12-1-2019

Controlled Vocabularies versus Social Tags: A Brief Literature Review

Arindam Sarkar
Jadavpur University, infoarindam83@gmail.com

Prof. Udayan Bhattacharya
Jadavpur University

Controlled Vocabularies versus Social Tags: A Brief Literature Review

Arindam Sarkar[1]and Prof. Udayan Bhattachrya[2]

[1]Research Scholar, Department of Library and Information Science, Jadavpur University
E-mail: infoarindam83@gmail.com
[2]Professor, Department of Library and Information Science, Jadavpur University

Abstract: *The present study is a literature review, based on the comparative study between controlled vocabularies and social tags from various perspectives. Critical comments of experts on similarities, co-relationship, uses, trends between controlled vocabularies and social tags have found their manifestation in this literature review. Through this study an effort has been made to portray the overall picture of previous research regarding this topic.*

Keywords: *Controlled vocabularies; Social tags; Literature review; Co-relationship*

Introduction

The controlled vocabulary is a set of preselected terms made by experts in controlled way for assigning in various applications. Generally they are used in subject indexing schemes, subject headings, thesauri, taxonomies and other forms of knowledge organization systems. In library and information science controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. Controlled vocabularies solve the problems of homonyms, synonyms etc.

Tagging, also known as social tagging or user tagging or collaborative tagging, has gained in mileage since the first social bookmarking system named del.icio.us was started in 2003. As an application of web 2.0 social tagging increases access points; more entry points, that are helpful for easy retrieve of resources i.e. documents. The annotated resources can be of any type or in any format, such as videos (e.g. YouTube), photos (e.g. Flickr), academic papers (e.g. CiteULike), books (e.g. LibraryThing) and so on. Social tagging can be viewed as a technique, by which many users add metadata in the form of keywords, genres, subjects to shared content or resources (Golder and Huberman, 2006).

VanderWal coined the term ‘folksonomy’ (a portmanteau of the words folk and taxonomy) to describe the conceptual structures generated by social tagging systems (Wal, 2007).

All social tagging systems or controlled vocabulary have the common purpose of helping users share, store, organize and retrieve the resources they are interested in.

Social tagging is different from controlled process, social tagging is done in a totally uncontrolled environment. Social taggers do not need to any expertise rules for tagging. They apply their own verbal descriptors to resources that interest them.

Now the question is what are the differences and connections between social tags and expert-created generated controlled terms especially in creation of book or other catalogue?

Development of the information retrieval concepts, knowledge organization and dissemination concepts in W3 environment is going under an osmosis process. This is the age of #tag, social tagging, folksonomy, social bookmarking etc. therefore it is a wonderful platform to test and compare between such type users centric concept with professional expert generated concept.

Literature review is necessary whenever a research work is to be done. It helps in understanding the present research problem and how to reach to the goals of the study. By over viewing the similar kinds of previous works, a researcher can avoid duplication of research work. Besides that it helps in various ways in different stages of the research. This study attempts to portray an overall picture of previous research on this topic.

Methodology

This study is simply based on a documentary analysis of previous research papers regarding comparisons between social tags and controlled vocabularies from different perspective.

Background works

Following few paragraphs describes some conceptions regarding tagging or social tagging and summary of some literature on comparison between controlled vocabularies and social tags.

“The move towards social software and what is generally known as Web 2.0” (O’Reilly 2005). It has generated interesting shared metadata and social tagging as an approach to resource description.

According to Rafferty and Hidderley (2007) “the development of folksonomies and social tagging moves resource description towards a more dialogic communicative practice where creators, readers, listeners and viewers of documents are encouraged to add their own tags”. There are a number of websites that use social tagging, and these include text based websites such as CiteULike, music based websites, such as lastfm.com, image based websites, such as Flickr, fan websites, for example Archive of Our Own and social websites such as Facebook and Twitter (Rafferty and Hidderley 2007).

Ma and Cahier (2013) described an interesting case-study on Twitter, a popular micro blogging platform. It uses the hashtag as the convention that allows users to describe and, increasingly, to comment on content. Twitter users can very easily create tweets and re tweet the initial tweet. Hashtags establish a bi-directional interaction between the user and the information resource, which on the one hand allows people to follow and acquire news, opinions and people's status updates, and on the other hand allows user participation in the creation of hashtags, facilitating the creation and propagation of content throughout the platform. Hashtags are user driven and serve as metadata to code and spread ideas and trends quickly and easily, however, it can be difficult to interpret hashtags and discover their relationships because of their free-form nature (Ma and Cahier, 2013).

