Making Sense of Information Scatter on the Web

Suresh K. Bhavnani[1], Lada A. Adamic[2]
School of Information, University of Michigan, Ann Arbor, MI 48109-1092
{bhavnani[1], ladamic[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 regularities underlying such skewed distributions, this paper presents the results of three analyses. (1) A distribution analysis describes 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 an important role in providing comprehensive information of a topic. (3) A network analysis reveals regularities of how particular facts co-occur in pages. The regularities underlying scatter revealed by the three analyses provide implications for the design of search systems and websites to help users find and make sense of scattered information, and implications for a model to explain the process through which information scatter occurs.

Author Keywords

Healthcare, searching, Webometrics, distributions.

ACM Classification Keywords

H.3.3 Information search and retrieval---Search process.

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. 22]. 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?") [2].

One clue to the above difficulty is provided by expert healthcare searchers who know which *combination* of sites to visit in which *order* [2, 7, 8]. A recent study suggests 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 [4]. From a searcher's perspective, such distributions mean that information is highly scattered, and therefore sophisticated search strategies are required 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 regularities underlying the information scatter based on three analyses: (1) A *Distribution Analysis* (reported previously) 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, and (3) A *Network Analysis* reveals regularities in how facts are linked together through the pages in which they co-occur. The three analyses together provide implications for the design of future search systems and websites to help users find and make sense of comprehensive information, and for a model to explain the process through which information scatter might be occurring.

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 [12], go online without a definite search plan so that they find most sites accidentally [15], and often end their searches prematurely with incomplete information [2, 7]. In contrast, expert searchers (such as healthcare reference librarians) know which sites to visit in which sequence. For example, in a recent study [2] 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 [2].

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

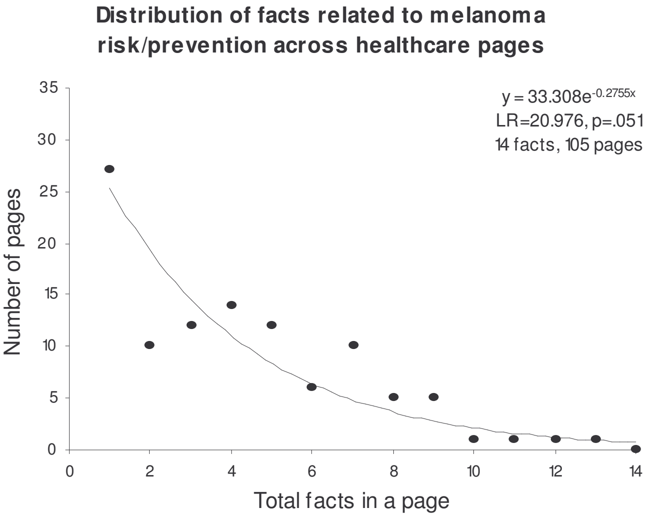


Figure 1. The distribution of risk/prevention facts across relevant pages in high-quality sites is highly skewed (best-fitted by a discrete exponential curve, Likelihood Ratio=20.967, p=.051 where significant fit is >.05), with no page containing all the facts.

DISTRIBUTION ANALYSIS: DATA COLLECTION AND PRIOR RESULTS

Our recent study [4] suggests why finding comprehensive information is difficult. The study consisted of two inter-rater experiments (whose data collection is 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. natural language statements such as “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 [6], similar to the taxonomy developed by Pratt et al. [22]. 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 for 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 [4] 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 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 [13] for a review). Additionally, the results describe how the information is distributed *across* pages and sites.

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 [2].

Because the retrieval of incomplete information can affect healthcare decisions and treatment compliance [22], 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 [9], and across databases [16]), few studies 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 that varied in the density of facts. 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 reveal regularities within the scatter that could help users find and make sense of scattered information?

CLUSTER ANALYSIS: REGULARITIES IN THE DENSITY OF FACTS WITHIN PAGES

Our goal was to analyze whether underlying the scatter of facts, were page profiles that had distinct densities of facts. We first describe the analysis of melanoma risk/prevention pages, and then show how those results generalize across the five melanoma topics.

