Expert Systems with Applications 41 (2014) 5780–5787

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A multi-document summarization system based on statistics and linguistic treatment

Rafael Ferreira [a,b,*], Luciano de Souza Cabral [a], Frederico Freitas [a], Rafael Dueire Lins [a], Gabriel de França Silva [a], Steven J. Simske [c], Luciano Favaro [d]

[a] Informatics Center, Federal University of Pernambuco, Recife, Pernambuco, Brazil
[b] Department of Statistics and Informatics, Federal Rural University of Pernambuco, Recife, Pernambuco, Brazil
[c] Hewlett-Packard Labs., Fort Collins, CO 80528, USA
[d] Hewlett-Packard Brazil, Barueri, São Paulo, Brazil

ARTICLE INFO

Keywords:
Multi-document summarization
Extractive summarization
Sentence clustering

ABSTRACT

The massive quantity of data available today in the Internet has reached such a huge volume that it has become humanly unfeasible to efficiently sieve useful information from it. One solution to this problem is offered by using text summarization techniques. Text summarization, the process of automatically creating a shorter version of one or more text documents, is an important way of finding relevant information in large text libraries or in the Internet. This paper presents a multi-document summarization system that concisely extracts the main aspects of a set of documents, trying to avoid the typical problems of this type of summarization: information redundancy and diversity. Such a purpose is achieved through a new sentence clustering algorithm based on a graph model that makes use of statistic similarities and linguistic treatment. The DUC 2002 dataset was used to assess the performance of the proposed system, surpassing DUC competitors by a 50% margin of f-measure, in the best case.

1. Introduction

According to Kunder (2013), the estimated size of the web in 2013 was around 3.82 billion pages. This number grows every day at a fast pace, particularly regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.). Thus, it has become humanly unfeasible to efficiently sieve useful information from such a huge mass of documents. Automatic methods are needed to process the Internet data efficiently, scavenging useful information from it.

Text summarization (Wang, Li, Wang, & Deng, 2010) (TS) is a method that aims to create a compressed version of one or more documents, extracting the essential information from them. In other words, the goal of a summary is to present the main ideas in a document in less space (Radev, Hovy, & McKeown, 2002). A TS system is supposed to (i) identify the relevant contents from texts; (ii) eliminate redundant information; and (iii) keep a high level of coverage (He, Qin, & Liu, 2012) of contents.

TS can also be classified according to the number of documents simultaneously analyzed as *single* and *multi-document* summarization (Nenkova & McKeown, 2012). Multi-document summarization addresses the problems of text overload, as many documents share similar topics (Atkinson & Munoz, 2013).

In general, multi-document summarization is either *generic* (also termed *extractive*) (Alguliev, Aliguliyev, & Hajirahimova, 2012a) or *query-based* (Luo, Zhuang, He, & Shi, 2013). Generic summarization systems extract the main ideas from a text collection, while query-based ones select sentences related to a specific query performed by the user.

The same techniques used in single document summarization systems apply to multi-document ones; in multi-document summarization some issues as the degree of redundancy and information diversity increase, however. In a collection of texts on the same subject or a single topic (or a few topics), the probability of finding similar sentences is significantly higher than the degree of redundancy within a single text. Hence, anti-redundancy methods are crucial in multi-document summarization (Atkinson & Munoz, 2013). This issue is a well-known problem in TS: a good summary should avoid repeated information. Redundancy may

* Corresponding author at: Department of Statistics and Informatics, Federal Rural University of Pernambuco, Recife, Pernambuco, Brazil. Tel.: +55 8197885665.
E-mail addresses: rflm@cin.ufpe.br (R. Ferreira), lscabral@gmail.com (L. de Souza Cabral), fred@cin.ufpe.br (F. Freitas), rdl@cin.ufpe.br (R.D. Lins), gfps.cin@gmail.com (G. de França Silva), steven.simske@hp.com (S.J. Simske), luciano.favaro@hp.com (L. Favaro).

http://dx.doi.org/10.1016/j.eswa.2014.03.023
0957-4174/

be perceived as a kind of "noise" that affects the quality of the final summary. On the other hand, summaries are supposed to encapsulate the maximum amount of information from texts (Goldstein, Mittal, Carbonell, & Kantrowitz, 2000) making possible the understanding of the main ideas from the original texts.

This paper proposes a new sentence clustering algorithm to deal with the redundancy and information diversity problems. The central assumption is that building a joint model of sentences and connections yields a better model to identify diversity among them (Cohn, Verma, & Pfleger, 2006). Based on that, the proposed algorithm uses the text representation proposed in Ferreira et al. (2013) to convert the text into a graph model containing four types of relations between sentences: (i) similarity statistics; (ii) semantic similarity; (iii) co-reference; and (iv) discourse relations. Such representation encapsulates the traditional approaches found in state of art systems (statistics similarity and semantic similarity) and linguistic treatment that improve the performance of redundancy elimination (Lloret & Palomar, 2013).

The proposed algorithm works as follows:

1. It converts the text into a graph model.
2. It identifies the main sentences from graph using Text Rank (Mihalcea & Tarau, 2004).
3. It groups the sentences based on the similarity between them.

The DUC 2002 conference dataset (NIST, 2002) was used to evaluate the algorithm presented, assessing it against the systems submitted to that conference. Two different experiments were conducted following the DUC 2002 guidelines: for each collection of documents, summaries with 200 words (first task) and 400 words (second task) were generated. The proposed system achieves Results 50% (first task) and 2% (second task) better than its competitors in terms of f-measure (Baeza-Yates & Ribeiro-Neto, 1999). It is important to stress that the DUC 2002 dataset is still the most widespread benchmark used today for multi-document summarization analysis and that to the best of the knowledge of the authors of this paper no other summarization system surpassed the performance figures published in NIST (2002).

The rest of this paper is organized as follows: Section 2 describes the main related work. Section 3 introduces the proposed system, its architecture and implementation. Section 4 presents an evaluation using the DUC 2002 conference dataset. Finally, some conclusions and discussion of lines for further work are presented in Section 5.

2. Related work

As already mentioned, multi-document summarization can be classified into *generic* and *query based* summarization. Currently, query based summarization has drawn a higher degree of interest in the community, due to its immediate applicability in commercial systems such as automatic customer services.

Traditional methods rely on statistics to create summaries. For instance, PRCN (Luo et al., 2013) is statistical framework to find *relevance*, *coverage* and *novelty* in multi-document summarization. It applies probabilistic latent semantic analysis (Hofmann, 1999) and probabilistic hyperlink-induced topic search (Cohn & Chang, 2000). PRCN attains good results regarding relevance (i.e., the identification of the key ideas of the text) and coverage (dealing with redundancy by excluding similar sentences). Canhasi and Kononenko (2014) models the query and the documents as a graph in order to increase the variability and diversity of the produced query-focused summary. It uses terms, sentences and documents as sets of vertices and the similarities among them as edges. The clusters are built based on the weight of the edges. Both works (Canhasi & Kononenko, 2014; Luo et al., 2013) are only suitable for query-based multi-document summarization.

Another example of this kind of summarizer is presented by Gupta and Siddiqui (2012). It combines single document summaries using sentence clustering techniques to generate multi-document summaries. It works as follows: (i) First, it creates a single document summary (using sentence scoring method); (ii) Then, it clusters the sentences using both syntactic and semantic similarities among sentences to represent the parts of the texts to be introduced in the summary; (iii) Finally, it generates the summary by extracting a single sentence from each cluster.

Canhasi and Kononenko (2014) proposed incorporating graphs to represent terms, sentences, documents and a query to improve coverage in query-based multi-document summarization. Goldstein et al. (2000) tries to minimize redundancy and to maximize both relevance and diversity. It first segments the documents into passages, and indexes them using inverted indices. After identifying the text passages which are relevant to the query using the cosine similarity, a number (depending on the compression rate) of sentences is selected. Finally, it reassembles the selected sentences into the final summary.

In the context of generic summarization, some systems must be highlighted. Radev, Jing, Stys, and Tam (2004) uses a cluster-based method to determine the relevance of sentences eliminating redundancy. His approach employs sentence scoring methods to select sentences for the summarization, achieving good results in redundancy detection. DESAMC + DocSum (Alguliev, Aliguliyev, & Isazade, 2012b) relies on genetic algorithms to create a summary taking into account several aspects of the text, namely *relevance*, *information coverage*, *diversity* and *length limit*. Atkinson and Munoz (2013) combines discourse-level knowledge and corpus-based semantic analysis to create summaries. Atkinson and Munoz (2013) claims that by employing rhetorical knowledge one obtains better quality summaries. The aforementioned approaches present valid contributions to the field and display good performance, in general. Chen, Jin, and Zhao (2014) uses a two-layer graph structure model to summarize documents. It uses the concept of a *phrase* as *related words* that appear together in a sentence. For example, "semitic western religion" and "Christian Philosophy" convey the same basic idea. The similarities among sentences are measured using cosine and co-occurrence of phrases similarities. The HITS algorithm (Wan, 2008) is used to identify the relevant sentences. All of these techniques use supervised learning, requiring a pre-annotated dataset, however.

Other works in generic summarization apply clustering methods to achieve larger information diversity, eliminating redundancy. Alguliev and his collaborators (Alguliev, Aliguliyev, & Mehdiyev, 2013) propose a generic document summarization method which is based on sentence-clustering. In their approach, sentences are represented as a *bag of words* and statistical and semantic similarities measure the dissimilarity among sentences. Such similarities are not combined. The authors use only one kind of similarity to perform the clustering process. A ranking-based sentence clustering framework is reported in Yang, Cai, Zhang, and Shi (2014). Differently of all previous work, it uses the information in documents, sentences and words to create clusters. In addition, that work proposes two different ranking functions (*simple* and *authority* ranking) to extract the main sentences from each cluster.

3. A new multi-document summarization algorithm

The multi-document summarization system proposed in this paper is based on statistical methods and linguistic treatment to increase information diversity of summaries also dealing with