Golder and Huberman (2006) defined "tagging as a process of labelling and categorizing information through which meaning emerges for individual users". Furner (2010) considered tagging a kind of indexing: "Tagging is the activity of assigning descriptive labels to useful (or potentially useful) resources".

Cattuto et al. (2007) described "social tagging generally means the practice whereby internet users generate keywords to describe, categorise or comment on digital, content. Tagging allows users to record their individual responses to the information objects. Tagging tools are generally formed of a triplet of user, information object and keyword. Tags, documents and users form a tri-partite graph, which means that tags are also connected" (Cattuto et al., 2007). In this environment, users as well as documents are connected.

In the early days, the emerging concepts and vocabulary relating to social tagging were still to be fixed, for example, in 2007, Zauder, Lazic and Zorica wrote that "collaborative tagging is also frequently called social tagging and distributed classification, used as a synonym for folksonomy and even confused with social bookmarking". They emphasized that the "term folksonomy should be used for the totality of tags produced by users through the collaborative tagging process, not used to refer to the process itself". They also explored that "Social bookmarking, while often using collaborative tagging is not synonymous with collaborative tagging. Collaborative tagging is the process by which users of a Web service add natural language keywords to information resources, creating a personalised collection which can be made available to all users" (Zauder, Lazic and Zorica, 2007).

Trant (2009), also distinguishes between tagging as a "process with a focus on user choice of terminology", while a folksonomy is the "resulting collective vocabulary (with a focus on knowledge organization)" and social tagging is the "socio technical context within which tagging takes place (with a focus on social computing and networks)" (Trant, 2009).

The strengths and weakness of tagging as a kind of indexing can partly be inferred from its characteristics relative to other forms of indexing. Furner (2010) wrote that "tagging can be characterized as a form of (1) manual, (2) ascriptive [assigned as opposed to derived], (3) natural language[as opposed to controlled vocabularies], (4) democratic indexing, which is typically undertaken by (5) resource creators and (6) resource users who have (7) low levels of indexing expertise, (8) high levels of domain knowledge, and (9) widely varying motivations, and which is commonly used to represent (10) non- or quasi-subject-related properties, and frequently (but far from exclusively) applied to (11) resources such as images that do not contain verbal text"(Furner, 2010).

Martinez-Avila (2015) summarized that Disadvantages or weaknesses in social tagging have long been recognized in the literature. Doctorow's (2001) arguments on the weak side of social tagging as follows, "people lie in a competitive world, common people are too lazy to do something they do not understand, people refuse to exercise care and diligence in their tag creation, people do not know themselves, schemata are not neutral, metrics influence results, and there is more than one way to describe something".

Kroski (2005) wrote that “amongst the weaknesses of social tagging are a lack of synonym and homonym control, a lack of precision and hierarchy, a basic level problem where broad and narrow terms are used interchangeably, and a susceptibility to unethical gaming” (Kroski 2005).

Feinberg (2006) draws attention to the limitations of social tagging in relation to the notion of social intelligence with reference to examples drawn from Surowiecki. She argues that “while social tagging systems might be democratic in allowing anyone to tag, there is no sense of a community coming together to determine how are source should be indexed”. She suggests that “if a political metaphor is to be used to characterise the attitude regarding authority in social tagging systems, then ‘social classification’, as Feinberg calls it, should be likened to libertarianism, "where everyone's whims are allowed to flourish"”.

In case of comparison between tags and controlled vocabularies Bogers and Petras (2017) pointed out the effectiveness of controlled vocabulary vs. tags in book search. Their study was based on over 2 million book records and over 330 real-world book search requests, with a highly controlled and in-depth analysis of topical metadata, comparing controlled vocabularies with social tags. At the end tags perform better overall in this setting, but controlled vocabulary terms provide complementary information, which will improve a search. In addition, they investigated the possible causes of search failure. form this study they concluded that neither tags nor controlled vocabularies are wholly suited to handling the complex information needs in book search, which means that different approaches to describe topical information in books are needed (Bogers and Petras, 2017).