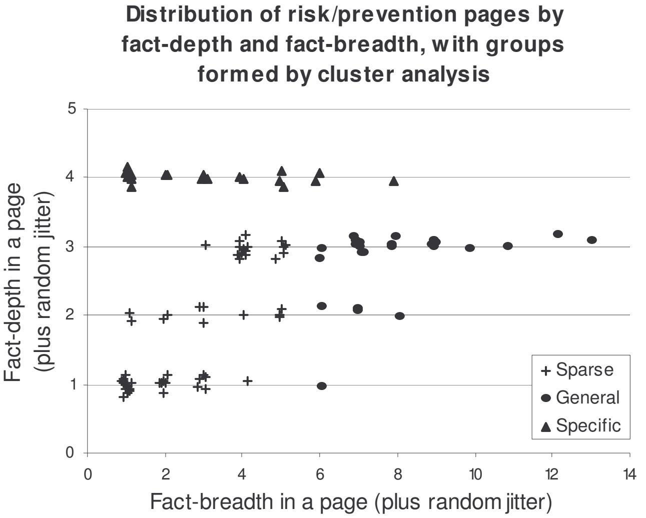


Figure 2. A cluster analysis, with three clusters as input, shows the existence of three page profiles. Sparse pages (+) having fact-breadth<=5 and fact-depth=1-3. General pages (•) having fact-breadth>5 and fact-depth=1-3. Specific pages (▲) having fact-depth=4 and any number for fact-breadth.

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 [4]). *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 [14] 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 first cluster in the lower left hand corner (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 second cluster on the right hand side of the figure (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.

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).

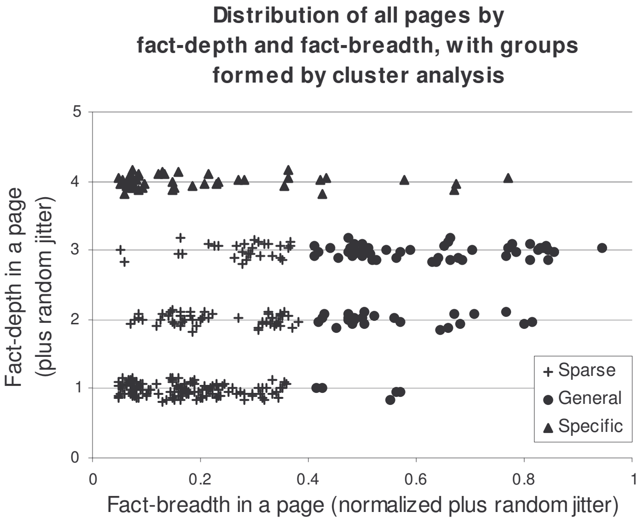


Figure 3. A cluster analysis (with three clusters as input) of pages from all five topics shows boundaries for three page profiles. Sparse pages have fact-breath<=40% and fact-depth=1-3. General pages have fact-breath>40% and fact-depth=1-3. Specific pages have any number of facts and fact-depth=4.

[1] The MML [14] 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.

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 [3] 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 phenomenon that generalizes across the melanoma topics, and across domains.

