Scientific Metadata Quality Enhancement for Scholarly Publications

Chun Guo
School of Library
and Information Science
Indiana University Bloomington
chunguo@indiana.edu

Jinsong Zhang
Dalian Maritime University
Dalian, China
zjs.dlmu@gmail.com

Xiaozhong Liu
School of Library
and Information Science
Indiana University Bloomington
liu237@indiana.edu

Abstract

Keyword metadata is very important to the access, retrieval, and management of scientific publications. However, author-assigned keywords are not always readily available in digital repositories. In this study, in order to enhance metadata quality, we explore different automatic methods to infer keywords from scholarly articles, including supervised topic modeling, language model, and mutual information. Evaluation results showed that the linear combination of mutual information and topic modeling with full text outperform other methods on MAP, while language model with abstract performed better than other methods on the measure of precision@10.

Keywords: keyword inference, topic modeling, language model, mutual information

Introduction

Keyword metadata is a very important access point for digital libraries. It provides a brief summary of the topics discussed in an academic publication. Although it is usually strongly recommended or required that an author provide keywords when they submit a paper for review, only a small number of publications have author-assigned keywords.

Take the distribution of publications with a certain number of keywords in the ACM digital library as an example (shown in Figure 1). For the 248,893 publications within the database, the number of keywords for a single publication ranges between 0-49. The average number of keywords per publication is 2.11. More than half of the publications (52.3%) do not have any author-assigned keywords, while 4.9% of them only have one or two keywords.

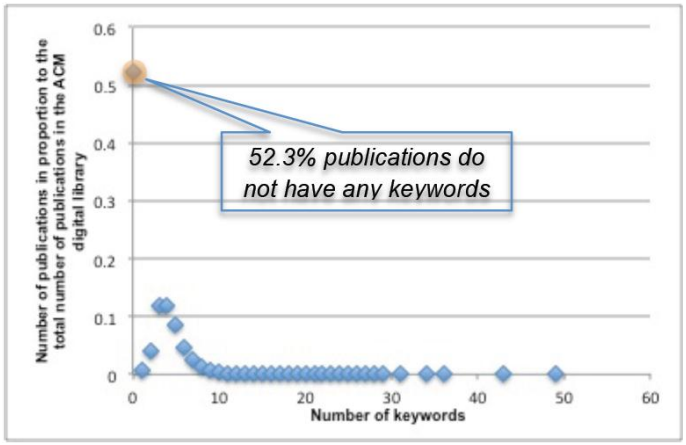

Figure 1. Distribution of publications with a certain number of keywords in the ACM digital library.

Guo, C., Zhang, J., & Liu, X. (2013). Scientific metadata quality enhancement for scholarly publications. *iConference 2013 Proceedings* (pp. 777-780). doi:10.9776/13382

It is hard to ensure the quality of keyword metadata, since little guidance is provided to the authors when they assign keywords to their submissions. Although some publishing agencies have domain experts manually assign keywords for publications, it is very laborious work and can hardly be applied to very large collections. In this study, we explore inferring keywords for scholarly publications utilizing different automatic approaches.

Literature Review

Previous studies used two methods to infer keywords for documents: keyword assignment and keyword extraction. Keyword assignment, a.k.a. text categorization (Dumais et al., 1998), assumes that all potential keywords come from a predefined controlled vocabulary. Then machine learning model is trained to classify publications into those categories, while each publication is assigned with one or more category labels. However, this approach ignores the dynamic nature of author keywords. For instance, new concepts emerges everyday, a static controlled vocabulary is immediately out of date the moment it is created. Meanwhile, controlled vocabularies are usually created by domain experts, which, in most cases, may not reflect the interests of authors and readers.

The other approach for keyword inference is keyword extraction. It is not restricted to a set of candidate keywords from a selected vocabulary. Instead, any phrase in a new document can be extracted as a keyword. Tomokiyo and Hurst (2003) used language model to extract keywords from newsgroups. Kea (Frank et al., 1999) uses TFIDF (term frequency $*$ inverse document frequency) and normalized word position as machine learning features to extract keywords from documents. While extraction-based methods generate a more diverse set of keywords, some of the keywords extracted are not reasonable from a human perspective (Witten, Paynter, Frank, Gutwin, & Nevill-Manning, 1999).

In this study, we use author assigned keywords as predefined labels for keyword inference, and each keyword is represented by a topic model or a language model. We used keywords that frequently appear within a domain, which makes them more author/user centric.

Method

Keyword Inference

LLDA. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a widely used method for topic modeling. Labeled LDA (Ramage, Hall, Nallapati & Manning, 2009) further constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user labels. Given a set of labeled documents, machine can learn the word-topic probability distribution, while each topic is represented by a label. As a result, unlike classical LDA, it is not necessary to interpret each topic or to set empirical topic number for a model. If documents used in the training process are well representative of the topic domain, then we can use the derived model to predict labels for other documents from the same domain.

