

Exploring the Context of User, Creator and Intermediary Tagging

Margaret E. I. Kipp
Faculty of Information and Media Studies
University of Western Ontario
mkipp@uwo.ca

Abstract

This paper examines the results of a study of the three groups involved in creating index keywords or tags: users, authors and intermediaries. Keywords from each of the three groups were compared to determine similarities and differences in term use. Comparisons suggested that there were important differences in the contexts of the three groups that should be taken into account when assigning keywords or designing systems for the organisation of information.

Introduction

The substantial increase in access to information afforded by the Internet has only strengthened the importance of being able to, at once, distinguish between similar documents and locate relevant documents. The topics of navigability, findability, and relevance, under the guise of information retrieval and information seeking, have been of importance in library and information science for some time and supporting these tasks is one of the important tasks of information architects.

Searching a large document space for information is a difficult problem, due to the sheer size of the space, as well as the ambiguities inherent in natural languages. This problem is only exacerbated by the increasing use of digital databases consolidating masses of data. Classification and indexing via controlled vocabularies is a common method of attempting to resolve this problem by using controlled vocabularies to rationalise natural languages by removing ambiguities and consolidating similar items. A solidly designed classification system using terms and keywords appropriate to the context of the intended user can help to reduce the difficulty inherent in searching large document spaces for information.

While the creation of generic hierarchical classification systems or subject specific taxonomies has a long history, the design of these systems has largely been left to professional intermediaries such as librarians and information architects. Because of the increasing amount and specialisation of information being collected and user requests for more fine grained access, these systems tend to be too generic for user needs. While full text search can provide fine grained access to supplement controlled vocabularies, this access tends to be at the expense of precision due to the use of differing terminology. Increasingly, user tagging and folksonomies

created in a distributed fashion through social bookmarking sites are being touted as a potential solution to these problems. (Mathes 2004; Hammond et al 2005; 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)

Mathes (2004) notes that there are three common groups involved in the assignment of keywords to documents. These groups are authors, intermediaries (such as librarians), and users. Author keywords have received relatively little attention, perhaps because they can be difficult to collect. And, while intermediaries have been indexing documents for some time, large scale user created collections of tagged documents are relatively new. Like the hierarchical thesauri created by intermediaries to organise knowledge formally, the new user created folksonomies allow the user to navigate from one topic to another using related links (related terms in a thesaurus). However, relationships in the world of folksonomies include relationships that would never appear in a thesaurus, such as the identity of the user (or users) who used the tag. (Morville 2005, 137) This phenomenon adds a new contextual dimension to the act of organising information that is not present in intermediary assigned keywords.

This paper examines the differences in the context of user, author and intermediary assigned keywords or tags using the social bookmarking site citeulike (<http://citeulike.org/>), which has been specialised for use by academics who wish to store links to academic articles. Using a set of journals which had been found to request author keywords, a sample of journal articles tagged in citeulike were selected for analysis. The chosen articles were manually restricted to those which were indexed in INSPEC (Institution of Engineering and Technology, Hertfordshire, UK), Library Literature (H.W. Wilson Company, New York) or both. Both INSPEC and Library Literature provide intermediary assigned controlled vocabulary subject headers for searchers. Thus each article selected for this study has 3 sets of keywords assigned by three different classes of metadata creators. Term comparison via the INSPEC and Library Literature thesauri, and descriptive statistics were used to examine differences in context and term usage between the three classes of metadata creators.

Related Studies

Voorbij (1998) studied the correspondence between words in the titles of monographs in the humanities and social sciences and the librarian assigned descriptors existing in the online public access catalogue of the National Library of the Netherlands. His study used a 7 point scale of comparison between the title keywords and these descriptors, comparing the descriptors to the title words selected by the author. Voorbij used the different relationships in a thesaurus as an indication of closeness of match, beginning with an exact (or almost exact match), continuing to synonyms, broader terms, narrower terms, related terms, relationships not formally in the

thesaurus, and terms which did not appear in the title at all. (Voorbij 1998, 468)

A similar study by Ansari (2005) examined the degree of exact and partial match between title keywords and the assigned descriptors of medical theses in Farsi. She found that the degree of match was greater than 70 per cent. (Ansari 2005, 414) Both studies suggest that title keyword searching alone and controlled vocabulary searching alone lead to failure to find some articles.

Methodology

Since the purpose of the study was to compare all three forms of indexing, articles for this study were chosen from scholarly journals whose instructions for authors request author keywords, which was determined by examining sample articles and journal webpages. Journals included in this study are all in the field of library and information science including the Journal of Documentation, Information Processing and Management and the Journal of the American Society for Information Science and Technology. (See table 1 for the full list.)

