The Distribution of Online Healthcare Information: A Case Study on Melanoma

Suresh K. Bhavnani
School of Information, University of Michigan, Ann Arbor MI, 48109-1092

To understand the difficulties users face when retrieving comprehensive healthcare information, this paper analyzes how facts related to a widely available healthcare topic are distributed across high-quality webpages. An inter-rater experiment with two skin-cancer physicians helped identify 14 facts necessary for a comprehensive understanding of melanoma risk and prevention. A second inter-rater experiment analyzed how those facts were distributed across 189 relevant webpages from high-quality sites. The analysis revealed that the distribution of facts is highly skewed, where few pages have many facts, many pages have a few facts, and no single page or site provides all the facts. A more detailed analysis suggests that the distribution is being caused by a trade-off between depth and breadth, leading to the existence of general, specialized, and sparse pages. Furthermore, the analyses reveal patterns and complexities in the relationships between facts, pages, and websites. These distribution results pinpoint the difficulties faced by searchers, and provide insights for the design of future systems that guide users in retrieving comprehensive healthcare information.

INTRODUCTION

A synergistic relationship between healthcare organizations, and the rapid growth in the number of healthcare information seekers [1], has resulted in the development of huge repositories of healthcare information. For example, the National Cancer Institute's (NCI) website currently provides information, related to 118 different cancers, distributed across hundreds of pages. Given such vast resources, one might expect that users could obtain comprehensive information about a healthcare topic by visiting one webpage, or even one large website like NCI. However, this is counter to the conclusions reached by many information scientists. These scientists have argued that as the number of information sources about a specific topic increases, the information across the sources follows a power-law distribution [e.g. 2], where a few sources have a lot of information about the topic, and a large number of sources have very little information. Such a distribution can make the retrieval of complete information about a topic a difficult, if not an impossible task [3].

Because the incomplete retrieval of healthcare information can have dangerous consequences, we believe the analysis of how such information is distributed across sources deserves closer inspection. Previous distribution studies of information include how articles are distributed across journals [4], how words are distributed within a book [5], and more recently how incoming web links are distributed across webpages [6]. However, much less is known about how *facts* related to a search topic are distributed across relevant webpages.

This paper presents two experiments to understand how facts related to a common healthcare topic are distributed across relevant webpages in high-quality sites. In Experiment-1, two skin cancer physicians independently rated the importance of facts related to melanoma risk and prevention. The high inter-rater agreement enabled our research team to identify a set of facts necessary for a comprehensive understanding of melanoma risk and prevention at different levels of importance. In Experiment-2, a different judge rated the degree of detail that each fact occurred within 189 relevant pages from high quality sites. These ratings were subsequently verified through the ratings of another independent judge. The analysis of the ratings revealed the relationship between facts of the same healthcare topic, between facts across different types of pages, and between facts, webpages, and websites. The analysis also helped to pinpoint the complexities involved in finding accurate and comprehensive information related to a healthcare topic, and suggested a distribution-conscious approach to the development of future search systems.

EXPERIMENT-1: IDENTIFICATION OF FACTS

The goal of Experiment-1 was to identify a set of facts that skin cancer physicians agreed was necessary for a user to have a comprehensive understanding of *descriptive information* related to *melanoma risk and prevention*[1] (which will henceforth be referred to as *melanoma risk/prevention*).

Our research team chose to focus on the distribution of melanoma risk/prevention for two reasons: (1) questions related to this topic were the most frequent in an empirical study [7] of user questions related to skin cancer, and (2) research related to this topic is well known, and guidelines for the general public are widely available on the Web [8].

[1] In an earlier study [7], skin cancer physicians developed a hierarchical taxonomy of real-world user questions, where one of the high-level nodes was *risk/prevention*, and whose sub-nodes included *descriptive information*, and *statistical information*.

