Towards a Model of Information Scatter: Implications for Search and Design

Suresh K. Bhavnani[1], Frederick A. Peck[2]
School of Information, University of Michigan, Ann Arbor, MI 48109-1092
{bhavnani[1], peckf[2]}@umich.edu

ABSTRACT
Recent studies suggest that users often retrieve incomplete healthcare information because of the complex and skewed distribution of facts across relevant webpages. To understand the *causes* for such skewed distributions, this paper presents the results of two analyses: (1) A distribution analysis discusses how facts related to healthcare topics are scattered across high-quality healthcare pages. (2) A cluster analysis of the same data suggests that the skewed distribution can be explained by the existence of three page profiles that vary in information density, each of which play in important role in providing comprehensive information of a topic. The above analyses provide clues towards a model of information scatter which describes how the design decisions by individual webpage authors could collectively lead to the scatter of information as observed in the data. The analyses also suggest implications for the design of websites, search algorithms, and search interfaces to help users find comprehensive information about a topic.

Author Keywords
Healthcare, searching, Webometrics, distributions.

INTRODUCTION
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). The development of extensive healthcare portals such as MEDLINEplus, coupled with powerful search engines like Google, suggests that finding such comprehensive healthcare information is relatively easy. However, while users of search engines and domain portals 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 (e.g. "What are the risk and prevention factors for melanoma?") (Bhavnani, 2001).

One clue to the above difficulty is provided by expert healthcare searchers who know which *combination* of sites to visit in which *order* (Bhavnani, 2001; Bhavnani et al., 2006). A recent study suggests that such expert behavior emerges because the distribution of facts related to common healthcare topics is skewed: a large number of sources have very few facts, while a few sources have many, but not all, facts about a topic (Bhavnani, 2005). From a searcher's perspective, such distributions mean that information is highly scattered, and therefore requiring sophisticated search strategies to find comprehensive information. However, because little is known about the causes underlying such scatter of information, few solutions have been proposed.

Because the retrieval of incomplete healthcare information can lead to dangerous consequences, this paper attempts to shed light on understanding the nature and causes of information scatter, and its implications to designers of websites, search algorithms, and interfaces. The paper begins by discussing the results from two analyses: (1) A *Distribution Analysis* (reported previously) briefly describes how facts related to healthcare topics are scattered across high-quality healthcare pages, (2) A *Cluster Analysis* of the above healthcare pages suggests that webpage authors trade-off the depth and breadth of facts to create pages with three distinct information densities. The above results suggest a model of information scatter which could explain through future analyses, how the design decisions by individual webpage authors can collectively lead to the scatter of information as observed in the data. The analyses provide implications for the design of websites, search algorithms, and search interfaces to help users find comprehensive information in healthcare and other domains.

THE DIFFICULTY OF FINDING COMPREHENSIVE HEALTHCARE INFORMATION

Several studies have shown that novice healthcare searchers typically rely on general-purpose search engines to find relevant pages (Eysenbach and Kohler, 2002), go online without a definite search plan so that they find most sites accidentally (Fox and Fallows, 2003), and often end their searches early with incomplete information (Bhavnani, 2001; Bhavnani et al., 2006). In contrast, expert searchers (such as healthcare reference librarians) know which sites to visit in which sequence. For example, in a recent study (Bhavnani, 2001) an expert healthcare searcher 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 expert healthcare searchers to find comprehensive information quickly and effectively, compared to novices who were unable to infer such knowledge by just using Google (Bhavnani, 2001).

What motivates an expert to visit many different sites to find healthcare information, and why is it difficult for novices to do the same?

DISTRIBUTION ANALYSIS: DATA COLLECTION AND PRIOR RESULTS

Our recent study (Bhavnani, 2005) suggests why finding comprehensive information is difficult. The study consisted of two inter-rater experiments (whose data collection are briefly described here because of its relevance to the analyses in the rest of the paper). 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) based on a taxonomy of real-world questions (Bhavnani et al., 2002), similar to the taxonomy developed by Pratt et al. (1999). The facts were rated 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.

The second inter-rater experiment analyzed how the facts identified by the physicians were distributed across relevant pages from the top ten websites 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 graduate students were asked to independently rate on each page 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. 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. Both the above experiments had high inter-rater agreement (see (Bhavnani, 2005) for details).

