Webology, Volume 4, Number 2, June, 2007

Editorial
Folksonomies: Why do we need controlled vocabulary?

[Alireza Noruzi]

Introduction

The Web consists of diverse information collections in terms of the type of content, context, format and quality. However, this diversity, as good as it is, often brings challenges for users in their web information seeking activities. The technologies such as *wiki*, *blog*, *RSS*, and *folksonomy* that build *Web 2.0* present an opportunity to share knowledge and facilitate interactions between users and computers. One of the main challenges of *Web 1.0* was that users were not engaged in information organization. Currently folksonomy-based systems (e.g., *Del.icio.us*) engage users in bookmarking and introducing their favorites.

What is a Folksonomy?

A *folksonomy* is a user-generated taxonomy used to categorize and retrieve web content such as web resources, online photographs and web links, using open-ended labels called *tags*. A folksonomy is most notably contrasted from a *taxonomy* in that the authors of the labeling system are often the main users (and sometimes originators) of the content to which the labels are applied. The labels are commonly known as tags and the labeling process is called *tagging* (Folksonomy, 2007). Tags help to improve search engine effectiveness because content is categorized using a familiar, accessible, and shared vocabulary.

Folksonomy, a free-form tagging, is a user-generated classification system of web contents that allows users to tag their favorite web resources with their chosen words or phrases selected from natural language. These tags (also called concepts, categories, facets or entities) can be used to classify web resources and to express users' preferences (Noruzi, 2006). Folksonomy is a classification of the users, by the users and for the users. The most popular, widely used folksonomy-based systems are:

1. Del.icio.us: www.del.icio.us
2. CiteULike: www.citeulike.org
3. Connotea: www.connotea.org
4. Flickr: www.flickr.com
5. Furl: www.furl.net
6. LibraryThing: www.librarything.com
7. Scuttle: www.scuttle.org
8. Shadows: www.shadows.com
9. Simpy: www.simpy.com
10. TagCloud: www.tagcloud.com
11. Tagzania: www.tagzania.com
12. Technorati: www.technorati.com

13. Unalog: www.unalog.com
14. Yahoo's MyWeb: http://myweb.yahoo.com
15. YouTube: www.youtube.com

Why does a folksonomy-based system need a thesaurus?

The purpose of this note is to answer the following question: *what can a thesaurus do for a folksonomy-based system*? Properly developed and used, a thesaurus can play several roles:

- It can be a separate tool to which both folksonomy users and searchers refer in deciding how to tag documents and queries for indexing and retrieval.
- It can function behind a search interface, facilitating searches without requiring users to interact with it as a separate operation.
- It can be used to improve the retrieval results from a search engine (Milstead, 2000).

A folksonomy-based system should use a thesaurus:

- to provide a means by which the use of terms in a given subject field may be standardized.
- to locate new concepts in a way which makes sense to users of the system.
- to provide classified hierarchies so that a search can be narrowed or broadened systematically, if the first choice of search terms produces either too few or too many results/hits.
- to provide a choice between singular and plural forms. Some words have two different connotations. Many concepts cannot be adequately represented by single words, and compounds are necessary.
- to correct typographical errors made by folksonomy users.
- to provide a guide for folksonomy users and searchers of the system for choosing the correct term for a subject search; this highlights the importance of cross-references. If a folksonomy user uses more than one synonym for the same resource -for example, "man," "men," "male," and "human" - then that resource is liable to be indexed haphazardly under all of these tags; a searcher who chooses one and finds resources tagged there will assume that s/he has found the correct term and will stop his/her search without knowing that there are other useful resources tagged under the other synonyms.
- to provide guides to terms which are related to any tag in other ways. Similar terms (related terms) should be linked together by three types of relationships: (i) hierarchical relationships, (ii) associative relationships, and (iii) equivalence relationships. For example, a search for the word "employees" will find records with the word "employees" but not records with words "employee," "worker," "laborer," "laborers," etc. The thesaurus is a way around this problem.

