Compiling Keyphrase Candidates for Scientific Literature Based on Wikipedia

Hung-Hsuan Chen[1], Jian Wu[2], and C. Lee Giles[2]

[1] Computer Science and Information Engineering, National Central University
hhchen@ncu.edu.tw
[2] Information Sciences and Technology, Pennsylvania State University
{jxw394,giles}@ist.psu.edu

Abstract. Keyphrase candidate compilation is a crucial step for both supervised and unsupervised keyphrase extractors. The traditional methods are usually based on the lexical or frequency properties of the phrases to come up the list. However, terms collected based on these properties do not always semantically meaningful. We show that Wikipedia can be a great auxiliary resource to compile meaningful keyphrase candidates for scientific literature. We conducted empirical experiments on digital libraries of two disciplines, namely Computer Science and Chemistry. The results suggest that Wikipedia has a good coverage of the two disciplines and has the potential to be applied to other scientific disciplines.

Keywords: Keyphrase extraction, keyphrase candidate compilation, Wikipedia

1 Introduction

Extracting keyphrases from articles is essential for natural language processing and digital libraries. The extracted keyphrases can also be the foundation of other services, such as expert search [3], collaborator search [1], venue search, and algorithm search [10]. Although the problem has been investigated for decades, recent research suggested that automatic keyphrase identification is still challenging [4, 5].

Keyphrase extraction can be supervised or unsupervised. Supervised keyphrase extraction typically formulates the task as a binary classification problem in which a model M is trained to determine a phrase p to be a keyphrase or not. Such method is highly dependent on the training data. As a result, the model M could be biased toward a certain domain and less effective in others. In addition, it is not easy to obtain numerous articles with keyphrases of high quality for training. On the other hand, unsupervised keyphrase extractors rely on the characteristics of the words or the phrases to infer their likelihood of being keyphrases. Common techniques include TF-IDF and its variations, graph based ranking, cluster based ranking, etc. [4]

Both supervised and unsupervised keyphrase extractors usually require generating a list of potential keyphrases, called keyphrase candidates, before performing

keyphrase extraction. Since the final set of extracted keyphrases is a subset of the keyphrase candidates, the candidate list should include as many potential keyphrases as possible to achieve a higher recall. However, naïvely adding terms to the list may hurt the analysis efficiency and lower the precision. Several heuristics are commonly applied to compile the list. We list three possible methods below. First, allowing only terms of certain part-of-speech (POS), such as a noun or a noun phrase, to be included in the list [7]. Second, only n-grams conforming to certain conditions are collected [9]. Third, removing the stop words and treat the single-word terms as the candidates [6]. Although these approaches are widely used, they analyze only the lexical properties, not the semantic properties, of the terms in the article. As a result, it is very likely to include trivial terms, such as "experimental results" and "difficult problem", in the candidate list.

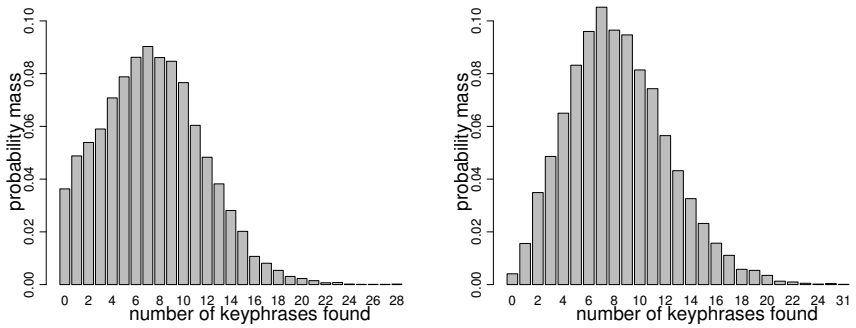
We propose to utilize Wikipedia as an auxiliary resource to compile the list of keyphrase candidates for scientific literature. Since Wikipedia is manually edited, the titles, the links, and the category structure are typically non-trivial terms. Experiments were performed on two scientific domains, namely Computer Science and Chemistry. The results suggested that Wikipedia is a promising resource for keyphrase candidate compilation and has a good coverage of the two disciplines.

2 Methodology

We collected the titles and the anchor texts (i.e., the visible and clickable text in a hyperlink) of Wikipedia pages to compile keyphrase candidates. Compared to the POS tagger and n-gram based approaches, using Wikipedia has three advantages, as described below.

