Hong Zhang, Linda C. Smith,
Michael Twidale, and Fang Huang Gao

Seeing the Wood for the Trees: Enhancing Metadata Subject Elements with Weights

Subject indexing has been conducted in a dichotomous way in terms of what the information object is primarily about/of or not, corresponding to the presence or absence of a particular subject term, respectively. With more subject terms brought into information systems via social tagging, manual cataloging, or automated indexing, many more partially relevant results can be retrieved. Using examples from digital image collections and online library catalog systems, we explore the problem and advocate for adding a weighting mechanism to subject indexing and tagging to make web search and navigation more effective and efficient. We argue that the weighting of subject terms is more important than ever in today's world of growing collections, more federated searching, and expansion of social tagging. Such a weighting mechanism needs to be considered and applied not only by indexers, catalogers, and taggers, but also needs to be incorporated into system functionality and metadata schemas.

Subjects as important access points have largely been indexed in a dichotomous way: what the object is primarily about/of or not. This approach to indexing is implicitly assumed in various guidelines for subject indexing. For example, the Dublin Core Metadata Element Set recommends the use of controlled vocabulary to represent subject in "keywords, key phrases, or classification codes."[1] Similarly, the Library of Congress practice, suggested in the *Subject Headings Manual*, is to assign "one or more subject headings that best summarize the overall contents of the work and provide access to its most important topics."[2] A topic is only "important enough" to be given a subject heading if it comprises at least 20 percent of a work, except for headings of named entities, which do not need to be 20 percent of the work when they are "critical to the subject of the work as a whole."[3] Although catalogers are aware of it when they assign terms, this weight information is left out of the current library metadata schemas and practice.

A similar practice applies in non-textual object subject indexing. Because of the difficulty of selecting words to represent visual/aural symbolism, subject indexing for art and cultural objects is usually guided by Panofsky's three levels of meaning (pre-iconographical, iconographical, and post-iconographical), further refined by Layne in "ofness" and "aboutness" in each level. Specifically, what can be indexed includes the "ofness" (what the picture depicts) as well as *some* "aboutness" (what is expressed in the picture) in both pre–iconographical and iconographical levels.[4] In practice, VRA Core 4.0 for example defines subject subelements as:

> Terms or phrases that describe, identify, or interpret the Work or Image and what it depicts or expresses. These may include generic terms that describe the work and the elements that it comprises, terms that identify particular people, geographic places, narrative and iconographic themes, or terms that refer to broader concepts or interpretations.[5]

Here again, no weighting or differentiating mechanism is included in describing the multiple elements. What is addressed is the "what" problem: What is the work of or about? Metadata schemas for images and art works such as VRA Core and CDWA focus on specificity and exhaustivity of indexing, that is, the precision and quantity of terms applied to a subject element. However, these schemas do not address the question of *how much* the work is of or about the item or concept represented by a particular keyword.

Recently, social tagging functions have been adopted in digital library and catalog systems to help support better searching and browsing. This introduces more subject terms into the system. Yet again, there is typically no mechanism to differentiate between the tags used for any given item, except for only a few sites that make use of tag frequency information in the search interfaces.

As collections grow and more federated searching is carried out, the absence of weights for subject terms can cause problems in search and navigation. The following examples illustrate the problems, and the rest of the paper further reviews and discusses the precedent research and practice on weighting, and further outlines the issues that are critical in applying a weighting mechanism.

Hong Zhang (hzhang1@illinois.edu) is PhD Candidate, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, **Linda C. Smith** (lcsmith@illinois.edu) is Professor, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, **Michael Twidale** (twidale@illinois.edu) is Professor, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, and **Fang Huang Gao** (fgao@gpo.gov) is Supervisory Librarian, Government Printing Office.

Examples of Problems

Exhaustive Indexing: Digital Library Collections

A search query of "tree" can return thousands of images in several digital library collections. The results include images with a tree or trees as primary components mixed with images where a tree or trees, although definitely present, are minor components of the image. Figure 1 illustrates the point. These examples come from three different collections and either include the subject element of "tree" or are tagged with "tree" by users. There is no mechanism that catalogers or users have available to indicate that "tree" in these images is a minor component.

