Webology, Volume 5, Number 3, September, 2008

Descriptor and Folksonomy Concurrence in Education Related Scholarly Research

Robert Bruce
Dr. Martin Luther King, Jr. Library, San JosÃ© State University, San JosÃ©, California, USA.
Email: Robert.Bruce (at) sjsu.edu

Received August 6, 2008; Accepted September 25, 2008

Abstract

Folksonomies are a decentralized yet collaborative form of classification based on user-defined keywords (also known as tags). Although this uncontrolled method of classification lacks rules for term standardization and usage, it has potential for organizational patterns and an emerging vocabulary (terminology). The objective of this research is to analyze the descriptors and tags from journal articles indexed in the Education Resources Information Center (ERIC) and the folksonomy-based website CiteULike to determine overlap between the controlled and uncontrolled vocabularies. Metadata from 2,786 journal articles indexed in ERIC and CiteULike was collected using Perl and MySQL. The total metadata was comprised of 2,899 unique ERIC descriptors, 3,176 unique CiteULike tags, and 1,083 unique CiteULike users. An analysis of this metadata revealed that 240 of the CiteULike tags matched ERIC descriptors. The low number of tag-descriptor matches in this research indicates that CiteULike users do not use the same terminology as subject specialists who maintain descriptors in the ERIC thesaurus.

Keywords

Collaborative tagging; Social tagging; Social classification; Knowledge organization; Taxonomies; Folksonomies; Controlled vocabulary; Descriptors

Introduction

Social networking and collaboration, key elements of the "Web 2.0" paradigm, have brought forth a new method for organizing information. Web 2.0 oriented websites such as Flickr, Delicious, and Library Thing allow users to share and associate keywords or "tags" with their favorite photographs, URL bookmarks, or books, respectively. Depending on context, a tag can be either a description of an entity or a process for categorization. Vander Wal (2007) defined the tagging classification process as a "folksonomy," a portmanteau of the words "folks" (in other words, the users who tag) and "taxonomy" (Â¶ 7).

Lambiotte and Ausloos (2005) describe folksonomies as "anarchic" and "democratic" (p. 4). This classification scheme lacks any form of governance or authority that oversees term standardization and usage criteria. As a result, anyone may contribute tags they have created to the folksonomy as a whole. In essence, the tags that comprise a folksonomy are an uncontrolled vocabulary. The uncontrolled nature of tags has two advantages: (1) they are forward compatible - expandable - to categorize future unforeseen subject matter (for example, emerging technologies), and (2) they may quickly evolve with language. However, the lack of control in tags is also an inherent disadvantage, particularly in regards to homonymity and synonymy. Homographic tags are ambiguous while synonymous tags are categorically inconsistent and redundant. A controlled vocabulary resolves issues such as homonymity and synonymy through qualifiers and equivalence relationships. Qualifiers differentiate the meaning of homographs while equivalence relationships link synonymous terms to a preferred (or controlled) term. The principle motivation for utilizing controlled vocabularies is to establish a consistent framework for defining term meaning, scope, and relationship for purposes of knowledge representation, maintenance, and retrieval (National Information Standards Organization [NISO], 2005, pp. 10-11).

Developing and maintaining a controlled vocabulary may be a costly and time consuming process, particularly in regards to vocabulary design, backwards compatibility (to preserve semantic relationships), and expandability (to describe future terminology). For these reasons, folksonomies may serve as a useful resource for term choice or maintenance. Cattuto, Loreto, and Pietronero (2007) note that a frequently

occurring tag can become a community-driven, preferred term to describe or categorize a given concept (p. 1461). Golder and Huberman (2006) observed this phenomenon statistically when analyzing tag usage over time on the website Delicious.

In the first of three studies, Lin et al. (2006) compared controlled vocabularies and folksonomies that were assigned to medical-related journal articles. The research involved empirically measuring the similarity between tags from the folksonomy-based website Connotea and the medical-related controlled vocabulary MeSH (Medical Subject Headings) from PubMed for 45 documents indexed on both sites (Lin et. al., 2006, pp. 5-6). The results indicated that approximately 11% of the 540 Connotea tags evaluated matched PubMed's MeSH terminology (Lin et. al., 2006, p. 6). Following these results, Lin et. al. (2006) noted that tags from Connotea tended to be less formal than the controlled vocabulary in MeSH (p. 7). Furthermore, Lin et. al. (2006) observed a personalization effect in folksonomies: Connotea users limited their tags to certain facets of an article rather than a holistic description of the article's subject matter (Lin et. al., 2006, p. 6).

