Mining Keywords from Short Text Based on LDA-Based Hierarchical Semantic Graph Model

Wei Chen, Kunming University of Science and Technology, China
Zhengtao Yu, Kunming University of Science and Technology, China
Yantuan Xian, Kunming University of Science and Technology, China
Zhenhan Wang, Kunming University of Science and Technology, China
Yonghua Wen, Kunming University of Science and Technology, China

ABSTRACT

Extracting keywords from a text set is an important task. Most of the previous studies extract keywords from a single text. Using the key topics in the text collection, the association relationship between the topic and the topic in the cross-text, and the association relationship between the words and the words in the cross-text has not played an important role in the previous method of extracting keywords from the text collection. In order to improve the accuracy of extracting keywords from text collections, using the semantic relationship between topics and topics in texts and highlighting the semantic relationship between words and words under the key topics, this article proposes an unsupervised method for mining keywords from short text collections. In this method, a two level semantic association model is used to link the semantic relations between topics and the semantic relations between words, and extract the key words based on the combined action. First, the text is represented with LDA; the authors used word2vec to calculate the semantic association between topic and topic, and build a semantic relation graph between topics, that is the upper level graph, and use a graph ranking algorithm to calculate each topic score. In the lower layer, the semantic association between words and words is calculated by using the topic scores and the relationship between topics in the upper network allow a graph to be constructed. Using a graph sorting algorithm sorts the words in short text sets to determine the keywords. The experimental results show that the method is better for extracting keywords from the text set, especially in short articles. In the text, the important topics, the relationship between topics and the correlation between words can improve the accuracy of extracting keywords from the text set.

KEYWORDS
LDA, Semantic Association, TextRank, Word2Vector

1. INTRODUCTION

There are millions of academic documents in the current academic network. How to discover knowledge from the literature is a challenging problem. Keywords as a concise summary of the literature have been widely used in Natural Language Processing and information retrieval tasks. Such as scientific and technical literature summarization (Sarkar, Nasipuri, & Ghose, 2010), text clustering...
(Hammouda, Matute, & Kamel, 2005) recommendation (Pudota, Dattolo, Baruzzo, Ferrara, & Tasso, 2010), query(Jones & Staveley, 1999). The extraction methods of key words are mainly divided into two kinds: supervised learning and unsupervised learning (Hasan & Ng, 2010, 2011; Merrouni, Frikh, & Ouhbi, 2017). In the literature review, Hasan (Hasan & Ng, 2011) summarized the supervised and unsupervised learning methods in the extraction of keyword extraction.

In supervised learning, the core idea is to convert the task of extracting key words into a two-classification problem. Using positive and negative examples of artificial annotation as training corpus, we train models by acquiring the various feature values and then classify the correct keywords from the literature. i.e. decision tree algorithm (C4.5) (Turney, 2002), Bayesian method (Frank, Paynter, Witten, Gutwin, & Nevill-Manning, 1999), neural network(Sarkar et al., 2010), SVM (Jiang, Hu, & Li, 2009), maximum entropy (Yih, Goodman, & Carvalho, 2006). These algorithms use all kinds of features in the text and integrate the features into the classification model for solving. In early studies, Frank built a system called KEA (Frank et al., 1999). The two features: TF-IDF and phrase first appear in the relative position of the document are incorporated into naive Bayes formula to predict. The prediction quality of the system depends on whether the training set can match the document which need to handle. KEA make the researcher realize that word frequency information and position information is most important feature to judge the keyword. The GenEx (Turney, 2000) method developed by Turney extracts the three features TF-IDF, the position of the phrase which first appears in the document and the length of phrases, and integrates them into the decision tree generation algorithm based on heuristic rules, which improves the evaluation of F value. Nguyen (Nguyen & Kan, 2007) extends the KEA to integrate the phrase’s position information in the different chapters of the article into the model. Medelyan (Medelyan, Frank, & Witten, 2009) extends KEA and integrates the information of Wikipedia access into the model. Chuang (Chuang, Manning, & Heer, 2012) presents a model incorporating statistical information and language information to identify descriptive keywords from text. Caragea (Caragea, Bulgarov, Godea, & Gollapalli, 2014) proposed a supervised method of merging citation information with traditional feature information in the paper to extract key words from the scientific literature. Compared with previous methods, this method has improved the quality of extraction. However, when important words do not appear in important positions, such as the beginning of a sentence, or are not quoted, the prediction effect of the model will be affected.

In the unsupervised learning, the core idea is to transform the keyword extraction task into a kind of sorting work based on probability statistics or based on graphs. Use the various types of features in the text or the associations between features for statistics or sorting.

The unsupervised method based on probability statistics is earlier, and usually combined with the ti-idf_ENREF_1 value in this paper, language characteristics, and related statistical information to extract keywords. Beltagy (El-Beltagy & Rafea, 2009) expressed the text as the tf-idf word weight matrix form, using the matrix to calculate the correlation between the word and the word, and sorting keywords by other clustering algorithm. Liu (Liu, Li, Zheng, & Sun, 2009) combined the ti-idf value as one of the sorting features with the sorting algorithm to extract keywords. Frikh (Frikh, Djaanfar, & Ouhbi, 2011) used CHIR to obtain the co-occurrence distribution of words in text documents to extract key words ontology and achieved better results.

Although the unsupervised method based on probability statistics has achieved certain results, a large number of studies are still based on graph-ranking algorithms. Graph-ranking algorithm can integrate more features into the algorithm for comprehensive calculation and is considered to be the latest technology for current keyword extraction. The graph-based sorting method builds a graph of vertices and edges on the input text set. The vertices in the figure represent words or concepts of an entity. The edges represent the relationships between words or concepts, such as the co-occurrence of words, the semantic relationship, the integration of multiple features, etc. The value of each edge is a quantified representation of the specified relationship, using PageRank or an extended algorithm for iterative calculations. Gollapalli (Gollapalli & Caragea, 2014) uses the graph ordering method to
A Security Requirements Engineering Tool for Domain Engineering in Software Product Lines
Jesus Rodriguez, Eduardo Fernandez-Medina, Mario Piattini and Daniel Mellado (2011). Non-Functional Properties in Service Oriented Architecture: Requirements, Models and Methods (pp. 73-92). www.igi-global.com/chapter/security-requirements-engineering-tool-domain/52230?camid=4v1a

New Profession Development: The Case for the Business Process Engineer