Margaret E. I. Kipp
University of Wisconsin-Milwaukee, Milwaukee, WI

Comparing controlled vocabularies and tags: research methodologies and research goals = Une comparaison des vedette-matières et des étiquettes : les méthodologies et les buts de recherches

Abstract: Tags have been compared to controlled vocabulary terms and have been suggested as replacements or enhancements in indexing. This paper explores tagging and controlled vocabulary studies in the context of studies examining title and author keywords or user search terms and uses the results to analyse 236000 PubMed records tagged in CiteULike.
Résumé : Les étiquettes ont été comparées avec les vedette-matières et ont été suggérés comme remplacement ou comme addition à l'indexation conventionnelle. Cet article examine la recherche sur l'étiquetage et les vedette-matières en comparaison avec des études examinant les mot-clés de titre et d'auteur ou les mot-clés des requêtes d'usager et utilisera ses résultats pour analyser 236000 notices catalographiques de PubMed qui ont été étiquetées sur CiteULike.

1. Introduction

Social tagging is still a new phenomenon, but it has become extremely popular spreading beyond social bookmarking sites like delicious.com where it originated to sites such as Amazon.com and many next-generation library catalogues. Proponents suggest social tagging will offer subject indexing in areas where indexing was prohibitively expensive due to collection size or completely lacking such as on the web. Mathes (2004) noted the similarities between tagging and traditional indexing and suggested a call for action in studying terms used in indexing by professional indexers, authors and users (Mathes 2004). This paper will examine the early history of indexing term comparisons, make comparisons to later work in social tagging and then report on the preliminary results of a study examining title, author and MeSH keywords and tags from a set of PubMed articles bookmarked on CiteULike.

2. Background

Title Keywords

One of the earliest studies of title keywords for indexing was by Montgomery and Swanson (1962) who discovered that there was a high degree of concurrence between title keywords for entries in Index Medicus and assigned subject headings (86%), but found that 14% of articles were unindexable based solely on the title (Montgomery and Swanson 1962). O'Connor (1964) found that many indexes had much lower rates of match between title keywords and subject headings. Frost (1989) revisited these studies in the context of the introduction of machine-readable LCSH into catalogues and found that 73% of title keywords matched exactly or partially to subject headings, though this

varied substantially by field (Frost 1989). Voorbij (1998) found a similar degree of match using monographs from the humanities using a more extensive set of thesaural categories -- exact match, related term match, narrower term match, etc (Voorbij 1998).

Author Keywords

Schultz, Schultz and Orr (1965) compared author keywords to document titles and to indexing terms assigned by subject matter experts and found that author keywords matched subject terms more closely than title terms (Schultz, Schultz and Orr 1965). More recently, Kipp (2005; 2007) examined author keywords in comparison to tags and subject headings using a modification of Voorbij's (1998) categories and found a high degree of overlap between tags, author keywords and subject headings when partial or related term matches were considered (Kipp 2005; Kipp 2007). Gil-Leiva and Alonso-Arroyo (2007) examined author keywords from scientific articles and found a 46% overlap with subject headings when author keywords were normalised (Gil-Leiva and Alonso-Arroyo 2007). Heckner et al (2008) studied tags and author keywords and found an approximately 58% overlap in content. They also reported that taggers tended to use more general concepts than authors (Heckner et al 2008). Strader (2009) compared author keywords to LCSH terms assigned to electronic theses and found 65% of author terms matched exactly, partially or were variant forms of the headings (Strader 2009).

User Search and Query Terms for Indexing

Carlyle (1989) compared user vocabulary directly to LCSH and found a 47% exact match between user vocabulary and LCSH and up to a 70% match when using stemming and other matching algorithms to correct for plurals and punctuation (Carlyle 1989). Gross and Taylor (2005) examined user search terms from transaction logs and found that approximately one third of keyword searches conducted would have failed without controlled vocabulary terms. Garrett (2007) studied the use of subject headings to enhance eighteenth century documents and found that as many as 60% of searches would fail without the addition of keywords due to terminological drift over time (Garrett 2007).

In many of these studies, the authors concluded that title, author or user generated keywords added additional potential subject access points to a record and that some non-trivial number of searches would fail without them.

3. Methodology

The first part of the study compared the methodologies, data sets and results of a set of social tagging studies which compared tags to controlled vocabularies for the purpose of identifying whether: a) tagging could be used to enhance records already indexed by controlled vocabularies b) tagging could be used to enhance records not yet indexed or c) whether tags were not sufficiently useful as index terms to be worth adding to records.

