This is the accepted version of the following article: Lu, K., & Mao, J. (2015). An automatic approach to weighted subject indexing – An empirical study in the biomedical domain. *Journal of the Association for Information Science and Technology, 66*(9), 1776-1784. DOI: 10.1002/asi.23290 which has been published in final form at: http://onlinelibrary.wiley.com.ezproxy.lib.ou.edu/doi/10.1002/asi.23290/full

An automatic approach to weighted subject indexing – An empirical study in the biomedical domain

Kun Lu

School of Library and Information Studies, University of Oklahoma, 401 West Brooks, Norman, Oklahoma, USA, 73019

Email: kunlu@ou.edu

Jin Mao

Center for the Studies of Information Resource, Wuhan University, Luojiashan, Wuchang, Wuhan, Hubei, China, 430072

Email: danveno@163.com

Abstract

Subject indexing is an intellectually intensive process that has many inherent uncertainties. Existing manual subject indexing systems generally produce binary outcomes for whether or not to assign an indexing term. This does not sufficiently reflect the extent to which the indexing terms are associated with the documents. On the other hand, the idea of probabilistic or weighted indexing was proposed a long time ago and has seen success in capturing uncertainties in the automatic indexing process. One hurdle to overcome in implementing weighted indexing in manual subject indexing systems is the practical burden that could be added to the already intensive indexing process. This study proposes a method to infer automatically the associations between subject terms and documents through text mining. By uncovering the connections between MeSH descriptors and document text, we are able to derive the weights of MeSH descriptors manually assigned to documents. Our initial results suggest that the inference method is feasible and promising. The study has practical implications for improving subject indexing practice and providing better support for information retrieval.

Keywords: Subject indexing; Automatic subject term weighting; MeSH descriptors; Probabilistic model

Introduction

A large volume of information is embodied in natural language text. It has been well acknowledged that natural language has a loose structure that allows great variability. Different words can be used to refer to the same concept, and the same word can refer to different concepts depending on the context. This may lead to problematic information retrieval (IR) when user queries are matched with the terms from documents. Subject indexing is designed primarily to help control the variability in natural language. By normalizing both users' and authors' vocabularies via a controlled vocabulary, such as a thesaurus, it is expected a concept level match will be achieved to solve the vocabulary problem of IR (Furnas, Landauer, Gomez, & Dumais, 1987). MeSH is the primary thesaurus used to describe the content of the biomedical literature. Many efforts have been made to assign these carefully designed descriptors to the content in the hope of better information organization and retrieval. The assignment of an MeSH term to a document is a binary decision (i.e. assign or not assign) made by professionals based on their interpretation of the content, use of the thesaurus, and understanding of the anticipated users. Although MeSH terms have shown effectiveness in many IR applications (Shin & Han, 2004; Meij, Trieschnigg, de Rijke, & Kraaij, 2010; Jalali & Borujerdi, 2011), the current binary model of description using MeSH terms is inadequately reflects the inherent uncertainties in the subject indexing process (Mai, 2001). It has been noted that a piece of work can be related to multiple concepts and each concept can have variable importance depending on whether it is a major or minor point (Kent, Lancour, & Daily, 1978). However, existing subject indexing systems are not capable of reflecting the importance of subject terms. One way is to distinguish the primary and secondary subject descriptors, such as to assign asterisks to the MeSH descriptors that reflect the major points of the article by NLM (National Library of Medicine)

(http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_030.html); to distinguish primary and secondary subjects through an indicator in the MARC records in United Nations catalog (http://www.un.org/depts/dhl/unbisref_manual/indexpolicy/650.htm#weight); and similar effort by Library of Congress to specify the levels of subject as primary, secondary, and others (http://www.loc.gov/marc/bibliographic/bd650.html). The use of this weighting mechanism is still limited when compared with the probabilistic models that have been very successful in automatic indexing and free-text searching. In fact, the idea of applying probabilistic indexing to manual subject indexing was introduced in early experiments by Kent et al. (1978) where a guideline for manually assigning probabilistic weights was outlined. A more recent study by Zhang et al. (2011) further advocated the importance of weighting mechanisms to subject indexing. This paper is based on the findings from previous literature that a probabilistic model is more advantageous for subject indexing and will better capture the inherent uncertainties in the indexing process. However, it is also noted that the burden of manually distinguishing the extent to which the subject terms are associated with documents can be prohibitive. The purpose of this study is to propose an automatic approach to infer the associations between the manually assigned MeSH descriptors and documents based on text mining algorithms. These associations are implicitly created by indexers when they assign index terms to documents. The study aims to uncover the associations for weighted subject indexing. The method introduced in the study has practical implications for implementing weighted subject indexing systems which will provide better support for information organization and retrieval. This study extends our previous work (Lu & Mao, 2013) by further refining and validating the proposed method. More specifically, this study poses the following research questions:

1. Can probabilistic inference methods be used to derive the weights of subject terms in the metadata automatically?

2. How accurate are the results of the automatically inferred weights of manually assigned subject terms when compared to human indexer judgments?