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

These results are in agreement with numerous studies that show that many online healthcare sites provide inaccurate or incomplete information (see (Eysenbach et al., 2002) for a review). Additionally, the results describe how the information is distributed *across* pages and sites.

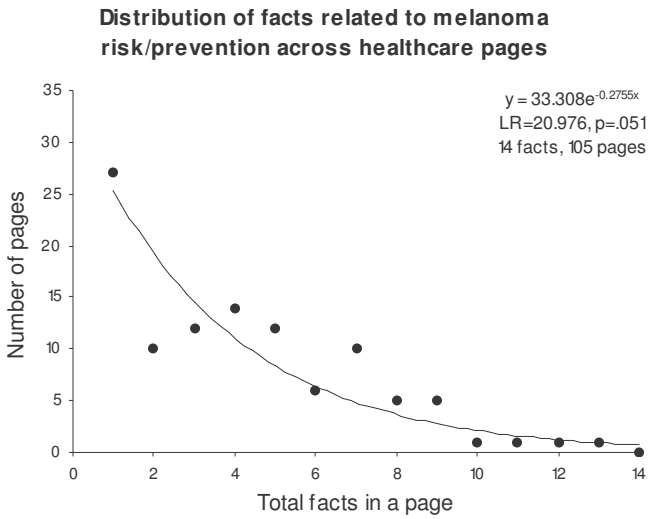

Figure 1. The distribution of risk/prevention facts across relevant pages in high-quality sites is highly skewed (best-fitted by a discrete exponential curve, Likelihood Ratio=20.967, p=.051 where significant fit is >.05), with no page containing all the facts (Bhavnani, 2005).

Repercussions of the skewed distribution of healthcare information

The above study sheds light on the complex environment often encountered when searching for comprehensive information. Searchers must visit a combination of pages and websites to find all the facts about a topic. Furthermore, because there are many more pages that contain only few facts, there is a high probability that users will find such pages and end their searches early. As neither search engines, nor domain portals address this problem, users have difficulty knowing when they have found all the relevant information, and often terminate their searches early with incomplete information (Bhavnani, 2001).

Because the retrieval of incomplete information can affect healthcare decisions and treatment compliance (Sturdee, 2000), the above scatter of information needs careful investigation. While several studies have analyzed the scatter of information at different levels of granularity (articles of a topic across journals [Bradford, 1948], and across databases [Hood and Wilson, 2001]), none have analyzed the scatter of facts across webpages, and little is known about the possible causes for such scatter. An informal analysis of the pages suggested the existence of different page profiles in the distribution. For example, some pages had few facts with a lot of detail, while others had many facts in a little detail. Could such page profiles explain the skewed distribution of facts across pages?

CLUSTER ANALYSIS: EXPLORING THE CAUSES OF INFORMATION SCATTER

Our goal was to rigorously analyze whether different page profiles could explain the skewed distribution of facts across webpages. We first describe the analysis of melanoma risk/prevention pages, and then show how those results generalize across the five melanoma topics.

Method

To identify the page profiles, we used cluster analysis to automatically cluster the webpages according to depth and breadth of fact coverage. *Fact-depth* of a page was defined as the maximum depth of any relevant physician-identified fact on that page (as rated by the judges in our previous study [Bhavnani, 2005]). *Fact-breadth* of a page was defined as the total number of facts for a topic that occurred on that page (also determined from the data collected from our earlier study).

The cluster analysis was done in the following two steps:

1. Estimate number of clusters. We used the Minimum Message Length[1] (MML) criterion (Figueiredo and Jain, 2002) to estimate the optimum number of clusters based on fact-depth and fact-breadth. 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 was done by selecting a random selection of pages from our dataset, and calculating the mean number of words devoted to the relevant facts at each level of detail (1 mapped to 23.93 words, 2 mapped to 66.07 words, 3 mapped to 119.73 words, and 4 mapped to 513.57 words). MML was run for 1-5 clusters.

2. Identify boundaries of clusters. We used the K-means algorithm (SPSS, version 11.5) to determine the cluster boundaries. Inputs to K-means were the same two variables used for MML (fact-depth and fact-breadth), with number of clusters provided by MML.