Language model. Language Model is a topic dependent Information Retrieval Model. A document is treated as if it is generated by an unknown language model and we can estimate that model using statistical methods. If a document is relevant to a query, they are likely to be generated from similar language models. Based on this assumption, documents are ranked based on the likelihood of their background language model generating the given user query.

Vice versa, if we have a list of potential keywords for a topic domain, we can predict a publication's matching keywords by ranking all the potential keywords based on their likelihood of being generated by the background language model of the given publication.

Mutual information. Some publication venues require authors to provide category information for their paper during the submission process. For example, most ACM submissions require authors to pick categories from a controlled category list. Since both category words and keywords provide topical information, category words and keywords related to the same topic would highly likely co-occur in the same publication. Therefore, we hypothesize that combining mutual information with LLDA would further improve LLDA's performance. We calculate mutual information score (MI_score) for each category-keyword pairs appearing in the corpus. Then we derive a new ranking score using a linear combination of mutual information score and LLDA score (LLDA_score) based on the formula below:

$$new\ score = \alpha \cdot MI\_score + (1 - \alpha) \cdot LLDA\_score$$

α is the parameter used for linear combination and different α values were tested to achieve best performance.

Evaluation

The inferred keywords for each publication are evaluated against the original author assigned keywords. We use two indicators to measure keyword inference performance: mean average precision (MAP), and precision at 10.

We use greedy matching as the baseline method in the evaluation. We search each potential keyword from the full text of that article by using greedy matching. For example, if "music information retrieval" existed in the title, we wouldn't use the keyword "information retrieval". Matched keywords are ranked based on the position of their initial appearance in the article's full text.

Experiment

Data

We used 41,370 publications from 111 journals and 1,442 conference/workshop proceedings on computer science (mainly from the ACM digital library) for the experiment, for which full text was extracted from the PDF files. The selected papers were published between 1951 and 2011. From these we extracted 20,394 publications' text (accounting for 67.7% of all the sampled publications), including titles, abstracts, and full text. For the other publications, we used the title, abstract, and information from a metadata repository to represent the content of the paper.

LLDA Model Training

We sampled 20,394 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. For instance, if a paper has 6 author-provided keywords, our LLDA training would have assumed that this paper is a multinomial distribution over these 6 topics. During pre-processing we also clustered similar keywords if the edit distance between them were very small, e.g., "k-means" and "k means", or if two keywords shared the same stemmed root, e.g., "web searches" and "web searching".

Finally, we trained a LLDA model with 1,239 topic labels (keywords). These topic labels were used as potential keywords to be assigned to publications.

Experimental Results

Table 1
Experimental Results

	llda_abstract	llda_fulltext	lm_abstract	lm_fulltext	greedy matching	mi_llda_full_0.2	mi_llda_full_0.5	mi_llda_full_0.7
MAP	0.1877	0.2632	0.2372	0.2133	0.2195	**0.2714**	0.2632	0.2632
P@10	0.104	0.1453	**0.1784**	0.166	0.1573	0.1457	0.1453	0.1453

The best performing method on MAP is the combination of mutual information and LLDA with full text (mi_llda_full_0.2). The best performing method on precision@10 is language model with abstract. Another interesting finding is that LLDA performs better with full text, while language model performs better with abstract, on both MAP and precision@10. It indicates that all terms within a document are important for LLDA inference, whereas language model favors terms that provide strong topical information. Apparently, paper full text contains more words than abstract, which helps the LLDA model to make more accurate inference. But it also brings more noisy terms, which hampers the performance of

language model. One the other hand, abstract, as a succinct summary of the publication, has a high concentration of topic terms, and works better for language model based keyword inference.

Conclusion

Keyword metadata is an important access point for publications in a digital library. However, most articles do not have any keywords assigned by their authors or only have one or two. Besides, manually assigning keywords to publications is a tedious work. In this paper, we explored automatic keyword inference using topic modeling techniques and full text data. The linear combination of mutual information and LLDA with full text outperformed other methods on MAP, while language model with abstract performed better than other methods on precision@10.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In E. K. Makki & E. L. Bouganim (Eds.), *Proceedings of the seventh international conference on Information and knowledge management* (pp. 148-155). Bethesda, Maryland, United States: ACM.

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction, *Proceedings of the 16th international joint conference on Artificial intelligence* (Vol. 2, pp. 668-673). Stockholm, Sweden: Morgan Kaufmann Publishers Inc.

Tomokiyo, T., & Hurst, M. (2003). A language model approach to keyphrase extraction. In L. Levin, T. Tokunaga & A. Lenci (Eds.), *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment* (Vol. 18, pp. 33-40). Sapporo, Japan: Association for Computational Linguistics.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Vol. 1, pp. 248-256). Singapore: Association for Computational Linguistics.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: practical automatic keyphrase extraction. In N. Rowe & E. A. Fox (Eds.), *Proceedings of the fourth ACM conference on Digital libraries* (pp. 254-255). Berkeley, California, United States: ACM.