Note that we are not calling this out as an error in the professionally developed subject terms, nor indeed in the end user generated tags. Although particular images may have an incorrectly applied keyword, we want to talk about the vast majority where the keyword quite correctly refers to a component of the image. Furthermore, such keywords referring to minor components of the image are extremely useful for other queries. This kind of exhaustive indexing of images enables the effective satisfaction of search needs, such as looking for pictures of "buildings, people, and trees" or "trees beside a river." With large image collections, such compound needs become more important to satisfy by combinations of searching and browsing. To enable them, metadata about minor subjects is essential.

However, without weights to differentiate subject keywords, users will get overwhelmed with partially relevant results. For example, a user looking for images of trees (i.e., "tree" as the primary subject) would have to look through large sets of results such as a photograph of a dog with a tiny tree out of focus in the background.

For some items that include rich metadata, such as title or description, when people look at a particular item's record, with the title and sometimes the description, we may very well determine that the picture is primarily of, say, a dog instead of trees. That is, the subject elements have to be interpreted based on the context of other elements in the record to convey the "primary" and "peripheral" subjects among the listed subject terms. However, in a search and navigation system where subject elements are usually treated as context-free, search efficiency will be largely impaired because of the "noise" items and inability to refine the scope, especially when the volume of items grows.

Lack of weighting also limits other potential uses of keywords or tags. For example, all the tags of all the items in a collection can be used to create a tag cloud as a low cost way to contribute to a visualization of what a collection is "about" overall.[6] Unfortunately, a laboriously developed set of exhaustive tags, although valuable for supporting searching and browsing within a large image collection, could give a very distorted overview of what the whole collection is about. Extending our example, the tag "tree" may occur so frequently and be so prominent in the tag cloud that a user infers that this is mostly a botanical collection.

Selective Indexing: LCSH in Library Catalogs

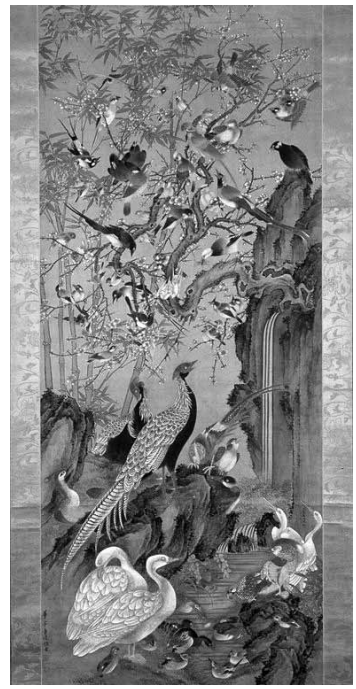
Although more extreme in the case of images in conveying the "ofness," the same problem with multiple subjects also applies to text in terms of "aboutness." The following example comes from an online library catalog in a faceted navigation web interface using Library of Congress Subject Headings in subject cataloging.[7]

The query "psychoanalysis and religion" returned 158 results, with 126 in "psychoanalysis and religion" under the Topic facet. According to the *Subject Headings Manual*, the first subject is always the primary one, while the second and others could be *either* a primary or nonprimary subject.[8] This means that among these 126 books, there is no easy way to tell which books are "primarily" about "psychoanalysis and religion" unless the user goes through all of them. With the provided metadata, we do know that all books that have "psychoanalysis and religion" as the first subject heading are primarily about this topic, but a book that has this same heading as its second subject heading may or may not be primarily about this topic. There is no way to indicate which it is in the metadata, nor in the search interface.

As this example shows, the Library of Congress manual involves an attempt to acknowledge and make a distinction between primary and nonprimary subjects. However in practice the attempt is insufficient to be really useful since apart from the first entry, it is ambiguous whether subsequent entries are additional primary subjects or nonprimary subjects. Consequently, the search system and, further on, the users are not able to take full advantage of the care of a cataloger in deciding whether an additional subject is primary or not.