Facts related to *descriptive information* for *melanoma risk and prevention*	**Judge-1 ratings**	**Judge-2 ratings**	**Final ratings**
1. **Having fair skin** [or type I or II skin; or white skin; or tendency to burn, not tan; or green or blue eyes, or red or blond hair] **increases your risk of getting melanoma** [or skin cancer]	5	5	5
2. **High UV exposure** [or sunburn] **increases your risk of getting melanoma** [or skin cancer]	5	5	5
3. **Having many moles** [or more than 50 moles] **increases your risk of getting melanoma**	5	5	5
4. **Having dysplastic nevi** [or atypical moles] **increases your risk of getting melanoma** [or skin cancer]	5	5	5
5. **Having a giant** [or >20 cm] **congenital mole** [or mole present at birth] **increases your risk of getting melanoma** [or skin cancer] [must mention "giant" and "congenital" or "mole present at birth"]	1	3	2
6. **Having a family history of melanoma** [or members of your family who have had melanoma] **increases your risk of getting melanoma** [or skin cancer]	5	5	5
7. **Having a personal history of melanoma increases your risk of getting melanoma** [or skin cancer]	5	5	5
8. **Having a weakened immune system** [or immune deficiencies] **increases your risk of getting melanoma** [or skin cancer]	3	1	2
9. **Having Xeroderma Pigmentosum increases your risk of getting melanoma** [or skin cancer]	4	2	3
10. **Calculate your personal risk of getting melanoma** (source of calculator is provided)	4	4	4
11. **Wearing protective clothing can help to prevent melanoma**	5	5	5
12. **Wearing UV-protective sunglasses can help to prevent melanoma**	1	1	1
13. **Wearing sunscreen can help to prevent melanoma**	4	5	4.5
14. **Avoiding UV Rays** [or avoiding peak sunlight hours; or seeking shade] **can help to prevent melanoma**	5	5	5
15. **Examining your body for suspicious moles** [or changing moles, or itching moles, or moles that match the ABCDs] **can help to prevent melanoma from spreading**	5	5	5

Figure 1. Fifteen facts related to descriptive information for melanoma risk and prevention, and how two judges (who were skin cancer physicians) independently rated their importance on a 5-point Likert scale. The final importance rating for each fact was calculated by averaging the scores given by each judge.

Method: Two experienced skin cancer physicians were asked to independently rate the importance of 15 facts[2] related to melanoma risk/prevention using a 5-point Likert 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 physicians were told that they should rate the importance of each fact keeping in mind a concerned user looking for melanoma risk/prevention information on the Web. Furthermore, they were free to modify the wordings of the facts, or to add new facts. After they had completed their ratings, the physicians independently discussed their ratings with the researcher to make any clarifications.

Results: Only one of the physicians made minor changes in the wordings of 3 facts (none of which changed the original meaning of the fact) and neither of them added any new facts. Figure 1 shows the high agreement between the two physicians for the list of facts. As shown, the physicians agreed completely on 11 facts (73%), but disagreed on 1 fact (7%) by 1 point, and 3 facts (20%) by 2 points[3]. The judges did not disagree by more than 2 points for any fact, which therefore represents very high agreement between the judges.

As shown in the last column of Figure 1, a final rating for each fact was calculated by averaging the two judge's scores. This resulted in one fact that both judges rated as unimportant (Fact 12), and was excluded from the analysis. The analysis therefore enabled us to identify a set of 14 facts related to a comprehensive understanding of melanoma risk/prevention at different levels of importance. This set of facts was used in the next experiment designed to understand the distribution of these facts across relevant webpages.

EXPERIMENT-2: ANALYSIS OF DISTRIBUTION

The goal of Experiment-2 was to understand not only how melanoma risk/prevention facts were distributed across relevant webpages, but also the amount of such information in each page and site.

Material: Given that there exists a large number of healthcare sources that are unreliable, we focused our survey on sites that were known to contain reliable melanoma information. A set of reliable melanoma sites was defined as the union of all the sites pointed to by the melanoma page in MEDLINEplus (a leading healthcare portal), and the top 5 most comprehensive

[2] The list of facts was derived by studying relevant pages from melanoma sites pointed to by MEDLINEplus.

[3] Neither Cohen's Kappa, nor Cohen's weighted kappa are relevant for these data because of the very high skew in the agreements. The data in Figure 1 is therefore shown to provide direct evidence of the high inter-rater agreement.