First, the title or the anchor text of a Wikipedia page typically represents one concept, such as a person, an algorithm, a molecule, etc. Thus, it is usually appropriate to assume the entire title or the entire anchor text as exactly one keyphrase candidate, no matter how long or how short the phrase is. On the other hand, when using only lexical properties, it is sometimes challenging to automatically decide which terms should be joined together to represent one concept. For example, the term "Barnes & Noble" should be one phrase to represent the giant book corporation, but it is very likely to be treated as two separated terms "Barnes" and "Nobel" by a lexical-based analyzer; the term "Markov chain Monte Carlo" should be one term, although both "Markov chain" and "Monte Carlo" are valid concepts by themselves. Several languages, such as Thai, Chinese, and Japanese, can be even more challenging in determining a set of characters as a meaningful concept, because these languages exhibit no space boundaries between words and therefore difficult to tokenize and identify a valid term.

Second, the title or the anchor text of a Wikipedia page is usually written as a commonly represented format. Therefore, we do not need to worry about converting a term into its normally used type, such as converting a plural noun into a singular noun. Traditionally, format conversion is accomplished by stemming. However, not every term should be expressed in the stemmed format.



(a) 10,000 randomly selected documents

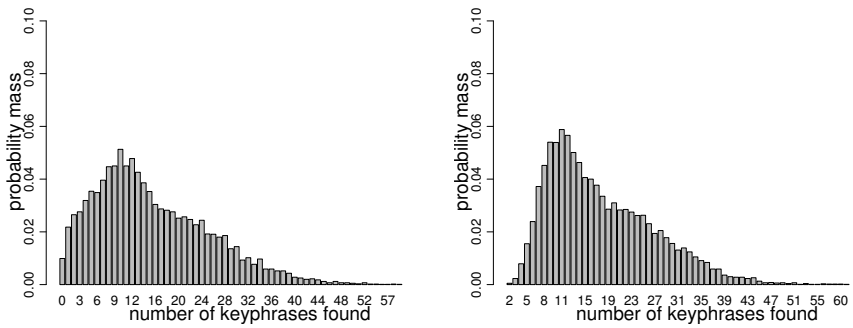
(b) 10,000 randomly selected documents with at least 4 words in title and 20 words in abstract

Fig. 1. Empirical probability mass function of number of keyphrases found in the title and the abstract for a document in CiteSeerX.

For example, we mostly say "social media" rather than "social medium", and we use "data analysis" instead of "datum analysis". In addition, a stemmer may make mistakes, such as over-stemming or under-stemming, because natural languages are not always regularly constructed. The stemming problem can be more severe in other languages, such as Hebrew and Arabic, which have much more complex rules than English.

Third, Wikipedia can be helpful in identifying the ambiguous terms or the acronym of many possible candidate terms. Given the targeted documents are within a certain domain, say Computer science, we could crawl only the pages related to the topic. In practice, we utilize the category structure of Wikipedia to perform focused crawling. A disambiguated term, such as SVM, may refer to Saskatchewan Volunteer Medal, a civil decoration for volunteers in Canada, Schuylkill Valley Metro, a proposal for a railway system linking Philadelphia and Reading in Pennsylvania, or Support Vector Machine, a powerful machine learning technique. When crawling Wikipedia pages of Computer Science domain, SVM would naturally be determined as Support Vector Machine, since the other alternatives do not of fall in the Computer Science category.

To identify the keyphrases from a document, we compared the context with the candidate list and claimed a phrase to be one keyphrase if it is in the candidate list. To efficiently search the candidate list and perform the longest prefix matching lookup, we created a trie (a prefix tree) for the keyphrase candidates, as suggested in [9].



(a) 10,000 randomly selected documents

(b) 10,000 randomly selected documents with at least 4 words in title and 20 words in abstract

Fig. 2. Empirical probability mass function of number of keyphrases found in the title and the abstract for a document in RSC

Table 1. Statistics of the number of keyphrases found per document in CiteSeerX and RSC.

Set ID	Desc.	Min	Q1	Q2	Mean	Q3	Max.
A	10,000 randomly selected CiteSeerX documents	0	4	7	7.409	10	28
B	10,000 randomly selected RSC documents	0	8	13	15.413	22	66
C	10,000 CiteSeerX documents whose titles have at least 4 words and abstracts have at least 20 words	0	5	8	8.313	11	31
D	10,000 RSC documents whose titles have at least 4 words and abstracts have at least 20 words	2	11	16	17.741	24	67

3 Experiments

3.1 Experimental Data

Wikipedia is edited manually and therefore the title or the anchor text typically represents a meaningful topic. However, the coverage of Wikipedia in scientific domain, such as Computer Science or Chemistry, is unknown. To answer the question, we conducted empirical study on two digital libraries of different discipline: (1) CiteSeerX, a digital library currently focused on Computer Science and several related fields, and (2) the publicly available metadata of documents from Royal Society of Chemistry (RSC), a professional chemistry society in UK.

