MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1055

Enhancing a Taxonomy for Health Information Technology: An Exploratory Study of User Input Towards Folksonomy

Brian E. Dixon[a,b], Julie J. McGowan[a,b,c]

[a] *Medical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA*
[b] *National Resource Center for Health IT, Agency for Healthcare Research and Quality, Rockville, Maryland, USA*
[c] *Department of Knowledge Informatics and Translation, Indiana University School of Medicine, Indianapolis, Indiana, USA*

Abstract

The U.S. Agency for Healthcare Research and Quality has created a public website to disseminate critical information regarding its health information technology initiative. The website is maintained by AHRQ's National Resource Center (NRC) for Health Information Technology. In the latest continuous quality improvement project, the NRC used the site's search logs to extract user-generated search phrases. The phrases were then compared to the site's controlled vocabulary with respect to language, grammar, and search precision. Results of the comparison demonstrate that search log data can be a cost-effective way to improve controlled vocabularies as well as information retrieval. User-entered search phrases were found to also share many similarities with folksonomy tags.

Keywords:

Classification, Information retrieval, Internet, Health informatics

Introduction

Since September 2004, the U.S. Agency for Healthcare Research and Quality (AHRQ) has invested over $266 million in its health information technology (health IT) initiative. The goal of AHRQ's investment is to develop and disseminate health IT evidence and evidence-based tools to support AHRQ's overall mission of improving the quality, safety, efficiency, and effectiveness of health care for all Americans. A major component of AHRQ's health IT initiative is the National Resource Center (NRC) for Health IT, initially created by AHRQ to assist its grantees and contractors. Today the NRC is a public resource for those interested in implementing and using health IT. The main point of interaction with the AHRQ NRC is through its website[1].

Since its launch in 2006, the NRC has strived to continuously improve the website. To accomplish its goal, the NRC routinely captures usage metrics and user feedback, considered a best practice in the industry [1]. The metrics and feedback are used to identify website performance and usability issues. In addition, the data are used to identify content gaps. The NRC website supports a large, heterogeneous group of informatics practitioners and researchers, including novice providers just starting down the path towards implementation, adoption, and usage of health IT applications.

During the past two years, the NRC has focused on supporting these novice users in their quest to find knowledge resources to support local implementation and adoption of health IT. Part of this support involved the creation of a controlled vocabulary to describe the diverse content available on the site. The taxonomy of health IT terminology was used to organize web pages and index items made available through a search function. However, initial usability testing revealed that the taxonomy was confusing to novice users who did not use the same language and grammar as that used by the experts who created the taxonomy [2]. Based on the results of the usability testing, the NRC team removed the taxonomy from the site's outward facing information architecture but continued to use it for categorizing knowledge resources on the back-end.

The initial emphasis on users' ability to effectively browse the website, or click through the various pages to find information and knowledge resources, resulted in a lower priority for improvements to the search function. This changed in 2008 when usability testing revealed that many users, both novice and experienced, were frustrated with the website's search function. Furthermore, usability testing results showed that experienced site users tend to use search first, rather than browse through a site's information architecture.

To improve the site's search function, the NRC focused on enhancing its taxonomy. Best practice in information retrieval calls for the use of domain specific, controlled vocabularies to index the content made available through the search function [3]. Since the NRC already employed a controlled vocabulary to index its content, efforts focused on enhancing the taxonomy to improve search queries.

Recent information science literature has described the use of a folksonomy, in addition to or in place of, a controlled vocabulary to improve the user experience for searching. A folksonomy involves the use of open-ended, collaboratively generated metadata (or tags) for categorizing a site's content [4]. Folksonomies are further referred to as social bookmarking, social tagging, and social classification based on the fact that users typically create metadata tags for themselves and then

[1] http://healthit.ahrq.gov

share the content and tags with others. Common folksonomy websites include Digg, Delicious, and CiteULike.

Whereas taxonomies are top-down, controlled vocabularies, created and maintained primarily by librarians or domain experts, folksonomies are bottom-up, uncontrolled vocabularies that utilize familiar, accessible, and shared concepts created and maintained by a community of users [3,4]. In addition, folksonomies may have several advantages over taxonomies. First, folksonomies have been described as dynamic and forward looking with the capacity to categorize unforeseen subject matter, including emerging technologies [5]. Second, folksonomies may be a less expensive alternative to the development, maintenance, and enforcement of a tightly controlled vocabulary [5,6]. Finally, folksonomies may have a gentler learning curve for novice users [6]. These advantages may be attractive to a publicly funded program with limited resources for long term site development and maintenance.