As per Vaidyaa and Harinarayanab (2016) “social tags are user generated metadata and play vital role in Information Retrieval (IR) of web resources”. Their study was an attempt to determine the similarities between social tags extracted from LibraryThing and Library of Congress Subject Headings (LCSH) for the titles chosen for study by adopting Cosine similarity method. Their result shows that social tags and controlled vocabularies are not quite similar due to the free nature of social tags mostly assigned by users whereas controlled vocabularies are attributed by subject experts. “In the context of information retrieval and text mining, the cosine similarity is most commonly adopted method to evaluate the similarity of vectors as it provides an important measurement in terms of degree to know how similar two documents are likely to be in relation to their subject matter. The LibraryThing tags and LCSH are represented in vectors to measure Cosine similarity between them”. The study of cosine similarity technique is one of the most important issues in the context of information retrieval process. This research work prominently tries to highlight the relation between social tags and controlled vocabularies by representing them in vector space to determine the cosine value for them. The cosine score reveals similarity or dissimilarity between tags and vocabularies which is expressed in mathematical value and not by semantic meaning of the words chosen. Hence meaning of the word has no role in determining the cosine value for these set of terms. This study of social tags proves the fact that they could not replace the value of controlled vocabularies in the context of information retrieval (IR). The controlled indexing has greater IR value than social tags for efficient retrieval results. The future studies need to be carried out by increasing the number of social tags and descriptors from controlled vocabularies to test if there is any variance in the cosine similarity score for better understanding of the complementary and supplementary relation between user warrant and literary warrant (Vaidyaa and Harinarayanab, 2016).

Dash (2015) explores the representation of bias in social tags and Library of Congress Subject Headings, with a particular focus on the motivations of the layperson (the tagger) and the expert (the cataloguer). A mixed methodological approach was adopted for this study. A framework for measuring bias was defined and constructed and this was applied via a simple coding scheme to a total of 500 social tags from LibraryThing and 175 Library of Congress Subject Headings from the Library of Congress online catalogue. These were harvested from a sample of 50 popular feminist fiction titles. The analysis demonstrated that, "although there were a higher proportion of unbiased social tags than unbiased LCSH, issues of bias were found in both systems. The two systems displayed very distinct issues of bias, given the differing motivations of the tagger (personal) and the cataloguer (to allow subject access)". The research demonstrated the idea that "the concepts of bias and interpretation are inseparable; and (regardless of system and language), one cannot interpret anything without applying personal, cultural and leaned biases based on a particular worldview" (Dash, 2015).

Bogers and Petras (2015) explored a large-scale comparison of the contributions of individual metadata elements like core bibliographic data, controlled vocabulary terms, reviews, and tags to the retrieval performance. They used a test collection of over 2 million book records with metadata elements from Amazon, the British Library, the Library of Congress, and LibraryThing. They concluded that tags and controlled vocabulary terms do not actually outperform each other consistently, but seem to provide complementary contributions: some information needs are best addressed using controlled vocabulary terms whereas other are best addressed using tags (Bogers and Petras, 2015).

Most of the work comparing tags to CVs for book search has remained theoretical. Few exploratory studies have focused on the potential of these metadata elements for retrieval. The only notable exception is a large-scale empirical comparison by Koolen (2014), who found that UGC (User-generated content), in particular reviews, outperformed professionally assigned metadata. This paper delve deeper into this problem: which (combination of) metadata elements can best contribute to retrieval success, and how does the retrieval performance of tags and CVs compare under carefully controlled circumstances?

This study presented an empirical comparison in the book search domain using LibraryThing (LT), Amazon, the Library of Congress (LoC), and the British Library (BL) as data providers. This study was used a large-scale collection from the INEX Social Book Search Track, filtered to allow a fair comparison between tags and CVs. A substantial set of requests representing real information needs is used. The analysis focuses on the differences in using tags or CVs overall and distinguished by different book types or request types.

The most important conclusion was that Tags and CVs achieve similar retrieval effectiveness in book search. Those results were found after levelling the playing field for both as much as possible, by requiring both CV and Tag content to be present in every document. Still, significant differences exist in the distribution of CV terms and Tags. The average number of types is much larger for the CV than the Tags element set, whereas the average number of tokens is much larger for the Tags element set. This means that there were more unique terms in CV, but more repetition of them in Tags (Koolen, 2014).

According to Lee and Schleyer (2012) “the increasing popularity of social tagging and the limitations of controlled indexing (primarily cost and scalability), it is reasonable to investigate to what degree social tagging could substitute for controlled indexing”. In this study, they compared CiteULike tags to Medical Subject Headings (MeSH) terms for 231,388 citations indexed in MEDLINE. Their study was consisted of 1,087,524 social tags, 24,121 of them distinct, to 2,822,934 MeSH terms, 21,129 of them distinct, for a set of 231,388 biomedical papers using increasingly sophisticated text processing methods. Their result shows that “CiteULike tags and MeSH terms are quite distinct lexically, reflecting different viewpoints/processes between social tagging and controlled indexing” (Lee and Schleyer, 2012).