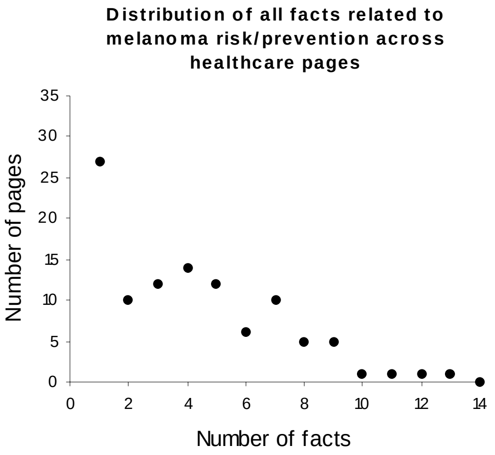

Figure 2. The distribution of risk/prevention facts across relevant pages in high-quality sites is highly skewed, with no page containing all the facts.

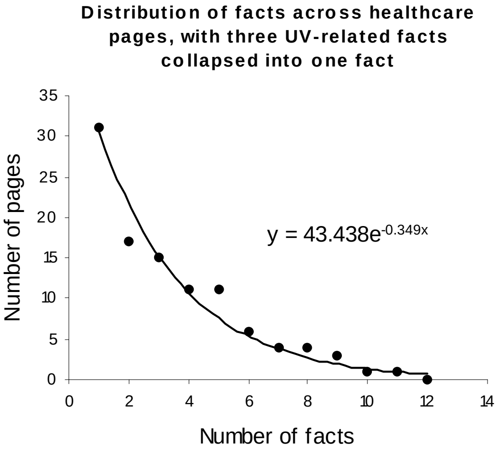

Figure 3. The distribution of risk/prevention facts with three facts related to UV protection collapsed into one. This caused the prominent bumps in the distribution shown in Figure 2 to smoothen out. A discrete exponential curve provided the best fit to the resulting distribution.

sites identified in a recent study of online melanoma information [8]. This union resulted in 10 sites.

To compensate for the widely varying quality of internal search engines provided by these sites, we used Google to search *within* each of the 10 sites for pages related to the 14 melanoma risk/prevention facts (identified from Experiment-1), and for general melanoma risk/prevention. We therefore generated 160 Google queries (e.g. melanoma risk UV OR ultraviolet OR sun OR sunlight OR sunburn site:cancer.gov), each of which was iteratively tested by a group of 3 search experts until the best set of pages showed up in the top 10 hits. It is important to note that this query generation process was used to provide a best-case scenario for identifying the most relevant pages within each site, and goes far beyond the kind of search that a typical user would perform.

The highly targeted queries were used to retrieve the top 10 hits from each site. Subsequently, duplicates, news items, pages for health professionals, non-English pages, dictionary pages, personal homepages, and broken links were removed. This resulted in 189 unique webpages, a set which we believe had a high probability of containing all the pages from each site containing melanoma risk/prevention information.

Method: A printed version of the 189 webpages was given to a rater who judged the extent to which the 14 facts related to melanoma risk/prevention were covered in each page, using a 5-point Likert scale (0=Fact not covered on page, 1=Fact covered in less than one paragraph, 2=Fact covered in one paragraph, 3=Fact covered in more than one paragraph, 4=Webpage mostly devoted to fact, although other facts could also be covered on the same page). The reliability of the above rater was assessed by requesting a second rater to perform the same evaluation on a random selection of 25% of the 189 webpages.

Analysis and Results: The raters had high agreement on whether or not a fact was present in a page (96.2% agreement, Cohen's kappa=.81), and the extent to which the fact was covered on that page (92.9% agreement, Cohen's weighted kappa=.73).

Figure 2 shows a plot of the number of webpages that contain an ascending number of melanoma risk facts (84 pages with no facts were dropped in order to limit our analysis to only relevant pages). As shown, the distribution is skewed to the left where there are many pages that contain a few facts, and very few pages (toward the right tail) contain many but not all the facts. While this skewed distribution is similar to the results of other information distribution studies [e.g. 4, 5, 6], it does not explain why over 75% of the pages from *reliable* sites contained less than half of the facts. Furthermore, the skewed distribution has two prominent bumps, one at 4 facts, and another at 7 facts. The question that arose was whether these bumps were caused by outliers in an otherwise smooth distribution, or caused by some other underlying phenomenon.

To probe the above questions, we performed analyses to understand the relationship between: (1) facts and other facts, (2) facts and webpages, and (3) facts, webpages, and websites.