Results

The lowest MML value was obtained for three clusters. This meant that three clusters best characterized the data. This result was used to determine the subsequent cluster boundaries using K-means.

Page Profiles. Figure 2 shows the results from the K-means cluster analysis for *only* the 105 melanoma risk/prevention pages (the same dataset shown in Figure 1). As shown, the lower left-hand cluster (shown as plus signs) is bounded by fact-breadth=1-5, and fact-depth=1-3. These pages are labeled *sparse* as they have few facts relevant to the topic in mostly low levels of detail. The right-hand cluster (shown as dots) has fact-breadth=6-13 and fact-depth=1-3. These pages are labeled *general* as they have many more facts relevant to the topic in mostly medium levels of detail. The top cluster (shown as solid triangles) fact-breadth=1-8, but is limited to fact-depth=4. These pages are labeled *specific* as they have a lot of detail about at least one fact.

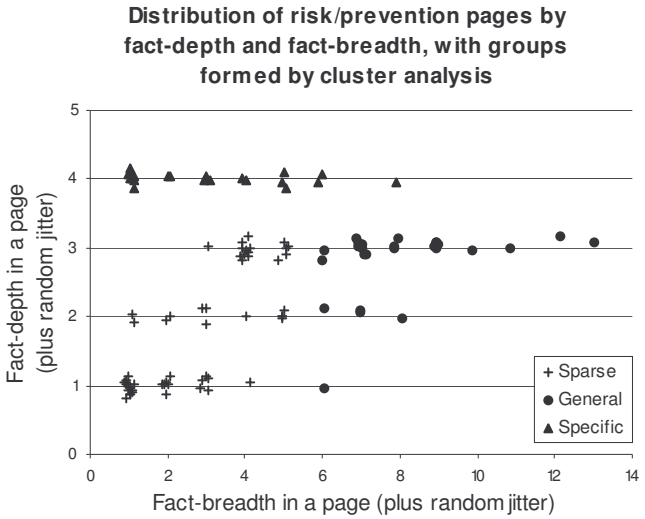

Figure 2. A cluster analysis, with three clusters as input, shows the existence of three page profiles. Sparse pages having fact-breadth<=5 and fact-depth=1-3. General pages having fact-breadth>5 and fact-depth=1-3. Specific pages having fact-depth=4 and any number for fact-breadth.

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

Explaining the Skewed Distributions. Figure 2 shows that there is a higher percentage (74%) of specific and sparse pages (both of which contain relatively low number of facts). These pages constitute the left part of the distribution shown in Figure 1. In comparison, there is a smaller percentage (26%) of general pages (which contain a high number of facts). The distribution of facts across the risk/prevention pages is skewed towards few facts because there are many more specific and sparse pages, each of which have a low mean number of facts (sparse: μ=2.78, specific: μ=2.87).

Generality of the Page Profiles. To test the generality of the cluster analysis results, we repeated the distribution analyses for all the 336 pages retrieved across the five melanoma topics mentioned earlier, and then repeated the cluster analysis for all the topics collapsed. The overall distribution was also skewed (best fitted by a discrete exponential curve, $y=142.736e^{-3.54x}$). Because the number of facts for each topic ranged from 6-14, fact-breadth for each topic was normalized from 0-1. As shown in Figure 3, the analysis revealed clusters that were virtually identical in proportion and boundaries to those identified for the risk/prevention pages (61.6% sparse pages with fact-breath<=40% of total facts, and fact-depth=1-3; 23.8% general pages with fact-breath>40% and fact-depth=1-3; 14.6% specific pages have any number of facts and fact-depth=4). Furthermore, a more recent study (Bhavnani, 2003) of architectural images across high-quality image databases found a similar pattern of pages across those sites. The existence of general, specific, and sparse pages therefore appears to be a broad phenomenon across the five melanoma topics, and across two domains.

Insights into the Role and Creation of Page Profiles

In addition to providing a plausible explanation for the skewed distributions, the cluster analysis also provided insights into the *role* of each page profile, and the *process* through which they might be created.

