

Convergence and divergence in tagging systems: An examination of tagging practices over a four year period

Abstract: This paper analyses the tagging patterns on delicious.com over a 4 year period using informetrics methods to assess how collaborative tagging supports and enhances traditional document indexing. Patterns in tag usage also highlighted practices related to personal and collective information organisation which conventional systems are unable to facilitate.

Résumé : Cette communication analyse les modèles d'étiquetage sur delicious.com sur une période de quatre ans au moyen de méthodes informétriques dans le but d'évaluer comment l'étiquetage collaboratif appuie et augmente l'indexation traditionnelle. Les modèles d'utilisation des étiquettes mettent également en évidence des pratiques d'organisation individuelle et collective que les systèmes conventionnels ne peuvent soutenir.

1. Introduction

Traditionally, information organisation has involved creating a representation (e.g. a bibliographic entry) and organising that representation as a substitute for the document itself. In the traditional card catalogue, this representation appeared at multiple places in alphabetic sequences of title, author and subject. When bibliographic records migrated to the computer environment (first with online databases, then with online catalogues), the subject access points (or descriptors) became important means of navigating the information archive, by enabling collocation, and the use of hierarchical and associative syndetic links to other useful terms. Thus indexing was the main means of representing the aboutness of the document without the full document being available. In the creation of bibliographic records, cataloguers and indexers were tasked with the creation and application of index terms and classification numbers to represent the aboutness of the document.

With the creation of digital archives and databases of articles, free text searching became a viable method for the location of documents in addition to the use of index terms. Free text search of the full text of an article was seen as a potential replacement for index terms. Both options, indexing and free text search, are useful and important, but neither is completely successful. Indexing is expensive due to the need for trained indexers and subject specialists and full text is not always available, whether due to copyright issues or the fact that it is not available in digital form. Even with access to the full text, issues of synonymy, homonymy and vocabulary choice make searching difficult and users still cannot be sure if they have found the most relevant documents about a particular subject.

Social tagging and folksonomies created in a distributed fashion through social bookmarking sites are being touted as a potential solution to some of the problems inherent in the organisation of a rapidly expanding collection of information (Morville 2005). This use of user tags, combined with topic maps and tag clusters, may have the potential to provide the benefits of a controlled vocabulary which controls for terminological differences, while still allowing the use of natural language vocabulary

(Shirky 2005). Additionally, the user created nature of these organisational schemes suggests a new method for resolving the gap between a user's information need and its translation into a search query by increasing the user's involvement in the categorisation process and combining it with elements of personal information management.

Early work by Kipp (2005), Hammond et al (2005), Sen et al. (2006), Golder and Huberman (2006), and Kipp and Campbell (2006), determined that tagging has many similarities to conventional indexing and also substantial differences; however, early studies of tagging were unable to examine a body of tagging data over a longer period of time. This paper attempts to shed light on this debate by looking closely at the patterns of tagging which have developed over a 4 year period associated with a set of URLs bookmarked in the delicious.com bookmarking system, and assessing how the tags applied by individual users in this popular tool both resemble and differ from descriptors that professional indexers would apply to a set of documents.

2. Methodology

This case study examined the changes over time in the tags applied to a set of delicious.com bookmarks first selected for their popularity and for being highly tagged in 2006. Examining the delicious.com data by bookmark or URL allows for the thorough examination of a set of tags assigned to specific URLs, mimicking the study of a set of index terms assigned to documents in a library, because we are interested in how tagging compares to traditional library indexing of books and other documents. An examination of tagging patterns after a 4 year period allows for the discovery of longer term trends in tagging that may not have been apparent with only a year or two of data. If tagging is to provide useful patterns over the long term it is essential that researchers examine data over a longer period of time to look for evidence of stabilisation of terms or patterns (or the lack thereof) in the tags and tag clouds.

Data for this study was collected in 2006 and again in 2008. In each case, data collected included all posts (a single instance of a bookmark being posted and tagged by one user) since the initial posting of the item, so this study was able to examine up to 4 years of data per URL. The data was examined using standard informetrics and statistical measures, inter indexer consistency measures and co-occurrence analysis.

3. Results

Results of the study show that there is still a mix of consensus and divergence in tagging term use and tagging patterns. While some of the chosen URLs maintained or even increased their popularity, others experienced a severe drop in popularity. This is, of course, similar to the popularity of documents in a library which are seasonal or event specific. Many patterns in the studied URLs have remained relatively stable. The number of tags per person per article continued to hover around 1-3 tags and many tag frequency graphs showed roughly the same sets of popular tags as had been present in 2006.

