A Citation-based approach to automatic topical indexing of scientific literature

Arash Joorabchi and Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Republic of Ireland
{Arash.Joorabchi, Hussain.Mahdi}@ul.ie

Abstract

Topical indexing of documents with keyphrases is a common method used for revealing the subject of scientific and research documents to both human readers and information retrieval tools, such as search engines. However, scientific documents that are manually indexed with keyphrases are still in the minority. This work proposed a new unsupervised method for automatic keyphrase extraction from scientific documents which yields a performance on a par with human indexers. The method is based on identifying references cited in the document to be indexed and, using the keyphrases assigned to those references for generating a set of high-likelihood keyphrases for the document. We have evaluated the performance of the proposed method by using it to automatically index a third-party testset of research documents. Reported experimental results show that the performance of our method, measured in terms of consistency with human indexers, is competitive with that achieved by state-of-the-art supervised methods. The results of this work is published in [1].

1. Citation-based keyphrase extraction

A significant portion of electronic documents published on the Internet become part of a large chain of networks via some form of linkage that they have to other documents. In relation to scientific literature which is the subject of our work, the citation networks among scientific documents have been successfully used to improve the search and retrieval methods for scholarly publications, e.g., see [2]. These studies have successfully shown that citation networks among scientific documents can be utilized to improve the performance of three major information retrieval tasks; namely, clustering, classification, and full-text indexing. In our opinion, the results of these studies indirectly suggest that the content of cited documents could also potentially be used to improve the performance of keyphrase indexing of scientific documents. In this work, we have investigated this hypothesis as a new application of citation networks by developing a new Citation-based Keyphrase Extraction (CKE) method for scientific literature and evaluating its performance. The proposed method can be outlined in three main steps:

1. Reference extraction: this comprises the process of identifying and extracting reference strings in the bibliography section of a given document and parsing them into their logical components.
2. Data mining: this is a three-fold process. In the first stage, we query the Google Book Search (GBS) to retrieve a list of publications which cite either the given document or one of its references. Then, in the second stage, we retrieve the metadata records of these citing publications from the GBS database. Among other metadata elements, these records contain a list of key terms extracted from the content of the citing publications. In the final stage of the data mining process, we extract these key terms along with their numerically represented degree of importance from the metadata records of the citing publications to be used as primary clues for keyphrase indexing of the given document.
3. Term weighting and selection: this process starts by searching the content of the given document for the set of key terms collected in the data mining process (step 2 above). Each matching term would be assigned a keyphraseness score which is the product function of seven statistical properties of the given term, namely: frequency among the extracted key terms, frequency inside the document, number of words, average degree of importance, first occurrence position inside the document, frequency inside reference strings, and length measured in terms of the number of characters. After computing the keyphraseness scores for all the candidate key terms, a simple selection algorithm is applied to index the document with a set of most probable keyphrases.

2. Experimental Results

Our CKE algorithm clearly outperforms its unsupervised rival, Grineva et al. algorithm [3]. In comparison to its supervised rivals, the CKE algorithm significantly outperforms KEA [4] under all conditions. However, it yields a slightly lower averaged inter-consistency score ($\leq$1.1%) compared to Maui [5].

3. References

[1] Mahdi A. E. and Joorabchi A., A Citation-based approach to automatic topical indexing of scientific literature, *Journal of Information Science* 2010; 36, 6.
[2] Aljaber B., et al., Document clustering of scientific texts using citation contexts, *Information Retrieval* 2009; 13, 2: 101-131.
[3] Grineva M., et al, Extracting key terms from noisy and multi-theme documents. In: 18th international conference on World wide web; 2009; Madrid, Spain
[4] Witten I. H., et al., KEA: practical automatic keyphrase extraction. In: fourth ACM conference on Digital libraries; 1999
[5] Medelyan O., *Human-competitive automatic topic indexing* (Ph.D Thesis, University of Waikato, 2009).