Towards High Performance Text Mining: A TextRank-based Method for Automatic Text Summarization

Shanshan Yu, College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou, China
Jindian Su, College of Computer Science and Engineering, South China University of Technology, Guangzhou, China
Pengfei Li, College of Computer Science and Engineering, South China University of Technology, Guangzhou, China
Hao Wang, Norwegian University of Science and Technology in Aalesund, Aalesund, Norway

ABSTRACT

As a typical unsupervised learning method, the TextRank algorithm performs well for large-scale text mining, especially for automatic summarization or keyword extraction. However, TextRank only considers the similarities between sentences in the processes of automatic summarization and neglects information about text structure and context. To overcome these shortcomings, the authors propose an improved highly-scalable method, called iTextRank. When building a TextRank graph in their new method, the authors compute sentence similarities and adjust the weights of nodes by considering statistical and linguistic features, such as similarities in titles, paragraph structures, special sentences, sentence positions and lengths. Their analysis shows that the time complexity of iTextRank is comparable with TextRank. More importantly, two experiments show that iTextRank has a higher accuracy and lower recall rate than TextRank, and it is as effective as several popular online automatic summarization systems.

KEYWORDS

Automatic Summarization, High Performance Computing, Similarity Calculation, Text Mining, TextRank, Unsupervised Method

1. INTRODUCTION

It is commonly agreed that we are in the era of big data (Wang et al. 2015). Among various types of data, texts are the most common and pervasive all over the network. Although many effective technologies such as distributive or parallel computations have been proposed, e.g., MapReduce (Slagter et al. 2013; Salgter, et al. 2015; Salgter, et al. 2015), the information overload problem is getting worse as the quantity of data keep increasing rapidly. Automatic text summarization arises as an effective technology for producing a concise and fluent summary conveying the key information in the original text document (Nenkova & McKeown, 2012). Currently, high performance automatic summarization has already become a very important topic in the area of machine learning and data mining, and it is widely used in a large number of industrial sectors, especially in search engines such as Google, Baidu, Yahoo and news portals such as BBC, CNN and NBC News. Many researchers have developed various word-based, sentence-based and graph-based summarization methods. Among them, graph-based methods have attracted a lot of attentions. For example, Ferreira et al. (2013) proposed a four-dimension (including similarity, semantic similarity, co-reference and discourse
information) graph model by taking co-reference resolution and the role of pronouns in connecting the sentences into consideration. See (Gupta & Lehal, 2010) and (Joshi & Sonawane, 2015) for more detailed surveys of extractive summarization techniques and graph-based methods.

Among various graph-based summarization methods, the highly-cited TextRank (Mihalcea & Tarau, 2004) algorithm is a typical unsupervised extractive one, which was originally inspired by Google's famous PageRank (Page, 1998) by taking the similarities between sentences as a type of recommendation or voting to build the corresponding graphs. The output summary will then consist of those selected sentences that are ranked highly according to their converged weights after a number of iterative calculation steps. Based on the original proposal of TextRank, some further improved algorithms for single-document summarization have been proposed. For example, Barrera (2012) introduced an improved algorithm by combining syntactic, semantic and statistical methodologies. Barrios (2015) proposed new alternatives to the similarity functions for computing the distances between sentences in TextRank, including longest common substring and cosine distance. Several researchers also applied TextRank in the fields of keyword extractions. For example, Wan (2012) proposed a novel iterative reinforcement approach to simultaneously extracting summary and keywords by fusing structural information about articles, including word-to-word, sentence-to-sentence and word-to-sentence. Li (2012) introduced tag information into TextRank to estimate the weights of edge and term importance for Chinese documents. Palomino (2011) presented a comparison of three unsupervised algorithms for keyword extractions with respect to Belga News Archive and showed that TextRank was the most successful one compared with the other two algorithms through information radius and chi-square test. Wu (2015) also proposed an unsupervised method named TR-LDA for summarizing microblog by cascading two key-bigram extractors based on TextRank and Latent Dirichlet Allocation (LDA), where two sentence ranking strategies were used based on the key-bigram sets. Some successful applications about TextRank includes multi-document summarization (Wan, 2007) and information retrieval (Blanco & Lioma, 2007; Blanco & Lioma, 2012), which all achieved good performance results. In summary, the above-mentioned works mainly use statistical information (such as term frequency or co-occurrence) and similarity functions (such as mutual information, cosine distance, Dice coefficient, Pearson's χ² statistic) to calculate the similarities between terms or sentences and build the corresponding unweighted or weighted TextRank graphs. They didn't consider linguistic information about the texts, especial the structures and semantics, which often play an important role in many supervised methods.

In this paper, we propose an improved TextRank method considering structural and semantic information of the input text, such as sentence-to-title similarities, paragraph structures, special sentences, sentence positions and lengths. Our analysis shows that the time complexity of iTTextRank is comparable with TextRank. More importantly, two experiments show that iTTextRank has a higher accuracy and lower recall rate than TextRank, and it is as effective as several popular online automatic summarization systems.

The organization of the rest of this paper is as follows: Section 2 briefly introduces the TextRank algorithm. Section 3 presents the steps of building a weighted TextRank graph in details. In Section 4, we present our improved method iTTextRank and analyze its time complexity. Section 5 describes our two experiments to show the effectiveness and efficiency of iTTextRank. Section 6 concludes the paper and points out some future directions.
Applying Security to a Big Stream Cloud Architecture for the Internet of Things
www.igi-global.com/article/applying-security-to-a-big-stream-cloud-architecture-for-the-internet-of-things/143902?camid=4v1a

Event Processing in Web Service Runtime Environments
www.igi-global.com/chapter/event-processing-web-service-runtime/44404?camid=4v1a