User-based tagging as information retrieval: possibilities and limitations.

Journal: Aslib Proceedings

Purpose: The purpose of the paper is to examine the value of collaborative tagging as a method of organising information material.

Design/methodology/approach: The paper examines the advantages and disadvantages of collaborative tagging compared to traditional controlled vocabularies. Conclusions are drawn from an analysis of how items are tagged on Delicious and an assessment of their usefulness for the end user.

Findings: Despite a lack of consistency in collaborative tagging, there is a currency and affordability that controlled vocabularies lack. Furthermore, given the analysis of tagging on Delicious, it is proposed that a combination of the consistency of controlled vocabularies and the immediacy and relevance of collaborative tagging would provide a suitable method for managing the flow of information.

Research limitations/implications: More thorough research should be conducted by the profession on a wide-range of user-based tagging systems to assess the effectiveness of collaborative tagging as an information retrieval tool.

Practical implications: User-based tagging should be encouraged amongst the library profession and means found to combine user-based tagging with controlled vocabularies in order to improve information retrieval for the end user.

Originality: This paper contributes to the debate about the value of user-based tagging as an information retrieval tool.

Paper type: Research paper

Keywords: Classification, Internet, Tagging, Information Retrieval

Background

The organisation of materials into a searchable information retrieval system has traditionally relied on the experience of trained information professionals. As indexing was traditionally the responsibility of professionals, it was down to the indexer to undertake a conceptual analysis of the information source to determine its 'aboutness'. This required the indexer to develop an understanding of the subject matter of the material as well as an awareness of the needs of those who wish to use it (Lancaster, 1986, p.3). Once a conceptual analysis had been completed, a controlled vocabulary was consulted that provided a limited set of terms which could be applied to classify the material. As controlled vocabularies provided a specific number of index terms, they ensured an element of consistency in the indexing process, as well as theoretically making the job of the end user easier. By only using terms from a pre-defined list, it ensured a consistency that was believed would not have been apparent had indexers been able to freely apply their own index terms. This, as Lancaster argues, should provide the end user with all the terms required to perform a comprehensive search of relevant materials (Lancaster, 1986, p.8). However, despite the long term use of controlled vocabularies they are not without their problems.

The biggest problem faced by any organisation seeking to implement a controlled vocabulary is cost. The more sophisticated (and therefore more accurate) the vocabulary, the more expense it becomes to apply and maintain (Lancaster, 1986, p.239). As a result of this, organisations are faced with a dilemma between high cost and high precision or low cost and low precision. It would be extremely costly to conceptually analyse every information resource before determining which index items should be applied, particularly in the modern era. It is not only cost that creates problems in relation to controlled vocabularies, the experience of the end user also suggests problems with this method of index control.

Due to the reliance on a thesaurus when indexing, there is an expectation that the end-user understands the index terms that would be applied. A user who is unaware of the appropriate terminology will not necessarily find the materials they are looking for and would, therefore, require 'educating' on utilising the retrieval system. Although this may be desirable, it is certainly not practical. Furthermore, controlled vocabularies rely on the assumption that information documents have stable meanings that are universally accepted (Rafferty and Hidderley, 2007). Of course, as every reader is individual, different documents will mean varying things to different people. The interpretation of a document's

'aboutness' may vary greatly between the indexer and the user. Ultimately, this form of organising information is unnecessarily complex for the end-user in the internet age. As the phenomenon of web 2.0 has emerged, so too has the concept of 'aboutness' being determined by the end user, rather than a 'professional'. It is this shift towards user-based tagging that this paper seeks to analyse and identify the possibilities and limitations of this form of indexing.

The emergence of a new vocabulary?

The growth of the internet in recent years has also seen the emergence of new ways of indexing information materials. Social networking and Web2.0 have becoming increasingly popular with internet users. For the information professional, perhaps the most intriguing development of the social networking phenomenon has been the emergence of 'tagging' based services. Sites such as *Flickr* (www.flickr.com/), a web based tool for storing and organising photographs, and *Delicious* (http://delicious.com/), an online tool for managing favourite web pages, enable users to apply their own index terms to items using 'tags'. Unlike traditional forms of cataloguing, tagging does not rely on controlled vocabularies or thesauri, instead it relies on the user to perform a conceptual analysis to determine the 'aboutness' of what it is that they are indexing. Instead of referring to a trained 'professional', users determine the index terms that are most relevant to *them* and apply them accordingly. Once tagged, the item can be browsed by other users of the service using the terms that have been applied to the material. Often, these tags are manifested in tag clouds