The study provides a concrete implementation to infer the weights of subject terms--MeSH descriptors in this case--based on text mining algorithms and validates the results with the manually assigned major subject headings. It should be noted that although the current study examines the biomedical domain with MeSH as the indexing language, the same idea could also apply to other domains and indexing languages.

Related Work

Subject indexing is at the heart of information organization (Hjørland, 2008). The objective of subject indexing is to construct a surrogate of a document by providing a shorter description of its content. The indexing process generally includes the conceptual analysis of the item and the decision of index term assignment (Taylor, 2008). In this section, a brief review of the literature on subject indexing theories, binary indexing versus weighted indexing, and automatic subject indexing is outlined.

Subject indexing theories

This section briefly reviews the theories of indexing with a focus on probabilistic indexing theories. Indexing theory has been considered the central theoretical construct in information science (Travis & Fidel, 1982). Hjørland (2011) stated that a neglect of theories of indexing may be fatal for the field. Maron and Kuhns (1960) proposed the theory of probabilistic indexing that introduced probabilistic models to the indexing process. Cooper (1978) considered indexing as a

decision making process and developed the utility-theoretical indexing which suggests that the future use of index terms should be viewed as the criteria for term selection. Cooper and Maron (1978) explained the connection between probabilistic indexing and utility-theoretic indexing. Salton (1975) summarized the principles of automatic indexing and showed a preference towards those indexing terms with medium document frequencies. To determine the subjects of a document is an intellectually intensive task. Based on the analysis of a number of alternative theories, Hjørland (1992) concluded that philosophical epistemology is the foundation for the determination of the subjects of documents. Later, Hjørland (2005) further advocated the domain analysis approach for research on indexing and retrieving. More recently, Kleineberg (2013) argued that three dimensions (i.e. ontological, epistemological and methodological) need to be considered to represent the context-dependent nature of knowledge.

Subject indexing is closely related to the theory of aboutness as "aboutness" and "subject" can be considered synonyms in information science (Hjørland, 2001). Maron (1977) considered aboutness to be behavior correlated: "what a document is about is just the index terms that would be used to ask for that item" (p.40). It is interesting to note that the operational definition of aboutness put forward by Maron is probabilistic or weighted by nature (Maron, 1977). However, he did not comment on how to derive the weights, by human indexers or machines. Hutchins (1977) suggested a user-oriented 'aboutness' where a distinction between what has been known to users ('theme') and what is new ('rheme') should be made. Blair (1990) considered language and meaning as the fundamentals of indexing and argued that the theory of indexing should be based on the theory of language and meaning. Hjørland (2011) argued that theories of indexing should be viewed pragmatically, that is to index based on the goals such as "request oriented indexing". In reviewing the subject indexing theories, it has been found that most research has focused on the

issue of how to assign indexing terms to a document rather than how to distinguish the extent to which the indexing terms are associated with a document.

Binary indexing versus weighted indexing

The distinction between the relative dominance and subordination of different elements of a document has been discussed in indexing theories for a long time (Wilson, 1968). Cooper and Maron (1978) discriminated binary indexing from weighted indexing, where binary indexing only involves either assigning the index term or not without any intermediate choices, and weighted indexing allows specifying numeric indicators of the strength of assignments. They showed a strong preference to weighted indexing approach which allows users to consider the most promising documents first. In Maron (1979), a further recommendation was made against the binary thinking of subject indexing. Kent et al. (1978) applied manually assigned weights in subject indexing and found improved retrieval performance. The indexing process of the Cranfield tests incorporated manually weighted indexing to indicate the relative importance of each concept within the document (Cleverdon, 1991). More recently, Zhang et al. (2011) explored the problems in the current subject indexing systems due to the binary indexing and advocated for weighting mechanisms to subject indexing. On the other hand, weighted indexing has been employed in automatic indexing long before that (Salton, 1975).

Weighted indexing is also related to the concepts of exhaustivity and specificity of indexing. Both indexing characteristics, such as exhaustivity and specificity, and weighting methods can influence the retrieval effectiveness (Wolfram & Zhang, 2008), while Foskett (1996) noted that weighting can be used to counteract the inverse relationship between exhaustivity and specificity to some extent.

Automatic subject indexing

Another related research area is automatic subject indexing (Tzeras & Hartmann, 1993; Plaunt & Norgard, 1998; Ruch, 2006; Medelyan & Witten, 2008; Willis & Losee, 2013). Automatic subject indexing should be distinguished from automatic indexing (Willis & Losee, 2013). In automatic indexing (Salton, Wong, & Yang, 1975; Salton, Wu, & Yu, 1981), terms from the original items are generally used to represent the items. This type of indexing does not rely on controlled vocabularies. On the other hand, in automatic subject indexing, controlled vocabularies are used to index the items. The general approach is to learn the patterns and associations from already indexed items and predict the subject terms for new items. However, the accuracy of automatic subject indexing still remains to be improved. It should be noted that the essential techniques used for automatic subject indexing may also be applicable to inferring weights for subject terms in most cases. However, the goal is fundamentally different. Automatic subject term weighting does not replace manual indexing. Instead, it enhances the manual indexing by inferring the probabilities of the manually assigned terms so as to distinguish to which extent the document is about the terms. Another related concept that is in-between automatic subject indexing and manual subject indexing is the machine-aided indexing (Kingbiel, 1973) where machine outputs are used to support human decisions in indexing.