While consistency and convergence is increasingly present in tag frequency graphs, this study shows that divergence of tag term use is still present and may in fact be increasing. Approximately 2/3 of the tags collected in this study were unique (62% in 2006 and 68% in 2008). Therefore, it can be said that in the aggregate people tend to agree on certain tags which apply to bookmarked items while maintaining a substantial divergence of opinion in the area of personal opinion and differing terminology choice often referred to as the long tail (Anderson 2004).

An examination of tagging patterns also shows that previous taggers may influence the selection of tags by new users. Examining the intersection of each individual user's tag list with the set of all tags assigned by the previous users shows a small, but significant, positive correlation suggesting that users are being influenced by previous users' tags. Despite this influence, there is also evidence of substantial and continuing divergence of opinion as shown by consistently low inter tagger consistency scores. However, inter tagger consistency scores are not substantially lower than inter indexer consistency scores reported by studies examining indexers, which suggests that tagging may provide sufficient convergence to be useful.

Over a 4 year period, the proportions of each tag as compared to the total number of tags also fluctuate although islands of stability do exist. The fluctuations are greater than might be expected given previous studies of shorter term use of tags. These fluctuations may be a sign of tag decay over time (Russell 2007) and may also indicate changes in user perspectives over time or the beginnings of terminology changes over time.

4. Discussion

This study does not definitively answer the question of whether collaborative tagging systems can function as an effective entry vocabulary for controlled vocabularies and a replacement where controlled vocabularies have never been deployed. Convergence (as examined by frequency analysis of tags) and divergence (as examined by inter indexer consistency and co-word analysis of tags) measures, however, provide some illuminating insights into the way tagging patterns emerge. They reveal that closely-related terms are not necessarily revealed through co-occurrence; they also reveal that users employ a wide variety of conventions in constructing tags: conventions which they apply inconsistently. While clustering is present in many cases, it is not always clearly marked, suggesting that tagging, like many other indexing methods, resorts to multiple terms to describe the aboutness of documents. Additionally, while early studies of tagging reported that convergence and stability were simply a matter of having sufficient users tag an item, this study suggests that tagging may always show fluctuations in term usage based on fluctuations in user interests in everyday life. These results suggest there is continuity between conventional indexing and user tagging: a continuity that could form the basis for a complementary system of subject access that will enrich conventional indexing and provide strong support for its continued utility. Additionally, the differences suggest that user tagging extends beyond the traditional objectives of subject access, and expresses a dynamic relationship between document and user, and between subject and task, which may lead to new ways of modelling subject access. Fluctuations in the proportions of tags and in terminology use suggests that tagging may even have the possibility of remaining current with fluctuations in terminology use in popular literature.

5. References

Anderson, Chris. 2004. The long tail. *Wired* 12.10.

<http://www.wired.com/wired/archive/12.10/tail.html>

Golder, Scott A., and Bernardo A. Huberman. 2006. The structure of collaborative tagging systems. *Journal of Information Science* 32, no. 2:198-208.

Hammond, Tony, Timo Hannay, Ben Lund, and Joanna Scott. 2005. Social bookmarking tools (I): A general review. *D-Lib Magazine* 11, no. 4.

<http://www.dlib.org/dlib/april05/hammond/04hammond.html>

- Kipp, Margaret E.I. 2005. Complementary or discrete contexts in on-line indexing: A comparison of user, creator and intermediary keywords. *Canadian Journal of Information and Library Science* 29, no. 4:419-436. <http://eprints.rclis.org/8379/>
- Kipp, Margaret E.I., and D. Grant Campbell. 2006. Patterns and inconsistencies in collaborative tagging practices: An examination of tagging practices. *Annual General Meeting of the American Society for Information Science and Technology, Austin, TX, USA, November 3-8, 2006*. <http://eprints.rclis.org/8315/>
- Morville, Peter. 2005. *Ambient findability: What we find changes who we become*. Sebastopol, CA: O'Reilly.
- Russell, Terrell. 2007. Tag decay: A view into aging folksonomies. Proceedings of the Annual Meeting of the American Society for Information Science and Technology, Milwaukee, WI, USA, October 19-24. <http://weblog.terrellrussell.com/2007/11/tag-decay-poster-from-asist-is-online/>
- Sen, Shilad. Shyong K. (Tony) Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Ried. 2006. Tagging, communities, vocabulary, evolution. *CSCW Conference 2006*.
- Shirky, Clay. 2005. Ontology is overrated: Categories, links, and tags. *Shirky.com*. http://shirky.com/writings/ontology_overrated.html