Lin, Beaudoin, Bui, and Desai's (2006) comparison of folksonomies and Medical Subject Headings (MeSH) brings forth further research to consider: the level of similarity between folksonomies and controlled vocabularies in a discipline other than medicine. The objective in this research is to determine the extent that a controlled vocabulary (descriptors) from the Educational Resources Information Center (ERIC), an education-based online index sponsored by United States Department of Education's Institute of Education Sciences, matches the "uncontrolled" vocabulary (tags) created by users of CiteULike for journal articles indexed by both sites.

CiteULike is a folksonomy-oriented website by which users create and share bibliographic citations of their favorite scholarly research ("CiteULike: Frequently asked questions", n.d.). Users may categorize their references by assigning one or more user-defined tags to each bibliographic citation. CiteULike tags may contain letters, numbers, hyphens or underscores. CiteULike tags may not contain a space character; thus a multi-word tag such as "information retrieval" would need to be entered as something like "information_retrieval", "information-retrieval", or "InformationRetrieval" instead. CiteULike tags are not case-sensitive so "InformationRetrieval" is the same as "informationretrieval".

The Educational Resources Information Center (ERIC) is a searchable online index of bibliographic citations and abstracts (and occasional full-text) of education-based research ("About ERIC", n.d.). Citations within ERIC are classified by descriptors (a controlled vocabulary) to denote the subject matter of the research ("About the ERIC Thesaurus", n.d.). Each citation contains one or more descriptors. Each descriptor may contain one or more words separated by space (for example, "Cognitive Development"). ERIC descriptors may contain letters, numbers, hyphens, parenthesis, and the space " " character.

Method

Three Perl-based programs were written to retrieve, store (in a MySQL relational database), and analyze metadata and associated bibliographic citations of journal articles indexed in ERIC and CiteULike. Data retrieved from ERIC and CiteULike were comprised of bibliographic journal article citations (article title, author, journal title, volume, issue, year, and page) and their associated metadata. The metadata collected consisted of ERIC numbers (unique identifier for each article indexed on ERIC), ERIC descriptors (the specific terms in which a given article citation is indexed), CiteULike usernames (to uniquely identify a given user) and the CiteULike tags they created.

The data collection process occurred in three phases. In the first phase, a snapshot of the entire CiteULike metadata (including CiteULike usernames and their associated tags) was downloaded as a data dump file from CiteULike ("Available datasets", n.d.). In the second phase, a Perl-based program automatically traversed the CiteULike website for each tag listed in the metadata dump file. The data collected in the second phase included the bibliographic journal article citations (article title, author, journal title, volume, issue, year, and page) associated with each tag listed in the dump file retrieved in phase one. In the third phase, a second Perl-based program automatically searched the ERIC website by article title for every journal article collected from CiteULike in phase two. If an exact match (case-insensitive) by article title occurred, a subsequent verification of the remaining bibliographic citation on ERIC (author, journal title, volume, issue, and year) was compared to the remaining CiteULike bibliographic citation data (retrieved in phase two) for that article title. If a journal title was abbreviated (in ERIC or CiteULike), that title was manually searched in PubMed journals database and the California Institute of Technology Library's journal abbreviations webpage to determine the full (unabbreviated) title of the journal (for example, "*J Inf Sci*" is an abbreviation for "*Journal of Information Science*"). (This second comparison was performed to avoid journal articles with the same article title yet published in a different journal by different authors). If

the bibliographic journal citation in ERIC and CiteULike matched, the ERIC descriptors associated with that journal article were retrieved and stored in a MySQL database.