Figure 1 A 'tag cloud' on *Delicious*

that reproduce all the tags created by the user and emphasises their frequency of use through the size of the font (see Figure 1). This enables the browser to

quickly identify the most common tag and find the material with which that term has been applied. This low-cost method of indexing is particularly useful in terms of indexing information materials accessible via the internet. With over a trillion unique URLs on the internet (Google, 2008), indexing all the information materials on the internet would be prohibitively expensive for any institution, particularly if a controlled vocabulary was utilised. Conceptual analysis of every page on the internet would be an endless task as new pages emerge continuously. As Golder and Huberman argue (Golder and Huberman, 2005), user-based tagging is most useful when there is 'nobody in the librarian role' or when there is too much content for a single authority to classify, and the internet certainly fits that description. Certainly, many view the emergence of user-based tagging as a way to supersede the indexing role of the information professional and to facilitate resource discovery over the Web (MacGregor and McCulloch, 2006).

The development of user-based tagging presents a number of possibilities for the end user. As the indexing does not rely on a controlled vocabulary and an information 'professional' to determine the 'aboutness' of the item, users are able to apply language that is relevant to them. This means that terms that are not necessarily appropriate from a professional perspective can be utilised by the user who will have no such qualms about what may be deemed as correct terminology. For example, the term 'credit crunch' is a relatively recent addition to our common vocabulary, and not one that is likely to be employed by a professional indexer given its colloquial nature. A search on *The Guardian*'s website shows that in the period 2003-6, the phrase was only used sixteen times (Guardian, 2008) and had yet to enter common usage. However, in 2008 the term was used over six thousand times, reflecting how the phrase has now become a popular term to describe the economic situation. In fact, an entry for the term 'credit crunch' only appeared on *Wikipedia* as recently as February 2007 (Wikipedia, 2007). This is also reflected in the use of the tag 'credit_crunch' on *Delicious*, the bookmarking website. To date, there are nearly two thousand web pages tagged with the term 'credit_crunch' and the first link with that tag was added as recently as August 2007 (Delicious, 2008a). For an information professional, the application of colloquial terms to an item would not be considered an appropriate method of indexing. The use of such terminology underlines one of the great advantages of user based tagging. Although an information professional may not think a colloquial term is suitable for indexing an item, it might be the most appropriate term for the end user. As a result, the

application of user-based tagging has greater relevance for the end user compared to terms identified by a professional indexer. Furthermore, as these index terms are generated independently, there is no requirement to expend vast amounts of time trying to reach agreement to ensure that the 'correct term' is applied.

However, it is not just in the currency of vocabulary that this new 'uncontrolled vocabulary' is advantageous in its application. Collaborative tagging has numerous advantages for the casual browser. Unlike controlled vocabularies where index terms are applied according to rules, collaborative tagging allows a certain amount of flexibility in the way items are indexed. This means that items may not be indexed 'correctly', but they may still prove useful for the casual browser. As Mathes argues, there is a serendipitous nature to collaborative tagging that enables the end user to find materials that they would otherwise not have been exposed to if they had simply relied on a controlled vocabulary (Mathes, 2004). As 'aboutness' is determined by an individual indexer, it is possible for many indexers to evaluate 'aboutness' in different ways. What is the correct index term for one person is not necessarily the right one for another. This allows users to discover items that they may not think are relevant initially, but may turn out to be highly relevant to their search. With a wide variety of tagging, there follows an increased number of access points for the end user leading to a number of ways of exploring that which is available and relevant.

Collaborative tagging

Although collaborative tagging has a number of advantages for cataloguing information, it is not without its limitations. Perhaps its greatest strength is also its biggest weakness. By allowing anyone to tag anything as they wish, some peculiar tagging habits have developed that would make it virtually impossible for the casual browser to find what they are looking for. For example, an article about the 'One Laptop Per Child' program on the *BBC* website (BBC, 2008) was indexed on Delicious with the following terms:

- 'oplc'
- 'One_laptop_per_child'
- 'one', 'laptop', 'per' and 'child' (all individual tags)

(Delicious, 2008b)

The latter form of tagging is a particularly common method of indexing items on *Delicious*. Many users decide that they will only apply single word tags to any item so that when an item such as this one occurs, the user applies each individual word as a tag. Whilst this may be relevant for the individual users (three users use this format), it is not particularly practical for those conducting a comprehensive search on the subject (Lancaster, 1986, p.8). Similarly, although users familiar with the abbreviated form might well find the materials they require, those who are not will probably be unable to locate such information. This isn't so much of an issue with materials that have been bookmarked by a variety of users, but it does create problems if the item is only bookmarked once (or if only one of the less obvious tag terms had been applied). If a controlled vocabulary were applied to the article, only one of these terms would be applied to the resource which would ensure that all the appropriate materials are located, but the question is then which of these terms should be used? Although the above tagging may make an advocate of controlled vocabularies wince, there are clearly users who find these terms useful (or else why would they use them?).

