Enhancing Access to the Web: Vocabulary Analysis on Users' Tags and Professionals' Index Terms

Yunseon Choi
Graduate School of Library and Information Science
University of Illinois at Urbana Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211, USA
ychoi10@illinois.edu

ABSTRACT

This ongoing research aims to answer whether user-generated tags through social tagging could be used to enhance access to web resources and provide additional access points beyond professionally-generated ones. This study conducts qualitative vocabulary analysis of both users' tags and professionals' index terms.

Categories and Subject Descriptors
H.3. **[Information Storage and Retrieval]:** Content Analysis and Indexing – *indexing methods, linguistic processing, thesauruses*

General Terms
Performance, Human Factors, Standardization, Languages, Verification

Keywords
Controlled vocabulary, Digital Libraries, Folksonomy, Organization, Subject gateways, Subject indexing, Social tagging, Tags, Vocabulary analysis, Web

1. INTRODUCTION

A growing number of web resources have required new tools for organizing and providing access to the web. Subject gateways are such tools, designed to provide access to quality resources selected and indexed by specialists. However, a problem with these approaches is that most of them use traditional library schemes based on controlled vocabulary for subject access. Controlled vocabularies impede continuous development due to the rapid growth of digital libraries, so traditional indexing methods face the challenge in dealing with web resources. Furthermore, current systems are organized and indexed by professional indexers. Despite efforts to involve users in developing organization systems, these systems are not necessarily based on users' real languages.

iConference 2010, February 3–6, 2010, University of Illinois at Urbana-Champaign, Illinois, U.S.A.

Social tagging has received significant attention since it helps organize contents by user-generated tags. Social tagging allows users to add their tags to reflect their interests. Several researchers have discussed social tagging behavior and its usefulness for classification or retrieval. Nevertheless, further research is needed to qualitatively investigate social tagging and to verify its efficacy and benefit.

This paper is part of an ongoing research study which aims to answer whether and how social tagging could enhance access to web resources. In this paper, we provide the preliminary analysis of the following points: (1) whether tags have attributes beyond describing subjects of a document, (2) whether professional indexers have various or alternative interpretations of the same web document, and (3) whether tags would provide additional access points beyond index terms or keywords.

2. BACKGROUND

2.1 Organization of the Web

2.1.1 Subject gateways as organizing tools for the web

Subject gateways can range from "loosely collated commercial directories" such as Yahoo! subject categories, to "collections of quality assessed web resources compiled by the academic or research community" [1]. This study will refer to the concept of the latter for further discussion. Examples of such subject gateways include BUBL [2] and Intute [3]. BUBL describes itself as 'Free User-Friendly Access to selected internet resources covering all subject areas, with a special focus on Library and Information Science' [4]. It offers broad categorization of subjects based on the Dewey Decimal Classification (DDC) scheme [2]. Intute is a free web service aimed at students, teachers, and researchers in UK further education and higher education [3]. It is reported that Intute mainly uses the Universal Decimal Classification (UDC) and DDC for classification and has adapted them for in-house use. Intute also uses several thesauri for its subject relevance and comprehensiveness [5].

2.1.2 Challenges of controlled vocabulary for the web

For effective indexing, the indexing process needs to be controlled by using a so-called controlled vocabulary [6]. Yet, as there are more and more resources available on the web, existing controlled vocabularies have been challenged in their ability to index the range of digital web resources, e.g., slowness of revision, expensive indexing, and terms limited to topics found in physical and traditional library collections.

2.2 Social Tagging

Social tagging is described as "user-generated keywords" [7]. Since tags indicate users' perspectives in indexing resources, they have been suggested as a means to improve search and retrieval of resources on the web. Social tagging is a promising way to compensate for the disadvantages of traditional professional indexing because it is low-cost with a great number of users from everywhere contributing to the creation of tags. Thus, users' tags might be alternative terms for additional entry points of retrieval which are not easily attained using controlled vocabularies [8][9][10].

