Cluster Computing (2019) 22:S8669–S8675
https://doi.org/10.1007/s10586-018-1937-1

Concepts recommendation for searching scientific papers

Yang Chi[1]· Jinchao Zhu[1] · Lan Huag[1,2,4] · Hao Xu[1,2,3,4]

Received: 21 December 2017 / Revised: 18 January 2018 / Accepted: 30 January 2018 / Published online: 28 March 2018

Abstract

Scientific retrieval systems need to be given domain search terms for searching publications, however, as natural language, search terms provided by users are often fuzzy and limited and some relevant terms are always overlooked in searching. Meanwhile, users always desire to be given domain related keywords to enlighten themselves what other terms can be used for their searching. This paper presents a concepts recommendation model in scientific paper retrieval, in which concepts are extracted from keyword in scientific papers, and some data mining algorithms are used to calculate the similarity between search terms and concepts and do recommendation for users. This model is simple and can be used with small dataset, in which all training data used is from meta data of papers that is easy to acquired. Experimental result hold good precision, which shows that this research not only simplifies searching step and improves the searching quality for users, but also lays the foundation for semantic search.

Keywords Concepts recommendation · Data mining · Information retrieval · Scientific papers

1 Introduction

The exponential growth of electronic scientific papers with unstructured nature has made finding and searching rather difficult [1]. The aim of searching scientific papers is to map a natural language query, which describes information that seeker needs by several keywords called search terms, to a set of papers in given database. There are three principally paths called syntactic information retrieval, semantic enabled syntactic information retrieval and semantic information retrieval, all of them need provided search terms for search engine in common firstly [2]. Search term plays an important role in query process affecting search quality, but as fuzzy natural language, if they are used to do a search directly, especially in searching scientific papers, some relevant terms are overlooked, because related research concepts are syntactically irrelevant [3].

There is a small, but notable, set of related scholarly work committed to studying keywords in scientific papers. For several years, many researchers have done very impressive jobs related to this research. The presence of a long history of statistics has been indicated through analysis of keyword counts in an early scientific journal, and the research also found the functional approach and model-based curve clustering turned out to be better at tracing and comparing individual temporal evolution of keywords, despite computational and theoretical limitations [4]. Some structural analysts indicated that entropic and clustering approaches are used to improve statistical keyword detection in short texts, which concluded the clustering approach is able to better discriminate the degree of relevance of low frequency words than the entropic approach. In this experiment, the clustering method is used to divide scientific papers, which are completely short texts, into different classes, and measure the weight of keywords in different classes [5]. Extracting keywords has been achieved for scientific paper retrieval and the research used a novel, unsupervised cascade learning scheme to extract a ‘cluster-theme keyword’ structure from related papers of a research topic, helping users quickly locate research

✉ Hao Xu
xuhao@jlu.edu.cn

[1] College of Computer Science and Technology, Jilin University, Changchun, China

[2] School of Management, Jilin University, Changchun, China

[3] Department of Computer Science and Technology, Zhuhai College of Jilin University, Guangdong, China

[4] Symbol Computation and Knowledge Engineer of Ministry of Education, Jilin University, Changchun, China

interests. However, it did not consider that although different keywords are not in the same subject, there still has a relationship between them. Consequently, the results of research interests were incomplete [6]. To identify the most-used keywords in the title and content of articles published over time, a methodology has been developed and providing insight for traceability of future research directions, which have important influence on attaining publication content, as well as quality information for readers and authors [7]. Some systems could refine user interest profiling, focusing on extending scientific subject ontology via keyword clustering and improving the accuracy and effectiveness of recommendations of electronic academic publications in online services [8]. A research used identified keywords of published literature to explore major subjects and evaluate the quality of the Korean Journal of Womens' Health Nursing from 2007 to 2009 [9]. Some keyword search engines could enable users to easily access XML data without the need of learning XPath or XQuery and study possibly complex data schemas [10]. Some researchers used two methods, one based on online communities and another based on the set of nouns obtained by morphological analysis of webpages to cluster keywords of a search engine [11]. Leverage and citation-network and meta data information are also used to recommend relevant keywords for papers [12]. Many other researchers were dedicated to keywords of scientific papers, and many methods were used for keyword extraction, keyword search and keyword analysis [13].