In reviewing existing literature, it has been found that the idea of incorporating weights into subject indexing was proposed decades ago (Kent, et al., 1978). However, very few existing subject indexing systems apply it. One possible reason is that the assignment of weights would add extra effort to the already labor-intensive subject indexing process. On the other hand, the performance of automatic subject term assignment is yet to be satisfactory. In this study, we propose a method that automatically infers the probabilities/weights of subject terms after the manual indexing

process. It should be noted that this study is not attempting to develop a new automatic subject indexing algorithm. Instead, the method just adds weights to the manually assigned index terms by exploring the inter-document and intra-document associations. These associations are implicitly created when professional indexers assign subject terms to documents. The proposed method intends to uncover the associations and develop the weighted subject indexing system.

Proposed Method

This section introduces a novel approach to automatically estimate weights for the manually assigned subject terms. The method is based on the Mutual Information theorem (Manning, Raghavan, & Schutze, 2008). Mutual Information measures the mutual dependence of two random variables. In this paper, the mutual dependence can be interpreted as the associations between subject terms and documents. Therefore, the mutual information between the document and the subject terms indicates the extent to which the document is about the subject terms. Frequency-based weighting, TF-IDF, was employed to distinguish the importance of the terms and subject headings (Salton, Wu, & Yu, 1981). It should be noted that there are other possible approaches to estimate the associations between documents and subject terms, such as Dice coefficient, Jaccard coefficient, or Cross Entropy (Trieschnigg, Meij, De Rijke, & Kraaij, 2008). This method only represents a first attempt to implement the idea of weighted subject indexing.

Inference process

The essential idea of this method is to infer the weights from the associations between subject terms and document text. As in the language modeling approach (Ponte & Croft, 1998), a document is viewed as a probability distribution of terms, denoted as θ_d. To quantify the associations between documents and subject headings, we calculate the weighted mutual

information between a document d and the assigned subject heading h. The formula is represented as:

$$\mathrm{I}(\theta_{\mathrm{d}};\mathrm{h}) = \sum_{t\in\theta_d} w(t,h)p(t,h)\log\frac{p(t,h)}{p(t)p(h)} \quad (1)$$

where t represents a term in the document, h is a subject heading associated with the document, and $w(t,h)$ is the weight of the pair $<t,h>$. TF-IDF weighting is calculated as follows:

$$\mathrm{w}(\mathrm{t},\mathrm{h}) = \mathrm{w}(\mathrm{t}) * \mathrm{w}(\mathrm{h}) = (\mathrm{tf}+0.5) * log\frac{N+0.5}{df_t+0.5} * \frac{N+0.5}{df_h+0.5} \quad (2)$$

where tf is the term frequency of the term t in the document, N is the total number of documents in the collection, df_t is the document frequency of the term t (i.e. number of documents that contain the term t), and df_h is the document frequency of the subject term h (i.e. number of documents that contain the subject term h). The logarithm function of the MeSH IDF component is intentionally removed to place more emphasis on the IDF component of the MeSH subject headings. With respect to $p(t,h)$, $p(t)$ and $p(h)$, Maximum Likelihood Estimation (MLE) is used to obtain the probabilities. If the document frequency of an object ι in the corpus is $\#(\iota)$, the probability can be calculated as:

$$\mathrm{p}(\iota) = \frac{\#(\iota)}{N} \quad (3)$$

where N is the total number of documents in the collection. Finally, we obtain the ultimate weight of subject heading h in document d by normalizing the weighted mutual information of all subject headings in the document.

$$\mathrm{I}(\theta_{\mathrm{d}};\mathrm{h}) = \frac{\mathrm{I}(\theta_{\mathrm{d}};\mathrm{h})}{\sum_{h'\epsilon\mathrm{d}}\mathrm{I}(\theta_{\mathrm{d}};h')} \quad (4)$$

The results computed from equation (4) are then used as the weights of subject terms in a document to indicate to which extent the subject terms are associated with the document.

Qualifiers

In the indexing process, indexers may assign one or more qualifiers to a main heading to further specify which aspects of the concept the document is about (http://www.nlm.nih.gov/mesh/indman/chapter_19.html). This allows users to narrow down their searches to a specific aspect of the concept if they are not interested in all aspects of it. When inferring the weights of the MeSH subject headings, a distinction is made among different qualifiers of the same main heading as not all the aspects should have the same weight. Therefore, if a document is indexed with "Electric Countershock/IS/MT", different weights will be inferred for "Electric Countershock/IS" and "Electric Countershock/MT" respectively.