We randomly selected $10,000$ documents from CiteSeerX as Set A and $10,000$ documents from RSC as Set B. Using the title and the abstract, we counted the number of terms appeared in the keyphrase candidate.

3.2 Results

Figure 1(a) and Figure 2(a) show the empirical density function of the number of matched terms per document in CiteSeerX and RSC respectively. As shown, only less than 4% of the documents in CiteSeerX and 1% of the documents in RSC have no keyphrase match.

To further study the documents with 0 matched keyphrases, we scrutinized 100 of them and found that the titles and the abstracts for most of them are extremely short, mainly due to parsing error. To eliminate the confounding parsing factor, we randomly selected $10,000$ documents whose titles have at least 4 words and abstracts have at least 20 words from CiteSeerX as Set C and from RSC as Set D. The empirical density functions for the new samples are shown in Figure 1(b) and Figure 2(b). Only less than 0.5% of the sampled papers in CiteSeerX and none of them in RSC have no keyphrase match. Statistical summaries of the number of matched keyphrases per sampled document are shown in Table 1. The result demonstrated that Wikipedia has a good coverage of the two disciplines, and very likely to be a helpful resource in compiling keyphrase candidate for documents of other scientific disciplines as well.

4 Deployment

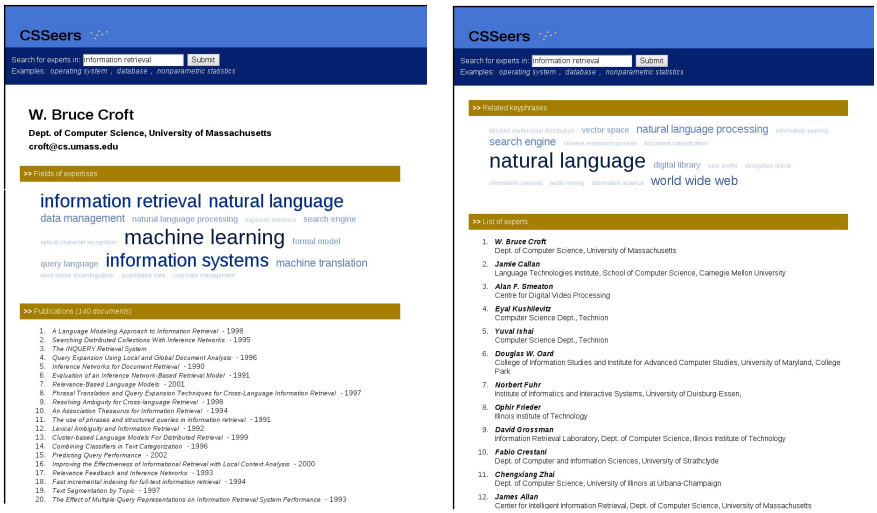
We have utilized the discovered keyphrase candidates to support several systems. Here we introduce some of them.

4.1 CSSeer and CollabSeer

CSSeer[1] is an expert recommender system built on top of four million academic documents in the fields related to Computer Science and Information Science [2,3]. To efficiently return a list of experts of the specified sub-domain (e.g., information retrieval), CSSeer preprocesses the texts in the title and the abstract of each document to extract the keyphrase candidates as the input texts for more complex algorithms. Since most interesting keyphrases are preprocessed and indexed, CSSeer can effectively return a list of experts within seconds. On the other hand, if a user submits a query term which is not included in the preprocessed keyphrase list, calculating the expert score of a user to the query term *in real time* is impractical [2]. Alternatively, we probably need to approximate the expert score by considering only the top related important documents (instead of the full four million documents). However, the approximation considers at most hundreds of documents, which inevitably ignores most of the available information. As a result, the keyphrase candidate extracting method forms an essential component in the CSSeer recommendation service.

Figure 3 shows two snapshots of the CSSeer system. On the left (i.e., Figure 3(a)), the list of expertise of Dr. W. Bruce Croft is compiled based on the

[1] http://csseer.ist.psu.edu/



(a) W. Bruce Croft's expertise list and publication list

(b) Related keyphrases and experts of "information retrieval"

Fig. 3. Screenshots of CSSeer.

Table 2. Statistics of the increase ratio of the keyphrase candidates of the 1,000 sampled CiteSeerX documents.