A further problem with tagging comes with the issue of homonyms. Whereas a controlled vocabulary can direct the user to the appropriate term using a qualifier, at present user-based tagging does not have that capability. For example, a user searching for information on nails using a controlled vocabulary might be re-directed with a qualifier such as 'Nails (fasteners)' (Lancaster, 1986, p.6). However if the same search terms were applied to a user-based tagging system, the user would also discover results that related to beauty products. It is quite unlikely that the user would be interested in both varieties of nail products. This problem can only be addressed on *Delicious* by investigating the other tags that have been applied to the items, which could be particularly time consuming and would most likely exclude relevant material.

Although tagging does present some problems regarding consistency, there is the opportunity to develop tagging as an effective method for classifying information. A combination of a controlled vocabulary and user-generated tagging has the potential for developing a simple to use method of categorising information. Even though *Delicious* can appear to be a little inconsistent in the application of tags,

commonalities do emerge, where a large number of users tend to agree on the 'aboutness' of a particular resource. In a study conducted by Golder and Huberman

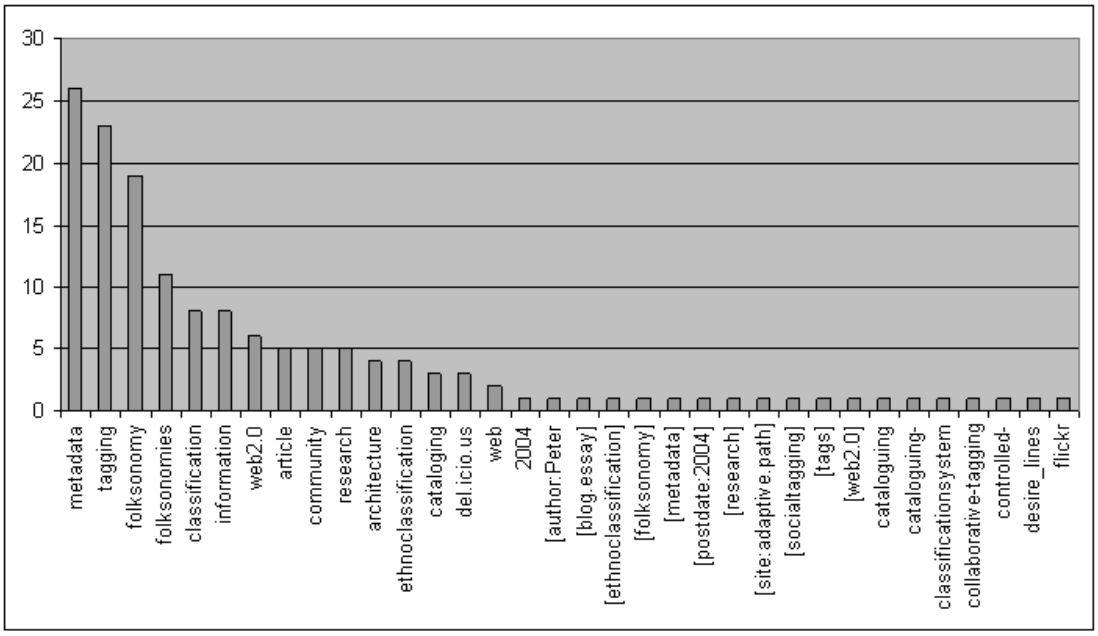

Figure 2 Tags applied to Merholz's article

(Golder and Huberman, 2006), it emerged that after a small number of bookmarks a 'nascent consensus' emerged that is not affected by additional tags. For example, an article by Peter Merholz entitled 'Metadata for the masses' has been bookmarked on *Delicious* by 56 members (Delicious, 2008c). Whilst there is a wide variety of tags that have been applied (there are 61 different tags for this item), there is a common sense amongst users regarding the 'aboutness' of the article. As one can see from figure 2, there is a general agreement about what tags should apply to this item. After the first four tags, there is a steep drop-off. In fact, should *folksonomies* and *folksonomy* be combined (the only difference being that the former is the plural form of the latter), there is an even bigger drop-off after the first three tags (*metadata*, *tagging* and *folksonomy*). As each of these tags are utilised by 50% of those that have bookmarked the article, this suggests that there is a high degree of agreement between users regarding the subject of the article (or it's 'aboutness'). Consequently, it would appear that there is indeed a 'nascent consensus' amongst tagged item as Golder and Huberman suggest.