Record types

It is also noted that there are different MeSH record types (http://www.nlm.nih.gov/mesh/intro_record_types.html) and not all record types are related to the subject of documents. The five record types of MeSH terms are: Descriptors (describe the subject of an indexed item), Publication Types (describe the genre of an indexed item), Geographics (describe the geographic characteristics), Qualifiers (describe a particular aspect of a subject), and Supplementary Concept Records (index chemicals, drugs, and other concepts such as rare diseases). The five record types can be broadly categorized into subject-related types (Descriptors, Qualifiers, and Supplementary Concept) and non-subject-related types (Publication Types, and Geographics). In this study, only the subject-related MeSH subject headings are included in the tests as the focus is on subject indexing. The non-subject-related headings are removed during the pre-process.

Results

The proposed method was applied to the Ohsumed collection as a pilot test. Ohsumed is a clinically-oriented Medline subset with 348,566 documents over a five-year period (1987-1991) (http://ir.ohsu.edu/ohsumed/ohsumed.html). Each document consists of seven metadata fields: title, MeSH, author, publication type, abstract, source, and record identifier. The study uses three fields: title, MeSH, and abstract. Out of the 348,566 documents, 23 documents have empty MeSH fields. In total, there are 348,543 documents with manually assigned MeSH descriptors. The purpose of the test is to use the method proposed earlier to automatically derive the weights for the MeSH descriptors. A standard list of English stopwords and Porter stemming were applied to the terms in titles and abstracts during the pre-process. MeSH descriptors with different qualifiers were separated and non-subject-related MeSH descriptors were excluded as mentioned earlier. Some descriptive statistics of the Ohsumed collection are provided in Table 1. After tokenization of the field contents, 58,431,800 tokens, or individual words, were identified. The number of unique tokens, or distinct words, was 269,611. The average document length was 167.63 words (after removing stopwords).

Table 1. Descriptive statistics of the Ohsumed collection.

No. of documents	No. of tokens	No. of unique tokens	Avg. doc. length (in words)
348,566	58,431,800	269,611	167.63

Table 2 lists the descriptive statistics of MeSH descriptors in the collection. Each document in our collection was assigned 12.02 MeSH descriptors on average with a standard deviation of 5.22. The

range is from 0 to 49 MeSH descriptors per document. After excluding non-subject-related MeSH descriptors, the average number of subject-related MeSH descriptors per document is 9.91 with a standard deviation of 4.54. The range of the number of subject-related MeSH descriptors per document is from 0 to 47. Out of the subject-related MeSH descriptors, there are on average 2.95 major MeSH subject headings and 6.96 non-major MeSH subject headings per document. The major subject headings were identified by professional indexers as the major points of the documents.

Table 2. Descriptive statistics of MeSH descriptors in Ohsumed collection

	Avg. # MeSH per doc	Std. # MeSH per doc	Min # MeSH per doc	Max # MeSH per doc
All MeSH	12.02	5.22	0	49
Subject-related MeSH	9.91	4.54	0	47
Subject-related major MeSH	2.95	1.37	0	13
Subject-related non-major MeSH	6.96	4.13	0	41

We applied automatic indexing to the titles and abstracts of the documents and then computed the mutual information between the MeSH descriptors and the documents as proposed in the previous section. In this way, we automatically derived the weights for the already assigned MeSH

descriptors in the collection. To provide an example of our results, document ID 89000261, titled "Tumor necrosis factor in middle ear effusions", was assigned seven MeSH descriptors: "Child", "Child, Preschool", "Female", "Human", "Male", "Otitis Media/*ME", and "Tumor Necrosis Factor/*AN". Among them, "Otitis Media/ME" and "Tumor Necrosis Factor/*AN" were annotated as major MeSH descriptors by NLM indexers and the rest as non-major descriptors. "Female" and "Male" are also non-subject-related terms, and thus were excluded from the inference process. With the method proposed in the study, we are able to assign weights to the five subject-related MeSH descriptors automatically. Figure 1 provides the weights of the subject headings that are assigned to the document.

