Automatically infer subject terms and documents associations through text mining

Kun Lu
School of Library and Information Studies
University of Oklahoma
401 West Brooks, Norman, OK, 73019
kunlu@ou.edu

Jin Mao
Wuhan University
Center for the Studies of Information Resource
Wuhan, China
danveno@163.com

ABSTRACT

Subject indexing is an intellectual intensive process that bears many inherent uncertainties. Existing subject index systems generally produce binary outcomes on whether assigning an indexing term or not, which does not sufficiently reflect to which extent the indexing terms are associated with documents. On the other hand, probabilistic models have seen great success in capturing the uncertainties in the automatic indexing process. One hurdle to achieving weighted indexing in manual subject indexing process is the practical burden that could be added to the already intensive indexing process. In this study, we propose a method to automatically infer the associations between subject terms and documents through text mining. By uncovering the connections between MeSH terms and document text, we are able to derive the weights of MeSH terms in documents. Our initial results suggest that the new method is feasible and promising. The study has practical implications for improving subject indexing practice.

Keywords

Subject indexing; weighted indexing; text mining

INTRODUCTION

Natural language has a loose structure that allows great variability. Many words can be used to refer to the same concept and the same word can refer to different concepts depending on the contexts. This may lead to problematic retrieval when user queries are matched to terms from documents. Controlled vocabularies are the major tools to help overcome the variability in natural language. By normalizing both users' and authors' vocabularies via controlled vocabulary thesaurus, it is expected to achieve a concept level matching and solve the vocabulary problem (Furnas et al., 1987). MeSH is the primary thesaurus for describing the content of biomedical literature. Tremendous efforts are invested to assign these carefully designed descriptors to the content in hope for better organization and retrieval. The assignment of MeSH terms is a binary decision by professionals based on their interpretation of the content and use of the thesaurus. While MeSH terms have shown great effectiveness in many IR applications (Shin & Han, 2004; Meij, et al., 2010; Jalali & Borujerdi, 2011), the current binary model of description using MeSH terms is insufficient in reflecting the inherent uncertainties in the subject indexing process (Mai, 2001). It has been noted that a piece of work can be related to multiple facets and each facet could have different importance depending on whether it is the major or minor point. However, the importance of the MeSH terms is not represented in this model. The only effort to this end is to assign an asterisk to the MeSH terms that reflect the major points of the article[1]. On the other hand, probabilistic models have seen great success in automatic indexing and free-text searching. Therefore, we hypothesize that a weighted indexing model would be more advantageous for manual subject headings and better capture the inherent uncertainties in subject indexing. However, it would be impractical to add further burden to indexers and ask them to make the judgment. The purpose of this study is to propose an automatic approach to infer the associations between MeSH terms and documents based on text mining algorithms. It should be noted that although we are using MeSH terms in this study, the same idea could also be applied to other manual indexing systems.

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

PROPOSED METHOD

In this section, we introduce a novel approach to estimate weights for the manually assigned subject headings. Our method is based on the mutual information theorem. In information theory, mutual information measures the mutual dependence of the two random variables. In our research, this mutual dependence can be interpreted as

[1] http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_030.html

content relatedness. Thus, the mutual information of the document as well as the subject heading indicates the extent to which the document is about the subject heading. It should be noted that this method only represents a first attempt to implement the idea.

As in the language modeling approach, we view a document 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:

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

where $w(t,h)$ is the weight of the pair <t,h>, *t* represents a term in the document and *h* is a subject heading associated with the document. We use TF-IDF weighting to calculate the weight:

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

where *N* is the total number of documents in the collection, df_t is the document frequency of term *t*, df_h is the document frequency of subject heading term *h*.

With respect to $p(t,h), p(t)$ and $p(h)$, maximum likelihood estimation can be applied. If the document frequency of the object ι in the corpus is $\#(\iota)$, the probability can be calculated as:

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

Finally, we obtain the ultimate weight for subject heading h in document d by normalizing all the obtained weighted mutual information for each document.

$$w(\theta_d;h) = \frac{I(\theta_d;h)}{\sum_{h \in d} I(\theta_d;h)} \quad (4)$$

.

RESULTS

We applied our method 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)[2]. Each document consists of seven fields: title, MeSH, author, publication type, abstract, source and record identifier. Out of 348,566 documents, 23 have empty MeSH field. In total, we have 348,543 documents with manually assigned MeSH terms. Our purpose is to use the method proposed above to automatically derive weights for the MeSH terms.

