Expert Systems with Applications 41 (2014) 535–543

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization

Ercan Canhasi *, Igor Kononenko

Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Query-focused document summarization
Weighted archetypal analysis
Multi-element graph
Matrix factorization

ABSTRACT

Most existing research on applying the matrix factorization approaches to query-focused multi-document summarization (Q-MDS) explores either soft/hard clustering or low rank approximation methods. We employ a different kind of matrix factorization method, namely weighted archetypal analysis (wAA) to Q-MDS. In query-focused summarization, given a graph representation of a set of sentences weighted by similarity to the given query, positively and/or negatively salient sentences are values on the weighted data set boundary. We choose to use wAA to compute these extreme values, archetypes, and hence to estimate the importance of sentences in target documents set. We investigate the impact of using the multi-element graph model for query focused summarization via wAA. We conducted experiments on the data of document understanding conference (DUC) 2005 and 2006. Experimental results evidence the improvement of the proposed approach over other closely related methods and many of state-of-the-art systems.

1. Introduction

Document summarization is an automatic procedure aimed at producing a generic or a query-focused compressed summary of a document or a set of documents, sharing the same or similar topics, by reducing the document(s) in length. According to the number of documents to be summarized, the summary can be a single document or a multi-document. Single-document summarization can only distill one document into a shorter version, while on the contrary, multi-document summarization can compress a set of documents. Multi-document summarization can be seen as an enhancement of single-document summarization and can be used for outlining the information contained in a cluster of documents. Since multi-document summarization combines and integrates the information across documents, it performs data synthesis and data mining. Based on the purpose, the summaries can be categorized into generic and query-based summaries.

Query-focused multi-document summarization is a special case of multi-document summarization. Given a query, the task is to produce a summary which can respond to the information required by the query. Different from generic summarization, which needs to preserve the typical semantic essence of the original document(s) (Mani, 1991; Ricardo & Berthier, 1999), query-focused summarization purposely demands the most typical (archetypal) summary biased toward an explicit query.

The continuing growth of available online text documents makes research and applications of query-focused document summarization very important and consequently attracts many researchers. Since it can produce brief information corresponding to the users queries, it can be applied to various tasks for satisfying different user interests. Queries are mostly real-world complex questions (e.g., "Track the spread of the West Nile virus through the United States." is a query example). Such complicated questions make the query focused summarization task quite difficult. The real problem is how to model the question jointly with the documents to be summarized and thus bias the answer, i.e. summary, towards the provided question.

Most existing research on applying matrix factorization approaches to Q-MDS explores either low rank approximation or soft/hard clustering methods. The former have a great degree of flexibility but the features can be harder to interpret. The latter extract features that are similar to actual data, making the results easier to interpret, but the binary assignments reduce flexibility. These techniques can be jointly seen as a factor analysis description of input data exposed to different constraints. Inadequately, most of these methods does not directly incorporate the query information into summarization process, thus the summarization is general about the document collection itself. Moreover, most existing works assume that documents related to the query only talk about one topic. Even thought query-focused summarization, by its definition, is biased toward a given query, in our understanding it does not mean

* Corresponding author. Tel.: +37745542501.
E-mail addresses: ercan.canhasi@uni-prizren.com (E. Canhasi), igor.kononenko@fri.uni-lj.si (I. Kononenko).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.079

that the produced summary should not show the diversity in content as much as possible.

In the paper, we try to overcome limitations of the existing algebraic methods and study a new setup of the problem of query-focused summarization. Since the archetypal analysis completely assembles the advantages of clustering and the flexibility of matrix factorization we propose using the AA in Q-MDS. Consequently, the main concerns of the paper are: (1) how to incorporate query information in its own nature of an archetypal analysis based summarizer; and (2) how to increase the variability and diversity of the produced query-focused summary. For the first concern, we propose a weighted version of archetypal analysis based summarizer able to directly use the query information. The second one is answered by the nature of the archetypal analysis itself, which clusters the sentences into distinct archetypes.

The main contributions of the paper are three-fold:

1. A novel query-focused summarization method wAASum is proposed.
2. Modeling the input documents and query information as a multi-element graph is introduced.
3. The effectiveness of the proposed approach is examined in the context of Q-MDS.

To show the efficiency of the proposed approach, we compare it to other closely related summarization methods. We have used the DUC2005 and DUC2006 data sets to test our proposed method empirically. Experimental results show that our approach significantly outperforms the baseline summarization methods and the most of the state-of-the-art approaches.