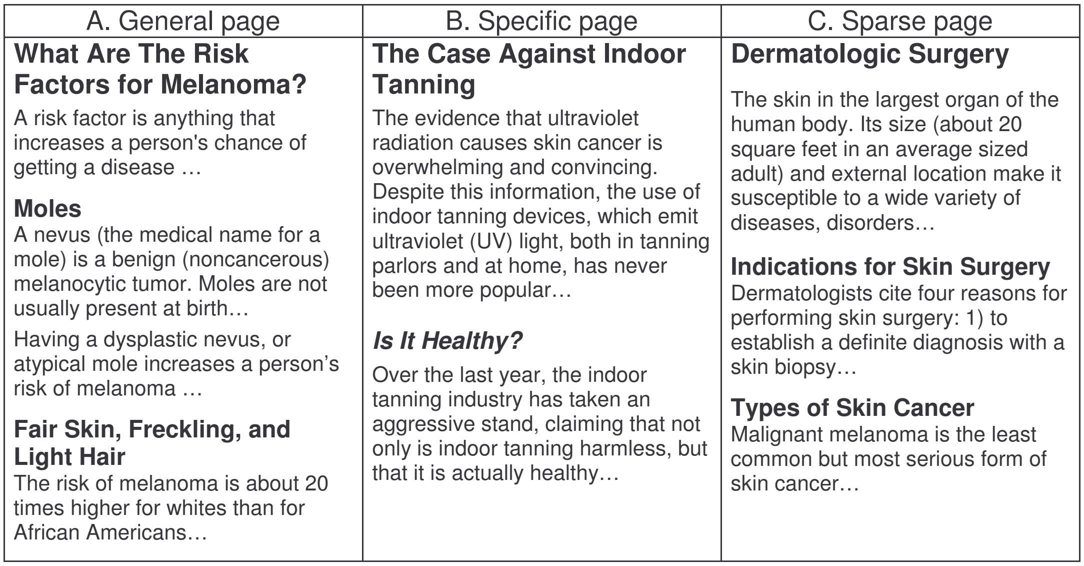
Insights into the Role of Page Profiles

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.

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.

While the cluster analysis revealed the density of facts within pages, it did not reveal how particular facts occurred in particular pages. For example, while the distribution and cluster analyses revealed that there are many pages with few facts, they did not reveal if such fact-poor pages contain rare or common facts. Furthermore, the analyses could not reveal if subsets of facts tended to co-occur in subtopic pages. Such results could suggest effective ways to find rare facts about a topic and to make sense of subtopics.



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.

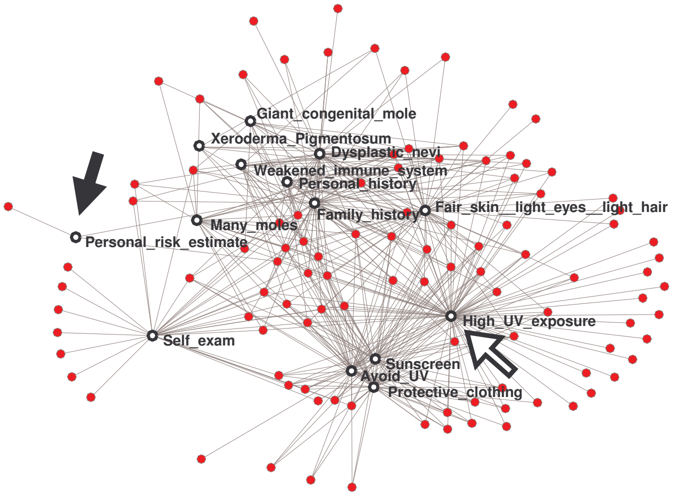


Figure 5. The scatter network for risk/prevention showing how 14 facts (white labeled nodes) occur in 108 relevant pages (solid nodes). The white arrow points to a common fact, and the black arrow points to a rare fact. (The fact labels have been abbreviated for visual clarity.)

NETWORK ANALYSIS: REGULARITIES IN FACT CO-OCCURRENCE WITHIN PAGES

To understand how particular facts co-occurred in particular pages, we conducted a network analysis on the five topics. A network is a graph consisting of nodes and edges; nodes represent one or more types of entities (e.g. facts and pages), and edges between the nodes represent a specific relationship between the entities (e.g. a fact is contained within a page). Figure 5 shows a *bipartite network* (where edges exist only between two different types of entities) of how facts about melanoma risk/prevention occur in relevant pages.