Creating the ability for users to develop and share folksonomy tags on the NRC site would be difficult. U.S. Government policies restrict agencies from collecting users' names and other identifying information without strong oversight [7]. Before the NRC could ask for permission to enable users to login to a personal profile, create folksonomy tags, and view other users' tags, the Web team desired to explore the use of user-generated language and grammar to enhance the search function. In this paper, we outline our methods for approximating a folksonomy with user-generated search queries and present the results of an exploratory study in which user-centered language and grammar is compared with the expert-created taxonomy. We further suggest how folksonomies and other forms of user-entered concepts can be used to improve taxonomies as well as search functionality.

Materials and Methods

The AHRQ National Resource Center website utilizes the social networking plug-in AddThis (www.addthis.com). This application enables users to share web pages and content items with others via third-party Web 2.0 applications, including Twitter, Delicious, and Digg. Although these third-party applications make sharing easy for users by leveraging existing infrastructure, they do not allow AHRQ to easily review the content tags assigned to the items shared by users. Attempts to retrieve public folksonomy tags via Delicious and Digg for AHRQ pages and content did not yield substantive results. Nearly all of the users utilizing these services are storing the links and tags as private, perhaps sharing them with a limited number of peers or using them as personal bookmarks. This prevented a direct evaluation of folksonomy tags associated with NRC information and knowledge resources.

Therefore we approximated user-generated concept tags by utilizing an available data source, the maintenance logs of the NRC website. These logs contain many data on anonymous users' interactions with the site, including search phrases and keywords automatically captured each time users perform a query. The logs are comprehensive, and they are routinely used for other performance and usability monitoring.

Our hypothesis was that user-entered search phrases and keywords, extracted from queries, would exhibit the same characteristics as folksonomy tags. When testing the NRC site's taxonomy, we observed that users typically searched for information using concepts and phrases from their language and grammar. These concepts did not always overlap with the highly controlled vocabulary used in the first version of the site's taxonomy [2]. So in our search to identify an alternative, practical source for pilot data to evaluate the potential use of a folksonomy, we hypothesized that user-entered search terms may reflect users' language and grammar in a similar way to that of folksonomy tags.

We examined twelve months worth of search logs ranging from July 1, 2008 through June 30, 2009. The logs from December 2008 were corrupt, so they were excluded from the final analysis. A total of 34,816 user-entered search phrases were extracted from the logs and analyzed.

Three analytical methods were employed to review the search phrases and determine their appropriateness as a source of quality improvement data. First, the occurrence of each search phrases was counted, and the top 100 phrases were analyzed for patterns and trends. Our belief was that the search phrase patterns and trends would be similar to those observed of folksonomies by previous information science researchers.

Second, the top 100 phrases were mapped to the National Resource Center's taxonomy [1] to qualitatively evaluate its robustness and identify gaps. The mapping was also performed to examine the search phrases. We believed that the phrases would represent health IT concepts in the natural language and grammar of the end users.

Third, a non-random sample of five search phrases from the top 100 was selected for additional qualitative review. Each original search phrase and its mapped taxonomy concept were used to execute independent searches of the website. The search results were then examined for relevance. The search results were evaluated using 10-*Precision* method as described by Pera [8] and defined in Equation (1). This 10-*Precision* equation produces a precision value for the top 10 search results for a given query (*Q*).

$$10\text{-}Precision = \frac{\#_{of_{Retrieved_{Relevant_{Records}}}}}{10} \quad (1)$$

Precision values for the five user-generated search phrases and repeated, independent searches using the mapped taxonomy concepts were calculated. The values were then compared and contrasted. We hypothesized that the mapped concept precision values would be higher for each of the 5 paired queries. We further hypothesized that precision values for the mapped concept queries would be 1.0 since the taxonomy was engineered to facilitate precise information retrieval.

Results

Top 100 User-Entered Search Phrases

The 34,816 log records contained 8,574 unique search phrases. The number of occurrences for each unique phrase was counted for analysis. When sorted in descending order, the search phrases reveal an inverse logarithmic relationship as

shown in Figure 1. The curve begins to level off after the fifth most popular phrase, and just 30 phrases were entered more than 100 times in the eleven month period. These top 30 phrases were entered 12,707 times, which represents 36.5% of the total phrases observed over the 11 month period. The top search phrase, "health information technology," was entered 2,650 times and accounted for 7.6% of the total queries.