The second part of the study examined 236 000 bibliographic records collected from PubMed articles bookmarked on CiteULike. A script, linkouts.py, was used to automatically collect XML formatted Medline records using Entrez queries for each PubMed ID on CiteULike. These records were then enhanced with the CiteULike data associated with that PubMed ID specifically: the tags, CiteULike ID and number of users

who posted the article. The index terms, or potential index terms, in these records were then analysed using methods outlined in previous studies.

4. Results

Tagging and Controlled Vocabularies

A number social tagging studies have explored comparisons of tags and controlled vocabularies in the context of prior research into the use of title keywords, author keywords and end-user search terms for indexing items in an OPAC or journal database. A representative selection of these studies is presented below.

Bruce (2008) analysed tags assigned to articles indexed in ERIC and found a very small number of exact matches to ERIC terms, but did not analyse partial matches. Trant (2009) studied tags assigned to museum artifacts through the steve.museum tagger. Preliminary analysis showed that 70% of tags did not match terms in the museum documentation leading Trant to suggest that these terms should be compared to terms used in searches, especially failed searches (Trant 2009).

Kipp (2005; 2011) adapted Voorbij's (1989) thesaural analysis method to compare tags, author keywords and descriptors assigned to LIS articles tagged on CiteULike. Tags were more likely to match author keywords exactly (33%) than descriptors (16%), but author keywords were as likely to match exactly or be related terms of descriptors (19%) (Kipp 2011). Thomas et al (2009) adapted this thesaural comparison for books tagged in LibraryThing with associated Library of Congress Subject Headings finding that 6% of tags matched LCSH exactly while 31% matched thesaural categories and 35% were related to LCSH but not subject headings (Thomas et al 2009).

Good et al (2009) examined tags from CiteULike and Connotea associated with PubMed citations. They compared the tags to MeSH using normalised strings (9-10% match), concepts (20-30% match) and semantic groups (80% match) (Good et al 2009).

While exact matches were less common in all studies, partial matches were much more common and many authors suggested that partial matches or matches to failed search terms should be examined in order to discover the potential of non-matching but relevant tags to improve searching and browsing of collections.

Early Results from the PubMed Study

A random sample of articles was selected from the set of all CiteULike articles with PubMed IDs. Since these articles are tagged on CiteULike, they may have a number of possible associated index terms for study including: tags, MeSH descriptors, title keywords and author keywords. Two examples (which include author keywords) are shown here to illustrate the differences between these terms.

1: Frank PL Lai, et al. (2008). Arp2/3 complex interactions and actin network turnover in lamellipodia. EMBO J. 2008 April 9; 27(7): 982–992. PMCID: PMC2265112 (11 authors)

Tags: **arp**

Author Keywords : **Arp2/3 complex**; cofilin; FRAP; lamellipodium; migration

Title Keywords : **Arp2/3 complex;** interactions; actin network; turnover; lamellipodia

MeSH Descriptors : Actin Capping Proteins/metabolism; Actin Depolymerizing Factors/metabolism; **Actin-Related Protein 2-3 Complex/metabolism**; Actins/metabolism; Adaptor Proteins; Signal Transducing/metabolism; Animals; Cell Line, Tumor; Cortactin/metabolism; Fluorescence Recovery After Photobleaching; Mice; Models, Biological; Protein Binding; Pseudopodia/metabolism; Rabbits; Wiskott-Aldrich Syndrome Protein Family/metabolism

Figure 1: Keywords for Article 1

Article one (Figure 1) was tagged by only one person on CiteULike. In this case, the tag is an acronym for Actin-Related Protein, a term which appears in the Author and Title Keywords and the MeSH headings.

2: Hongbo Xie, et al. (2007). Functional Anthology of Intrinsic Disorder. III. Ligands, Postranslational Modifications and Diseases Associated with Intrinsically Disordered Proteins. J Proteome Res. 2007 May; 6(5): 1917–1932. PMCID: PMC2588348
(7 authors)

Tags: 2007; **disorder**

Author Keywords: Intrinsic **disorder**, protein structure, protein function, intrinsically disordered proteins, bioinformatics, **disorder** prediction

Title Keywords: Intrinsic **Disorder**; Ligands; Postranslational Modifications; Diseases; Intrinsically **Disordered** Proteins

MeSH Descriptors: Animals; Computational Biology; Databases, Protein; Humans; Ligands; Protein Conformation; Protein Processing, Post-Translational; Proteins/chemistry; Proteins/genetics; Proteins/metabolism; Sequence Analysis, Protein

Figure 2: Keywords for Article 2

Article two was also tagged by one person. In this case we see an example of a tag which does not match and a tag which matches to parts of other keywords, but is ambiguous

enough that we cannot be sure if the tagger intended this tag to mean intrinsic disorder or disordered protein structures or something entirely different.