Networks have two advantages for analyzing complex regularities within scatter. (1) They represent a particular relationship between different nodes and therefore can reveal complex regularities in how specific facts occur in specific pages. (2) They can be rapidly visualized and analyzed using a toolbox of network algorithms. For example, Figure 5 shows how the *Spring* layout algorithm [11] helps to visualize scatter. The algorithm simulates placing springs between connected nodes, and a weakly repulsive force between nodes that are not connected. As shown, the result is that facts which co-occur in many of the same pages are placed close to each other, and close to the pages that mention them.

Method

To understand how rare facts co-occurred within pages, we analyzed the five melanoma topics (one of which was risk/prevention) in three-steps.

1. Each dataset was visualized as a bipartite graph using the Spring algorithm described earlier. We refer to these graphs as *scatter networks.*

2. Each of the five scatter networks was then analyzed using *degree correlation*. The degree of each node is the number of edges connected to it, and the degree correlation compares the degree of each node with the degree of its neighbors (i.e. it compares the degree of each fact with the degree of each page in which that fact occurs). A positive degree correlation means that common facts (fact-nodes with many edges) tend to occur in fact-rich pages (page-nodes with many edges), and rare facts (fact-nodes with few edges) occur in fact-poor pages (pages with few edges). A negative degree correlation would produce the opposite result. (Networks in a wide range of domains such as biology, technology, and sociology typically exhibit degree correlations ranging from -0.4 to +0.4 [17].)

3. Because there were common pages between the topics, we combined the facts and pages for the five topics to create a sixth *inter-topic* scatter network. This was done to understand how the facts across the topics related to each other.

Results

The analysis revealed that for four of the five topics, there was a negative degree correlation between the fact degree and page degree: *risk/prevention* (-0.26***)[1], *doctor's exam* (-0.35***), *diagnostic tests* (-0.17*), and *disease stages* (-0.56***). For example, in the risk/prevention scatter network (Figure 5), the common fact *high UV radiation* (close to the white arrow) occurs in many fact-poor pages, presumably because of its importance to preventing melanoma. Furthermore, the rare fact *personal risk estimate* (close to the black arrow) occurs in two fact-rich pages in the center of the graph. In contrast, for the topic *self-exam,* there was a positive degree correlation (0.14*). Here, fact-rich pages tended to contain the same common facts related to *mole appearance*, and rare facts such as *resources for locating a dermatologist*, were located on pages containing few or no other facts.

When rare facts tend to occur in fact-rich pages (negative degree correlation), it implies that users should be pointed to fact-rich pages that are likely to include comprehensive information, including the rare facts. However, when the opposite is true, pointing a user to multiple fact-rich pages without consideration of distribution of those facts would produce redundant and incomplete results.

Figure 5 also shows two clusters of co-occurring facts, corresponding to two subtopics: melanoma risk (e.g. *weakened immune system*, *family history*) at the top and melanoma prevention (e.g. *sun screen, avoid UV*) at the bottom. The facts therefore co-occurred based on meaningful subtopics reflecting the structure in the topic.

Next we analyzed how the facts co-occurred in pages across the topics using the inter-topic scatter network. While a visual analysis of small networks (e.g. the network in Figure 5) can reveal co-occurring facts, larger networks need to be analyzed using a *community-finding* algorithm [10]. This algorithm aggregates nodes into groups

[1] *** and * denote significance at the 0.001 and .05 levels respectively.

(communities) and stops when it has achieved maximum modularity, meaning that the number of edges within communities compared to the number of edges between communities is much higher than if the communities had been randomly assigned.

If the five topics had mostly dedicated pages for each topic, then the algorithm would identify 5 communities in the connected part of the graph. However, as shown in Figure 6, the algorithm found only three communities consisting of facts and pages from: (1) *doctor's exam, diagnostic tests,* and *disease stages*, (2) *self-exam*, and (3) *risk/prevention*. The algorithm revealed that the three topics in the first community are strongly related through their co-occurrence on pages, and indeed, all three topics deal with diagnosis. In contrast, the other two communities tend to have more dedicated pages. Besides grouping three topics together, the algorithm also assigned three facts to a different topic than how they had been categorized by physicians. Two self-exam facts *checking the entire skin surface* and *checking for irregular moles* (shown as solid white circles in the right-hand-side fact group in Figure 5) are placed in the community with the *risk and prevention* facts. This probably occurs because self-exams are a basic part of *risk and prevention*. Similarly, one fact about the *doctor's exam* topic, *having regular doctor's exams* (shown as a solid black circle) is also grouped under *risk and prevention*. This is probably because a doctor's exam is a basic part of risk and prevention. Note that these groupings would not be apparent without analyzing all the topics together.