Kipp (2011) mentioned that “social tagging has become increasingly common and is now often found in library catalogues or at least on library websites and blogs. Tags have been compared to controlled vocabulary indexing terms and have been suggested as replacements or enhancements for traditional indexing”. This paper explored tagging and controlled vocabulary studies in the context of earlier studies examining title keywords, author keywords and user indexing and applied these results to a set of bibliographic records from PubMed which are also tagged on CiteULike. Preliminary results were found that “author and title keywords and tags are more similar to each other than to subject headings, though some user or author supplied terms do match subject headings exactly. Author keywords tend to be more specific than the other terms and could serve an additional distinguishing function when browsing” (Kipp, 2011).

Matthews, Jones, Puzon, Moon and Tudhope (2010) thought that “traditional subject indexing and classification are considered infeasible in many digital collections. Automated means and social tagging are often suggested as the two possible solutions”. Their study investigates ways of enhancing social tagging via knowledge organization systems, with a view to improving the quality of tags for increased information discovery and retrieval performance. They concluded that to improve the information retrieval in the digital environment social tags are really helpful (Matthews and others, 2010).

According to Rorissa (2008) “web 2.0 and social or collaborative tagging have altered the traditional roles of indexer and user”. In today’s web environment, end users create, organize, index, and search for images and other information sources through social tagging and other collaborative activities. “Social tagging of images such as those on Flickr, an online photo management and sharing application, presents an opportunity that can be seized by designers of indexing tools and systems to bridge the semantic gap between indexer terms and user vocabularies”. This study pointed out the differences and similarities between user-generated tags and index terms. For this study a random sample of Flickr images and the tags assigned by users were analyzed and compared with another sample of index terms from a general image collection using established frameworks for image attributes and contents. The result was found that “there is a fundamental difference between the types of tags and types of index terms used”. This work provides an insight into the differing natures of Flickr tags and traditional index terms assigned to images in a general-purpose collection. Findings of the study suggested that user generated tags and professionally assigned index terms have different underlying structures (Rorissa, 2008).

Lu, Park and Hu, (2010) thought that social tagging, as a recent approach for creating metadata, has caught the attention of library and information science researchers. Many researchers recommend incorporating social tagging into the library environment and combining folksonomies with formal classification. However, some researchers are concerned with the quality issues of social annotation because of its uncontrolled nature.
In this study, they compare social tags created by users from the LibraryThing website with the subject terms assigned by experts according to the Library of Congress Subject Headings (LCSH). The purpose of this study was to examine the difference and connections between social tags and expert-assigned subject terms and further explore the feasibility and obstacles of implementing social tagging in library systems. In this paper, they investigated the differences and connections between social tags and expert-created subject terms, which are an integral part of library metadata. To conduct the study, they collected a sample of MARC bibliographic records provided by the Library of Congress. Using the ISBNs in the records, they also collected a sample of social tags created by users on the LibraryThing website. Although the bibliographic metadata record contains a range of elements representing various attributes of a book entity, in their study they focus on the comparison between the value of subject access fields and LibraryThing tags. They found that experts and users agree on some terms for describing the resources. The results of their study show that it is possible to use social tags to improve the accessibility of library collections. However, the existence of non-subject-related tags may impede the application of social tagging in traditional library cataloguing systems (Lu, Park and Hu, 2010).

By Trant (2008), methods of researching the contribution of social tagging and folksonomy were described, and outstanding research questions were presented. This is a new area of research, where theoretical perspectives and relevant research methods are only now being defined. This paper provides a framework for the study of folksonomy, tagging and social tagging systems. Three broad approaches are identified, focusing first, on the folksonomy itself (and the role of user tags in indexing and retrieval); secondly, on tagging (and the behaviour of users); and thirdly, on the nature of social tagging systems (as socio-technical frameworks). This paper provides a framework for both the study of social taggingand folksonomy and the analysis of their contribution to the on-line information landscape (Trant, 2008).

Other researchers e.g., “Chung and Yoon (2009); Jorgensen (1998, 1999, 2003); Trant (2006) have also found differences between users’ tags and descriptions of images by professional indexers. Their recommendation was that social tagging and traditional/professional indexing should be used together to complement each other. The importance of tagging as complementary activity to indexing is not a new idea”(Lu, Park and Hu, 2010).

Mathes (2004) examined 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. This paper was based on the comparisons title, author and MeSH keywords and tags from a set of PubMed articles bookmarked on CiteULike (Mathes 2004).

According to Shiri, Revie & Chowdhury (2002) while many studies have compared social tagging terms to controlled vocabularies, this paper was 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. They suggested that “tagging does not replace controlled vocabularies, but instead provides an added dimension to subject access” (Shiri Revie and Chowdhury 2002).