Min	Q1	Q2	Mean	Q3	Max.
0%	42.86%	56.52%	60.73%	72.73%	600%

keyphrase candidates extracted from his publications. On the right (i.e., Figure 3(b)), the phrases that are most relevant to the query phrase "information retrieval" is also generated based on the keyphrase candidates compiled by our introduced method.

CollabSeer is another system that was leveraged on the keyphrase candidate compiled based on the introduced method. Essentially, CollabSeer recommends potential collaborators to a researcher's interested area *within her academic social circle*. Like CSSeer, we identify each user's research interest and expertise based on the keyphrase candidates discovered from her previous publications. Figure 4 shows a snapshot of the expertise list of an author.

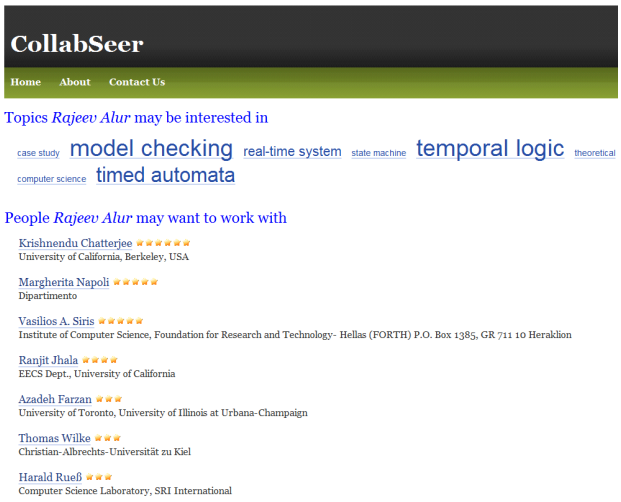


Fig. 4. A snapshot of the expertise list

Table 3. A comparison of the average recalls based on the 100 sampled CiteSeerX documents.

Method	POS-tagging	Wikipedia matching	A combination of both
Avg. Num. of Keyphrase Candidates	15.95	11.39	24.96
Average Recall	73.06%	48.00%	91.67%

4.2 CiteSeerX

CiteSeerX[2] is an autonomous digital library for scientific literature. For each document, CiteSeerX provides a summary tab that shows the abstract and the keyphrases extracted from the abstract, as shown in Figure 5.

The current online version of the keyphrase list is compiled based on an unsupervised method which tags the nouns and the noun phrases by the Stanford POS Tagger and noun phrase rules [8,11,12] and naïvely treats these noun phrases as the keyphrase candidates. However, we found that the recall of such a method is only about 70%. Since the final extracted keyphrases are only a subset of the keyphrase candidates, we would like the keyphrase candidates to include many potential keyphrases to achieve a higher recall. We plan to update this keyphrases candidate generating process by a mixture of the original method (POS-tagging-based) and the method introduced in this paper (Wikipedia-based) to increase the recall.

[2] http://citeseerx.ist.psu.edu/

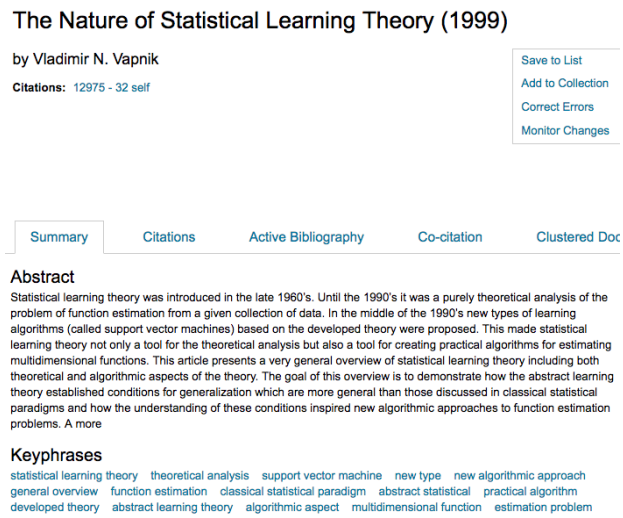


Fig. 5. A snapshot of the CiteSeerX summary tab

As an initial study, we randomly selected 1,000 papers whose abstract contains at least 20 words, and compile the keyphrase candidates by a mixture of the original and the new method (i.e., we merge the keyphrase candidates returned by the two methods). We found that, on average, the mixture approach increases the number of keyphrase candidates per document from the original 14.49 to 23.29. The increase ratio is $(23.29 - 14.49)/14.49 = 60.73\%$ on average. Table 2 shows the summary of the increase ratio of the 1,000 sampled documents, and Figure 6 displays the empirical cumulative density function (ECDF) of the increase ratio of these documents.