DESIGN AND MODEL IMPLICATIONS

The distribution, cluster, and network analyses provide implications for the design of search systems and websites, and for a model to explain the process through which information scatter might be occurring.

Implications for the Design of Future Search Systems

When users search for information about a broad topic, they also attempt to construct representations of that information to assist them in achieving real-world goals in an iterative process referred to as *Sensemaking* [20]. The three analyses together suggest opportunities not just to help users find comprehensive information, but also to make sense of that information. The distribution analysis provided the primary observation that even within high-quality websites that were highly relevant to a topic, there were many pages with few facts about a topic, and few pages with a many facts. Furthermore, there was no page which had all the facts. This result suggests that searchers need a *portfolio* of pages, which together provide all relevant information about a search topic.

The cluster analyses revealed that underlying the skewed distribution of facts across pages, there were three types of pages which differed in information density. General pages provided many facts about the topic, mostly in medium amount of detail. On the other hand, specific pages tended to provide detailed accounts of a few facts, and sparse pages tended to provide a brief mention of a few relevant facts in the context of related topics. These results suggest that while users might benefit from getting a portfolio of pages about a topic to *find* comprehensive information, they could also benefit from an organization of those pages to help *make sense* of that information. For example, search results could be organized such that users first read a combination of general pages (to get an overview of all the facts), followed by a combination of specific pages (to enable users to systematically understand details about particular facts), and followed by sparse pages (to understand how their search topic is related to other topics). Such an organization going from general to specific to sparse pages could help users make sense of highly scattered information.

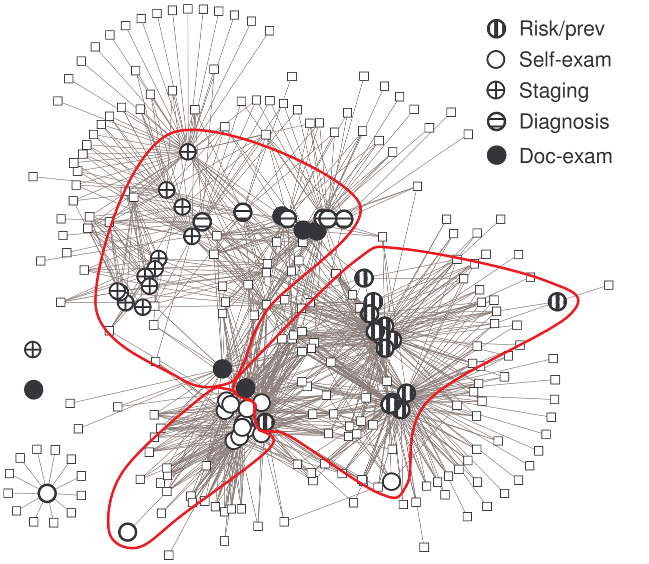


Figure 6. The inter-topic scatter network showing how 53 facts (circles) about 5 melanoma topics occur in 336 relevant pages (squares). The curved shapes show groupings of facts created by the community-finding algorithm.

The network analyses revealed other regularities related to fact co-occurrence which could be used to design how search results are presented to users. The results revealed a hierarchy of subtopics ranging from groupings of topics related to melanoma (e.g. *doctor's exam, diagnostic tests,* and *disease stages*, all related to diagnosis) and subtopics within a topic (e.g. melanoma risk and melanoma prevention within melanoma risk/prevention). Such a decomposition reflects the semantic structure of a topic, and could help users make sense of the topic. Furthermore, for four of the five topics analyzed, rare facts co-occurred in fact-rich pages. This suggests that for many topics visiting a combination of general pages might suffice to find all facts. However, when rare facts occur mainly in fact-poor pages, then an organization that highlights the occurrence of rare facts might be needed.