A total of 205 entries were collected from citeulike.org. Each had been tagged by users of citeulike with at least one tag. Data were collected from citeulike.org with a python script (citeulike.py) and parsed to exclude articles which had not yet been tagged by users. Entries for which author keywords and database descriptors could not be found were excluded manually leaving 176 entries. Articles were located in online databases to extract author keywords and either INSPEC or Library Literature for intermediary descriptors. Identification was done via exact title match or digital object identifier (<http://www.doi.org/>). Where database descriptors were available, but author keywords were not, author keywords were replaced by significant words from the title of the article.

Journal	Article Count
Journal of the American Society for Information Science and Technology	68
Journal of Documentation	18
Information, Communication and Society	9
Information Processing and Management	52
International Journal of Geographical Information Science	8
Information and Organization	4
The Information Society	17

Table 1: Journals with author assigned keywords

To compare data, the controlled vocabularies of two on-line databases were used for the comparison: INSPEC and Library Literature as both databases index articles from information science journals. Controlled vocabulary terms from the thesauri used by these two databases also formed the third set of terms, referred to as intermediary descriptors.

Two forms of analysis were used: descriptive statistics and term comparison. User tags, author keywords, and intermediary assigned descriptors were compared based on a 7 point scale, similar to that used by Voorbij (1998). While Voorbij examined descriptor correspondence to title keywords, this study examines the correspondence between all three sets of tags using the structured thesaurus to generate similarity comparisons. Where possible, comparisons have been done across all three sets of terms, but where the term (or any related term) is lacking from one set, the other two sets were compared against the 7 categories. The following are the categories as modified.

1. Same - the descriptors and keywords are the same or almost the same (e.g. plurals, spelling variations, acronyms and multiword terms split into facets)
2. Synonym - the descriptors and keywords are synonyms (corresponds to USED FOR in a thesaurus)
3. Broader Term - the keywords or tags are broader terms of the descriptors
4. Narrower Term - the keywords or tags are narrower terms of the descriptors
5. Related Term - the keywords or tags are related terms of the descriptors
6. Related - there is a relationship (conceptual, etc) but it is not obvious to which category it belongs or it is not formally in the thesaurus
7. Not Related - the keywords and tags have no apparent relationship to the descriptors, also used if the descriptors are not represented at all in the keyword and tag lists

An initial sample of 10 entries was examined to determine if additional categories would be necessary. An additional sample of 50 entries was selected from the total number of entries for this term comparison with entries from each journal being selected proportionately to the total number of articles in the sample using a random number table.

Preliminary Findings

The largest number of tags provided by users for a single article was 13, by authors: 14, and by intermediaries: 7. Over 60% of tagged articles had between 1 and 3 tags, 4-6 author keywords and 2-4 intermediary descriptors assigned. Despite the potential for a large number of tags assigned by different users, articles did not tend to have a substantially larger number of tags. This may be due to the small volume of highly tagged articles in the sample set. The majority of articles had been tagged by 2 users, although a few articles had been tagged by as many as 10.

Using the modified version of Voorbij's scale, it was found that the most common relationship discovered in the groups of user, author and intermediary keywords examined was category 6 or related but not formally in the thesaurus. This form of relationship occurred in 35% of match cases. The next most common relationships were Broader Term and Narrower Term combined

at 22% and Equivalence (category 1) at 21%. Related Terms, in the thesaural sense, occurred in 14% of matches and only 8% of matches were synonyms. Non matching items were not included in this count, but occurred from 1-3 times per article.

Thesaural Relations

Unsurprisingly, some terms used by users did fall into the traditional thesaural relationships of related terms, broader terms and narrower terms. These relationships were less common than the first, sixth and seventh categories, covering equivalence (or near equivalence) of terms, and the related and not related categories respectively. In total, the thesaural relations accounted for 65% of all matches. This includes the equivalence category, synonyms, broader terms, narrower terms and related terms.

As could be expected for a free text tagging effort, user tag lists did tend to contain both spelling variants and plurals of the author keywords and intermediary descriptors, for example; 'communities-of-practice' and 'communities_of_practice' used as tags for the same article. Also, as expected, this phenomenon did not occur in the author or intermediary keywords.

Acronyms and abbreviations are extremely common in user tags, as are spelling variations. Some users have even helpfully provided spelling variations and both long forms and abbreviations in their tag sets. This also occurs when one user tags with abbreviations and the other uses long forms. This linkage of terms, which are then all displayed on the articles page, is extremely useful. INSPEC provides a similar service with its controlled and uncontrolled terms, where the controlled terms will tend to contain the full form of the term and the uncontrolled terms will contain the acronym. For example, the term GIS is used by both users and authors, while INSPEC provides Geographic Information Systems in its controlled terms and GIS in the uncontrolled terms. This apparent duplication would be extremely useful to newcomers to the field or interdisciplinary researchers.

