment
54. What you need to know about moles and dysplastic nevi
55. Melanoma (PDQ) treatment
56. Skin cancer (PDQ) screening
57. What you need to know about skin cancer
58. Skin cancer (PDQ) prevention
59. Taking part in clinical trials: Cancer prevention studies: What participants need to know
60. At risk for melanoma
61. Summer: open season for melanoma
62. Number, size, and type of moles are key risk factors for melanoma
63. [Melanoma > Other melanoma resources and articles] > All in the family
64. Guidelines for melanoma-prone families
65. [Melanoma] > What to look for
66. Bad to worse: melanoma increases again
67. [The skin cancer foundation] > Children
68. The case against indoor tanning
69. What you really need to know about moles and melanoma
70. What you really need to know about moles and melanoma
71. [Melanoma and skin cancer information] > Treatment
72. Occupational health and safety policy for outdoor workers exposed to ultraviolet radiation and seasonal heat
73. [Melanoma and skin cancer information] > Frequently asked questions
74. What are the risk factors for cancer of unknown primary?
75. [Melanoma prevention] > Reducing your risk
76. Mature skin
77. The darker side of tanning
78. Mature skin
79. The sun and your skin
80. [AcneNet > Treatment] > Over the counter products
81. What is aging skin?
82. Skin cancer
83. [MelanomaNet] > Medical diagnosis

84. Self-examination for melanoma
85. What causes melanoma skin cancer?
86. Detecting skin cancer
87. Sunlight and Ultraviolet exposure
88. Cancer prevention and early detection facts and figures 2002
89. Ultraviolet light
90. Types of cancer: Melanoma: The basics
91. [Melanoma prevention] > Sun safety
92. Melanoma prevention
93. [Melanoma prevention] > Sun safety for kids
94. [Diagnosing melanoma] > Self-examination
95. Facing forward series: Ways you can make a difference in cancer: resources to learn more
96. Skin cancer treatment
97. Unusual cancers of childhood (PDQ) treatment
98. Taking part in clinical trials: Cancer prevention studies: What are cancer risk factors?
99. [The skin cancer foundation] > Frequently asked questions
100. A meeting of the minds: Highlights of the melanoma program at the Seventh world congress on cancers of the skin
101. Low socioeconomic status may increase melanoma mortality
102. Sunscreens and prevention of malignant melanoma
103. Increased melanoma risk from indoor UVA tanning
104. [The skin cancer foundation > Prevention] > Dress for sun protection
105. [The skin cancer foundation] > Self-examination

DOI: 10.1002/asi