Role of Page Profiles

Analysis of the page contents in each cluster provided insights about the role of each profile. General pages typically provided overviews of topics in the form of bulleted descriptions of different facts, or frequently asked questions (FAQ) that covered the facts. Figure 4A shows an example of a general page, which was rated by the judges as having 12 facts with a maximum of 2 paragraphs about each fact. In contrast, specific pages typically provided detailed elaborations of a few facts. Figure 4B shows an example of a specific page, which was rated by the judges as having one fact (high UV exposure) elaborated for most of the page.

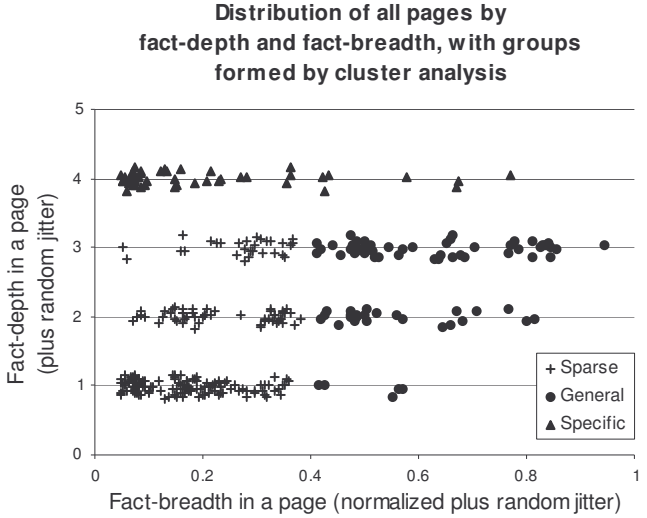

Figure 3. A cluster analysis (with three clusters as input) of pages from all five topics shows boundaries for three page profiles. Sparse pages have fact-breath<=40% and fact-depth=1-3. General pages have fact-breath>40% and fact-depth=1-3. Specific pages have any number of facts and fact-depth=4.

A. General page	B. Specific page	C. Sparse page
What Are The Risk Factors for Melanoma? A risk factor is anything that increases a person's chance of getting a disease … **Moles** A nevus (the medical name for a mole) is a benign (noncancerous) melanocytic tumor. Moles are not usually present at birth… Having a dysplastic nevus, or atypical mole increases a person's risk of melanoma … **Fair Skin, Freckling, and Light Hair** The risk of melanoma is about 20 times higher for whites than for African Americans…	**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… ***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…	**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… **Indications for Skin Surgery** Dermatologists cite four reasons for performing skin surgery: 1) to establish a definite diagnosis with a skin biopsy… **Types of Skin Cancer** Malignant melanoma is the least common but most serious form of skin cancer…

Figure 4. Examples of the text in three different webpage profiles (graphics, links, and font variations have been removed for clarity). General pages (A) have many facts in one or two paragraphs each, specific pages (B) have only one fact covered across the entire page, and sparse pages (C) have only one fact covered in less than one paragraph.

While the purpose of general and specific pages was intuitively clear, we did not immediately understand the purpose of sparse pages. An analysis of the sparse pages revealed that they contained information about topics *outside* (e.g. non-cancerous skin problems) the topic of melanoma risk/prevention. Figure 4C shows an example of such a page, which was rated by the judges as having one fact described in one sentence. Such pages therefore appear to play the role of briefly mentioning a melanoma fact to enable readers make the connection between the main topic focus of the page (e.g. non-cancerous skin problems), and a melanoma fact (e.g. dangers of UV radiation).

Given that the general and specific pages tended to provide facts about a single topic (e.g., melanoma risk/prevention), and the sparse pages showed how those facts related to other topics, we expected these pages to be linked together within a site, to enable a user to easily navigate between them. However, our analysis has revealed that very few general pages provide direct links to specific and sparse pages. For example, while the Skin Cancer Foundation website had 3 general pages, 8 specific, and 5 sparse pages about melanoma risk/prevention, only 1 general page directly linked (either through content or menus) to 2 specific and 1 sparse pages.

