THE ROLE OF TEXT SUMMARIZATION IN DIGITAL LIBRARIES

Róbert MÓRO, Mária BIELIKOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{moro,bielik}@fiit.stuba.sk

Abstract. Automatic text summarization plays important role in helping users to assess relevancy of the documents on the Web, or more specifically in the domain of digital libraries. We proposed a method of personalized text summarization which improves the conventional automatic text summarization methods by taking into account terms relevant in the domain as well as interests of the users reflected by what fragments of text they annotate. We experimentally evaluated the proposed method, obtaining better summaries compared to generic variant.

1. Introduction

The aim of automatic text summarization is to provide users short and concise summaries of the most important information contained in the documents which then serve as surrogates that help users to quickly assess the documents' relevancy. It plays, therefore, a crucial role in the search systems and during navigation sessions by speeding the whole process up and diminishing the overall information load. Although most of the research in the field of text summarization concerns with generic approaches, the existing works on personalization of summarization suggest that the users can significantly benefit when their interests [4], goals or context are taken into account. The reason is that the users are more interested in finding out, how the document matches their information need and less in having a general overview of all its topics.

Traditionally, the articles' abstracts play the role of summaries in the domain of digital libraries. Because they are created manually, they have better quality than automatic summaries, however, they usually provide only one perspective (from the author's point of view) and lack any means of personalization. In order to remedy the former, the *citation summaries* were proposed [7]. It is a specific type of summaries consisting of the sentences

citing the given article. The idea is that citation sentences convey information deemed relevant or interesting by researchers other than its authors, thus providing more perspectives on the given article. *Impact-based summarization* also utilizes the citation sentences [6]; in this case in order to identify relevant passages in the original article which are then summarized. However, neither of the discussed approaches that uses citations as a source of metadata, considers personalization of the resulting summaries, thus leaving room for improvement.

2. Method of personalized text summarization

We proposed a method of personalized summarization [1] that is able to take into account information from various sources and modify the relevancies of the documents' terms in order to extract information that is the most important or most interesting for the user (in the given context). It is based on the method of latent semantic analysis [5, 8].

The method consists of the following steps: The document is pre-processed; the terms are extracted from the document and the document's text is segmented to sentences during this step. Next we construct a matrix of terms and sentences which represents an input to singular value decomposition (SVD). Finally, the sentences are selected based on the matrices outputted from SVD using the approach proposed in [8].

It is the step of matrix construction that we identified as suitable for extension in order to personalize the summaries. We construct a personalized matrix of terms and sentences using our proposed weighting scheme which extends the conventional weighting scheme based on TF-IDF method by linear combination of multiple raters [1, 2]:

$$w(t_{ij}) = \sum_k \alpha_k R_k(t_{ij}), \tag{1}$$

where $w(t_{ij})$ is a weight of a term t_{ij} in the matrix and α_k is a linear coefficient of a rater R_k. Both the weights $w(t_{ij})$ and the linear coefficients α_k can be any real number. The rater R_k is a function, which assigns each term from the extracted keywords set T its weight. We proposed a set of generic and personalized raters, the most prominent being *relevant terms rater* and *annotations rater*. The former assigns higher relevancy to the terms identified as relevant generally in the domain [9] or specifically for the user's interests [3] or current context. The latter utilizes annotations added by users, e.g. highlights. It is based on an assumption that users annotate parts of the documents which are interesting or important for them.

3. Evaluation and conclusions

We evaluated our proposed method in the domain of learning on the dataset of around 300 learning objects (educational texts) from the web-based learning system ALEF[1]. We conducted two experiments: in the first one we compared summaries considering the relevant domain terms to the generic variant [1]; in the second we evaluated summaries considering the highlights of the users [2]. In order to evaluate the quality of the summary variants we used an expert group of 5 people. They were presented with two different summary variants at the same time (without knowing which method was used to generate which one) together with the original document and their task was to decide which variants was better.

[1] http://alef.fiit.stuba.sk

Both variants (i.e. the one considering the relevant domain terms as well as the one considering the user-added highlights) achieved better results in comparison with the generic one: they were judged as better or the same 69% of time in the first case and 77% in the second. Moreover, the summaries that considered not only general domain knowledge, but were personalized to the interests of the users, were overall judged better.

Although we evaluated the method in the domain of learning, it should be sufficiently general and extensible to work in the domain of digital libraries of research articles as well. Nevertheless, the latter has its specifics that have to be accounted for, especially during preprocessing, such as documents' length or presence of non-textual data. On the other hand, we can use knowledge of the documents' structure in order to adapt weights of the terms and utilize other types of metadata, such as citations.

Acknowledgement: This work was partially supported by the grants VG1/0971/11/, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

to other papers publishing the results that are summarized here

[1] Móro, R., Bieliková, M.: Personalized text summarization based on important terms identification. In: *Proc. of the 23rd Int. Workshop on Database and Expert Systems Applications*, IEEE CS, (2012), pp. 131–135.

[2] Móro, R.: Combinations of different raters for text summarization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, (2012), vol. 4, no. 2, pp. 56–58.

Other references

[3] Barla, M.: Towards social-based user modeling and personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, (2011), vol. 3, no. 1, pp. 52–60.

[4] Díaz, A., Gervás, P.: User-model based personalized summarization. *Information Processing and Management*, vol. 43, no. 6, (2007), pp. 1715–1734.

[5] Gong, Y., Liu, X.: Generic text summarization using relevance measure and latent semantic analysis. In: *SIGIR'01: Proc. of the 24th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, Springer, (2001), pp. 19–25.

[6] Mei, Q., Zhai, C.: Generating impact-based summaries for scientific literature. In: *Proc. of ACL-08*: HLT, (2008), pp. 816–824.

[7] Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: *COLING'08: Proc. of the 22nd Int. Conf. on Computational Linguistics*, Association for Computational Linguistics, (2008), pp. 689–696.

[8] Steinberger, J., Ježek, K.: Text summarization and singular value decomposition. In: *ADVIS'04: Proc. of Advances in Information Systems*, Springer, (2004), pp. 245–254.

[9] Šimko, M.: Automated acquisition of domain model for adaptive collaborative web-based learning. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, (2012), vol. 4, no. 2, pp. 1–9.