It should be noted that not everyone agrees on the need for a controlled vocabulary or thesaurus in folksonomy-based systems. However, there is no way to maintain consistency over time or across folksonomy users without a thesaurus. An additional folksonomy problem arises when dealing with foreign languages. It is obvious that different languages, such as Chinese, Persian, Arabic, German, and French use different words. Moreover, within any given language, different fields use differing vocabularies.

In folksonomy-based systems, there are at least two vocabularies: (i) the users' vocabularies; and (ii) the searchers' vocabularies. Thus, there is a possibility of mismatch in any transition between vocabularies, a dissonance in meaning. If the searcher asks for A and the user tags B, they might be expressing the same meaning in different ways (synonyms), or they might both write A and be meaning different things (homographs)

(Buckland, 1999). A folksonomy-based system should control the use of synonyms and homonyms through the use of *preferred terms*. It should be able to relate synonyms and suggest a preferred option, popular synonym.

Language is hardly a precise method of communication. Often, a concept can be expressed in a multitude of different ways (e.g. "retrieve," "search," or "query"), and unless there is some mechanism to recognize them as pointing to the same concept, interpretation of words to their intended meaning can be rather difficult. In order to reduce such ambiguity in interpretation, some IR systems employ the use of "controlled vocabulary", a set of agreed-on terms that represent specific concepts, thereby reducing the number of possible values for meaning. Controlled vocabulary can be used to generate a list of preferred terms as tags representing key concepts of documents (Yang, 2005).

The other question which needs to be answered is: *when does a folksonomy-based system need a thesaurus*?

- If it has unstructured information, and needs to control and provide access to that information.
- If it is using a search engine for its information, and has found that the engine does not provide adequate results: Â¬ perhaps too much irrelevant information for some queries, while missing useful information that is in the file for other queries.
- If the customers of the information system are demanding better access (Milstead, 2000).

Main problems of folksonomy tagging

Four main problems of folksonomy tagging are **plurals**, **polysemy**, **synonymy**, and **depth (specificity) of tagging**.

Plurals: Plurals and parts of speech and spelling can undermine a tagging system. For example, if tags Cat and Cats are distinct, then a query for one will not retrieve both, unless the intelligent search system has the capability to perform such replacements built into it.

Polysemy: Polysemy refers to a word that has two or more similar meanings. "Poly" means 'many', and "semy" means 'meanings'.

Synonymy: Synonymy, different words with similar or identical meanings, presents a greater problem for tagging systems because inconsistency among the terms used in tagging can make it very difficult for a searcher to be sure that all the relevant items have been found.

Depth (specificity) of tagging: Specificity means how specific should the user (classifier) be in translating a concept into tag(s)? Web resources can be tagged to varying levels of specificity, from very broad subjects taken only from the title and abstract to the paragraph level. The depth of tags refers to how many tags there are, relative to a web resource in the system.

Differences between Folksonomy and Library Classification Systems

In library classification systems (e.g., *Dewey Decimal Classification: DDC*, and *Library of Congress Classification: LCC*), each book is in one unambiguous category which is in turn within a yet more general one. For example, *Lions* and *Tigers* fall in the genus *Panthera*, and domestic *Cats* in the genus *Felis*, but *Panthera* and *Felis* are both part of family *Felidae*, of which *Lions*, *Tigers* and *Domestic Cats* are all part. Similarly, books on Asia's

geography are in the Dewey Decimal system category 915 and books on *General history of Asia Iran*'s in 955, but both are subsumed by the 900 category, covering all topics in geography and history.

A library classification system assumes that for any new book, its logical place already exists within the system, even before the book was published (Shirky, 2005). In contrast, folksonomy tagging is neither exclusive nor hierarchical and therefore can in some circumstances have both advantages and disadvantages compared with hierarchical taxonomies.