Towards a Model of Information Scatter

The cluster analysis also provided three clues about the *process* through which the actions of web authors might be creating such scatter:

1. *Few pages with high depth and breadth*. The right-hand cluster in both diagrams shows that there exist very few pages that have both high breadth and high depth. One plausible explanation is that a rational web author avoids creating pages that have both high depth and breadth because such pages tend to be very long. For example, a page with 14 risk prevention facts, each discussed in detail, can far exceed the space provided by a computer screen, leading to the need of scrolling to read all the facts.

2. *Many pages that trade-off depth and breadth*. The existence of general pages that have more breadth than depth, and the existence of specific pages with the opposite profile suggest that authors create pages through a trade-off between fact depth and fact breadth. For example, an author might add facts with detail to a page until it exceeds a threshold, at which point detail of facts is removed from the long page (resulting in shorter general pages due to the shorter description of each fact), and moved into new pages where the elaboration can occur (resulting in specific pages containing high detail of a few facts).

3. *Many pages about related topics*. The existence of sparse pages, (which contain information about related topics interspersed with a few melanoma facts in low detail) suggests that authors introduce important and well-known facts into existing pages of other topics to show key relationships between topics. For example, a page author might feel compelled to mention the commonalities between preventing melanoma and preventing non-cancerous skin diseases such as acne.

The cluster analysis therefore suggests that the scatter of information might not be a random process. 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 pages and sites that we have observed in the data. We therefore propose the *Information Scatter Model* to begin to explore whether the collective actions of rational webpage authors could result in the scatter of information as observed in the data.

Our approach is similar to other models that attempt to explain the high degree of self-organization through the decentralized actions of many actors. For example, the *rich gets richer* model (Barabasi and Albert, 1999) explains how the actions of many authors of new websites, each choosing to link to existing popular sites, results in a highly skewed distribution of incoming links: a few websites have a large number of incoming links, and a large number of websites have few incoming links. While the above research has revealed much about the *structural* qualities of the web, our model focuses on explaining the self-organization of *content* across webpages.

We based the main sub-processes of our model on inferences from the data. However, (as is common in the development of most process models) these inferences were combined with intuitions about the decision-making process of a rational webpage author in order to fully operationalize the model. The following model is proposed as a hypothesis which needs to be rigorously tested computationally in future research.

As shown schematically in Figure 5, we hypothesize that the Information Scatter Model consists of five processes: (1) Facts about a topic are introduced into the world through the process of *generation,* which includes various social processes such as clinical trials. (2) A webpage author in each site decides whether or not to add a fact in high detail to pages through the process of *accumulation*. This process results in pages that vary in fact depth and breadth in each site. (3) When any page in a site exceeds a threshold for the number of facts on a page, the author decides to remove detail of each fact from that page through the process of *abstraction.* This process creates pages that have high fact-breadth and low fact-depth. (4) Concurrent to the process of abstraction, authors add facts in high detail to new pages through the process of *specialization*. This process creates new specific pages. (5) Authors add facts in low detail to existing pages on a site, through the process of *permeation*, to create sparse pages.