Given that there is a degree of commonality between different users' tags, it would be possible to control these tags and present a usable method for categorising information materials. Once an item had been bookmarked a certain

number of times, the software can direct users to what are the most appropriate tags. *Delicious*, for example, already makes suggestions for useful tags when a popular item is bookmarked. Merholz argues that classifications could emerge rather like 'desire lines'. These are paths that landscapers develop after they have let people create their own paths through the park. Once these 'desire lines' have emerged, it should be possible to use the most common tags to develop a controlled vocabulary that is more representative of the language of the user (Merholz, 2008). This combination of user-generated content and a controlled vocabulary could be very useful for the categorisation of the internet, as well as for other information materials. This combination of the consistency of controlled vocabularies and the currency of user-based tagging, would ensure a system that is user-friendly and relevant.

Conclusion

User-based tagging does have a number of advantages over controlled vocabularies. It has the advantage of being an economical alternative to the costly and time consuming processes required of a controlled vocabulary. It also has the advantage of creating information records that are in the language of the end user rather than that of the information professional. On the other hand, the lack of consistency in tagging causes a number of problems. For the end user it is relatively easy to miss items that may be relevant for their search strategy. If tags are incorrectly spelt, or utilise a format that is relevant for the creator of that record but not to anyone else, they will not be particularly useful for the searcher wishing to conduct a comprehensive search.

Despite these limitations, there is a way forward for tagging that is beneficial for the information professional. As was demonstrated by the analysis of tags applied to a particular web address, there was a level of agreement of the 'aboutness' of the article. The vast majority of users appeared to settle on three particular tags that were considered particularly relevant. Thereafter, there were a number of tags that were rarely applied and often represented individual perspectives. This tagging behaviour suggests that it would be possible to combine the consistency of a controlled vocabulary with the immediacy and relevance of user-based tagging. If the software was able to direct users to tags that were deemed particularly suitable due to their popularity, it would ensure consistency whilst also allowing the subject matter to still be determined by the end user. This combination of controlled vocabularies and user-based tagging is

not only useful for indexing information materials on the internet, it could also be use to allow library users to search photos within library collections, as well as books that are available through the OPAC. User-based tagging does have its limitations, but in combination with the principles of a controlled vocabulary, the possibilities of providing a service more relevant to the end user should outweigh any concerns about its application.

References

BBC (2008), "One Laptop signs up with Amazon", available at http://news.bbc.co.uk/1/hi/technology/7599652.stm (accessed 6 September 2008).

Delicious (2008a), "Recent credit_crunch bookmarks", available at http://delicious.com/tag/credit_crunch?detail=2&page=20 (accessed 23 Novemeber 2008)

Delicious (2008b), "Everyone's bookmarks for BBC NEWS/Technology/One Laptop signs up with Amazon", available at http://delicious.com/url/e9 4fc5b9ddeae02a639ad4bca35366a0?show=all (accessed 6 September 2008).

Delicious (2008c), "Everyone's bookmarks for adaptive path/metadata for the masses", available at http://delicious.com/url/e76fc6 b1303afcbb68e32969d4d84249 (accessed 5 Janury 2009)

Golder, S.A. and Huberman, B.A. (2006), "Usage patterns of collaborative tagging systems", *Journal of Information Science*, Vol 32 No 2, pp 198-208. Available from Sage at http://jis.sagepub.com/cgi/content/abstract/32/2/198 (accessed 11 November 2008).

Google (2008), "We knew the web was big…", available at http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (accessed 15 October 2008).

Guardian (2008), Search for the term "credit crunch", available at http://browse.guardian.co.uk/search?search_target=/search&fr=cb-guardian&search="credit+crunch"&N= (accessed 23 November 2008).

Lancaster, F.W. (1986), Vocabulary Control for Information Retrieval, Information Resources Press, Arlington, VA.

Macgregor, G. and McCulloch, E. (2006), "Collaborative tagging as a knowledge organisation and resource discovery tool", *Library Review*, Vol 55 No 5, pp 291-300. Available from Emerald at http://www.emeraldinsight.com /10.1108/00242530610667558 (accessed 23 November 2008).

Mathes, A. (2004), "Folksonomies – Cooperative Classification and Communication Through Shared Metadata", available at http://www.adammathes.com/ academic/computer-mediated-communication/folksonomies.html (accessed 11 November 2008).

Merholz, P. (2004), "Metadata for the Masses", available at http://www.adaptivepath.com/ideas/essays/archives/000361.php (accessed 25 November 2008)

Rafferty, P. and Hidderley, R.(2007), "Flickr and Democratic Indexing: dialogic approaches to indexing", *Aslib Proceedings*, Vol 59 No 4/5, pp 397-410. Available from Emerald at http://www.emeraldinsight.com /10.1108/00012530710817591 (accessed 14 October 2008).

Wikipedia (2008), "Credit Crunch", available at http://en.wikipedia.org/ w/index.php?title=Credit_crunch&dir=prev&action=history (accessed 23 November 2008).