Information Processing and Management 48 (2012) 767–778

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Mutual-reinforcement document summarization using embedded graph based sentence clustering for storytelling

Zhengchen Zhang [a,b], Shuzhi Sam Ge [a,b,c,*], Hongsheng He [a,b]

[a] Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117576, Singapore
[b] Social Robotics Lab, Interactive Digital Media Institute, National University of Singapore, Singapore 119613, Singapore
[c] Robotics Institute and School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, PR China

ARTICLE INFO

Article history:
Received 15 March 2011
Received in revised form 22 October 2011
Accepted 21 December 2011
Available online 31 January 2012

Keywords:
Document summarization
Sentence ranking
Space embedding
Sentence clustering
Storytelling

ABSTRACT

In this paper, a document summarization framework for storytelling is proposed to extract essential sentences from a document by exploiting the mutual effects between terms, sentences and clusters. There are three phrases in the framework: document modeling, sentence clustering and sentence ranking. The story document is modeled by a weighted graph with vertexes that represent sentences of the document. The sentences are clustered into different groups to find the latent topics in the story. To alleviate the influence of unrelated sentences in clustering, an embedding process is employed to optimize the document model. The sentences are then ranked according to the mutual effect between terms, sentence as well as clusters, and high-ranked sentences are selected to comprise the summarization of the document. The experimental results on the Document Understanding Conference (DUC) data sets demonstrate the effectiveness of the proposed method in document summarization. The results also show that the embedding process for sentence clustering render the system more robust with respect to different cluster numbers.

1. Introduction

Storytelling is an ancient and universal human method for education and entertainment. Recently much research has been down on storytelling by computers or robots. When a story is very long, it is necessary to provide a concise version for preview. Document summarization can generate a summary that contains the most important points of a story, which has been applied to many specific domains including biomedical Ling et al. (2007), email threads summarization Zajic et al. (2008) and patent document analysis Tseng et al. (2007). This technology may also benefit text processing such as document classification Shen et al. (2004) and question answering Demner-Fushman and Lin (2006).

The research of document summarization has been conducted in two parallel streams: abstract based summarization and extract based summarization. Abstract based summarization employs new words, phrases and sentences to express the same semantic meaning of the original document, whereas extract based method selects existing sentences from the document. The abstract based method is more expressive than the extract based method Knight and Marcu (2002). However, it depends on the natural language generation technology which is still immature. In general, the task of document summarization covers generic summarization and query-oriented summarization. The query-oriented method generates summaries of documents according to given queries or topics, and the generic method summarizes the overall sense of the document without any additional information.

* Corresponding author at: Robotics Institute and School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, PR China. Tel.: +65 6516 6821; fax: +65 6779 1103.

E-mail addresses: zhangzhengchen@nus.edu.sg (Z. Zhang), samge@nus.edu.sg (S.S. Ge), hongshenghe@nus.edu.sg (H. He).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.12.006

Extracting central sentences from a document can be formulated as a sentence ranking problem. In Erkan and Radev (2004), a graph for the document is constructed where the vertexes of the graph denote sentences in the document, and the weight of each edge in the graph is the similarity between two connected sentences. To calculate the similarity, each sentence is represented by a vector, and the value of each word in the vector is the product of the number of occurrences of the word in the sentence and the inverse document frequency of the word. The PageRank algorithm is implemented to rank all the sentences based on the graph model. In this method, the sentences that are similar to other important sentences will obtain high scores. The similarity can be considered as the mutual-effect between sentences. The iterative ranking methods considering mutual-reinforcement between different items have also been proved to be effective for document summarization Wan et al. (2007), Wei et al. (2008), and Zha (2002). A mutual reinforcement method that considers relationship of terms and sentences was proposed in Zha (2002). In this work, the saliency score of a term is determined by the sentences that it appears in, and the saliency score of a sentence is determined by the terms in it. The sentences in the document are partitioned into clusters and the ranking method is performed within each cluster. The algorithm is extended in Wan et al. (2007) where the homogeneous relationship between words, the homogeneous relationship between sentences, and the heterogeneous relationship between words and sentences are all taken into account for calculating the scores of sentences. Moreover, three granularities were considered in Wei et al. (2008) including document, sentence and term.

It has been proved that the sentences of a set of documents can be clustered into different groups, which were not considered in the above methods, to represent subtopics of the documents Harabagiu and Lacatusu (2010). In a long story, a lot of subtopics are contained and these information should be counted in the summarization. In this paper, we propose a framework which considers the mutual effect between clusters, sentences and terms instead of the relationship between documents, sentences, and terms to employ the cluster level information and the latent theme information in the clusters for document summarization. A matrix and a weighted undirected graph model are first constructed for a document, where column vectors of the matrix and the vertexes of the graph represent sentences of the document. The sentences are then clustered into different groups according to the distance between two sentences. In order to alleviate the influence of unrelated sentences in sentence clustering, we employ an embedding process to optimize the graph vertexes. A mutual-reinforcement algorithm is performed at last to calculate the ranks of all sentences. The sentences with high rankings are extracted as the summarization. The contributions of this paper are summarized as follows:

(i) an embedded graph based sentence clustering method is proposed for sentence grouping of a document, which is robust with respect to different cluster numbers;
(ii) an iterative ranking method is presented which considers the mutual-reinforcement between terms, sentences and sentence clusters; and
(iii) a document summarization framework considering sentence cluster information is proposed and the framework is evaluated using DUC data sets.

The remainder of this paper is organized as follows: Section 2 describes related work regarding document summarization including matrix factorization methods, graph ranking methods and machine learning methods. The proposed mutual-reinforcement ranking algorithm is presented in Section 3. Section 4 describes the experiments on DUC corpora and analyzes the influence on system performance of different parameters. We conclude our work in Section 5.

2. Related work

There are many methods to summarize documents by finding topics of the document first and scoring the individual sentences with respect to the topics. The centroid-based summarization method Radev et al. (2004) employs the vectors of statistically important words that are called centroid to express topics of a document cluster. The value of each word in the centroid is calculated by $TF \times IDF$, where TF is the term frequency, and IDF indicates the Invert Document Frequency. The weight of a sentence is the sum of the centroid value of words in the sentence. The sentence position is also considered and the first sentence in a document is assigned the highest weight. The topic signature Lin and Hovy (2000) is a more complex feature for representing the topics of documents. It is composed by a set of weighted terms including words, bigrams and trigrams highly correlated to the target topic. The score of a sentence is the sum of all the scores of content-bearing terms in the sentence. Seven representations of topics and eight different methods of generating multi-document summaries were presented in Harabagiu and Lacatusu (2010), where more features such as main verbs and their arguments were employed to represent the document topics.

Some methods store the lexical information into a matrix and employ matrix factorization technology to rank sentences instead of employing linguistic knowledge and statistical approach to summarize documents. Latent semantic analysis was introduced into document summarization in Gong and Liu (2001), where the authors constructed a matrix D for the given document. Each column of the matrix was the weighted term-frequency vector of a sentence. Singular value decomposition (SVD) $D = U\Sigma V^T$ was performed to the document matrix, which projects each column vector in matrix D to the column vectors of matrix V^T. From semantic point of view, each column of V^T represents a sentence of a document, and each row of V^T (the singular vector) denotes a set of word combination. The item V_{ij} denotes the importance of sentence j in the word combination i. The sentences with highest values corresponding to the word combinations were selected as the summary.