As noted by Heckner et al (2008) users do use much more general terms in some cases, although in other cases their terms do match subject headings or are variant forms of the headings (Kipp 2005; Strader 2009). Preliminary results of this study show that tags, author keywords, title keywords and descriptors (MeSH) all provide slightly different subject access terms, with some element of overlap and some elements of extension of the keywords available for search and browsing.

5. Conclusions

While many studies have compared social tagging terms to controlled vocabularies, this paper is the first to begin to compare these studies and analyse their methodologies and results. The majority of the tagging and controlled vocabulary studies have examined tagging from the point of view of creating end-user terms which could be used to enhance search in the catalogue or in article databases, a similar goal to that of end-user thesaurus research (Shiri and Revie 2002). Research suggests that tagging does not replace controlled vocabularies, but instead provides an added dimension to subject access. Early research into using tagging to enhance information retrieval supports the idea that tags can be used support controlled vocabularies by providing early access to emerging terminologies (Peters 2009; Lu and Kipp 2010). Overall, tagging has proven to be a useful addition to research into the effectiveness of subject indexing and provides us with strong support for the importance of subject access in addition to full text search.

6. References

Bruce, R. 2008. Descriptor and Folksonomy Concurrence in Education Related Scholarly Research. *Webology* 5(3). http://www.webology.ir/2008/v5n3/a59.html

Carlyle, A. 1989. Matching LCSH and User Vocabulary in the Library Catalog. *Cataloging & Classification Quarterly* 10(1):37–63.

Frost, C. O. 1989. Title Words as Entry Vocabulary to LCSH – Correlation Between Assigned LCSH Terms and Derived Terms from Titles in Bibliographic Records with Implications for Subject Access in Online Catalogs. *Cataloging & Classification Quarterly* 10(1):165–179.

Garrett, J. 2007. Subject Headings in Full-Text Environments: The ECCO Experiment. College & Research Libraries 68(1): 69-81.

Gil-Leiva, I. and Alonso-Arroyo, A. 2007. Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology* 58(8):1175–1187.

Good, B. M., Tennis, J. T., and Wilkinson, M.D. 2009. Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*. 2009 Sep 25;10:313. [PMC]

Gross, T. and Taylor, A. G. 2005. What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. *College & Research Libraries* 66(3): 212–30.

hlbacher, S., and Wolff, C. 2008. Tagging tagging. Analysing user keywords in scientific bibliography management systems. *Journal of Digital Information* 8(2). http://journals.tdl.org/jodi/article/view/246

Kipp, M. E. I. 2005. Complementary or Discrete Contexts in Online Indexing : A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science* 29(4):419–436.

Kipp, M. E. I. 2007. Tagging for health information organisation and retrieval. *North American Symposium on Knowledge Organization (NASKO), Toronto, June 14-15, 2007*, pages 63–74. http://eprints.rclis.org/handle/10760/10384

Kipp, M. E. I. 2011. User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike. *Canadian Journal of Library and Information Science* 35(1):17-48. (in press)

Lu, K. and Kipp, M. E. I. 2010. Can Collaborative Tagging Improve Retrieval Effectiveness? -- An Experimental Study. Proceedings of the Annual Meeting of the American Society for Information Science and Technology, October 24-27, Pittsburgh, Pennsylvania, USA.

Mathes, A. 2004. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Montgomery, C. and Swanson, D. R. 1962. Machinelike indexing by people. *American Documentation* 13(4):359–366.

O'Connor, J. 1964. Correlation of indexing headings and title words in three medical indexing systems. *American Documentation* 15(2):96–104.

Peters, I. 2009. *Folksonomies. Indexing and Retrieval in Web 2.0* (Knowledge & Information: Studies in Information Science). De Gruyter.

Schultz, C. K., Schultz, W. L., and Orr, R. H. 1965. Comparative indexing: Terms supplied by biomedical authors and by document titles. *American Documentation* 16(4):299–312.

Strader, C. R. 2009. Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations. *Library resources & technical services* 53(4):243–250.

Thomas, M., Caudle, D. M., and Schmitz, C. M. 2009. To tag or not to tag? *Library Hi Tech*, 27(3):411-434.

Trant, J. 2009. Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing

Research. *Journal of Digital Information* 10(1). http://journals.tdl.org/jodi/article/viewArticle/270

Voorbij, H. J. 1998. Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. Journal of Documentation, 54(4):466–476.