A comparison of the use of single word and multi word indexing terms could be of interest, but is somewhat hampered by the requirement that a citeulike tag be a single word. Many users have chosen to use hyphens or underscores to allow the use of multiword tags in a single word and others have simply removed the spaces from multiword groupings. The frequency of occurrence of such multiword groupings is due to the lack of a single term in English to denote the subject, but may also be related to familiarity with library subject headings as opposed to faceted classification systems like tags where core concepts are assigned separately to an item, with the knowledge that they can be combined in an ad hoc fashion to fully describe the aboutness of a document.

Related Tags

Many relationships fell into the 6th category (35%), related but with some ambiguity in the relationship. This category included relationships that were ambiguous or difficult to fit into categories 1-5 as well as relationships that were not formally listed in the thesaurus, but suggested by user tags, author keywords, or INSPEC's uncontrolled terms. Common relationships included that between an object and its field of study, the relationship between two fields of study which examine different aspects of the same phenomenon, and the use of a methodology or form of inquiry in a new environment.

A frequent example of a relationship between two fields of study that examine different aspects of the same phenomenon, finding information, is the relationship between 'information seeking' and 'information retrieval.' In INSPEC's thesaurus, 'information seeking' is not a descriptor, but it is often used in the uncontrolled terms since these terms are taken from the document itself, including the title and abstract. (Institution of Electrical Engineers, 18) Since it is not a controlled term, 'information seeking' related articles tended to be tagged 'information retrieval' in INSPEC while authors and users would tag them as 'information seeking.'

Another example is the relationship between 'knowledge' and 'knowledge management.' Authors and users frequently used the term 'knowledge' in their keywords and tags while the intermediary descriptor 'knowledge management' would be used by INSPEC. This relationship is not equivalence, narrower or broader term, but there is a relationship between the two as knowledge management is the field of study concerned with the organisation and processing of organisational knowledge so that it can be located and reused.

An example of the use of a methodology or form of inquiry in a new environment is the use of the terms 'link analysis' and 'citation analysis' to describe the study of the relationships between web hyperlinks. While citation analysis has a long history in library and information science, and the term citation analysis is an INSPEC descriptor, link analysis or hyperlink analysis is a relatively newer field examining a similar phenomenon (references to other articles or sites) in a new environment. Combining the terms 'citation analysis' and 'Internet' or 'web' would serve the same function as the term 'link analysis' but the combined term allows users to be more specific without adding terms. This inclusion of newer terms in the user tags can happen faster than it would in a traditional thesaurus, as one of the goals of a thesaurus is to reproduce the accepted state of knowledge in a field, which leaves the leading edge of the field time to determine standard terminology that will eventually be added to the thesaurus.

Unrelated Tags

Tags, keywords and descriptors falling into the 7th category (Not Related) tended to fall into five basic types: time management tags, geographic descriptors, specific details and qualifiers,

generalities, and other. Since the author does not want to presume that the thesaurus is inherently superior in its indexing, descriptors that did not match any terms used by the author or users were also placed in this category. All geographical terms but two came from the descriptors. In the sample studied, there was only one user assigned geographic keyword and one author assigned geographic keyword.

Time management tags such as 'todo' and 'maybe' suggest that users wish to be reminded of the item, but have not yet read or not yet decided what to do with it. This is the electronic equivalent of the pile of articles to be read. This type of tag is not represented in either author keywords or intermediary descriptors because it is not thought to have value to anyone outside the individual assigning the tag. These tags also tend to have a short lifespan and so would require frequent updating of entries in a database or OPAC. However, amazon.com has shown that such tags can have value. Wishlists and “people who bought this book also bought” lists can help people to find new and interesting items by following the purchasing and viewing trails of people who read and enjoy similar material. This suggests that scholars might well find a todo or toread tag useful if they find another scholar who is reading similar material. It is worth noting here that a specific toread tag did not turn up in the sample, but this information is encoded in the stars located in the article entries and is requested separately on the article entry form using a scale ranging from “Top priority” to “I don't really want to read this.” (<http://citeulike.livejournal.com/6890.html>)

Another time management tag located in the unrelated category was “lis510” which looks like a course code. This is another example of a time or space sensitive tag which would presumably be of little use to anyone not teaching or taking the course. However, this tag could be extremely useful in an academic library where users could then search the catalogue for books and articles the professor has marked for the course.