Relationship between facts An exploratory analysis of how the facts occurred within the pages led us to hypothesize that a small set of facts in the pages co-occurred frequently. A correlation matrix confirmed our hypothesis. Three of the fourteen facts (Facts 11, 13, and 14 in Figure 1) were highly correlated with each other (r >. 8) compared to the other facts. These three facts all dealt directly with UV protection. We therefore collapsed these three facts into a single averaged fact (which was considered to be on a page if that page contained more than 50% of the original 3 facts). As shown in Figure 3, the bumps smoothened out in the resulting distribution. To determine the shape of this distribution, three curves (power, discrete exponential, and truncated Poisson) were fit to the data using maximum likelihood estimation (MLE), each of which was tested for goodness-of-fit using the likelihood ratio (LR) test. A discrete exponential curve ($y=43.438e^{-0.349x}$) provided the best fit (LR=8.227, p=.607)[4].

To understand the distribution when only very, and extremely important facts were included, we repeated the above analysis with only that subset (facts 1-4, 6, 7, 10, 11, and 13-15 in Figure 1). Surprisingly, while the best-fit equation changed ($y=36.871e^{-0.301x}$), there was no single page that contained all the facts.

While the above distribution analyses revealed that no single page had all the facts related to melanoma risk/prevention, it was not clear what was causing the underlying distribution. After all, the pages came from the top 10 melanoma sites. What was causing this uneven scatter of facts across those webpages?

Relationship between facts and webpages An exploratory analysis of pages at both ends of the distribution revealed that pages with many facts appeared to provide information in not much detail, while pages with a few facts appeared to provide a lot of detail about a few facts. A more rigorous analysis revealed that pages with a maximum detail level of 2 or 3 (on the Likert scale described earlier), had a significantly higher number of facts (p<.001, mean number of facts=5.89, SD=2.63) compared to pages that had a maximum detail level of 4 (mean=2.87, SD=2.12), or a maximum detail level of 1 (mean=1.86, SD=1.21). This suggests the existence of *general* pages that cover many facts in a medium amount of detail, *specialized* pages that cover few facts in a high level of detail, and *sparse* pages that contain few facts in very little detail. The analysis therefore suggests that the skewed distribution is being caused by pages that make a trade-off between depth and breadth of fact coverage, with no single breadth page providing 100% of the facts.

While the above analysis focused on facts within pages, we wondered if there existed sites that contained a combination of pages that would provide access to all the facts.

	All levels of importance	**Very & extremely important levels**
Pages containing all facts for:		
Risk & prevention	0	0
Risk	0	1
Prevention	16	16
Sites containing all facts for:		
Risk & prevention	0	3
Risk	0	3
Prevention	8	8

Figure 4. The distribution of facts along two different levels of source, topic, and fact importance.

Relationship between facts, webpages, and websites An analysis to probe whether a combination of pages within a site provided access to all the facts revealed that again there was no site that contained all the facts related to melanoma risk/prevention. This scatter of facts across websites presents a complex situation for a user searching for comprehensive information about melanoma risk/prevention. We therefore proceeded to investigate even further to understand which subsets of risk/prevention facts were 100% covered in single pages or websites.

Analysis of facts related to subtopics within the pages revealed that there were 42 (40.0%) pages with *only* risk facts, 11 (10.5%) pages with *only* prevention facts, and 52 (49.5%) pages that had facts about both subtopics (which provided independent verification for why the physicians combined risk and prevention into a single node in the skin cancer taxonomy [7]). We therefore analyzed the coverage of facts for the two subtopics of risk/prevention, within pages and sites, and at two levels of fact importance.

Figure 4 shows a summary of the analyses along the above three dimensions: (1) topic granularity at two levels (risk/prevention, and only risk, and only prevention), (2) source granularity at two levels (page and site), and (3) fact importance at two levels (all levels of importance, and only very and extremely important). As discussed earlier, when we consider all the facts (column 2 in Figure 4), there is no single page or site that has all the facts. This is also true for only risk facts, but not true for prevention facts. When we consider only very, and extremely important facts (column 3 in Figure 4), there still is no page that contains all the facts, and only one page that contains all risk facts. However, there are many pages that contain all the prevention facts, and many sites that contain all the facts in all combinations.

The above complexity in the distribution of facts across relevant pages and websites is not unique to risk/prevention. A pilot study of the above experiment [9], and our ongoing analysis of information related to different melanoma topics have revealed similar patterns and complexities in their distributions.