Other Information Retrieval Systems

The negative effect of current subject indexing without weighting on search outcomes has been identified by some researchers on particular information retrieval systems. In a study examining "the contribution of metadata to effective searching,"[9] Hawking and Zobel found that the available subject metadata are "of little value in ranking answers" to search queries.[10] Their explanation is that "it is difficult to indicate via metadata tagging the relative importance of a page to a particular topic,"[11] in addition to the problems in data quality and system implementation. The same problem



A. Subject: women; books; dresses; flowers; trees; . . . In: Victoria & Albert Museum (accessed Aug. 30, 2010), http://collections.vam.ac.uk/item/014962/oil-painting-the-day-dream

B. Tags: Japanese; moon; nights; walking; tree; . . . In: Brooklyn Museum (accessed Aug. 30, 2010), http://www.brooklynmuseum.org/opencollections/objects/121725/Aoi_Slope_Outside_Toranomon_Gate_No._113_from_One_Hundred_Famous_Views_of_Edo

C. Tags: Japanese; birds; silk; waterfall; tree; . . . In: Steve: The Museum Social Tagging Project (accessed Aug. 30, 2010), http://tagger.steve.museum/steve/object/15?offset=2

Figure 1. Example Images with "tree" as a Subject Item

of multiple tags without weights is described:

> In the kinds of queries we have studied, there is typically one page (or at most a small number) that is particularly valuable. There are many other pages which could be said to be relevant to the query—and thus merit a metadata match—but they are not nearly so useful for a typical searcher. Under the assumption that metadata is needed for search, all of these pages should have the relevant metadata tag, but this makes the particular page harder to find.[12]

A similar problem is reported in a recent study by Lykke and Eslau. In comparing searching by controlled subject metadata, searching based on automatic indexing, and searching based on automatic indexing expanded with a corporate thesaurus in an enterprise electronic document management system, the authors found that the metadata searches produced the lowest precision among the three strategies. The problem of indiscriminate metadata indexing is "remarkable" to the authors compared with the automatic indexing systems, because

> human indexers should be better at weighting the significance of subjects, and be more able to distinguish between important and peripheral compared with computers that base significance on term frequency.[13]

Indeed, while various weighting algorithms have been used in automatic indexing systems to approximate the distinguishing function, there is simply no such mechanism built in human subject

metadata indexing even though human indexers are able to do the job much better than computers.

Weighting: Yesterday, Today, and Future

Precedent Weighting Practices

Written more than thirty years ago, the final report of the Subject Access Project describes how the project researchers applied weights to the newly added subject terms extracted from tables of contents and back-of-the-book indexes. The criterion used in that project was that terms and phrases with a "ten-page range or larger" were treated as "major" ones.[14]

A similar mechanism was adopted in the ERIC database beginning in the 1960s, with indexes distinguishing "major" and "minor" descriptors as the result of indexing. While some search systems allowed differentiation of major and minor descriptors in formulating searches, others simply included the distinction (with an asterisk) when displaying a record. Unfortunately, this distinguishing mechanism is no longer included in the later ERIC indexing data.

A system using weighted indexing and searching and still running today is the MEDLINE/PubMed interface. A qualifier [majr] can be used with a Medical Subject Headings (MeSH) term in a query to "search a MeSH heading which is a major topic of an article (e.g., thromboembolism[majr])."[15] In the search result page, each major MeSH topic term is denoted by an asterisk at the end.