Our own explorations in how to present organizations of relevant pages to users has produced encouraging results. For example, we manually constructed a prototype domain portal, called the *Strategy Hub*, which enables users to select a search topic, and then guides them to a combination of general, specific and sparse pages from reputed healthcare websites. A pilot study [7], and a more extensive experiment [8] showed that such guided searches enabled novice searchers to be more effective in retrieving comprehensive information about a topic when compared to similar users of Google and MedlinePlus. Future search systems that attempt to aid users not in just collecting comprehensive information about a topic, but also aiding them to make sense of it could similarly take into account the regularities in information scatter such as those we have discussed in this paper.

Implications for the Design of Websites

The analyses also provide implications for the design of websites. For example, the cluster 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, web 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.

Finally, we also 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.

Implications for a Model to Explain Information Scatter

The analyses 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 upper right hand corner of both cluster 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 few general pages that have more breadth than depth, and the existence of many 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 analyses therefore suggest 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

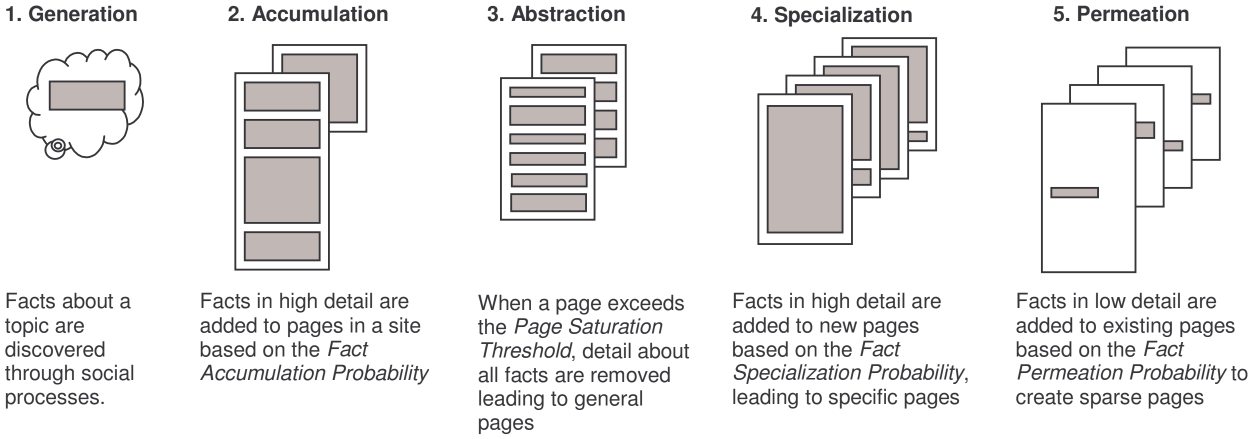

1. Generation
2. Accumulation
3. Abstraction
4. Specialization
5. Permeation
Facts about a topic are discovered through social processes.
Facts in high detail are added to pages in a site based on the *Fact Accumulation Probability*
When a page exceeds the *Page Saturation Threshold*, detail about all facts are removed leading to general pages
Facts in high detail are added to new pages based on the *Fact Specialization Probability*, leading to specific pages
Facts in low detail are added to existing pages based on the *Fact Permeation Probability* to create sparse pages

Figure 7. 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.

webpage authors could result in the scatter of information as observed in the data.

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.

As shown schematically in Figure 7, the Information Scatter Model is designed to consist of five subprocesses: (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.

The above 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 [1] 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. 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 are currently exploring a computational approach to test the above model, with the goal of exploring whether we can replicate the data, and make predictions of the scatter of information for new topics.