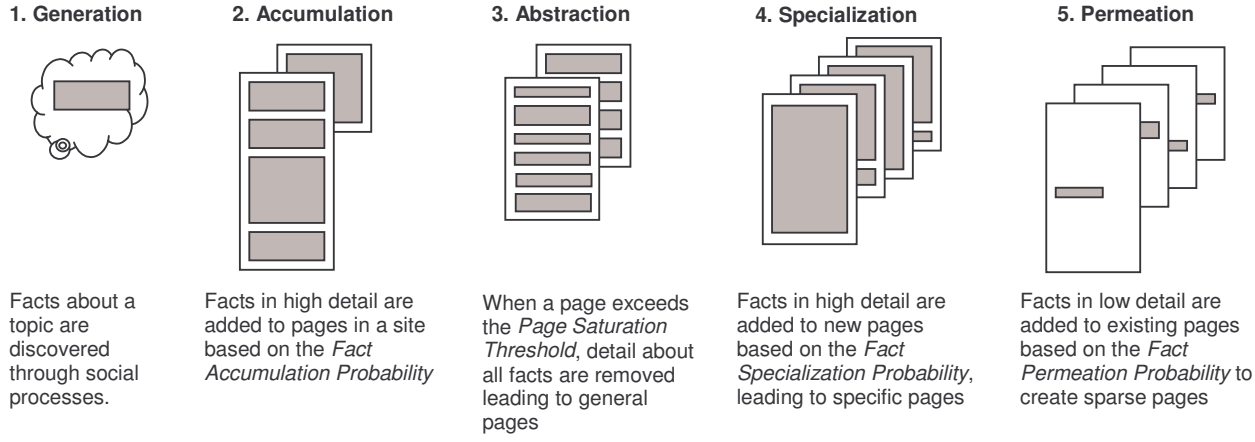

Figure 5. A schematic representation of the five processes in the Information Scatter Model, which generates general, specific, and sparse pages. Each gray box represents a fact in a webpage, and the area of the gray box represents amount of detail about that fact. The above processes occur over the total number of sites being modeled, and controlled by probability variables that determine how many facts occur on different pages.

We are currently exploring a computational model to test the above hypothesis. The goal of the model is to explore whether we can replicate the data, and make predictions of the scatter of information for new topics.

WHY SCATTER MATTERS: DESIGN IMPLICATIONS

The distribution and cluster analyses of information scatter provided an empirical foundation to understand the nature of information scatter, and suggested insights into the possible processes that result in that scatter. But why should understanding the nature and causes of scatter matter to users and designers?

We believe the answer lies in several studies, which show that many users end their searches prematurely (e.g. Eysenbach and Kohler, 2002) leading to the retrieval of incomplete information. The data suggests that because of the large numbers of specific and sparse pages, many users find them first, and terminate their searches early believing they have found all relevant facts. In contrast, studies of expert searchers (Bhavnani et al., 2006) show that they look for comprehensive information by first reading 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).

While the above general-specific-sparse search strategy appears useful when looking for comprehensive information, few websites organize their pages to guide uses to follow that strategy. As discussed earlier, even though reputed healthcare sites such as the Skin Cancer Foundation provide general, specific, and sparse pages about a topic, these pages are rarely linked, and therefore do not encourage users to browse easily across the page profiles. Furthermore, as discussed earlier, neither search engines nor domain portals provide such a search strategy. Because information scatter profoundly affects the search results of many users searching for comprehensive information in unfamiliar domains, such scatter needs to be addressed through the design of websites, search engines, and interfaces.

Implications for the Design of Websites

The following recommendations for the design of websites are aimed to guide users to follow the general-specific-sparse strategy to deal with information scatter. Although these recommendations might seem obvious, we have found many of the top-10 websites with melanoma information do not follow them, which directly impacts users searching for comprehensive information.

1. Reduce page length by abstracting and specializing. The analyses suggest that pages with high fact-breadth and high fact-depth will be abstracted. However, we found a few very long pages that had high fact-depth and fact-breadth. This suggests that page authors should pay close attention to the justification of providing such long pages, which are not conducive to online reading (Brinck et al., 2002). One approach is to reduce the page length by abstracting and specializing as represented by the model.

2. Consolidate general pages. The analyses suggest that there is typically more than one general page in a website, with no single page that contains all the facts. For a user this might not be the best way to present facts. Instead, authors should attempt to consolidate the facts (in medium detail) into one general page (or multiple well-connected pages) so that as many facts as possible about a topic are easily accessed. Furthermore, authors should clearly mark such pages as overviews, so users can get breadth information quickly and not miss important facts.

3. Link general pages to relevant specific pages. The analyses suggest the creation of many specific pages that contain detailed information about facts mentioned in the general pages. However, we found that the facts in the general pages were not linked to the respective specific pages. Page authors should attempt to make these links to encourage general to specific navigation.

4. Link general and specific pages to relevant sparse pages. The analyses suggest the generation of many sparse pages that mention facts in related topics. However, we found general and specific pages rarely linked to these pages. Page authors should provide links on general and specific pages, to the relevant sparse pages.