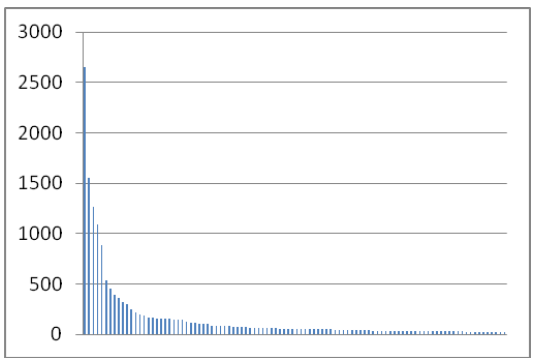

Figure 1- Distribution of the Top 100 User-Entered Search Phrases for http://healthit.ahrq.gov

Mapping User-Entered Search Phrases

Ninety-three of the top 100 search phrases were successfully mapped to existing taxonomy concepts. Thirty-three percent of the mapped terms pertained to just four political/administrative concepts: the Office of the National Coordinator for Health Information Technology, the Health IT Policy Committee, the Health IT Standards Committee, and the American Health Information Community.

The seven phrases that could not be mapped to existing taxonomy concepts are presented below in Table 1. The "knowledge library" concept refers to the area of the site where searches are executed, but the concept itself is not represented in the taxonomy. The concept "betaa" appears to be a misspelling. Two concepts (medical home, medical home model) are related, and they represent a health care delivery concept frequently discussed in the American medical community. The remaining three concepts appear to be terms invoked by robots scanning the site for downloadable content.

Table 1 – User-Entered Phrases for Which No Taxonomy Concept Existed

Search Phrase
knowledge library
default_collection
medical home
Saleslogix
xml_no_dtd
medical home model
betaa

User-Entered Search Phrases versus Taxonomy Concepts

Five unique search phrases from the top 100 were non-randomly chosen for additional analysis by the primary author. The author selected phrases that did not pertain to overly general concepts, such "health information technology," and phrases were further selected to represent a range of concepts.

For each phrase, two searches were performed. First, the original, user-entered phrase was used. Second, the mapped taxonomy concept was used. A total of 10 queries were independently performed. For each query, the authors examined the search results and calculated a 10-*Precision* value using Equation 1.

The selected search phrases and their 10-*Precision* values are summarized in Table 2. The original, user-entered phrases are listed first, followed by their 10-*Precision* values. Next the mapped taxonomy concepts are listed, followed by their 10-*Precision* values. The results show a general trend in which the taxonomy concepts performed as well or slightly better than the user-entered phrases.

Table 2 – Selected Search Phrases and Their Precision Scores

User-Entered Phrase	10-*Precision*	Mapped Taxonomy Concept	10-*Precision*
system implementation	0.9	Implementation	0.9
emr readiness assessment	0.2	Readiness Assessment	1.0
snomed	0.9	Coding Standards -> SNOMED	0.9
cpoe systems	0.6	Systems -> Computerized Provider Order Entry	0.9
time and motion study pdf healthcare	0	Workflow Impact -> Efficiency of Care	0.4

Discussion

Proponents of folksonomies suggest that they should replace traditional controlled vocabularies given the latter's limitations and expense [9]. Others in the information retrieval community view folksonomies as potential supplements to taxonomies, enhancing knowledge management practices with input from users [3,6]. Based on our review of the literature and exploratory study, we believe that folksonomies may be able to play a supporting role. We further assert that there are other sources of user input that can enhance information retrieval and knowledge management, namely search logs.

When we originally sought to enhance the NRC taxonomy, we looked towards a folksonomy. Because U.S. Government website policies do not favor private user accounts that collect

identifiable information [7], we turned to an alternative source of user input we believed would exhibit the same characteristics as folksonomy tags. Our exploration of this data source confirmed our hypothesis, revealing that user-generated search phrases indeed share characteristics with folksonomy tags.

Similarities between Folksonomy Tags and Search Phrases

User-generated search phrases are like folksonomy tags in the sense that they are uncontrolled. Folksonomies have been described as lacking rigor in their use of spelling, parts of speech, and use of plurals [4,6,10]. For example, the concepts "cat" and "cats" are typically unique concepts in a folksonomy, even though more structured vocabularies would relate the concepts to one another.

We found similar patterns within the search logs. For example, users varied in their use of "cost" versus "costs" when searching for information on the "typical cost" of a computerized provider order entry (CPOE) system. Thus one can imagine that if users were asked to tag a study on the "costs and benefits" of health IT [11], some users would use a "cost" tag while others would use a "costs" tag.