SUMMARY AND CONCLUSIONS

To understand the causes of information scatter in the healthcare domain, this paper presented results from three 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. Finally, the network analyses revealed regularities in how facts co-occurred in pages, resulting in pages specializing in subtopics, and an understanding of how rare and common facts co-occurred.

Because information scatter can cause the retrieval of incomplete information, which in turn can have dangerous consequences in domains such as healthcare, we explored the implications of the results for the design of future search systems and websites. Furthermore, the analyses also provided implications for a model that explains the process through which information scatter occurs.

While several studies [e.g. 1] have described the skewed distribution of in-links and page visits, relatively fewer studies have been done on 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. 21], 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 how 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 [18], and should lead to new approaches that enable more users to retrieve comprehensive information and make sense of it 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, F. Peck, T. Johnson, R. Little, S. Mathan, Y. Qu, F. Reif, V. Strecher, B. Suhm, R. Thomas, and G. Vallabha for their contributions. The cluster analysis and model presented in this paper have been recently published [5].

REFERENCES

1. Barabasi, A.-L., & Albert, R. Emergence of scaling in random networks. *Science* (1999), 286*,* 509-512.
2. Bhavnani, S.K. Important cognitive components of domain-specific search knowledge. *Proc. TREC 2001*, NIST (2001) 571-578.
3. Bhavnani, S.K. The Retrieval of Highly Scattered Facts and Architectural Images: Strategies for Search and Design. *Automation in Construction* (2005), 14, 724-735.
4. Bhavnani, S.K. Why is it Difficult to Find Comprehensive Information? Implications of Information Scatter for Search and Design. *JASIST* 56, 9, (2005), 989-1003.
5. Bhavnani, S.K., and Peck, F.A. Towards a Model of Information Scatter: Implications for Search and Design. *Proceedings of ASIST'2006* (2006).
6. Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., & Lu, X.. Getting patients to the right healthcare sources: From

real-world questions to Strategy Hubs. *Proc. AMIA 2002* (2002), 51-55.

7. Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., and Strecher, V.J. Strategy Hubs: Next-generation domain portals with search procedures. *Proc. CHI 2003*, ACM Press (2003), 393-400.
8. Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., and Strecher, V.J. Strategy Hubs: Domain Portals to Help Find Comprehensive Information. *JASIST* 57, 1 (2006), 4-24.
9. Bradford, S. C. *Documentation*. London: Crosby Lockwood, 1948.
10. Clauset, A., Newman, M.E.J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* 70, 066111 (2004).
11. Freeman, L. Visualizing Social Networks, *JoSS*, 1,1 (2001).
12. Eysenbach, G., & Köhler, C. 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, *BMJ 324*, (2002) 573-577.
13. Eysenbach, G., Powell, J., Kuss, O., & Sa, E-R. Empirical studies assessing the quality of health information for consumers on the World Wide Web: A systematic review. *Journal of the American Medical Association 287*, 20 (2002), 2691-2700.
14. Figueiredo, M.A.T., & Jain, A.K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 3 (2002), 381-396.
15. Fox, S., & Fallows, F. 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/reports/toc.asp?report=95 (Jul, 2003).
16. Hood, W., & Wilson, C. The scatter of documents over databases in different subject domains: How many databases are needed? *JASIST*, 52, 14, (2001) 1242-1254.
17. Newman, M. The structure and function of complex networks. *SIAM Review*, 45(2), (2003), 167-256.
18. Pirolli, P., & Card, S. K. Information Foraging. *Psychological Review 106*, (1999), 643-675.
19. Pratt, W., Hearst, M., & Fagan, L. A Knowledge-Based Approach to Organizing Retrieved Documents. Proc. AAAI ’99 (1999), Orlando, FL.
20. Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The Cost Structure of Sensemaking. *Proceedings of ACM INTERCHI'93*, 269-276.
21. Shneiderman, B., Byrd, D., & Croft, W.B. Clarifying search: A user interface framework for text searches. *DLIB Mag 3*, 1 (1997).
22. Sturdee, D.W. The importance of patient education in improving compliance. *Climacteric* 10, 2 (2000), 9-13.