In the meanwhile, we manually labeled the keyphrases of these 100 documents. We computed the recall of the keyphrase candidates generated from the following methods: (1) generating keyphrases based on the POS tagging; (2) generating keyphrases based on the Wikipedia terms; (3) a combination of (1) and (2). The average recall from this test dataset is shown in Table 3. By combining these two methods, we can achieve an average recall rate to 91.67% (increasing the number of keyphrase candidates by 9.01 on average).

5 Discussion

In this paper, we empirically validated that Wikipedia titles and the anchor texts are valuable resources to generate keyphrase candidates for scientific articles. We found that, based only on the abstract texts of the scientific documents, such a simple method can generate 8.3 keyphrase candidates for a typical paper in the field of Computer Science and Information Systems and 17.7 keyphrase candidates for a typical Chemistry paper. If we combine the Wikipedia resource

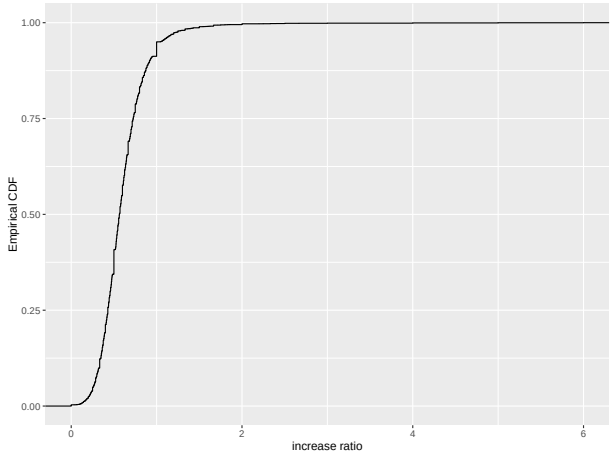


Fig. 6. The empirical cumulative density function of the increase ratio

and simple POS-tagging technique, the generated keyphrase candidates yield a very high recall rate (over 90% on average).

We built several systems partially based on the concept. Specifically, we generated each author's research expertise based on the keyphrase candidates of her previous publications and integrated the function into CSSeer (an expert recommender system for computer scientists) and CollabSeer (a collaborator recommender system for computer scientists). We generated the keyphrases for the documents collected by CiteSeerX and plan to update the current keyphrase list shown online.

For future work, we plan to apply similar concept to different domains. Finally, we are also in the process of releasing the title, abstract, and the extracted keyphrases of the 10 million academic documents collected by CiteSeerX. We hope that such a large dataset can benefit the research community in the digital library and information retrieval.

References

1. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: CollabSeer: a search engine for collaboration discovery. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. pp. 231–240. ACM (2011)
2. Chen, H.H., Ororbia, I., Alexander, G., Giles, C.L.: ExpertSeer: a Keyphrase Based Expert Recommender for Digital Libraries. arXiv preprint arXiv:1511.02058 (2015)
3. Chen, H.H., Treeratpituk, P., Mitra, P., Giles, C.L.: CSSeer: an expert recommendation system based on CiteseerX. In: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. pp. 381–382. ACM (2013)
4. Hasan, K.S., Ng, V.: Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In: Proceedings of the 23rd International Conference

on Computational Linguistics: Posters. pp. 365–373. Association for Computational Linguistics (2010)
5. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: A survey of the state of the art. In: ACL (1). pp. 1262–1273 (2014)
6. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to find exemplar terms for keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. pp. 257–266. Association for Computational Linguistics (2009)
7. Mihalcea, R., Tarau, P.: Textrank: bringing order into texts. In: Proceedings of EMNLP. vol. 4. Barcelona, Spain (2004)
8. Nguyen, T.D., Kan, M.Y.: Keyphrase extraction in scientific publications. In: Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, pp. 317–326. Springer (2007)
9. Treeratpituk, P., Teregowda, P., Huang, J., Giles, C.: SEERLAB: a system for extracting keyphrases from scholarly documents. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics (2010)
10. Tuarob, S., Mitra, P., Giles, C.: Building a search engine for algorithms. ACM SIGWEB Newsletter p. 5 (2014)
11. Williams, K., Chen, H.H., Choudhury, S.R., Giles, C.L.: Unsupervised ranking for plagiarism source retrieval. Notebook for PAN at CLEF (2013)
12. Williams, K., Chen, H.H., Giles, C.L.: Classifying and ranking search engine results as potential sources of plagiarism. In: Proceedings of the 2014 ACM symposium on Document engineering. pp. 97–106. ACM (2014)