Also, like folksonomy tags, user-generated search phrases tended to be broad and less specific than the NRC taxonomy. Previous studies have shown that when users tag content when contributing to a folksonomy, they often choose the cognitive path of least effort [12]. The example described by Munk and Mørk involved an article by Milton Friedman in the *New York Times*. Folksonomy users tended to label the article with very broad tags, such as "business," "economics," and "politics," whereas the article dealt primarily with the concept of corporate social responsibility.

We found that many of the top 100 search phrases exhibited similarly broad concepts. The top search phrase, accounting for 2,650 (7.6%) of the 34,816 total phrases, was "health information technology." Other broad phrases in the top 100 included: "hit standards," "it tools," and "data reporting." These labels would likely apply to many of the articles, white-papers, and other information resources found on the AHRQ Health IT website, making them generally unhelpful to users seeking more narrow concepts such as EHR adoption, standard clinical vocabularies, and project management tools.

Search phrases are like folksonomy tags in the sense they both follow inverse logarithmic or power law distributions. Folksonomy tag distributions are the subject of several detailed analyses [10,12]. Each analysis revealed a general trend whereby a handful of so-called "power tags" are very popular followed by the long tail of tags used sparsely.

We observed a similarly long tail of unique search phrases, accounting for nearly two-thirds of all search phrases. However, the caveat is that we did not apply any intelligence to the raw search phrases, matching them to similar concepts or attempting to relate them to each other in any fashion. It could be that many of the phrases overlap in semantic meaning and, as a result, there may be a shorter tail than these data would otherwise indicate.

Synonymy is another similarity between folksonomies and search phrases. Folksonomies are described as possessing large semantic overlap between tags [5,6,12]. Often this is a function of the folksonomy platform. Delicious, for example, does not permit spaces in tags. Therefore "New_York_City" and "NewYorkCity" are distinct tags with no relationship. In a controlled vocabulary, the synonymy of these two concepts would be managed by content experts.

When reviewing the top 100 search phrases, we observed quite a bit of semantic overlap. Twelve of the top 100 phrases conveyed the broad subject of "health information technology." If these variants were used as distinct, unrelated tags in a folksonomy, they would likely not improve the precision of the search function. Users would instead need to execute 12 queries, one for each variant of "health IT," to retrieve all items tagged with the various synonyms of "health IT."

Enhancing Taxonomies and Information Retrieval

Folksonomies have been described as forward thinking, meaning that they keep pace with changing language, grammar, and trends in society (e.g., emerging technologies and concepts) [6,9,10]. This benefit is one of many reasons that many web managers and information system designers are looking towards folksonomies to enhance or complement controlled vocabularies. Our exploratory study of user-entered search phrases revealed not only that search log data are similar to folksonomy tags but that search phrases can be used in a similar manner to enhance controlled vocabularies and improve information retrieval within a web site or application.

Proactive monitoring of search log data to enhance controlled vocabularies yielded three main benefits. First, the review of search phrases identified users' evolving language, grammar, and search behavior. Consider the 12 variants of "health IT." Using the search logs, additional variants of this concept were identified and mapped as synonyms of the general term "health information technology." This process expands and enhances the controlled vocabulary and will likely to lead to search function improvements, since users could enter any of the 12 variants and, once the synonyms are linked within the taxonomy, and receive a very similar list of precise search results.

We further identified a concept that was not yet in the taxonomy: medical home. Although medical home is not a core informatics concept, much of the discussion in the U.S. surrounding medical home development and maintenance has involved the use of health IT systems to enable efficient and effective coordination of care among a diverse set of providers. Therefore this concept is an important, related concept that should be represented in a controlled vocabulary designed to encompass the field of health IT.

We also noted unanticipated patterns of users' search behavior. Consider the search phrases for four government entities, the Office of the National Coordinator for Health Information Technology (ONC), the Health IT Policy Committee, the Health IT Standards Committee, and the American Health Information Community (AHIC), which accounted for a significant number of searches. Three of these four concepts are non-permanent government committees, which raises the issue of whether currently popular labels should be included within the controlled vocabulary or incorporated in other ways (e.g., folksonomy tags, related terms).

Second, enhancing the taxonomy using user-entered search phrases improved the precision of the site's search function.

The 10-*Precision* values associated with mapped taxonomy concepts were greater than or equal to the values associated with the original search phrases. If the controlled vocabulary was routinely enhanced to reflect users' behavior and language, then the search function should significantly improve over time. Furthermore, we learned that the fifth term, "Workflow Impact -> Efficiency of Care," yielded a low precision value and should therefore be modified. User data might therefore benefit taxonomies beyond just the identification of synonyms and new terms.

