Exploring the Distribution of Online Healthcare Information

Suresh K. Bhavnani, Renju T. Jacob, Jennifer Nardine, Frederick A. Peck

School of Information, University of Michigan, Ann Arbor, MI 48109-1092
{bhavnani, renjutj, jnardine, peckf}@umich.edu

ABSTRACT

Motivated by the importance of retrieving comprehensive healthcare information, we analyzed how information about 12 concepts related to a widely available healthcare topic is distributed across 145 high-quality webpages. The analysis reveals that the distribution of the concepts follows a power law where a few pages contain many concepts, while the majority contains less than half the concepts. The analysis also reveals the existence of general, specialized, and sparse pages, in addition to the large number of pages that users must visit before they have access to all the concepts. These results provide insights into expert search procedures, and motivate the design of future search systems that guide users in the retrieval of comprehensive information.

Keywords

Healthcare, web searching, distribution of information.

INTRODUCTION

Numerous organizations have invested huge resources to develop accurate and comprehensive healthcare sites. For example, the National Cancer Institute's website has 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 such source. However, information scientists have repeatedly argued that as the number of information sources about a specific topic increases, the information across the sources follows a Zipf [4] distribution 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 [1].

Because the incomplete retrieval of healthcare information can have dangerous consequences, we believe the distribution of healthcare information deserves close inspection. This paper describes the first step of our analysis to understand this distribution. We believe such analyses should lead to a deeper understanding of why expert healthcare searchers visit different types of sources in identifiable sequences [2], and how such knowledge can be made available to help users find comprehensive healthcare information.

CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-637-4/03/0004

ANALYSIS OF THE DISTRIBUTION OF HEALTHCARE INFORMATION ON THE WEB

Our analysis focused on the distribution of melanoma risk information, a healthcare topic that is well researched, and widely available on the Web [3]. The goal of our survey was to understand not only how melanoma risk concepts were distributed across relevant webpages, but also the amount of such information in each page.

Identification of concepts and pages related to melanoma risk A skin cancer expert identified 12 concepts that were necessary for a comprehensive understanding of melanoma risk. These consisted of 8 hereditary factors (e.g. fair skin), 1 lifestyle factor (high exposure to ultra-violet rays), 2 general statistics (e.g. general life time risk), and a personal risk estimate calculated from the above factors.

Given that there exist 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 by the union of all the sites pointed to by the melanoma page in MEDLINEplus (a leading healthcare portal), and the top 5 most comprehensive sites identified in a recent study of online melanoma information [3]. 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 12 melanoma risk concepts, and for general melanoma risk. We therefore used 130 Google queries (e.g. "melanoma fair skin site:cancer.gov"), and retrieved the top 10 hits returned from each query. Subsequently, duplicates, news items, pages for health professionals, non-English pages, and broken links were removed. This resulted in 145 unique webpages, which we believe represented a set with a high probability of containing reliable and comprehensive melanoma risk information.

Method A printed version of the 145 webpages was given to a rater who judged the extent to which the 12 concepts related to melanoma risk were covered in each page based on a 5-point Likert scale (0=concept not covered on page, 1=concept covered in less than one paragraph, 2=concept covered in one paragraph, 3=concept covered in more than one paragraph, 4=webpage mostly devoted to concept, although other concepts 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 total 145 webpages.

Analysis and Results The raters had high agreement on whether or not a concept was present in a page (Cohen's

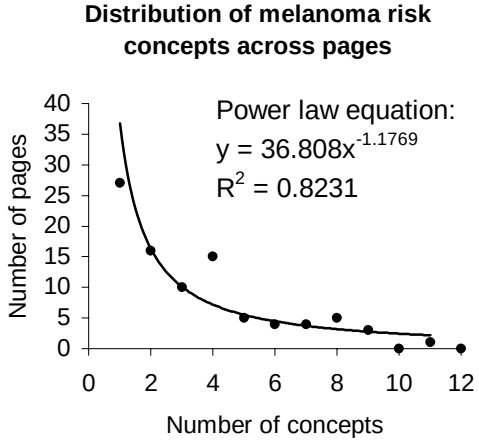

Figure 1. The occurrence of melanoma risk concepts in pages from reliable healthcare sites follows a power distribution (R^2=0.8231, p<.001). The distribution also shows that there were no pages that contain all the concepts related to melanoma risk.