Implications for the Design of Search Algorithms

While 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 search engines nor

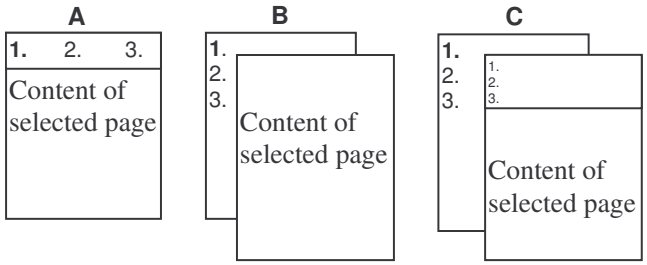

Figure 8. Schematic representations of three interface designs that attempted to encourage users to follow the general-specific-sparse search strategy.

domain portals guide users to deal with information scatter. To address this problem, we manually constructed a prototype domain portal, called the *Strategy Hub*, which guides users to a combination of general, specific and sparse pages from reputed healthcare websites. A recent study (Bhavnani et al., 2006) showed that such guided searches enabled novice searchers to be more effective in retrieving comprehensive information when compared to similar users of Google and MEDLINEplus.

The regularities in page profiles that we have observed suggest to us that the Strategy Hub could be automated using 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 (1) extract the corresponding list of facts for that topic from the database, (2) retrieve relevant pages for that topic using Google, (3) use content analysis tools such as latent semantic analysis (LSA) (Dumais et al., 1988) to dynamically determine the fact-depth and fact-breadth of each retrieved page, (4) use these calculated values to identify general, specialized, and sparse pages based on boundaries from the cluster analyses, and (5) present pages to the user in a specific order (such as from general to specific). Our initial studies have revealed that LSA performed reasonably well 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 Search Interfaces

While the information density algorithm attempts to automatically identify general, specific, and sparse pages given a topic and facts, do users actually follow that sequence when such pages are presented on the interface? Our development and testing of the Strategy Hub has revealed that it is not trivial to design an interface that guides users to follow the general-specific-sparse strategy.

As shown schematically in Figure 8A, our first attempt used a dual frame window. The top frame provided the three steps of the strategy in horizontal format. Each step had links for general, specific, or sparse pages. When a link was selected, the corresponding page was displayed in the bottom frame. This dual frame design is important because it is easy to forget the overall steps in a plan unless it is visible at all times. The dual frame design therefore provides a combination of a *context* view, which shows you where you are in the procedure, and a *focused* view of the content, an approach found to be critical for search interfaces (Egan et al., 1989).

However, a pilot study (Bhavnani et al., 2006) revealed that this design instantiation of the *focus plus context* concept resulted in two misunderstandings: (1) the steps of the strategy were often perceived as optional categories, rather than as recommended steps. This resulted in users not following the entire strategy. (2) The page displayed in the bottom frame was mistaken as a page that belonged to the Strategy Hub. When users selected links within the displayed page, it often took over the entire window and therefore removed the strategy from view. Many users therefore never used the Strategy Hub as intended.

Figure 8B shows how we attempted to deal with the above problems. First, we provided the strategy in a vertical layout to make the steps appear more like a process rather than categories. Next, when a link was selected in the strategy window, the relevant page opened in a new *content* window, which avoided

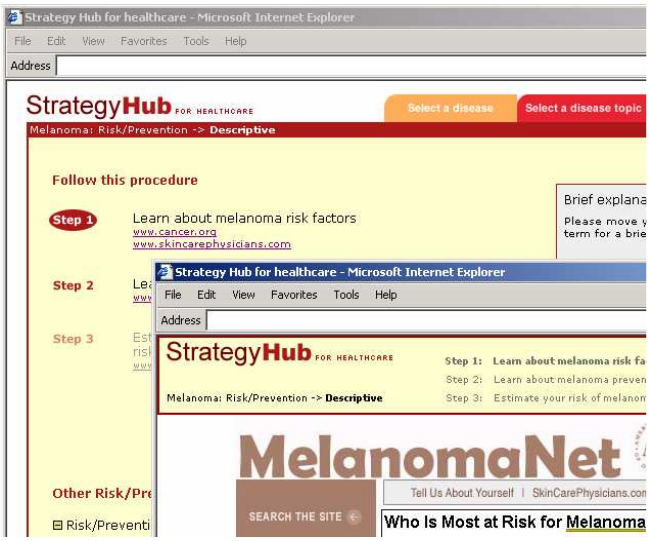

Figure 9. Dual window design, which encouraged more users to use the general-specific-sparse strategy.

the possibility of the strategy being removed from view. However, when users became engrossed in the content window, they often ignored the strategy window, and never returned to the Strategy Hub.