In this paper, a concepts recommendation model is used to provide related concepts before searching based on search terms, including three major steps: the first is to preprocess meta data and define concepts, the second is to acquire related terms based on search words and finally is to map these related terms to concepts that would be recommended. The metadata in this work contains title, keywords and abstract in papers. For scientific papers, the keywords defined by authors in papers, is perfect natural research concepts, which finding process does not need a lot of background knowledge and is easy, so this work extracts research concept from keywords in database. For mapping search terms to research concepts, this work regards single words in meta data of papers as intermediate node, in which select related single words are selected by calculating similarity value between candidate words and search terms, and then these related single words are matched to concepts that are recommended to users. There are two main contributions in this work. Firstly, different from other recommendation model using string to describe domain knowledge, this work proposes concept recommendation which extracts concept from keyword string. Secondly, semantic correlation is to compare meaning of the concepts not words, which need a great deal of background knowledge, in this part, this work puts forward related concepts recommendation model based on data mining algorithms, which only search for relatedness, do not need to deal with the relationship between concepts and only need metadata of papers which is easy to acquire.

This model recommends related concepts automatically, which not only solves the natural language fuzziness problem, simplifies the search steps, but also lays the foundation for the next step work of semantic search, which helps users quickly find accurate search concepts and improve retrieval efficiency.

2 Methodology

The proposed concept recommendation algorithm framework of this work is given in Fig. 1, which includes data acquiring, data preprocessing and concept extracting, and concept recommendation. Data of this work is acquired from Web of Science, which is the meta data of papers includes title, keywords, and abstract. Section 2.1 introduces data preprocessing method, Sect. 2.2 introduces concept recommendation model, and similarity calculation of concept recommendation is introduced in Sect. 2.3.

2.1 Data preprocessing and concept extracting

The goal of searching scientific papers is to map several search keywords called search terms to a set of papers in given database. All search engines need in common is provided search terms firstly. The accuracy of search terms affects search quality, but as natural language, which are fuzzy and always perform polysemy, synonyms, etc., so precise search terms are significant. In this paper, we propose a model to recommend related concepts for searching papers based on search terms. Data used in this work is the meta data of the paper, which includes title,

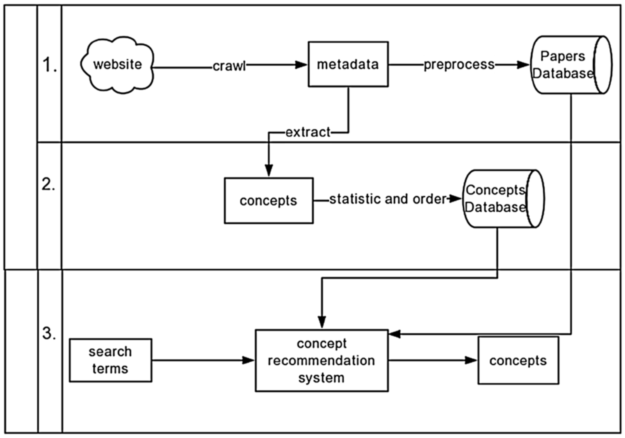

Fig. 1 Concept recommendation algorithm framework

keywords, and abstracts acquired in Web of Science stored at papers database.

For concepts recommendation implementation, it is advisable to further establish a concept database which contains two elements: concepts and the appearing frequency of every concept. So, according to this principle, this work extracts concepts from meta data of papers. The candidate terms of the concepts extracted from keywords. However, they are string, not concept. For example, “search engine” and “search engines”, “SVM” and “Support Vector Machine” are the different strings, but same concepts. So, this work merges synonyms by same NLP technologies and language rules defined. For example, if “a(b)” in our language rules, two terms appeared in context with this pattern, this work signs them as the same concept, and then, if they are found in another context with different language pattern, the new pattern can be added in language rules. Through this method, concepts are extracted from keywords and stored at concept database.