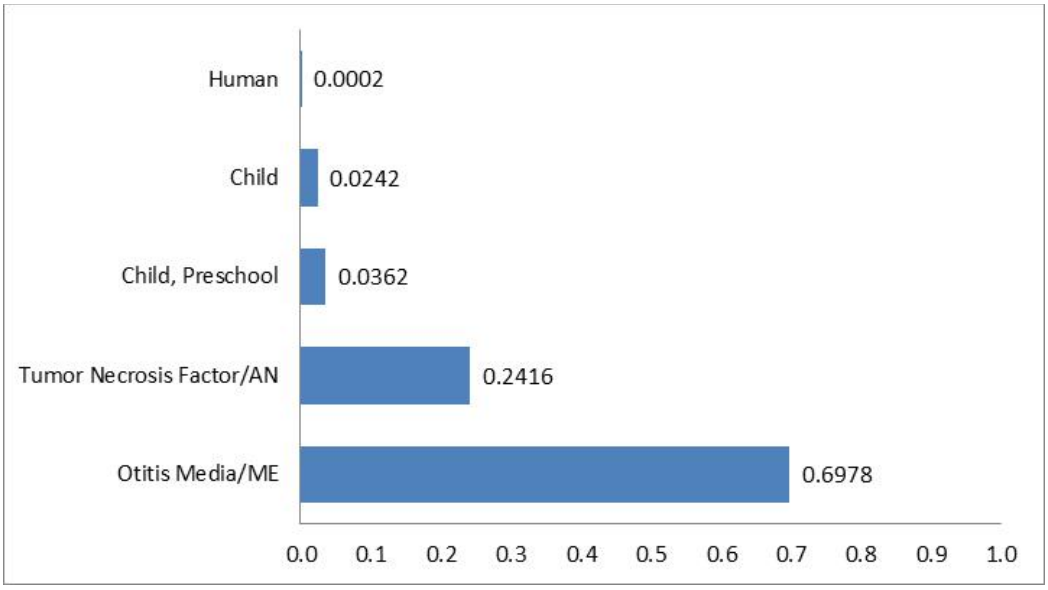

Figure 1. An example of weighted MeSH for the document titled "Tumor necrosis factor in middle ear effusions".

We can tell from Figure 1 that the article is mostly associated with the two major subject headings "Otitis Media/ME" (69.78%) and "Tumor Necrosis Factor/AN" (24.16%). The three non-major subject headings received much lower weights: "Child, Preschool" (3.62%), "Child" (2.42%), and "Human" (0.02%). A close examination of the associations between the subject headings and document text reveals that the two major subject headings "Otitis Media/ME" and "Tumor Necrosis Factor/AN" have higher associations with highly weighted terms in the document. For example, "Tumor Necrosis Factor/AN" has the highest association with the term "tnf" (an acronym for Tumor Necrosis Factor) that occurs 4 times in the document and has a relatively low document frequency (occurs only in 1187 documents in the collection). Similarly, "Otitis Media/ME" is highly associated with the terms "ear" and "effusion" that both occur 5 times in the document (document frequencies 2125 and 1480 respectively). On the other hand, the non-major subject headings "Human", "Child", and "Child, Preschool" have generally lower associations with the terms in the document and are generally associated with less important terms (i.e. terms with lower term frequency and higher document frequency.). Therefore, the proposed method assigns greater weights to major subject headings because they have higher associations with highly weighted terms in the document (i.e. terms with higher term frequency and lower document frequency as is defined in TF-IDF).

As this example shows, the automatic weighting method can distinguish major terms from non-major terms, and the weights indicate the subject relatedness of the MeSH descriptors to the document. With the weighted MeSH descriptors, we can not only tell what the document is about but also what the document is mostly about. The weighted subject indexing has more flexibility in capturing the uncertainties in the subject indexing process and provides better support for applications such as information retrieval and text mining. Retrieval algorithms may be developed

that take into account the weighted MeSH descriptors instead of treating them equally as in previous systems. In addition, the weighted MeSH descriptors can also be presented in the front-end interface to inform end-users the strength of the associations.

Validation

To further verify the method, we compared the automatically derived weights of the major subject headings (e.g. "Allied Health Personnel/*") to those of the non-major subject headings. There are in total 1,027,427 major subject headings and 2,426,870 non-major subject headings after splitting subject headings with multiple qualifiers and excluding non-subject-related ones. Table 3 provides descriptive statistics of the weights of the major subject headings and the non-major subject headings.

Table 3. Descriptive statistics of the weights of MeSH descriptors.

	N	Mean	Std.	Median
Major MeSH	1,027,427	0.173	0.140	0.133
Non-major MeSH	2,426,870	0.070	0.085	0.044

The average weight of the major subject headings is 0.173 in comparison with 0.070 of the non-major ones. As the distribution of the data is skewed, a nonparametric statistical test was employed. The difference is statistically significant with a Mann-Whitney's U-test ($p<0.01$). Provided that the major subject headings have averagely 147% higher weights than the non-major ones, we conclude that our method assigned significantly higher weights to the manually identified major subject headings than the non-major ones.

In addition to the weights, we also investigated the ranks of the major subject headings versus those of the non-major subject headings according to the inferred weights. The following table lists the distribution of the major subject headings and the non-major subject headings in the top 10 ranked subject terms by their weights:

Table 4: Distribution of major headings and non-major headings in the top 10 subject terms (Counts are in cells).