In library classifications, a book can be about several things at once. But the physical fact of the book has to be in one place, and if it is in one place, it cannot also be in another place. And this in turn means that a book has to be declared to be about some main thing. A book which is equally about two things breaks the 'be in one place' requirement, so each book needs to be declared to about one thing more than others, regardless of its actual contents. In folksonomy-based systems, there is no physical constraint that is forcing this kind of organization on users any longer. It is perfectly possible for any number of links to be in any number of places in a hierarchy, or in many hierarchies, or in no hierarchy at all (Shirky, 2005). Folksonomy, a user-generated classification, is in neither alphabetical nor hierarchical order. However, *Del.icio.us* gives each user an option to create hierarchy in their personal tags or to alphabetize their tags.

Folksonomy is a bottom-up approach where users themselves join the classification, compared to top-down taxonomy and library classifications. By this nature, folksonomy classification can reflect users' actual interest in real time (Niwa et al., 2006). In contrast to hierarchical library classifications (e.g., DDC or LCC) and thesauri, there is usually no limit for choice of tags in folksonomy; so many similar tags are generated. For instance, user A puts tag "mathematics" to a mathematical document, and user B puts tag "math" to the same document. Other examples: Plane (airplane), exam (examination), lab (laboratory), vet (veterinary surgeon), photo (photograph), fridge (refrigerator), phone (telephone), fax (facsimile), ad (advertisement), etc. This happens very often in folksonomy tagging, and is called "tag redundancy in folksonomy" (Niwa et al., 2006). Thus, there is no "*authority control*" on folksonomy entries.

Folksonomy-based systems can employ optional *authority control* of subject keywords, place, personal or corporate names and resource titles by connecting the system to established *authority control* files or controlled vocabularies using new techniques. A folksonomy-based system needs a controlled vocabulary and a suggestion-based system. Guy and Tonkin (2006) propose various system specific strategies for improving the quality of tags (e.g., spelling error checking, suggestion of synonyms, etc.) and encouraging users to observe certain collaborative tagging conventions (Macgregor & McCulloch, 2006).

Conclusion

Since the objectives of folksonomy are: (a) to make the process of saving and sharing links (URLs) as simple as possible; and (b) to let users easily share links (web resources) with all users throughout the world, a controlled vocabulary for a folksonomy-based system is essential to ensure tagging consistency across the database and between taggers. This may be a thesaurus or subject headings. By controlling the vocabulary using a thesaurus, tags are standardized and related resources are collocated for ease of discovery by the end-user (Lancaster, 1979). However, as Mika (2005) argues, it is not always possible for folksonomy users to express a complex concept with a single keyword and thus they may use more than one tag to express the concepts associated with a given item. The more tags available for a web resource, the better a folksonomy-based system can function.

In the future, it should be possible for search engine designers to design folksonomy-based engines with controlled vocabularies in different fields to improve web information retrieval.

Articles in This Issue

For this issue, we have four papers. Two papers deal with concepts related to "Web 2.0" and the last two papers deal with "use of web" and "ontology servers."

William F. Birdsall: *Web 2.0 as a social movement*. The study discusses Web 2.0 development and communications rights. Web 2.0 technologies try to find new ways to facilitate communication, collaboration, interaction and knowledge sharing between Internet users. In this study, the social movement for a right to communicate and the discourse surrounding Web 2.0 development are compared to demonstrate how Web 2.0 is a manifestation of an ongoing interaction between this human rights social movement and communication technology. This study concludes with a discussion on how a right to communicate serves as a conceptual framework for addressing a range of public policy issues arising out of the increasing use of the Web and as a framework for Web research and development. It is concluded that Web 2.0 development can be seen as part of a larger human rights movement.