Avg. # MeSH per doc	Std. # MeSH per doc	Min # MeSH per doc	Max # MeSH per doc
10.6	4.34	1	33

Table 1. Basic descriptive statistics of MeSH terms in Ohsumed collection

Some descriptive statistics is provided in Table 1. Each document in our collection was assigned 10.6 MeSH terms on average with a standard deviation of 4.34. The range is from 1 to 33 MeSH terms per document. We applied automatic indexing to each field in the documents and then computed the mutual information between MeSH terms and documents as proposed in the previous section. In this way, we automatically derived weightings for the already assigned MeSH terms in the collection. To provide an example of our results, document ID 22, titled "Emergency department thoracotomy", was assigned five MeSH terms: "Human", "Thoracic injuries/SU", "Transportation of Patients/MT", "Wounds, Penetrating/MO" and "Emergency Service, Hospital". With the method proposed in the study, we are able to assign weights to them automatically. Figure 1 provides the weights of the headings.

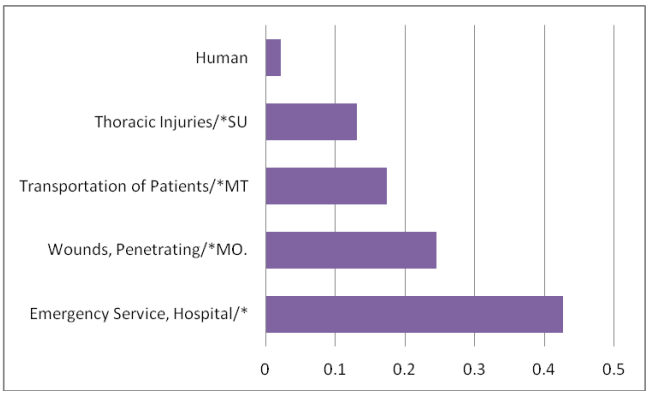

Figure 1. An example of weighted MeSH for document "Emergency department thoracotomy"

We can tell from Figure 1 that the article is mostly associated with the heading "Emergency Service, Hospital" (42.7%), following by "Wounds, Penetrating/MO" (24.6%), then "Transportation of Patients/MT" (17.4%) and etc. With the weighted MeSH terms, we have more options to represent the uncertainties in the subject indexing process. This could benefit the applications such as information retrieval and text mining. For example, we may develop retrieval algorithms that take into account the weighted MeSH terms instead of treating them equally as in previous systems. We can also present the MeSH weightings to end-users to inform them the strength of the associations.

To further verify our results, we compared the weights of the major headings assigned by NLM indexers (e.g. "Allied

[2] http://ir.ohsu.edu/ohsumed/ohsumed.html

Health Personnel/*") to those of the non-major headings. There are in total 1,070,533 major headings and 2,623,353 non-major headings in our collection. The average weight of major headings is 0.099 comparing to 0.092 of non-major headings. The difference is statistically significant with a two tailed t-test ($p<0.05$, df.=3,693,884). However, given that the major headings only have 7.6% higher weights than the non-major ones, we conclude that our method does not always weight the manually assigned major headings higher. One possible reason for this could be that MeSH also includes non-topical descriptors such as characteristics of the group being studied (e.g. the age group, human or other animal) and publication types (e.g. review, editorial). These descriptors may add noise to the text mining algorithm. Further exploration into this is needed in future study.

CONCLUSION

Subject indexing process employs subject analysis and controlled vocabulary to describe a document. Most existing subject index systems only produce binary outcomes on whether to assign indexing terms or not. This binary model does not adequately reflect the inherent uncertainties in subject indexing process. In this study, we proposed a method that automatically derives weightings 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 unnoticeably create connections between the MeSH terms and the document text. With a sufficient 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 for subject headings. This method does not add further burden to indexers. Additionally, with the new sample coming in, the method can also incorporate our dynamic understanding on the subject and adjust the weights accordingly. It should be noted that we are not aiming to replace the manual indexing process neither. The study in fact uses the results from manual indexing and derives weights for the subject terms. The initial results appear to be promising. And we are positive that there are better ways to estimate the weights. This study only serves as a pilot test. But what's more important is to pave the way for automatic subject term weighting system that helps to distinguish the extent to which the terms are associated with documents.

REFERENCES

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.

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

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

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.

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

.