Rank	1	2	3	4	5	6	7	8	9	10
Major	203816	188061	159358	128119	98877	75535	57263	42979	32414	23310
Non-major	144613	159273	183636	203398	213239	212753	205835	193639	176541	156352

A chi-square association test suggests the ranks of the major MeSH subject headings are significantly different from those of the non-major ones (d.f.=9, $p<0.01$). A further investigation of the residuals (Table 5) suggests that the major MeSH subject headings are more likely to appear in the top 4 positions, while the non-major subject headings are more likely to be ranked from 5 to 10 in the top 10 positions. These findings provide support that our method tends to rank the major subject headings higher than the non-major ones. The evidence further confirms the validity of the proposed method.

Table 5. Residuals of the chi-square association test from the data in Table 4.

Rank	1	2	3	4	5	6	7	8	9	10
Major	230	187	110	32	-34	-82	-117	-140	-152	-159
Non-major	-170	-138	-81	-24	25	61	86	104	113	118

Discussion

Subject indexing theory and practice have been around for a long time. Existing subject indexing systems employ binary decisions for subject term assignments, which cannot sufficiently capture the inherent uncertainties in the subject indexing process. A weighted subject indexing system will be more advantageous to reflect the uncertainties. However, it is impractical to add further burden to indexers and ask them to decide the weights. This study proposes an automatic approach to weighted subject indexing. As noted earlier, this study does not intend to replace the manual subject indexing process. Instead, our method relies on the implicit associations created by indexers when they assign subject terms and uncovers the associations by text mining algorithms. An empirical study in the biomedical domain suggests that our method can adequately distinguish the major MeSH descriptors from the non-major ones, and assign significantly higher weights to the major MeSH descriptors. The results confirm the validity of the method. The study is also in line with the continuous advocate for weighted subject indexing (Kent, et al, 1978; Zhang, et al., 2011).

There are a number of ways that the weighted subject indexing may contribute to information organization and retrieval:

First, subject terms have been widely used to improve information retrieval algorithmically, such as automatic query expansion (Lu, Kim, & Wilbur, 2009; Stokes, Li, Cavedon, & Zobel, 2009; Mu & Lu, 2010) and subject terms enhanced retrieval models (Meij et al., 2010). With the weighted subject indexing introduced in this study, the performance of the algorithms may be further improved since the weighted indexing allows more flexibility to capture the uncertainties.

Second, a richer front-end interface could be developed based on the weighted subject terms. Many current online databases that are equipped with subject terms only provide alphabetical lists of subject terms (e.g. PubMed). The alphabetical lists do not offer much insight into the importance of the subject terms. With the weighted subject indexing, a ranked list of subject terms could be presented to users in the front-end instead of the alphabetical list (Figure 2). Users will then be able to re-rank the retrieval results according their relatedness with the relevant subjects. It is hypothesized that the ranked lists provide richer information to users and may improve their retrieval effectiveness. Future study will further explore this issue.

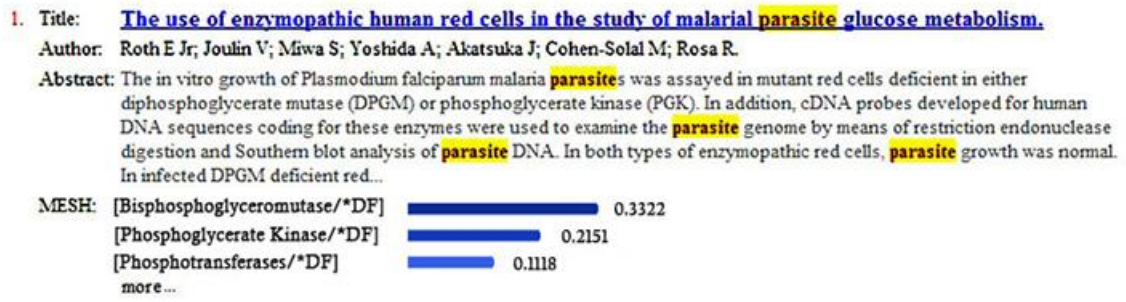

Figure 2. A front-end demo of weighted subject terms.

Besides the applications in information retrieval, the method proposed in this study can also be used to assist the manual indexing process. Our method can provide indexers with real time responses of the inferred weight of a subject term based on the inter-document and intra-document

associations in the collection. This information may be used to support their indexing decisions. Interestingly, the algorithm will also be able to incorporate the updates after they assign the subject terms and reflect the change in future inferences.

Weighted subject indexing can also benefit other analyses that are based on subject headings, such as co-word analysis (An & Wu, 2011).

Several issues need to be discussed to further understand the findings:

First, although automatic subject indexing and automatic subject term weighting may be based on similar techniques, the goals are very different. Automatic subject indexing attempts to learn from how indexers assign subject terms so that machines can conduct the subject indexing automatically. However, automatic subject term weighting does not intend to replace manual indexing as mentioned earlier. It helps to add weights to the already assigned subject terms. Subject indexing is an intellectually intensive process. It is difficult to fully automate the process before we can formally describe the process. Chung, Miksa and Hastings (2010) demonstrated that considering the indexing conceptions of human indexers is helpful for automatic subject indexing. Willis and Losee (2013) attempted to formalize the browsing aspects of the subject indexing process with a random walk model and improved our understanding of subject indexing. Although their model improves the accuracy of automatic subject indexing, the overall performance is still less than satisfactory. Further research that addresses the issue of formal description of the subject indexing process is still needed to reach the goal of effective automatic subject indexing.

