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, Limerick, Republic of Ireland

Introduction

Topical indexing of documents with keyphrases is a common method used for revealing the subject of scientific documents. However, scientific documents that are manually indexed with keyphrases are still in the minority. In this work we propose 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 iteself. 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].

Rationale

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. It has been 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 the document to be indexed and parsing them into their logical components.

2. Data mining: this is a three-fold process. In the first stage, we 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. 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.

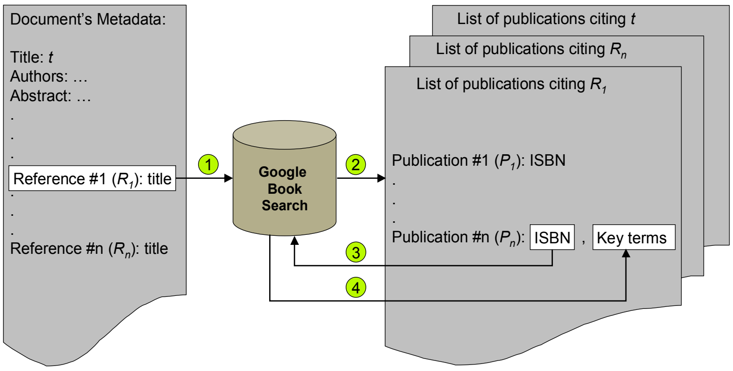
3. Term weighting and selection: this process starts by searching the content of the given document for the set of key terms collected in the data mining process (step 2). Each matching term would be assigned a keyphraseness score which is the product function of seven statistical properties of the term. After computing the keyphraseness scores for all the candidate key terms, a selection algorithm is applied to index the document with a set of most probable keyphrases.

Reference Extraction

The wide diversity of citation styles used in scientific literature makes automatic extraction and segmentation of references a non-trivial task. In this work, we use an open-source reference extraction software system called ParsCit [2], which is ML-based and uses Conditional Random Fields as its learning mechanism.

Data Mining

In the first stage of the data mining process, the Google Book Search (GBS) database is queried to retrieve a list of publications that cite either the document to be indexed or one of its references. The returned result for each query is an XML file containing metadata records of matching publications containing metadata elements such as title, author(s), ISBN/ISSN, etc. The ISBN/ISSN of each item is extracted to be used as its key identifier. In the second stage of the data mining process these key identifiers are used to retrieve the HTML pages containing the key terms extracted from their respective items. In the third and final stage of the data mining process a regular expression-based HTML parsing method is used to extract the key terms and their corresponding significance value from the retrieved HTML pages. The figure below illustrates the data mining process.


Document's Metadata:
Title: t
Authors: …
Abstract: …
Reference #1 (R_1): title
Reference #n (R_n): title
1
Google Book Search
2
3
4
List of publications citing t
List of publications citing R_n
List of publications citing R_1
Publication #1 (P_1): ISBN
Publication #n (P_n): ISBN , Key terms

Term weighting and Selection

The content of the document is searched for the key terms in the pool and each matching term *t* is assigned a keyphraseness score, *K*(*t*), using the following formula:

$$K(t)=\log_2(GF(t)+1)\times\log_2(LF(t)+1)\times 2^{RF(t)}\times\log_2(FO(t)+1)\times 2^{NW(t)}\times\log_2 NC(t)\times 2^{ADI(t)}$$

where,

- *GF*(*t*) is the Global Frequency of a given term *t*, and represents the occurrence frequency of the term in the pool of collected key terms.
- *LF*(*t*) is the Local Frequency of a given term *t*, and represents the occurrence frequency of the term in the document to be indexed.
- *RF*(*t*) is the Reference Frequency of a given term *t*, and is assigned the value 1 if *t* occurs inside any of the extracted reference strings, and 0 otherwise.
- *FO*(*t*) is the First Occurrence of a given term *t*. It represents the relative distance of *t*, where it occurs for the first time, from the beginning of the document.
- *NW*(*t*) is the Number of Words in a given term *t*.
- *NC*(*t*) is the Number of Characters, including spaces, in a given term *t*.
- *ADI*(*t*) is the Average Degree of Importance of a given term *t*.

The candidate terms are then sorted according to their corresponding score values and those with the highest likelihood scores are selected for the document.

System Evaluation & Experimental Results

we used the wiki-20 [3] test dataset to evaluate the performance of our CKE algorithm; and adopted the inter-indexer consistency formula to measure the quality of keyphrases assigned to the test documents by our algorithm, as compared to those assigned by human indexers. The table below shows the performance of our algorithm in terms of averaged inter-consistency with the human indexers and compares it with the performance of three competitive algorithms: the well-known KEA [4], an enhanced version of KEA known as Maui [3], and the work of Grineva et al. [5].

	Method	**No. of keyphrases assigned to each document**	**Inter-consistency (%)**		
			Min.	**Avg.**	**Max.**
Manual	Human indexing (gold standard)	Varied	21.4	30.5	37.1
Supervised	KEA (Naïve Bayes)	Static - 5	15.5	22.6	27.3
	Maui (Bagged Decision Trees & best features)	Static - 5	23.6	31.6	37.9
Unsupervised	Grineva et al.	Static - 5	18.2	27.3	33.0
	CKE (condition A)	Static - 5	22.7	30.6	38.3
	CKE (condition B)	Static - 6	26.0	31.1	39.3
	CKE (condition C)	Varied - the same as assigned by human indexers	22.0	30.5	38.7

References

[1] Mahdi A. E. and Joorabchi A., A Citation-based approach to automatic topical indexing of scientific literature, *Journal of Information Science* December 2010; 36, 6: 798-811.

[2] Councill I. G., Giles C. L. and Kan M.-Y. ParsCit: An open-source CRF reference string parsing package. In: Language Resources and Evaluation Conference (LREC 08)

[3] Medelyan O., Human-competitive automatic topic indexing (Ph.D Thesis, University of Waikato, 2009).

[4] Witten I. H., Paynter G. W., Frank E., Gutwin C. and Nevill-Manning C. G. KEA: practical automatic keyphrase extraction. In: fourth ACM conference on Digital libraries; 1999

[5] Grineva M., Grinev M. and Lizorkin D. Extracting key terms from noisy and multi-theme documents. In: 18th international conference on World wide web; 2009