2.2 Concepts recommendation model

This model studies concepts recommendation based the above preprocessed metadata. In this paper, a concepts recommendation model is proposed, including two main parts: similarity calculation for terms and term-concept matching, through which related concepts can be extracted and recommended. The flowchart of this concepts recommendation model is shown in Fig. 2. Based on search words, related terms are extracted by three kinds of calculations and cosine similarity is used to map those terms to concepts. After ranking, the most related concepts could be recommended.

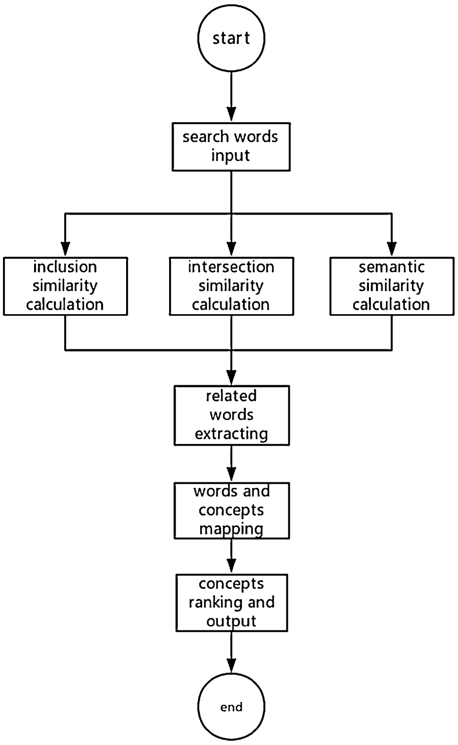
start
search words input
inclusion similarity calculation
intersection similarity calculation
semantic similarity calculation
related words extracting
words and concepts mapping
concepts ranking and output
end

Fig. 2 Flowchart of concepts recommendation model

The similarity is calculated between search term provided by users and terms in meta data of paper, which includes three factors: inclusion, intersection, and semantic relation. The method of similarity calculation could be introduced in 3.3.

After extracting terms with high similarity values with input term, this work maps these terms to concept, for example, maps ‘search’, ‘engine’, ‘web’, ‘optimization’ to ‘search engines’, ‘web search engines’, and ‘search engine optimization’. Cosine similarity is used in this work to accomplish this mapping. If n related words have been selected, the related words sequence can be seen as one vector W(w1, w2,…, wn), while each concept is another vector C(c1, c2,…, cn). The wi is the similarity weight value calculated above, which describes similarity degree of every word in search topic. And ci is the weight value of every word in concept. If the word is in concept phrase, the value is set with “1”, else, the value is set with “0”. Then cosine similarity algorithm can calculate the similarity of two vectors. In this way, this model figure out similarities between concepts vectors and related words vector and select those with higher value as related concepts. The cosine similarity can be seen in formula 1.

$$\cos\theta = \frac{\Sigma_{i=1}^{n}(W_i \times C_i)}{\sqrt{\Sigma_{i=1}^{n}(W_i)^2} \times \sqrt{\Sigma_{i=1}^{n}(C_i)^2}} \tag{1}$$

Two elements are considered in this work to rank the result of concepts, similarity (S) and frequency (F). This work ranks concepts with formula 2. W is the final importance weight value of each concept, the α in which is the parameter to regulate weight of S and F.

$$W = \frac{(\alpha^2 + 1)S \times F}{\alpha^2(S + F)} \tag{2}$$

2.3 Similarity calculation for terms

Three main relationships are constructed to measure the similarity between terms and search terms: inclusion, intersection, and semantic relation. All the word in the papers database selected from title, abstract, and keywords of paper could be confirmed as candidate words. The main task of semantic match is to extract the containing words, crossing words, and related words from candidate words according to the three relationships. The specific meanings of the containing words, crossing words, and related words are as follows.

Candidate words are selected from scientific paper titles and keywords which hold an inclusion relationship with search word. For instance, if ‘search engine’ is inputted by users as search word, the containing words and their similarity weight values with it could be calculated as follows. Table 1 reveals four keywords and their frequencies, and all of them contain ‘search engine’, so, every word composing these keywords is containing word, such as ‘web’, ‘optimization’, ‘marketing’, ‘search’ and so on. This model regulate the frequency of co-occurrence between every containing word and ‘search engine’ as the similarity value of the containing word. For example, the similarity value of containing word ‘optimization’ is 23, because the frequency of co-occurrence between ‘optimization’ and ‘search engine’ is 23. Similarly, ‘marketing’ is valued at 20, ‘search’ is valued at 165 (87 + 35 + 23 + 20). If the result value of a containing word is higher, the relationship between search word and it is closer.