Weighting Concept and the Purpose of Indexing

The weighting concept is connected with the fundamental purpose of indexing. The idea of weighting in subject indexing has been discussed in the research area of subject analysis for some time. Weighting gives indexing an increased granularity and can be a device to counteract the effect of indexing specificity and exhaustivity on precision and recall, as pointed out by Foskett:

> Whereas specificity is a device to increase relevance at the cost of recall, exhaustivity works in the opposite direction, by increasing recall, but at the expense of relevance. A device which we may use to counteract this effect to some extent is weighting. In this, we try to show the significance of any particular specification by giving it a weight on a pre-established scale. For example, if we had a book on pets which dealt largely with dogs, we might give PETS a weight of 10/10, and DOGS, a weight of 8/10 or less.[16]

Anderson also includes weighting as a part of indexing in the *Guidelines for Indexes and Related Information Retrieval Devices* (NISO TR021997):

> One function of an index is to discriminate between major and minor treatments of particular topics or manifestations of particular features.[17]

He also notes that a weighting scheme is "especially useful in high-exhaustivity indexing"[18] when both peripheral and primary topics are indicated. Similarly, Fidel lists "weights" as one of the issues that should be addressed in an indexing policy.[19]

Metadata indexing without weighting is related to the simplified dichotomous assumption in subject indexing—primarily about/of and not primarily about/of, which further leads to the dichotomous retrieval result—retrieved and not retrieved. Weighting as a mechanism to break this dichotomy is noted by Anderson in NISO TR021997.[20] In addition, researchers have noticed the limitations of this dichotomous indexing. In an opinion piece, Markey emphasizes the urgency to "replace Boolean-based catalogs with post-Boolean probabilistic retrieval methods,"[21] especially given the challenges library systems are faced with today. It is the time to change the Boolean, i.e., dichotomous, practice of subject indexing and cataloging, no matter whether it is produced by professional librarians, by user tagging, or by an automatic mechanism.

Indeed, as declared by Svenonius, "While the purpose of an index is to point, the pointing cannot be done indiscriminately."[22]

Needed Refinements in Subject Indexing

The fact that weighted indexing has become more prominently needed over the past decade may be related to the shift in the continuum from subject indexing as representation/surrogate to subject indexing as access points, which is consistent with the shift from a small number of subject terms to more subject terms. This might explain why the weighting practice is applied in the above mentioned MEDLINE/PubMed system. With web-based systems, social tagging technology, federated searching, and the growing number of collections producing more subject terms, to distinguish between them has become a prominent problem.

In reviewing information users and use from the 1920s to the present, Miksa points out the trend to "more granular access to informational objects" "by viewing documents as having many diverse subjects rather than one or two 'main' subjects," no matter what the social and technical environment has been.[23] In recognizing this theme in the future development of information organization and retrieval systems, we argue that the subject indexing mechanism

should provide sufficient granularity to allow more granular access to information, as demonstrated in the examples in the previous section.

Potential Challenges

While arguing for the potential value of weights associated with subject terms, it is also important to acknowledge potential challenges posed by this approach.

Human Judgment

Treating assigned terms equally might seem to avoid the additional human judgment and the subjectivity of the weight levels because different catalogers may give different weight to a subject heading. We argue that assigning subject headings is itself unavoidably subjective. We are already using professional indexers and subject catalogers to create value-added metadata in the form of subject terms. Assigning weights would be a further enhancement.

On the other hand, adding a weighting mechanism into metadata schemas is independent of the issue of human indexing. No matter who will do the subject indexing or tagging, either professional librarians or users or possibly computers, there is a need for weight information in the metadata records.

The Weighting Scale

In terms of the specific mechanism of representing the weight rating, we can benefit from research on weighting of index terms and on the relevance of search results. For example, the three categories of relevant, partially relevant, and nonrelevant in information retrieval are similar to the major, minor, and nonpresent subject indexing method in the examples above. Borlund notes several retrieval studies proposing more than three categories or using continuous scales instead of category rating.[24] Subject indexing involves a similar judgment of relevance when deciding whether to include a subject term. More sophisticated scales certainly enable more useful ranking of results, but the cost of obtaining such information may rise.

After the mechanism of incorporating weights into subject indexing/cataloging is developed, guidelines should be provided for indexing practice to produce consistent and good quality.