[4] The null hypothesis in a likelihood ratio test states that the distribution fits the curve being tested. A curve therefore has acceptable fit when p>0.05.

IMPLICATIONS FOR SEARCH AND DESIGN

The results of our study pinpoint the difficulties that users face when searching for comprehensive information about healthcare. Users must know that some pages have breadth information spanning many facts with medium levels of detail, while others have depth about a few. In addition, users also need to know that they have to visit more than one page that has breadth information to get all the relevant facts. Because conventional search tools like Google and MEDLINEplus do not provide this kind of information about relevant pages, the lack of such knowledge often leads users to end their searches early, leading to the retrieval of incomplete information [10].

The patterns and complexities in the distribution that we found should come as no surprise to search experts like healthcare librarians. Such experts have acquired deeply articulated knowledge to determine which pages to visit in what order when searching for comprehensive information about a topic [10, 11]. However, while much research has focused on identifying strategies for *finding* sources of information, far less is known about how experts select and order *known* and *relevant* sources of information. This paper suggests that a large part of search expertise must emerge from the complexities inherent in the types and distribution of the information within *relevant* pages. These complexities therefore need much more scrutiny than they have received in the past.

It is pertinent to note that even though we retrieved highly relevant pages from the top 10 sites, the pages varied widely along the depth and breadth of fact coverage, and in the scope of content that they provided. One might argue that content providers must strive harder to make sure that the information they provide on relevant pages is complete. However, we have come to believe that such an argument does not acknowledge the nature of information, especially as provided on the Web. Information on the Web (even in the best sites) is created by different authors, with different intentions, and targeted to different audiences resulting in high variability along many dimensions. While a small number of facts related to subtopics like melanoma prevention might co-occur in many pages, we believe that facts related to a vast number of topics will often have a scattered and complex distribution. This *is* the nature of most information on the Web, and we therefore must understand it, and design for it.

The above realization has led us to build a new kind of domain portal called a Strategy Hub [11] that embraces the complexity in the distribution of healthcare information. Strategy Hubs provide search procedures to guide users to different pages in a particular order to enable the retrieval of comprehensive information. Furthermore, we are exploring the development of new algorithms such as ones to automatically identify general vs. specific pages about a topic, and provide them to users in a particular order. An understanding of how information is distributed on the Web is therefore critical in understanding how to make such algorithms powerful and useful, and to ultimately assist users in getting comprehensive information when searching in unfamiliar and vast online domains such as healthcare.

ACKNOWLEDGEMENTS

The author thanks A. Abernathy, M. J. Bates, C. Bichakjian, G. Furnas, T. Johnson, R. Little, J. Nardine, F. Peck, F. Reif, V. Strecher, and R. Thomas for their contributions.

REFERENCES

1. Fox, S., & Rainie, L. The online health care revolution: How the Web helps Americans take better care of themselves. Pew Internet and American live project: Online life report. Available from: URL: http://www.pewinternet.org /reports/toc.asp?Report=26
2. Bates, M.J. Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors. JASIS, 49, 13 (1998), 1185-1205.
3. Bates, M.J. Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution. Proceedings of CoLIS4 (2002), 137-150.
4. Bradford, S. C. Documentation, London: Crosby Lockwood, 1948.
5. Zipf, G. K. Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley, Cambridge MA, 1949.
6. Barabasi, A.-L., & Albert, R. Emergence of Scaling in Random Networks, Science, 286, 509-512. (1999).
7. Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., & Lu, X. (2002). Getting patients to the right healthcare sources: From real-world questions to Strategy Hubs. Proceedings of AMIA'02 (2002), 51-55.
8. Bichakjian, C., Schwartz, J., Wang, T., Hall J., Johnson, T., & Biermann, S. Melanoma information on the Internet: Often incomplete-a public health opportunity? Journal of Clinical Oncology, 20, 1 (2002), 134-141.
9. Bhavnani, S.K., Jacob, R. T., Nardine, J., & Peck, F.A. Exploring the Distribution of Online Healthcare Information. Proceedings of CHI'03 (2003), 816-817.
10. Bhavnani, S.K. Important Cognitive Components of Domain-Specific Search Knowledge. Proceedings of TREC'2001 (2001), 571-578.
11. Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., Strecher, V.J. Strategy Hubs: Next-Generation Domain Portals with Search Procedures. Proceedings of CHI'03, (2003), 393-400.