3. DATA COLLECTION

In order to examine professional indexers' vocabulary and compare it with that of users', we investigate two major subject gateways: BUBL and Intute (see Table 1). Both cover various subjects, and this feature allows one-to-one comparison on each subject area. We also extract tags from a social bookmarking site, Delicious.com. Unlike other social bookmarking sites, which provide the number of votes or users' comments, Delicious.com provides tagging data since it allows users to add, organize and share tags. Additionally, Delicious.com consists of a broad range of web resources, not limited to scholarly documents (e.g., journal articles on CiteUlike.org) or specific types of resources (e.g., photos and videos on flickr.com).

Table 1. BUBL vs. Intute

Site characteristics	**BUBL**	**Intute**
Classification	DDC	UDC and DDC
Keywords	N/A	*Controlled*: Several thesauri for their subject relevance and comprehensiveness, e.g., SCIE for Social Welfare, the Hasset, IBSS, LIR for Law, and the NLM MeSH headings for Medicine *Uncontrolled*: terms from web sites' titles and descriptions the indexers provide
Subjects covered	Various subjects	Various subjects
Database	Searchable and browsable	Searchable and browsable

Sampling documents is based on 10 subject categories BUBL provides as top-level categories (see Table 2). Under each category, documents in alphabetical order will be searched in turn at the other two sites, Intute and Delicious.com. Tags that are assigned to the document at Delicious.com are extracted only if a web document is found at all three locations (BUBL, Intute, and Delicious.com). Furthermore, indexers' index terms of both BUBL and Intute are collected for the comparison with users' tags.

Table 2. BUBL subject categories

Top Categories	**Subjects covered**
000 Generalities	Computing, Internet, Libraries, Information Science
100 Philosophy and psychology	Ethics, Paranormal phenomena
200 Religion	Bibles, Religions of the world
300 Social sciences	Sociology, Politics, Economics, Law, Education
400 Language	Linguistics, Language learning, Specific languages
500 Science and mathematics	Physics, Chemistry, Earth Sciences, Biology, Zoology
600 Technology	Medicine, Engineering, Agriculture, Management
700 The arts	Art, Planning, Architecture, Music, Sport
800 Literature and rhetoric	Literature of Specific languages
900 Geography and history	Travel, Genealogy, Archaeology

4. PRELIMINARY DATA ANALYSIS

One example of the analysis to be undertaken for each web resource in the sample is provided in this section. The poster will present findings from several more cases. Vocabulary analysis is conducted on the following main points: (1) analysis on Delicious.com tags, (2) analysis on BUBL and Intute vocabularies, and (3) analysis on Delicious.com tags and Intute keywords

(1) Analysis on Delicious.com tags

The process of identifying bibliographic attributes of tags is based on the Functional Requirements for Bibliographic Records (FRBR) model. Since the attributes defined in the FRBR model were derived from "a logical analysis of the data that are typically reflected in bibliographic records" [11], it would support a more systematic and meticulous analysis on the attributes of tags. A preliminary analysis of pilot data has identified that tags have several types of attributes beyond describing subjects of documents. The identified tag attributes can be categorized by the attributes defined by FRBR as shown in Table 3.

Table 3. Identified tags and related FRBR attributes

Identified tags	**FRBR attributes**
References or resources, research paper (tagged as "researchpapers"), article, tutorial, magazine, books or e-books, journal etc.	Form of work
Kids, children, senior, older, K-12 etc.	Intended audience (Work)
Audio, images, text etc.	Form of expression

(2) Analysis on BUBL and Intute vocabularies

In order to examine different points of view on the same document between professional indexers, indexers' index terms from BUBL and Intute are analyzed. BUBL offers each document with the classification number based on DDC. For indexer's index terms from BUBL, this study analyzes index strings, which are category paths of classification. For example, regarding a document, *Amazon.com*, the following category paths can be recognized, and they will be collected for analysis:

- News media, journalism, publishing > Publishers and publishing > Booksellers and bookshops

The collection of an indexer's index terms from Intute is the same as BUBL. For a more accurate comparison based on an equal condition, only index strings of category paths in classification schemes are analyzed:

- Communication and Media Studies > New Media > Interactive Games and Gaming
- Music > Music Industry, Recording and Publishing
- Communication and Media Studies > Publishing > Bookselling

(3) Analysis on Delicious.com tags and Intute keywords

In order to inspect whether Delicious.com users' tags would provide additional access points beyond index terms or keywords that Intute professional indexers provide, the top ranked tags assigned to a document at Delicious.com are collected and normalized. This is done through the rules for vocabulary analysis such as checking spelling and word forms. The top 10 tags are compared with keywords (controlled or uncontrolled) from Intute. Intute's uncontrolled keywords are added if its indexers can find no suitable word in thesauri. The keywords provided by Intute are useful and are the most appropriate data in order to compare the professional indexer's point of view with the user's point of view in subject indexing on the same document.