Weights in Both Indexing and Retrieval System

Adding weights to subject indexing/cataloging needs to be considered and applied in three parts: (1) extending metadata schemas by encoding weights in subject elements; (2) subject indexing/cataloging with weight information; and (3) retrieval systems that exploit the weighting information in subject metadata elements. The mechanism will not work effectively in the absence of any one of them.

Conclusion

This paper advocates for adding a weighting mechanism to subject indexing and tagging, to enable search algorithms to be more discriminating and browsing better oriented, and thus to make it possible to provide more granular access to information. Such a weighting mechanism needs to be considered and applied not only by indexers, catalogers, and taggers, but also needs to be incorporated into system functionality.

As social tagging is brought into today's digital library collections and online library catalogs, as collections grow and are aggregated, and the opportunity arises for adding more metadata from a variety of different sources, including end user tagging and machine generated metadata, such weighting becomes more important than ever if we are to make productive use of metadata richness and still see the wood for the trees.

References

1. "Dublin Core Metadata Element Set, Version 1.1," http://dublincore.org/documents/dces/ (accessed Nov. 20, 2010).

2. Library of Congress, *Subject Headings Manual* (Washington, D.C.: Library of Congress, 2008).

3. Ibid.

4. Elaine Svenonius, "Access to Nonbook Materials: The Limits of Subject Indexing for Visual and Aural Languages," *Journal of the American Society for Information Science*, 45, no. 8 (1994): 600–606.

5. "VRA Core 4.0 Element Description," http://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf (accessed Mar. 31, 2011).

6. Richard J. Urban, Michael B. Twidale, and Piotr Adamczyk, "Designing and Developing a Collections Dashboard," In J. Trant and D. Bearman (eds). *Museums and the Web 2010: Proceedings*, ed. J. Trant and D. Bearman (Toronto: Archives & Museum Informatics, 2010). http://www.archimuse.com/mw2010/papers/urban/urban.html (accessed Apr. 5, 2011).

7. "VuFind at the University of Illinois," http://vufind.carli.illinois.edu (accessed Nov. 20, 2010).

8. Library of Congress, *Subject Headings Manual*.

9. David Hawking and Justin Zobel, "Does Topic Metadata Help with Web Search?" *Journal of the American Society for Information Science & Technology* 58, no. 5 (2007): 613–28.

10. Ibid.

11. Ibid.

12. Ibid, 625.

13. Marianne Lykke and Anna G. Eslau, "Using Thesauri in Enterprise Settings: Indexing or Query Expansion?" in *The Janus faced Scholar. A Festschrift in Honour of Peter Ingwersen,* ed. Birger Larsen et al. (Copenhagen: Royal School of Library & Information Science, 2010): 87–97.

14. Subject Access Project, *Books Are for Use: Final Report of the Subject Access Project to the Council on Library Resources* (Syracuse, N.Y.: Syracuse Univ., 1978).

15. "PubMed," http://www.nlm.nih

.gov/bsd/disted/pubmedtutorial/020_760.html (accessed Nov. 20, 2010).

16. A. C. Foskett, *The Subject Approach to Information*, 5th ed. (London: Library Association Publishing, 1996): 24.

17. James D. Anderson, *Guidelines for Indexes and Related Information Retrieval Devices*. NISO-TR02–1997, http://www.niso.org/publications/tr/tr02.pdf (accessed Nov. 20, 2010): 25.

18. Ibid.

19. Raya Fidel, "User-Centered Indexing," *Journal of the American Society for Information Science* 45, no. 8 (1994): 572–75.

20. Anderson, *Guidelines for Indexes and Related Information Retrieval Devices*, 20.

21. Karen Markey, "The Online Library Catalog: Paradise Lost and Paradise Regained?" *D-Lib Magazine* 13, no. 1/2 (2007).

22. Svenonius, "Access to Nonbook Materials," 601.

23. Francis Miksa, "Information Organization and the Mysterious Information User," *Libraries & the Cultural Record* 44, no. 3 (2009): 343–70.

24. Pia Borlund, "The Concept of Relevance in IR," *Journal of the American Society for Information Science & Technology* 54, no. 10 (2003): 913–25.