Conclusion

From the above literature review it has been clearly manifested that in spite of so many research works on this topic, focus has not been given on the metadata analysis of book catalogues of different domain like philosophy, history, mathematics, literature, political science etc. So a knowledge gap is seen in this regard. With the purpose of fulfilling the knowledge gap an effort has been made to pursue a research work on this area.

References

Bogers, T. & Petras, V. (2015). Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search? In Proceedings of iconference 2015.

Bogers, T. & Petras, V. (2017). An In-Depth Analysis of Tags and Controlled Metadata for Book Search. In Proceedings of iconference 2017.

Bogers, T. & Petras, V. (2017). Supporting Book Search: A Comprehensive Comparison of Tags vs. Controlled Vocabulary Metadata. Data and Information Management, 2017; 1(1): 17–34

Cattuto, C., Christoph, S., Andrea, B., Vito DP S., Vittorio, L., Andreas, H., Miranda, G., &

Gerd, S. (2007). Network Properties of Folksonomies. AI Communications, 20, 245-262.

Dash, C. G. J. (2015). A matter of context: An investigation into the representation of bias in social tags and Library of Congress Subject Headings. Department of Information Studies, AberystwythUniversity, Retrieved from https://pdfs.semanticscholar.org/c16a/5609399930a2643258e69b46437678a0132a.pdf

Doctorow, C. (2001). Metacrap: Putting the Torch to Seven Straw-Men of the Meta-Utopia. Retrieved from https://people.well.com/user/doctorow/metacrap.htm

Feinberg, M. (2006). An Examination of Authority in Social Classification Systems. *Advances in Classification Research Online*, 17(1), 1-11.

Furner, J. (2010). Folksonomies. *Encyclopedia of Library and Information Sciences*. New York: Taylor and Francis, 1858-1866.

Golder S. & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems, *Journal of Information Science,* 32(2), 198–208.

Kipp, M. E. I. (2011). Controlled vocabularies and tags: An analysis of research methods. *North American Symposium on Knowledge Organization (NASKO), Toronto, June 15-16, 2011*.

Koolen, M. (2014). User Reviews in the Search Index? That’ll Never Work!. In *ECIR ’14: Proceedings of the 36th European Conference on Information Retrieval*. 323–334.

Kroski, E. (2005). The Hive Mind: Folksonomies and User-Based Tagging. *Info Tangle Blog, December, 2005*

Lee, D. H. & Schleyer, T. (2010). A Comparison of MeSH Terms and CiteULike Social Tags As Metadata for the Same Items. In *IHI ’10: Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, New York, NY, USA, 445–448.

Lu, C., Park, J.-R. & Hu, X. (2010). User Tags versus Expert assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings. *Journal ofInformation Science,* 36(6), 763–779.

Ma, X. & Cahier, J. P. (2012). Visual Distinctive Language: Using a Hypertopic-based Iconic Tagging System for Knowledge Sharing. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) 2012 IEEE 21st International Workshop*. IEEE. 456-461.

Martinez-Avila, D. (2015). Knowledge Organization in the Intersection with Information Technologies. *Knowledge Organization,* 42(7), 486–498, Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=111299953&lang=pt-br&site=ehostlive.

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

Matthews, B., Jones, C., Puzon, B., Moon, J., Tudhope, D. et al. (2010). An evaluation of enhancing social tagging with a knowledge organization system. ASLIB Proceedings, 62(4/5), 447-465

Noruzi, A. (2006). Folksonomies: (Un) Controlled Vocabulary? *Knowledge Organization*, 33 (4), 199-203.

O'Reilly, T. (2005). What Is Web 2.0?: Design Patterns and Business Models for the Next Generation of Software". Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is- web-20.html.

Rafferty, P. M. (2017). Tagging. *Knowledge Organization*, 45(6), 500-516.

Rorissa, A. (2008). User-generated descriptions of individual images versus labels of groups and image: A comparison using basic level theory. Information Processing & Management. 44.5: 1741-1753.

Shiri A.A, Revie C, & Chowdhury G. (2002). Thesaurus-assisted search term selection and query expansion: a review of user-centered studies. *Knowledge Organization*, 29(1), 1–19.

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10(1), 1-44

Vaidya, P. & Harinarayan, N. S. (2016). The role of social tags in web resource discovery: an evaluation of user-generated keywords. *Annals of Library and Information Studies*, 63(Dec.), 289-297.

Wal, T. V. (2007). *Folksonomy coinage and definition, Retrieved from* http://vanderwal.net/folksonomy.html.

Zauder, K. Jadranka, L. L., & Mihaela, B. Z. (2007). Collaborative Tagging Supported Knowledge Discovery. *Information Technology Interfaces 2007. ITI 2007. 29th International Conference on*, IEEE. 437-442