Geographic tags, as previously indicated, were found mainly in the descriptors. This suggests that intermediaries are more likely to consider the geographic locations associated with the article to be relevant to the subject of the article. In the case of a copyright related article, tagged as “copyright, openaccess, romeo” the addition of the descriptor “Great Britain” would be extremely useful to a user searching for copyright related articles since copyright law varies greatly depending on country of origin. However, it is quite understandable that the users tagging this article did not consider this to be as important as the tags they actually used since this would presumably already be known to them. Another example of this phenomenon was a study of library students in Turkey in which the descriptor Turkey was not included in either the author or user tags. Two examples of geographic tags were found in user or author keywords, both referring to Internet policy in developing countries. Interestingly, these two tags were assigned where the descriptors failed to cover geographic location.

Another category of unrelated terms comprises specific details of the systems or user groups studied, qualifiers and methodologies. Surprisingly, the majority of these terms occurred only in the intermediary descriptors and did not appear in user or author keywords. Examples of these keywords included 'College and university students,' the specific group studied in the article, 'medical information systems,' the specific type of information system used in the information seeking study, and 'surveys,' representing the specific investigative method used in the tagged article. The lack of such identifiers in many user and author tagged studies suggests that, for example, both users and authors appear more interested in indicating that the article is about information seeking rather than about information seeking in a specific environment. Interestingly, the type of specific qualifiers used by users tended to refer to specific parts of the content of the article, for example, the use of the term 'web-graph' for a webometrics study to indicate that the article contains an application of graph theory to the topology of web links.

Comparable to the specifics category, another category of unrelated items was generalities. This category consisted of extremely general terms that could apply to almost any article in a field. Examples of this included the terms: computers, libraries/library, and information. This is not wholly unexpected as tagging systems lack a predesigned hierarchical thesaurus to provide access to broader or narrower terms. Users of tagging systems then have to provide any terms they consider relevant, including terms that might be considered too general to provide good distinguishability from other articles in the field.

The most commonly used tag in the final unrelated category, other, which occurred 18 times, was "no-tag". This turned out to be a system created default tag assigned to entries when the user has not assigned a tag; as such it does not provide any useful information about the contextual aboutness of the document for the user, although it does show interest in the document. It occurs in combination with other tags when multiple users have tagged the same document or if the original user neglects to remove it when editing the entry to add tags.

Discussion and conclusions

This study demonstrates that there are differences between the user, author and intermediary views of the concept space of the articles analysed. While intermediaries considered geographic location to be an important part of the description of the aboutness of an article, authors and users tended to assume it was somewhat less important than the other contexts of the articles. In many cases this may be true. For example, the difference between an information retrieval study performed in the United Kingdom and one performed in the United States is probably not significant due solely to the difference in geographic location.

Users considered time management information to be important as a tag for articles, wanting to encode information about their desire to read the article into the tags for easy access. This is

seen in the use of tags such as 'todo' and 'maybe' as well as in the use of the toread interface provided by citeulike when entering articles into the system.

Many user terms were found to be related to the author and intermediary terms, but were not part of the formal thesauri used by the intermediaries and thus were not formally linked to the intermediary terms in these thesauri. In some cases, this was due to the use of broad terms which were not included in the thesaurus such as information, knowledge, or computers. In many cases, this was due to the use of newer terminology or to differences in approach to a problem (information seeking vs information retrieval).

This study has implications for the design of systems for accessing, indexing and searching document spaces. The popularity of Google has demonstrated that users prefer to be able to search for items in a more natural way using one interface to locate items of a varied nature, but controlled vocabulary usage can be expensive. (Campbell and Fast 2004) User tagging, with its lower apparent cost of production, could provide the additional access points with less cost, especially if user tagging provides a similar or better search context.

References

- Ansari, M. (2005). Matching Between Assigned Descriptors and Title Keywords in Medical Theses. *Library Review*, 54(7), 410-4.
- Campbell, D. G., & Fast K. V. (2004). Panizzi, Lubetzky, and Google: How the Modern Web Environment is Reinventing the Theory of Cataloguing. *Canadian Journal of Information and Library Science*, 28(3), 25-38.
- Hammond, T., & et al. (2005). Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4). Retrieved January 11, 2005, from <http://www.dlib.org/dlib/april05/hammond/04hammond.html>
- Institution of Electrical Engineers. no date. Inspec on Engineering Village 2. Retrieved January 11, 2005, from <http://www.iee.org/publish/support/inspec/document/UserG/EV2UG.pdf>
- Mathes, A. (2004, December). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. *Adammathes.com*. Retrieved January 11, 2005, from <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>
- Morville, P. (2005). *Ambient Findability*. Sebastopol, CA: O'Reilly.
- Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. *Clay Shirky's writings about the internet*. Retrieved January 11, 2005, from http://shirky.com/writings/ontology_ouerrated.html
- 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), 466-76.