The remainder of the paper is organized as follows: Section 2 describes related work regarding document summarization and the archetypal analysis. In Section 3 the weighted archetypal analysis is introduced, whereas Section 4 presents the multi-element graph modeling. The details of the proposed summarization approach wAASum are presented in Section 5, where we give an overview of the new approach and an illustrative example of its use. Section 6 shows the evaluation and experimental results. Finally, we conclude in Section 7.

2. Related work

2.1. Query-focused multi-document summarization

Recently, algebraic methods, more precisely matrix factorization approaches, have become an important tool for query/topic focused document summarization. The exemplary methods used until now vary from low rank approximations, such as singular value decomposition (SVD) (Arora & Ravindran, 2008), latent semantic indexing (LSI/LSA) (Li, Li, & Wu, 2006; Yeh, Ke, Yang, & Meng, 2005), non-negative matrix factorization (NMF) (Lee, Park, Ahn, & Kim, 2009; Park, Lee, Ahn, Hong, & Chun, 2006) and symmetric-NMF (Wang, Li, Zhu, & Ding, 2008) to soft clustering approaches such as fuzzy K-medoids (Mei & Chen, 2012) and hard assignment clustering methods such as K-means (Wang et al., 2008). The graph based methods can also be categorized as a factorization methods since they are based on eigendecomposition which is closely related to the SVD.

Graph-based methods like LexRank (Erkan & Radev, 2004) and TextRank (Mihalcea. & Tarau, 2004) model a document or a set of documents as a text similarity graph, constructed by taking sentences as vertices and the similarity between sentences as edge weights. They take into account the global information and recursively calculate the sentence significance from the entire text graph rather than simply relying on unconnected individual sentences. Graph-based ranking algorithms were also used in query-focused summarization when it became a popular research topic. For instance, a topic-sensitive version of LexRank is proposed by Otterbacher, Erkan, and Radev (2009). It integrates the relevance of a sentence to the query into LexRank to get a biased PageRank ranking. Although the algorithm is proposed for sentence ranking in the context of question-focused sentence retrieval, it can be directly used for sentence ranking in the task of query-focused summarization. The recently proposed document-sensitive graph model (Wei, Li, Lu, & He, 2010) that emphasizes the influence of global document set information on local sentence evaluation, is shown to perform better than other graph models for multi-document summarization task where MDS is modeled as single combined document summarization.

Latent Semantic Analysis (LSA) is an approach to overcome problems of multiple theme coverage in summaries by mapping documents to a latent semantic space, and has been shown to work well for text summarization. The Q-MDS method using LSA applies singular value decomposition (SVD) to summarize documents. The method factorizes term-document matrix into three matrices, U, D, and V. Starting from the first row of V^T, the sentence corresponding to the column that has the largest index value with the right singular vector is selected to the next stage (Gong & Liu, 2001; Yeh et al., 2005). Then a query focus from a topic description is derived to be used for guiding the sentence selection (Li et al., 2006). However, LSA has a number of drawbacks, due to its unsatisfactory statistical foundations.

Lee et al. (2009) proposed a query based summarization method using NMF . This method is yet another successful algebraic method, which extracts sentences using the cosine similarity between a query and semantic features. This type of methods conduct NMF on the term-sentence matrix to extract sentences with the highest probability in each topic. Intuitively, the method clusters the sentences and chooses the most representative ones from each cluster to form the summary. NMF selects more meaningful sentences than the LSA-related methods, because it can use more intuitively interpretable semantic features and is better at grasping the innate structure of documents. As such, it provides superior representation of the subtopics of documents. Park et al. (2006) also proposed a query based summarization method using NMF. This method extracts sentences using the cosine similarity between a query and semantic features. However, the method might produce poor document summarization in the case that initial user query does not reflect the user's requirements.

The SNMF summarization framework for query focused summarization, as an extension by Lee et al. (2009), is based on sentence level semantic analysis (SLSS) and symmetric non-negative matrix factorization SSNF. SLSS can better capture the relationships between sentences in a semantic manner and SSNF can factorize the similarity matrix to obtain meaningful groups of sentences. However SNMF is unable to define closeness to the cluster center and closeness to the sentences in the same cluster, therefore it is incapable of considering both in defining the subtopic-based features.

A fuzzy medoid-based clustering approach, as presented by Mei and Chen (2012) is an example of soft clustering methods for Q-MDS. It is successfully employed to generate subsets of sentences where each of them corresponds to a subtopic of the related topic. The subtopic-based feature captures the relevance of each sentence within different subtopics and thus enhances the chance of producing a summary with a wider coverage and less redundancy.

A method, called MCLR (maximum coverage and less redundancy) (Alguliev, Aliguliyev, & Hajirahimova, 2012) models multi-document summarization as a quadratic boolean programming problem where objective function is a weighted combination of