Crossing words represent words selected from scientific papers’ titles and keywords which hold an intersection relationship with search word. In the same way of the calculation of containing words, regulate the frequency of co-occurrence between every crossing word and ‘search engine’ as the similarity value of that crossing word. The association rules algorithm is imported in this work to search for crossing words which hold a higher support degree with “search” or “engine’. The support degree of association rules algorithm represents the frequency of co-occurrence between two words.

The crossing words holding support degree with “search” are extracted by association rules and shown in Fig. 3. The size of the circle represents the support degree of the crossing word, and the thickness of lines between circles indicates the strengths of corresponding relationships.

Semantic related words hold a semantic relationship with search word, which appear frequently in the same context with search word. This work regards the word in all keywords (except stop word in English) appears in the same paper context (containing title, keywords and abstract) with search term as the semantic related word. And this work matches keywords to context by Aho–Corasick automation, and calculates tf–idf value of the semantic related word in the context as a, and the co-occurrence times with search term as b, and then a*b is the final similarity weight for related word.

Table 1 Keyword frequency

Number	Keywords	Frequency
1	Search engines	87
2	Web search engines	35
3	Search engine optimization	23
4	Search engine marketing	20

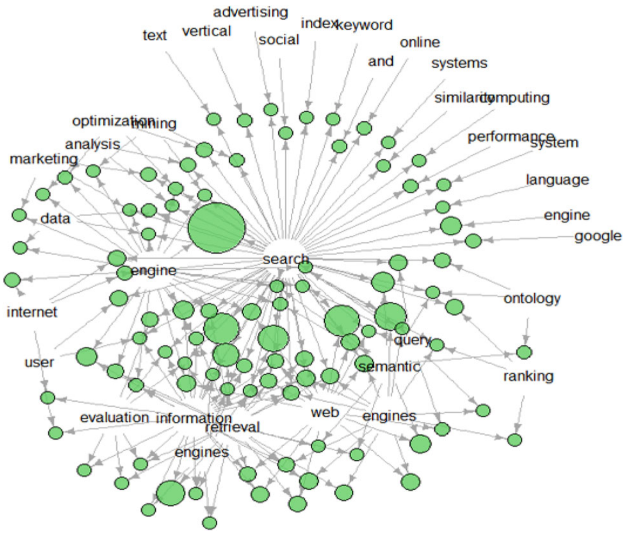

Fig. 3 Frequent words and their relationships

tf–idf, composed by Term Frequency (tf) and Inverse Document Frequency (idf), is a statistic technology to evaluate the significance of words in a document. The tf variable is the frequency of one word appearing in the document, while idf presents the level of the word appearing in other documents, which can be considered that if one word not often appears in other documents, it may hold more powerful capacity of classification and can get a higher idf value. As a result, tf–idf value is calculated as $tf \times idf$.

The final similarity between candidate word and search word is calculated by the formula which includes three aforementioned factors: normalized inclusion similarity as x, normalized intersection similarity as y, and semantic similarity as z. The parameter a, b, and c are weight of factors, which should be defined according to the application. The formula 3 is as follows:

$$Sim = a \times x + b \times y + c \times z \tag{3}$$

The dependent variable of formula 3, *Sim*, represents the similarity between search term and candidate word, which could be used to extract related words with higher *Sim* degree from candidate words. Then, the related words extracted in papers database are supposed to be mapped to some concepts in concepts database based on cosine similarity algorithm. Consequently, related concepts with higher similarities to search words could be selected.

3 Results and discussion

3000 articles in the field of search engine from Web of Science Core Collection are acquired in this work, which “search engine” is input as search topic, “article” is input as document types, and “computer science information systems or computer science theory methods or computer science artificial intelligence” as web of science categories. After pre-processing the text data, the metadata including title, abstract, and keywords of every paper is formed, which is extracted papers. Then, this work preprocesses the meta data and extracts concepts from keywords. Finally, this work acquires 4078 concepts from 7491 keywords.