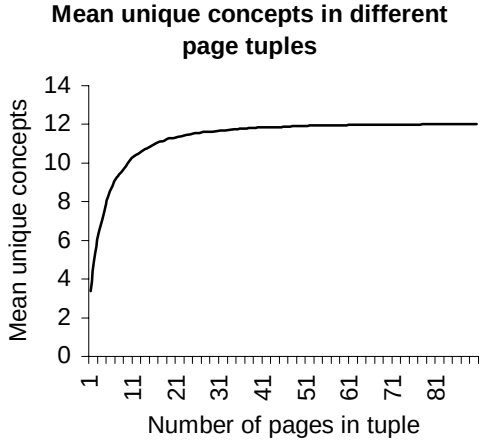

Figure 2. The distribution of the mean number of unique concepts contained in 1000 randomly selected pages for each n-tuple (page combinations of 1, 2, 3, etc.). The distribution estimates that on average users must visit about 25 pages before they have access to all 12 melanoma risk concepts.

kappa=.86), and the extent to which a concept was covered on a page (Cohen's weighted kappa=.80).

Figure 1 shows a scatter plot of the number of webpages that contain an ascending number of melanoma risk concepts (55 pages with no concepts were dropped). As shown, the distribution follows a power law, where a few pages contain many concepts, while many pages contain a few concepts.

Although the above distribution is similar in principle to the Zipf distribution [4], it does not explain why over 80% of the pages from *reliable* sites contained less than half of the concepts. A cursory analysis of pages at both ends of the distribution revealed that pages with many concepts appeared to provide information in not much detail, while pages with a few concepts appeared to provide a lot of detail about a few concepts. 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 concepts (p<.01, mean number of concepts=4.37, SD=2.60) compared to pages that had a maximum detail level of 4 (mean=2.56, SD=1.95), or a maximum detail level of 1 (mean=1.67, SD=.91). This suggests the existence of *general* pages that have medium amounts of detail, *specialized* pages that cover few concepts in a high level of detail, and *sparse* pages that contain few concepts in very little detail. The analysis of detail therefore provides an explanation for the distribution, which is a departure from typical distribution studies in healthcare that focus on concept occurrence and accuracy [e.g. 3].

The tail of the distribution in Figure 1 also shows that there were no pages that contained all the 12 concepts. To estimate how many pages users must visit before they can obtain comprehensive coverage of melanoma risk concepts, we calculated the mean number of unique concepts contained in 1000 randomly selected pages for each n-tuple (page combinations of 1, 2, 3, etc.). As shown in Figure 2, *assuming each page has an equal chance to be visited*, users on average must visit about 25 pages before they have access to all the concepts related to melanoma risk.

CONCLUSION

The analysis suggests that users seeking a comprehensive understanding, of even a common healthcare topic such as melanoma risk within high quality pages, have a fairly complex task. They must first visit more than one general page to get an overview of all the melanoma risk factors, and then visit specialized pages to get an in-depth understanding about specific concepts such as a personal risk assessment. Such search procedures are similar to what search experts have been observed to use, and because they are difficult to acquire just from using search engines like Google [2], motivate the design of new approaches to search systems that explicitly provide such guidance [2]. Furthermore, the regularities identified through such analyses should suggest automatic, or semi-automatic ways to identify search procedures that guide users in retrieving comprehensive information in critical domains such as healthcare.

REFERENCES

1. Bates, M.J. Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors. *JASIST 49*, 13 (1996), 1185-1205.
2. 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 the CHI'03,* (in press).
3. 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.
4. Zipf, G. K. Human behavior and the principle of least effort: *An introduction to human ecology*. Addison-Wesley, Cambridge MA, 1949.