Finally, the use of search phrases can be cost-effective. The search logs used to collect the data were already a foundational component of the website. They worked in the background, logging queries. The process of extracting the search phrases from the logs was quick and simple. Loading the phrases into an application for review and analysis also took less than a half a day. In total, less than one business day per quarter could be devoted to active review of search logs for new phrases and patterns of usage.

Other sites and applications that host controlled vocabularies could benefit from this study. For example, the Medical Subject Headings (MeSH) vocabulary maintained by the U.S. National Library of Medicine (NLM) currently does not contain concepts for health information exchange (HIE) and personal health records (PHRs). Reviewing search logs would probably reveal a number of queries for these concepts, which could prompt NLM to add them more rapidly to the MeSH tree than through the usual process of search term development. EHR and PHR applications might also benefit from collecting and analyzing user search data, which could aid in the retrieval of patients' health information or relevant evidence-based medicine knowledge.

Limitations of the Study

There are several limitations of this study. First, the study was exploratory in nature. The data were gathered from just one website, and they were not cross-checked with similar data from other websites. Second, only one expert was involved in reviewing the data and mapping search phrases to taxonomy concepts. Third, the sample used for analyzing precision was small. Finally, the study did not take advantage of natural language processing (NLP) techniques which could be utilized to perform automated scanning of the data to find new patterns and trends. NLP methods might also help us understand the structure of search phrases, which in turn may help to develop new methods for enhancing the search function interface and the way queries are formed by novice and intermediate users.

Conclusion

Website and application search logs contain user-generated phrases and keywords that exhibit similar characteristics to folksonomy tags. Using user-entered search phrases and keywords to enhance controlled vocabularies can be a cost-effective strategy for improving information retrieval. It may also be an effective complement to approaches, including folksonomies. Health informatics websites and applications should consider this technique to improve information retrieval and the overall usability of end-user products.

Acknowledgments

This paper is derived from work supported under a contract with the U.S. Agency for Healthcare Research and Quality (290-04-0016). The opinions expressed in this article are those of the authors and do not reflect the official position of AHRQ or the U.S. Department of Health and Human Services.

References

[1] Wood FB, Siegel ER, LaCroix E-M, Lyon BJ, Benson DA, Cid V, Fariss S. A practical approach to e-government Web evaluation. IT Pro. 2003; 5(3): 22-28.

[2] Dixon BE, Zafar A, McGowan JJ. Development of a taxonomy for health information technology. MEDINFO. 2007: 616-620.

[3] Gruber T. Ontology of folksonomy: A mash-up of apples and oranges. Intl J Semantic Web Info Sys. 2007. 3(1):1-11.

[4] Noruzi A. Folksonomies: (Un)controlled vocabulary? Know Org. 2006. 33(4): 199-203.

[5] Bruce R. Descriptor and folksonomy concurrence in education related scholarly research. Webology. 2008 Sep; 5(3): Article 59.

[6] Spiteri LF. Structure and form of folksonomy tags: The road to the public library catalogue. Webology. 2007 Jun; 4(2): Article 41.

[7] Wood FB, Siegel ER, Feldman S, Love CB, Rodrigues D, Malamud M, Lagana M, Crafts J. Web evaluation at the US National Institutes of Health: use of the American Customer Satisfaction Index online customer survey. J Med Internet Res. 2008 Feb 15;10(1):e4.

[8] Pera MS, Lund W, Ng YK. A sophisticated library search strategy using folksonomies and similarity matching. J Am Soc Info Sci Tech. 2009. 60(7):1392-1406.

[9] Shirky C. Ontology is overrated: categories, links, and tags. 2005 [cited 2009 Oct 9]. In: Clay Shirky's Writings About the Internet [Internet]. New York: Clay Shirky. Available from: http://www.shirky.com/writings/ontology_overrated.html

[10] Peters I, Weller K. Tag gardening for folksonomy enrichment and maintenance. Webology. 2008 Sep; 5(3): Article 58.

[11] Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton SC, Shekelle PG. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006 May 16;144(10):742-52.

[12] Munk TB, Mørk K. Folksonomy, the power law & the significance of the least effort. Knowl Org. 2007; 34(1): 16-33.

Address for correspondence

Brian E. Dixon, MPA, PhD (Cand)
Regenstrief Institute, Inc.
410 West 10th Street, Suite 2000
Indianapolis, IN 46202
U.S.A.
+1 (317) 423-5582

bdixon@regenstrief.org