This work calculates similarity of words and search terms and uses Analytic Hierarchy Process (AHP) to evaluate parameters—a, b, c—in the ‘Sim’ formula for three factors: inclusion similarity, intersection similarity, and semantic similarity. AHP is a combination of qualitative and quantitative analysis, which is systematic and hierarchical. The establishment of hierarchical structure of ‘Sim’ is represented in Formula 4, and the value of parameters in matrix is defined by experts based on the different importance of factors: containing words, cross words, and deviation words. In this work, the importance of intersection similarity is 1/3 of inclusion similarity, while the semantic similarity’s is 1/5.

$$\begin{array}{cc} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{pmatrix} 1 & 3 & 5 \\ 1/3 & 1 & 3 \\ 1/5 & 1/3 & 1 \end{pmatrix} \end{array} \tag{4}$$

APH results are shown in Table 2:

In order to evaluate this method, this work queries for six times with different search terms: “search engine”, “social network”, “machine learning”, “semantic search”, “natural language processing” and “information retrieval” and the result from the input ‘search engine’ are shown in Table 3.

The experimental result of this research has been evaluated and presented higher precision and lower recall in concepts recommendation. Precision Rate and Recall Rate are two basic indicators in the domain of information retrieval. The formulas of them are as follows:

$$Precision = \frac{\text{Retrieved related concepts}}{\text{All retrieved concepts}} \tag{5}$$

$$Recall = \frac{\text{Retrieved related concepts}}{\text{All related concepts}} \tag{6}$$

The evaluation of results of six queries experiment are shown in Table 3. The MAP is the average value of all the Precisions. The comparison of precision and recall value are also shown in Fig. 4.

Six queries experiment shows that this simple algorithm holds good precision, which suits the applications with small data set and need quick reaction. There are also some shortcomings, some keywords are incurrent to be recommended, because they have similar string with search term. So, the semantic relationship should be considered more in future work.

Table 2 Parameters value of formula 3

Parameter	a	b	c
Weight	0.8880	0.4121	0.1874

4 Conclusion

The exponential growth of electronic scientific papers with unstructured nature has made finding and searching rather difficult. In order to help users quickly search for accurate papers they need, this work proposes concept recommendation model to remind users related input keyword and expand their search term. Different from traditional string recommendation such as keywords recommendation, words recommendation and so on, this paper proposes concept recommendation, extracts concepts from keywords and designs one concept recommendation model for searching scientific papers, involving two elements—three similarity factors and a formula to calculate the similarity between words and search words and term-concept matching, which is the first attempt to recommend related concepts for scientific papers searching. Compared with other methods, this work uses scientific papers as dataset and keywords as resource to extract concepts, which is simple to realize and do not need much domain knowledge. The evaluation shows that this model holds good precision and recall.

However, this concept recommendation model only considers about string similarity and semantic similarity based on co-occurrence and machine learning technologies. The future work will focus on ontology learning to build an ontology of subject, through which the recommendation model could not only acquire related relationship between words and concepts, but also hypernym–hyponym relationships to improve the accuracy of recommendation.

Acknowledgements This work is supported by the Development Project of Jilin Province of China (Nos. 20170203002GX, 20160414009GH, 20170101006JC, 20160204022GX), the National Natural Science Foundation of China (No. 61472159), China

Table 3 Keyword recommendations result of "Search Engine"

Search engine		
Search engine	Information retrieval	Machine learning
Web search engines	Search engine optimization	Data mining
Web search	Search engine advertising	Meta search engine
World wide web	Semantic search	Electronic commerce
Semantic web	Web mining	Vertical search engine

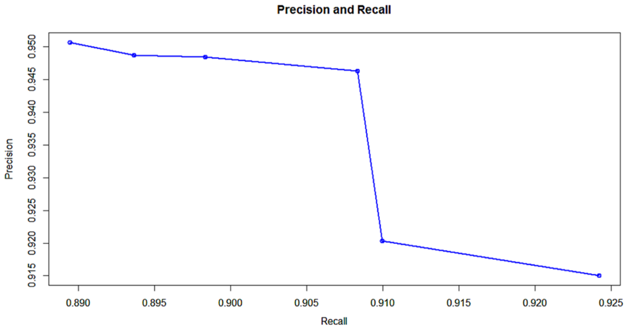

Fig. 4 Precision and recall for queries

Postdoctoral Science Foundation (No. 2016M601379) and MOE Research Center for Online Education Quantong Education Foundation (No. 2017YB131). Premier-Discipline Enhancement Scheme supported by Zhuhai Government and Premier Key-Discipline Enhancement Scheme supported Guangdong Government Funds.

