Evaluation and analysis of term scoring methods for term extraction

Published in Information Retrieval Journal, October 2016 [6]

Suzan Verberne
Radboud University
Nijmegen, the Netherlands
s.verberne@cs.ru.nl

Maya Sappelli
TNO
The Hague, the Netherlands
maya.sappelli@tno.nl

Djoerd Hiemstra
University of Twente
Enschede, the Netherlands
hiemstra@cs.utwente.nl

Wessel Kraaij
TNO and Leiden University
the Hague, the Netherlands
w.kraaij@liacs.leidenuniv.nl

Keywords or key terms are short phrases that represent the content of a document or a document collection. In this paper [6], we evaluated five term scoring methods for automatic term extraction on four different types of text collections: personal document collections, news articles, scientific articles and medical discharge summaries. Each collection has its own use case: author profiling, boolean query term suggestion, personalized query suggestion and patient query expansion.

Methods for term scoring were designed with a specific goal in mind, and are used in the literature for a range of diverse applications. It is as yet unclear how these methods compare to each other and how they perform on different types of collections (size, domain, language) than they were designed for. We therefore addressed the following research question: "What factors determine the success of a term scoring method for keyword extraction?"

In a series of experiments, we evaluated, compared and analysed the output of five unsupervised term scoring methods [2, 5, 4, 3, 1]. All have term frequency as central component and combine that principle with either of two additional principles: *informativeness* (specificity of a term for the collection) and *phraseness* (how tight the combination of words in a multi-word term is). We addressed the following subquestions in evaluating the quality of the extracted terms: (1) What is the influence of the collection size? (2) What is the influence of the background collection? (3) What is the influence of multi-word phrases?

We found that the most important factors in the success of a term scoring method are the size of the collection and the importance of multi-word terms in the domain. Larger collections lead to better terms; all methods are hindered by small collection sizes (below 1000 words). The most flexible method for the extraction of single-word and multi-word terms is Pointwise Kullback-Leibler Divergence for Informativeness and Phraseness [5].

Overall, we have shown that extracting relevant terms using unsupervised term scoring methods is possible in diverse use cases, and that the methods are applicable in more contexts than their original design purpose. Our final recommendation is that the choice of method and evaluation for term extraction should depend on the specific use case. It should always be taken into account that the use case poses specific requirements on the extracted terms: terms that are informative for author profiling are different from terms that are powerful for query expansion. Thus, not only the collection size, language and domain determine the success of a term scoring method, but also the context in which the terms are used – this context is not necessarily the purpose the method was designed for.

1. ACKNOWLEDGMENTS

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

2. REFERENCES

[1] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[2] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2004.

[3] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[4] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora*, pages 1–6. Association for Computational Linguistics, 2000.

[5] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 33–40. Association for Computational Linguistics, 2003.

[6] S. Verberne, M. Sappelli, D. Hiemstra, and W. Kraaij. Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval Journal*, 19(5):510–545, 2016.

DIR 2016, Delft, the Netherlands
© 2016 ACM. ISBN .
DOI: