A Comparative Analysis of User-generated and Author-generated Metadata for Web Resources

Bing He, Ying Ding
School of library and Information Science
Indiana University Bloomington, IN, USA
{binghe, dingying} @ indiana.edu

Staša Milojević, Elin K. Jacob
School of library and Information Science
Indiana University Bloomington, IN, USA
{smilojev, ejacob} @ indiana.edu

ABSTRACT
In this paper, we investigate the difference between metadata generated by users and authors. Delicious tags and HTML keyword META tags associated with the same set of web pages on topics related to semantic web are collected, forming two datasets (i.e., Delicious dataset and HTML dataset). Comparisons of the two datasets in micro and macro vocabulary overlap as well as classification of web pages are analyzed. The results show that (1) overlap between the two datasets exists; (2) non-overlapped tags in Delicious dataset reveal systematic deficiency of social tagging systems; non-overlapped tags in HTML dataset expose organization-oriented contents; and (3) Delicious dataset tends to cluster web pages according to their popularity and subject area while HTML dataset clusters the web pages according to different websites/authors.

Keywords
Metadata, social tagging system, HTML META tag, tag.

INTRODUCTION
Metadata is structured data about data, which provides powerful support for resource discovery and data use (Lynch, 1998). Before web 2.0, metadata has been generated in two ways: by professionals and by authors (Greenberg et al, 2002). Professional generation of metadata is usually used by catalogers in traditional libraries (Mathes, 2004). Author generation of metadata typically means that the creators of intellectual content (i.e., authors) provide information about their work (e.g., title, subject) according to Dublin Core. In age of Web 2.0, author described documents are promoted by HTML/XHTML. The keyword META tag is one of the most commonly used META tags (Craven, 2000), as it was supported and popularized by Internet search engines (e.g., Infoseek, AltaVista) in late 1990s and early 2000s.

Professional generation of metadata cannot keep up with the vast amount of content on the Web, which can be solved by author generation of metadata. Recently, social bookmarking websites have introduced and popularized user generation of metadata which can also address the scalability issue. They allow users to interactively annotate web resources using a set of unstructured and descriptive terms (i.e., tags). Frequently-occurred terms are assumed to represent the consensus among users.

While Semantic Web is closely related to RDF, ontologies, etc., every HTML document improves the understandability of the Web by the use of headings, lists, titles and other semantic markup wherever possible; the use of “tagging” creates folksonomies that may be equally meaningful (Shadbolt, Berners-Lee, & Hall, 2006). User generation and author generation of metadata provide metadata for web resources from quite different standpoint. In this paper, we compare users-generated tags and author-generated HTML keyword META tags of web pages with topics related to semantic web, so as to see what impact the different identities (i.e., users and authors) has been imposed on the syntax and semantics of generated metadata.

METHODOLOGY
All webpages that have been tagged with rdf on Delicious at least once between October 7, 2005, and January 30, 2008 were collected (George et al., 2008). 1,576 web pages (around 30% of all the webpages), which have both Delicious tags and HTML keyword META tags (referred to as Delicious dataset and HTML dataset in the rest of the paper), form the sample of this study.

RESULTS
Macro vocabulary overlap
The two datasets share 2,226 unique tags, which account for 7.8% of Delicious unique tags and 50.0% of HTML keyword unique tags. However, with occurrence of the shared tags into consideration, the percentages increase to 41.9% of Delicious tags and 82% of HTML tags, suggesting that those shared tags tend to have high frequencies in both datasets. Meanwhile, occurrences of shared tags in two datasets across web pages are highly correlated (the correlation coefficient is 0.73, \( p=0.0001 \)), indicating certain level of consensus between Delicious users and authors of websites.
The 10 most popular shared tags ranked by occurrences in Delicious dataset and HTML dataset include rdf, semanticweb, web, semantic, xml, web2.0, metadata, and ontology. Most of them are commonly used terms in the field of semantic web. The 10 most popular tags that are not shared by the two datasets are presented in Table 1. Two most frequently used tags in the Delicious dataset “semantic_web” and “semantic-web” are irregular forms of “semantic web”, which supports Mathes’ (2004) argument that not allowing space within one tag contributes to the inconsistency over social tagging vocabulary. Tags such as “toread”, “todo”, and “work” represent another significant group of Delicious tags that relate to individual users’ purposes and tasks. As for HTML keyword META tag dataset, “o’reilly” and “oreillynet.com” reflect the promotion motive of authors of web pages (o’reilly is a famous publisher); “tttxca”, “ttjca”, and “tttwaca” are particular abbreviations used within the scope of one website. “dumbill” and “edd” denote Edd Dumbill, who is a co-chair of the O’Reilly Open Source Convention. Authors of web pages pay additional attention to their own visibility. The Delicious tags for individual use and the META tags for ownership of websites expose the different motivations of users and authors.

CLASSIFICATION OF WEBPAGES

First, we measure the similarity between pairs of web pages within each dataset based on co-occurrence of tags and keyword META tags respectively (Cattuto et al., 2008):

\[ S_{t1,t2} = \frac{\sum_{t \in T_1 \cap T_2} \min(f_1(t), f_2(t))}{\sum_{t \in T_1} \max(f_1(t), f_2(t)) + \sum_{t \in T_2} \min(f_1(t), f_2(t)) - \sum_{t \in T_1 \cap T_2} \min(f_1(t), f_2(t))}, \]

where we denote with \( T_1 \) and \( T_2 \) the two sets of tags attached to a pair of web pages, with \( f_1(t) \) and \( f_2(t) \) the local occurrence of tag \( t \) in \( T_1 \) and \( T_2 \), and with \( f(t) \) the global frequency of tag \( t \).

We further apply hierarchical clustering analysis (HCA) to the similarity matrix obtained in last section. For comparison purpose, the web pages are reduced to four clusters for both datasets. Figure 1 shows permuted membership of each web page in the four clusters in the two datasets respectively. Each webpage has a unique color and blue lines are the boundaries between adjacent clusters. As shown in Fig. 1, every Delicious cluster contains web pages from all the four clusters in HTML dataset, which shows that HTML keyword META tag provides additional classification criterions that are different from users’ perspective. The most popular tags in each cluster for both datasets have shown those additional criterions, including: (1) organization dimension: “o’reilly”, “D-Lib” and “firefox” etc; (2) ownership of the website: “Berners-Lee”, “spivack”, and “nova” (Nova Spivack).

DISCUSSION AND CONCLUSION

The results show that for web pages with topics related to semantic web and RDF, user-generated metadata and author-generated metadata are quite different. A considerable portion of Delicious tags deals with particular individual tasks while HTML keyword META tags tend to emphasize on the ownership/authorship of web pages. Future work involves performing the analysis on a larger dataset and multiple subject areas.

REFERENCES