Louise F. Spiteri: *Structure and form of folksonomy tags: The road to the public library catalogue*. This study examines how the tags that constitute folksonomies are structured. It is argued that folksonomies have the potential to add much value to public library catalogues by enabling clients to: store, maintain, and organize items of interest in the catalogue using their own tags. In this study, tags were acquired over a thirty-day period from the daily tag logs of three folksonomy sites: Del.icio.us, Furl, and Technorati. The tags were evaluated against section 6 (choice and form of terms) of the *National Information Standards Organization* (NISO) guidelines for the construction of controlled vocabularies. It is revealed that the folksonomy tags correspond closely to the NISO guidelines that pertain to the types of concepts expressed by the tags, the predominance of single tags, the predominance of nouns, and the use of recognized spelling.

Wendy Aitken: *Use of Web in Tertiary Research and Education*. This study discusses the role of ICT among members of the world's oldest living culture, the Australian Aborigines, now officially called Indigenous Australians. Initially, students of Aboriginal Studies were discouraged from using the Web, on the grounds that they lacked experience in judging the reliability of academic sources. In recent times, this has changed dramatically in educational practice. In addition, Aboriginal communities are now fully embracing the Web with production of their own web sites for a variety of reasons, generally to share their unique knowledge and culture with the world (it is surprising how often you can hear the strangely earthy and haunting sound of the didgeridoo in the score of European orchestral music). In addition, they are using the Web to tell their story, including that of how the *Pitjantjatjara* people were moved from their country to make way for British nuclear tests in the 1940s. Other discussions concerned the appropriateness of the Noble Savage myth, and comparisons with Canada's indigenous/aboriginal people.

Mohammad Nazir Ahmad & Robert M. Colomb: *Overview of ontology servers research*. This study defines ontology as a complex information object, containing millions of concepts in complex relationships. If we want to manage complex information objects (i.e., ontologies), we need an information system called an "*ontology server*", "ontology repository" or "server repository." The ontology server is a tool that supports editing, browsing and creation of ontologies. The ontology server should be capable of mapping objects/ontologies/terms, relationships between terms and attributes associated with the terms. This study reviews and compares the main ontology servers that have been reported

in the literatures. This review reports the current technologies of ontology and the semantic web, particularly on ontology server development.

References

- Buckland, M. (1999). Vocabulary as a central concept in library and information science. In *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science* (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999). Ed. By T. Arpanac et al. Zagreb: Lokve, pp. 3-12.
- Folksonomy (2007, May 6). In *Wikipedia, The Free Encyclopedia*. Retrieved November 25, 2006, from http://en.wikipedia.org/wiki/Folksonomy
- Guy, M., & Tonkin, E. (2006), Folksonomies: tidying up tags? *D-Lib Magazine*, Vol. 12(1). Retrieved November 25, 2006, from http://www.dlib.org/dlib/january06/guy/01guy.html
- Lancaster, F.W. (1979). *Information retrieval systems: characteristics, testing and evaluation*. 2nd ed., John Wiley and Sons, Chichester.
- Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5): 291-300.
- Mika, P. (2005). Ontologies are us: a unified model of social networks and semantics. In *Proceedings of 4th International Semantic Web Conference* (ISWC2005), 2005, pp. 522-536.
- Milstead, J.L. (2000). About thesauri. Retrieved November 25, 2006, from http://www.bayside-indexing.com/Milstead/about.htm
- Niwa, S., Doi, T., & Honiden, S. (2006). Web page recommender system based on folksonomy mining. *Information Processing Society of Japan (IPSJ) Journal*, 47(5): 1382-1392.
- Noruzi, A. (2006). Folksonomies: (Un) Controlled Vocabulary. *Knowledge Organization*, 33(4), 199-203.
- Shirky, C. (2005). Ontology is overrated: Categories, links and tags. Retrieved November 25, 2006, from http://www.shirky.com/writings/ontology_overrated.html
- Yang, K. (2005). Information retrieval on the Web. *Annual Review of Information Science and Technology* (ARIST), 39 (1), 33-80.

Bibliographic information of this paper for citing:

Noruzi, A. (2007). "Editorial: Folksonomies: Why do we need controlled vocabulary?" *Webology*, **4**(2), editorial 12. Available at: http://www.webology.org/2007/v4n2/editorial12.html

---