Second, it should be noted that the weights inferred by the method are dynamic as new articles are added to the collection and new subject terms are assigned. This is important to reflect the development of our knowledge in understanding the subjects.

Third, other considerations such as user dynamics and domain features could be incorporated into the algorithm to trim the method for targeted groups (Hjørland, 2008). For example, user models could be built based on users' preferences to assign greater weight to the terms that are more important to the users. Similarly, domain specific terms can be emphasized in the term weighting when inferring the associations between subject descriptors and document text.

Last, subject metadata from other sources, such as user tagging, can also benefit from the weighting mechanism as is also pointed out in Zhang et al. (2011). However, further study will be needed to explore the applicability of the proposed method to other subject metadata sources.

Limitations

The limitations of the study should also be acknowledged:

First, the proposed method is based on the language modeling approach (Ponte & Croft, 1998) that suffers from the well acknowledged strong assumption of term independence: terms that appear in a document are independent of each other, which is obviously not true in the real world. Lavrenko (2009) argued that the model actually requires a much weaker assumption than term independence. Relieving the assumption of the model is still under discussion. Topic models (Blei, 2012) that are believed to have relieved the assumption are becoming increasingly popular. However, the assumptions of topic models are also questionable.

Second, the proposed method requires a sufficiently large sample to produce accurate results. As is the case in any other probabilistic model, an insufficient sample may lead to inferior results.

Third, the automatic subject term weighting method relies on the results of the manual subject indexing process. Therefore, the quality of the manual indexing process definitely has an impact

on the weighting outcomes. Factors such as inter-indexer consistency, experience of the indexers, and their domain knowledge may all affect the accuracy of the algorithm. The outcomes of the method introduced in this study represent the averaged patterns of different indexers that are involved in the subject indexing process.

Conclusion

The subject indexing process employs subject analysis and controlled vocabularies to describe a document. Most existing subject indexing systems only produce binary outcomes in deciding whether to assign an indexing term or not. This binary model does not adequately reflect the inherent uncertainties in the subject indexing process. In this study, we proposed a method that automatically derives weights for manually assigned subject terms through mining the implicit connections between subject headings and document text. When indexers assign MeSH terms to a document, they implicitly create connections between the MeSH terms and the document text. With a large sample size, these connections can be mined for patterns that help to evaluate the associations between MeSH terms and documents. The essential idea of our method is to uncover the connections and automatically assign weights to manually assigned subject terms. This method does not add further burden to indexers. Additionally, with new documents added to the collection, the method can also incorporate our dynamic understanding of the subject and adjust the weights accordingly. The study uses the results from manual indexing and derives weights for the subject terms. The initial results suggest the method can sufficiently distinguish major subject headings from non-major ones, which verifies the validity of the method. Theorists have long advocated for weighted subject indexing (Wilson, 1968; Cooper & Maron, 1978). The method proposed in the present study provides an avenue for weighted subject indexing systems. With the automatically inferred weights, new applications can be developed both in the front-end and back-end to better

support information organization and retrieval. The inferred weights may also be used in the subject indexing process to help indexers validate their judgments.

This study serves as a first step to advocate automatic approaches to weighted subject indexing. We believe that there are different ways to estimate the weights. What's more important in this study is to pave the way for an automatic subject term weighting system that helps to distinguish the extent to which the terms are associated with documents. We believe the weighted system can better serve the goals of information organization and retrieval. Future studies will develop new applications based on the findings from this study and validate the effectiveness of the weighted subject indexing system in different applications.

Acknowledgement

We thank Prof. Dietmar Wolfram and Prof. June Abbas for their comments on an earlier version of the manuscript. We also appreciate the constructive comments from the two anonymous reviewers.

Reference

An, X.Y., & Wu, Q.Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133-144.

Blair, D.C. (1990). Language and representation in information retrieval. Amsterdam: Elsevier.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Chung, E., Miksa, S., & Hastings, S.K. (2010). A framework of automatic subject terms assignment for text categorization: An indexing conception-based approach. *Journal of the American Society for Information Science and Technology, 61*(4), 688-699.

Cleverdon, C.W. (1991). The significance of the Cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 3-12)*, Chicago, Illinois, USA.

Cooper, W.S. (1978). Indexing documents by gedanken experimentation. *Journal of the American Society for Information Science, 29*(3), 107-119.

Cooper, W.S., & Maron, M.E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the ACM, 25*(1), 67-80.

Foskett, A.C. (1996). The subject approach to information. London: Library Association Publishing.

Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S.T. (1987). The vocabulary problem in human-system communication. *Communication of the ACM, 30*(11), 964-971.

Hjørland, B. (1992). The concept of 'subject' in information science. *Journal of Documentation, 48*(2), 172-200.

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content … and relevance. *Journal of the American Society for Information Science and Technology, 52*(9), 774-778.

Hjørland, B. (2005). Domain analysis: A socio-cognitive orientation for information science research. *Bulletin of the American Society for Information Science and Technology, 30*(3), 17-21.

Hjørland, B. (2008). What is knowledge organization? *Knowledge Organization, 2*(3), 86-101.

Hjørland, B. (2011). The importance of theories of knowledge: Indexing and information retrieval as an example. *Journal of the American Society for Information Science and Technology, 62*(1), 72-77.

Hutchins, W.J. (1977). The concept of 'aboutness' in subject indexing. *Aslib Proceedings, 30*(5), 172-181.

Jalali, V., & Borujerdi, M. (2011). Information retrieval with concept-based pseudo-relevance feedback in MEDLINE. *Knowledge and Information Systems, 29*(1), 237-248.

Kent, A., Lancour, H., & Daily, J.E. (1978). Probabilistic or weighted indexing. In *Encyclopedia of Library and Information Science: Volume 24 – Printers and Printing: Arabic Printing to Public Policy: Copyright, and Information Technology.* CRC Press.

Kleineberg, M. (2013). The blind men and the elephant: Towards an organization of epistemic contexts. *Knowledge Organization, 40*(5), 340-362.

Klingbiel, P.H. (1973). Machine-aided indexing of technical literature. *Information Storage and Retrieval, 9*(2), 79-84.

Lavrenko, V. (2009). A generative theory of relevance. Springer-Verlag: Berlin.

Lu, K., & Mao, J. (2013). Automatically infer subject terms and documents associations through text mining. In *Proceedings of the 76th annual conference of Association for Information Science and Technology (ASIST'2013)*. Montreal, Canada.

Lu, Z., Kim, W., & Wilbur, W. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval,* 12(1), 69-80.

Mai, J. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation, 57*(5), 591-622.

Manning, C.D., Raghavan, P., & Schutze, H. (2008). An introduction to information retrieval. Cambridge University: New York.

Maron, M.E., & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216-244.

Maron, M.E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science, 28*(1), 38- 43.

Maron, M.E. (1979). Depth of indexing. . *Journal of the American Society for Information Science, 30*(4), 224-228.

Medelyan, O., & Witten, I.H. (2008). Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology, 59*(7), 1026-1040.

Meij, E., Trieschnigg, D., de Rijke, M., & Kraaij, W. (2010). Conceptual language models for domain-specific retrieval. *Information Processing and Management, 46*(4), 448-469.

Mu, X., & Lu, K. (2010). Towards effective Genomic information retrieval: The impact of query complexity and expansion strategies. *Journal of Information Science, 36*(2), 194-208.

Plaunt, C., & Norgard, B.A. (1998). An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science, 49*(10), 888-902.

Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281), Melbourne, Australia.

Ruch, P. (2006). Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics, 22*(6), 658-664.

Salton, G. (1975). A theory of indexing. Philadelphia, Pennsylvania: Society for industrial and applied mathematics.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Salton, G., Wu, H., & Yu, C.T. (1981). The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science, 32*(3), 175-186.

Shin, K., & Han, S.Y. (2004). Improving information retrieval in MEDLINE by modulating MeSH term weights. Lecture Notes in Computer Science, *3136*, 388-394.

Stokes, N., Li, Y., Cavedon, L., & Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. Information Retrieval, 12(1), 17-50.

Taylor, A.G. (2008). The organization of information (3rd Ed.). Westport, Conn: Libraries Unlimited.

Travis, IL., & Fidel, R. (1982). Subject analysis. *Annual Review of Information Science and Technology, 17*, 123-157.

Trieschnigg, D., Meij, E., De Rijke, M., & Kraaij, W. (2008). Measuring concept relatedness using language models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 823-824),* Singapore.

Tzeras, K., & Hartmann, S. (1993). Automatic indexing based on Bayesian inference networks. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 22-35),* Pittsburgh, Pennsylvania, USA.

Willis, C., & Losee, R.M. (2013). A random walk on an ontology: Using thesaurus structure for automatic subject indexing. *Journal of the American Society for Information Science and Technology, 64*(7), 1330-1344.

Wilson, P. (1968). Two kinds of power. An essay on bibliographical control. Berkeley, CA: University of California Press.

Wolfram, D., & Zhang, J. (2008). The influence of indexing practices and weighting algorithms on document spaces. *Journal of the American Society for Information Science and Technology, 59*(1), 3-11.

Zhang, H., Smith, L.C., Twidale, M., & Gao, F.H. (2011). Seeing the wood for the trees: Enhancing metadata subject elements with weights. *Information Technology and Libraries, 30*(2), 75-80.