Figure 8C and Figure 9 shows our final attempt, where the content window was designed as a dual frame. The top frame contained a copy of the strategy where the step that had just been selected was highlighted. The bottom frame contained the contents of the selected link. This design significantly increased the number of subjects who followed the recommended strategy (Bhavnani et al., 2006). These results therefore suggest how to design interfaces which guide users to visit a sequence of webpages with the goal of finding comprehensive information about a topic.

SUMMARY AND CONCLUSIONS

To understand the causes of information scatter in the healthcare domain, this paper presented results from two analyses. The distribution analysis provided a first glimpse into the complex scatter of relevant information faced by novice searchers. Because there were many pages with few facts, and few pages with many but not all the facts, novice searchers require sophisticated strategies to find all the facts about each of the five melanoma topics that were analyzed. Next, the cluster analysis revealed that the above skewed distribution of facts across pages could be explained by the existence of page profiles with different information densities, each of which played an important role in the structuring of information. Furthermore, the nature of the page profiles suggested that their creation might not be a random occurrence, but rather the result of design decisions made by a rational webpage author. The above analyses suggest that the collective actions of many webpage authors could provide an explanation for the processes underlying the scatter.

Because information scatter can cause the retrieval of incomplete information, which in turn can have dangerous consequences in domains such as healthcare, we explored how the results lead to implications for the design of websites, search algorithms, and search interfaces.

While several studies (e.g. Barabasi and Albert, 1999) have described the skewed distribution of in-links and page visits, the distribution of facts across pages has received little attention. Furthermore, while there have been advances in understanding how to design search interfaces (e.g. Shneiderman et al., 1997), there has been far less attention on how the nature of information distributions could affect design. The main contributions of this paper are therefore (1) to bring attention to why information about topics is scattered across relevant pages, and (2) how that understanding can benefit the design of future systems. This understanding also sheds more light on how to design from a *Human-Information Interaction* perspective (Pirolli and Card, 1999), and should lead to new approaches that enable more users to retrieve comprehensive information when searching in vast and unfamiliar domains like healthcare.

ACKNOWLEDGEMENTS
This research was supported in part by NSF Award# EIA-9812607. The authors thank M. Bates, M. Cervone, G. Furnas, T. Johnson, R. Little, S. Mathan, F. Reif, V. Strecher, B. Suhm, R. Thomas, and G. Vallabha for their contributions.

REFERENCES

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

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

Bhavnani, S.K. (2005). The Retrieval of Highly Scattered Facts and Architectural Images: Strategies for Search and Design. *Automation in Construction, 14,* 724-735.

Bhavnani, S.K. (2005). Why is it Difficult to Find Comprehensive Information? *Journal of the American Society of Information Science and Technology*, *56*, 9, 989-1003.

Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., & Lu, X. (2002). Getting patients to the right healthcare sources: From real-world questions to Strategy Hubs. *Proc. AMIA 2002*, 51-55.

Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., and Strecher, V.J. (2006). Strategy Hubs: Domain Portals to Help Find Comprehensive Information. *Journal of the American Society for Information Science and Technology, 57*, 1, 4-24.

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.

Dumais, S.T., Furnas, G.W., Landauer, T.K. & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. *Proc. CHI 1988*, 281-285.

Egan, D.E., Remde, J.R., Landauer, T.K., Lochbaum, C.C., & Gomez, L.M. (1989). Behavioral evaluation and analysis of a hypertext browser. *Proc. CHI 1989,* 205-210.

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*. Avail: http://www.pewinternet.org/report_display.asp?r=95 (Accessed July, 2006).

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.

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. *Proc. ASIST 2004*.

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

Pratt, W., Hearst, M., & Fagan, L.A. (1999). Knowledge-Based Approach to Organizing Retrieved Documents. *Proc. AAAI 1999*.

Shneiderman, B., Byrd, D., & Croft, W.B. (1997). Clarifying search: A user interface framework for text searches. *DLIB Magazine 3*, 1.

Sturdee, D.W. The importance of patient education in improving compliance. *Climacteric, 10*, 2, 9-13.