References

1. Poelmans, J., Ignatov, D.I., Viaene, S., Dedene, G., Kuznetsov, S.O.: Text mining scientific papers: a survey on FCA-based information retrieval research. Workshop Adv. Data Min. Ind. Conf. **7377**, 273–287 (2012)
2. Gupta, Y., Saini, A., Saxena, A.K.: A new fuzzy logic based ranking function for efficient Information Retrieval system. Expert Syst. Appl. **42**(3), 1223–1234 (2015)
3. Kobayashi, I., Mai, S.: A study on an information recommendation system that provides topical information related to user's inquiry for information retrieval. ACM Int. Conf. Intell. Agent Technol. Workshops **26**(1), 39–48 (2006)
4. Trevisani, M., Tuzzi, A.: A portrait of JASA: the history of statistics through analysis of keyword counts in an early scientific journal. Qual. Quant. **49**(3), 1287–1304 (2015)
5. Carretero-Campos, C., Bernaola-Galván, P., Coronado, A.V.: Improving statistical keyword detection in short texts: entropic and clustering approaches. Phys. A **392**(6), 1481–1492 (2013)
6. Ren, F.: An unsupervised cascade learning scheme for 'cluster-theme keywords' structure extraction from scientific papers. J. Inf. Sci. **40**, 167–179 (2014)
7. Valerică, G.S.: Analysis method of research papers published for audit domain, based on titles and keywords. Ann. Econ. Ser. **4**(8), 53–60 (2015)
8. Xiaoyu, T., Qingtian, Z.: Keyword clustering for user interest profiling refinement within paper recommender systems. J. Syst. Softw. **85**(1), 87–101 (2012)
9. Kim, J.I., Lee, E.H., Kang, H.S., Oh, H.E., Lee, E.J., Jun, E.M.: Analysis of published papers by keywords and research methods in the korean journal of women health nursing (2007–2009). Korean J. Women Health Nurs. **16**(3), 307–316 (2010)
10. Liu, Z., Walker, J., & Chen, Y.(2007). XSeek: A semantic XML search engine using keywords. In: International Conference on Very Large Data Bases, pp. 1330–1333
11. Otsuka, S., Kitsuregawa, M.: Clustering of search engine keywords using access logs. Database Expert Syst. Appl. **4080**, 842–852 (2006)
12. Blank, I., Rokach, L., Shani, G.: Leveraging metadata to recommend keywords for academic papers. J. Assoc. Inf. Sci. Technol. **67**(12), 3073–3091 (2016)
13. Li, J., Dong, S.C., Li, Z.H.: Discussion on evolution of focal points of China's urbanization: based on keywords analysis of papers from CNKI during 1992–2011. Appl. Mech. Mater. **522–524**, 1656–1664 (2014)

Yang Chi is currently a Master Student of College of Computer Science and Technology, Jilin University, China. She received her B.S. from Jilin University, China in 2016. Her research interests include data analytics and knowledge graph.

Jinchao Zhu received her M.S. from Jilin University, China in 2017 and received her B.S. from Changchun University of Technology, China in 2014. Her research interests include data analytics and knowledge graph.

Lan Huag is currently a Professor and Vice Dean of College of Computer Science and Technology, Jilin University, China. She received her Ph.D., M.S. and B.S. from Jilin University, China, in 2003, 1999 and 1994. Her research interests include data mining and business intelligence.

Hao Xu is currently an Associate Professor at College of Computer Science and Technology, Jilin University, China. He received his Ph.D. from University of Trento, Italy, in 2012 and received his M.S. and B.S. from Jilin University, China, in 2008 and 2005. His research interests include knowledge management, big data analytics and human engagement.