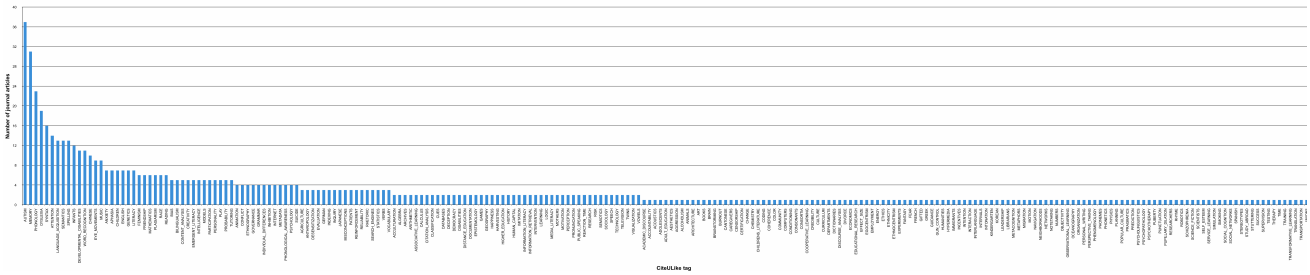
In the data analysis process, a Perl-based program counted the number of CiteULike tags that matched ERIC descriptors associated with the same journal article. Every CiteULike tag that matched an ERIC descriptor was counted only once, regardless of the number of articles that the tag and descriptor were associated with or the number of CiteULike users of that tag. For a match to occur, an ERIC descriptor and CiteULike tag had to exactly match (case-insensitive) and be identical in length (characters) with the exception of left and right parentheses in ERIC descriptors and underscore characters in CiteULike tags. ERIC descriptors that contained left and right parentheses were ignored in the comparison process; furthermore, left and right parentheses were not counted in determining the length of the descriptor. While the underscore character was counted in determining the length of the CiteULike tag, for matching purposes, this character was interpreted as a space " " character (for example, "electronic_resources" would be equivalent to "electronic resources"). No exceptions were utilized when hyphen characters were encountered in ERIC descriptors or CiteULike tags. The hyphen was processed like any other character.

Results

Data for this research was collected on CiteULike from July 1, 2007 to July 4, 2007 and from the ERIC website from July 13, 2007 to July 18, 2007. A sample of 2,786 journal articles indexed in CiteULike and ERIC formed the basis for this research. The total metadata associated with this sample included 2,899 unique ERIC descriptors, 3,176 unique CiteULike tags, and 1,083 unique CiteULike users. The number of CiteULike tags that literally matched ERIC descriptors on a matching per article basis was 240 (see Figure 1). This is about 7.6% of the total CiteULike tags (240 / 3,176).

Since the CiteULike tags and ERIC descriptors were compared literally (no semantic tag evaluation), factors such as grammar (roots of words), misspelling, and synonymy were not considered in determining the number of tag-descriptor matches; furthermore, the CiteULike tags and ERIC descriptors needed to be associated with the same journal article. This could explain why only 240 of the CiteULike tags matched ERIC descriptors on a matching per article basis.

Figure 1. CiteULike tags that matched ERIC descriptors ranked by usage



This usage count reflects the number of distinct journal articles with which a given tag was associated by one or more CiteULike users.

Conclusion

Folksonomies (tags) are useful supplements to controlled vocabularies since the former provide a means for personal organization outside the framework of the latter. The low number of tag-descriptor matches in this research indicates that CiteULike users do not use the same terminology as subject specialists who maintain descriptors in the ERIC thesaurus. The tag- descriptor matching criterion in this research was limited due to the complexity of language meaning. Further research involving semantic analysis of CiteULike tags may reveal an emerging vocabulary suitable for inclusion in the ERIC thesaurus as a controlled vocabulary.

References

- About ERIC. (n.d.). Retrieved August 5, 2008, from http://www.eric.ed.gov/ERICWebPortal/resources/html/about/about_eric.html
- About the ERIC Thesaurus. (n.d.). Retrieved August 5, 2008, from http://www.eric.ed.gov/ERICWebPortal/resources/html/thesaurus/about_thesaurus.html
- Available datasets. (n.d.). Retrieved August 5, 2008, from http://www.citeulike.org/faq/data.adp

- Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 1461-1464. doi:10.1073/pnas.0610487104
- CiteULike: Frequently asked questions. (n.d.). Retrieved August 5, 2008, from http://www.citeulike.org/faq/all.adp
- Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32, 198-208. doi:10.1177/0165551506062337
- Lambiotte, R., & Ausloos, M. (2005). Collaborative tagging as a tripartite network. Retrieved August 5, 2008, from http://arxiv.org/pdf/cs/0512090
- Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. *Proceedings of the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, 1-19. Retrieved August 5, 2008, from http://dlist.sir.arizona.edu/1790/01/lin.pdf
- National Information Standards Organization. (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. ANSI/NISO Z39.19-2005. Bethesda, MD: NISO.
- Vander Wal, T. (2007). Folksonomy coinage and definition. Retrieved August 5, 2008, from http://vanderwal.net/folksonomy.html

---

Bibliographic information of this paper for citing:

Bruce, Robert (2008). "Descriptor and folksonomy concurrence in education related scholarly research." *Webology*, **5**(3), Article 59. Available at: http://www.webology.org/2008/v5n3/a59.html

---