Table 4. Intute Keywords vs. Delicious Top 10 tags

Keywords at Intute		**Tags at Delicious.com**
Keywords - controlled	Amazon.com (Firm); books; publishing; publishers; bookselling; booksellers; electronic publishing; bookstores; motion pictures (visual works); videotapes; video games; digital versatile discs; music; software	shopping, books, amazon, online, bookstore, music, web, internet, fun, deals
Keywords - uncontrolled	online; electronic commerce; on-line; book stores; bookshops; e-publishing; films; movies; motion pictures; video tapes; digital video discs; DVDs; compact discs; CDs	

5. DISCUSSION

Table 3 illustrates that tags provide essential bibliographic attributes, which have not been identified in previous research. This provides a helpful understanding of features and patterns of tags in describing web documents.

Moreover, the preliminary analysis has revealed that there were some various or alternative interpretations in new subject areas, for instance, internet-related areas. There were different perspectives on the same document, *Amazon.com*, even between groups of professional indexers. BUBL places it at the category of *070.5 Publishers and Publishing* under the category of *070 News media, Journalism, Publishing*. Intute classifies it as the similar subjects with BUBL, e.g., *New media* or *Publishing*. However, Intute also categorizes it at the category of *Music industry, recording and publishing* under the category of *Creative and performing arts.*

Table 4 indicates that among the top 10 tags at Delicious.com, a term "shopping" which is ranked first is not included in the Intute keywords. However, it is worthwhile to note that the tag "shopping" might be an additional helpful access point for those who are interested in purchasing books or other related goods online.

6. CONCLUDING REMARKS

As part of an ongoing research study, this paper focuses on bridging the gap of insufficiency of studies on vocabulary analysis by comparing user-generated tags with professional-generated index terms regarding web resources. Current work will be complemented by quantitative measures performed on a large data set. The research also will evaluate indexing consistency of tagging and professional indexing in order to systematically verify the efficacy and quality of tags. This will provide a way of improving the organization of web resources by increasing the utilization of social tagging data.

7. ACKNOWLEDGEMENTS

This research has been conducted under the direction of Professor Linda C. Smith at the University of Illinois at Urbana-Champaign. I wish to express my deepest respect and gratitude to her.

8. REFERENCES

[1] University of Kent. *Library Services Subject Guides*. Retrieved from http://www.kent.ac.uk/library/subjects/healthinfo/subjgate.html. 2009

[2] BUBL Link Home. http://www.bubl.ac.uk

[3] "Intute." Wikipedia, The Free Encyclopedia. 2009. Wikimedia Foundation, Inc. 10 Sep. 2009.

[4] "BUBL." Wikipedia, The Free Encyclopedia. 2009. Wikimedia Foundation, Inc. 10 Sep. 2009.

[5] *Personal Communication via email with Intute.* May 21and June 2, 2009.

[6] Lancaster, F. W. 1972. *Vocabulary control for information retrieval*. Washington, D.C.: Information Resources Press.

[7] Trant, J. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10(1).

[8] Maltby, A. 1975. *Sayers' Manual of Classification for Librarians* (5th Ed.), Andre Deutsch, London.

[9] Hayman, S. 2007. Folksonomies and tagging: New developments in social bookmarking. *Ark Group Conference:*

Developing and Improving Classification Schemes 27-29 June, Rydges World Square, Sydney
[10] Quintarelli, E. 2005. “Folksonomies: power to the people”, *Proceedings of the 1st International Society for Knowledge Organization* . UniMIB Meeting, June 24, Milan, Italy, ISKOI, Italy.
[11] IFLA Study Group. 1998. *Functional requirements for bibliographic records: final report.* München: K.G. Saur.