Ramiz M. Aliguliyev [1], Makrufa Sh. Hajirahimova [2] DOI: 10.25045/jpit.v06.i2.10
[1,2]Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a.ramiz@science.az, [2]makrufa@science.az

AN OPTIMIZATION MODEL FOR AUTOMATIC TEXT SUMMARIZATION

In this paper, an unsupervised approach to automatic document summarization is proposed. This approach is based on sentence selection. In the proposed approach, sentence selection is modeled as an optimization problem. The model generally attempts to optimize three properties: relevance – summary should contain informative sentences that carry the main topics of the source text; redundancy – summaries should not contain multiple sentences that convey the same information; length – summary is bounded in length.

Keywords: *information overload, text mining, text summarization, redundancy, coverage, optimization model.*

Introduction

The rapid development of information and communication technologies has significantly changed the means of data processing, transfer, and storage. The main requisites (signatures and etc.) that provide the status legal document for the data are also subject to changes. A new form of data (document), which is an electronic document (e-paper), is available now. These types of documents have become the main form of data exchange between business bodies, citizens and authorities. Electronic document management systems (EDMS) have turned into the most important component of emerging electronic government (e-government). This is due to the fact that the clerical activities of government agencies are electronized, and the government departments are linking in the network environment. The state services are delivered online via the Internet to the citizens and the business sector regardless of place and time [1, 2]. Due to the promotion of electronic services a large number of e-documents (citizens' appeals, petitions and complaints, documents submitted by the business sector, correspondence between government agencies, etc.) are circulated and processed in these systems. Taking into account the dynamic and large document flow, most important elements, which ensure intellectualization of the system, facilitate the human labor, and increase the efficiency of the system, should be available in the implementation of EDMS of state authorities. The documents should be analyzed for various purposes for the effective implementation of the main functions of e-government. As the text is the major a carrier of information (80-90%), thus automatic classification of this type documents to be read by the leaders and officials rapidly, acquisition of certain ideas, and making the right decision emerges serious problems. Obviously, text type e-documents are impossible to be analyzed only by clerical EDMSs or common data management systems. Compressed form of text documents, i.e. summarization seems to be more effective in terms of information load caused by excessive volume of data [1, 2]. Since, presenting abbreviated summarization of documents people get familiar with the main contents of the document as soon as possible. Summarization is a shortened version of the document while retaining the main content. Its main problems are as follows:

- Length;
- Detecting informative sentences and thematic sections;
- Redundancy – preventing recurrence of the sentences with the same synopsis;
- Selection of the length proximity
- Summarization is the selection of with more characteristic information, i.e. sentences from the document (or documents) and joining them consistently.

The inclusion of an alternative sentence into the summary causes recurrence of the sentences with the same synopsis, i.e. redundancy, and the mechanisms should be developed to overcome it. Therefore, most candidate sentences are developed taking into account the length of

the given summary, and the best summarization selection strategy is more important than the selection of the best sentences. Comparing to the best sentence selection procedure, selection of the best sentences is a global optimization problem [3]. In connection with the problem solution, we propose an optimization model, which controls expanded coverage and minimum redundancy for automatic text summarization.

Analysis of available summarization methods and algorithms

Availability of the information on the Internet has intensified the studies in the field of automatic text summarization. In [4-9], automatic summarization, its challenges and state-of-the-art has been widely interpreted. Summarization aims at the identification of the informative sections (sentence, paragraph) reflecting the content of the document (or documents). Position, the frequency of keywords, title-keywords, syntactic criteria, the features of the sentence as indicating expressions denote the relevance of each sentence. Various extractive methods proposed so far rank the sentences by their weights and choose only one with the highest weight. Ouyang and et.al. implemented summarization by applying SVR (Support Vector Regression) in sentence-ranking issues [10]. Multilingual MEAD platform was developed for the first time for the document summarization with the extractive method for the document sets (http://www.summarization.com/mead/). MEAD uses to "centroid" features, namely the information in the clusters' centroid for the extraction of the sentences. It calculates three properties for each sentence (centroid value; positional value; first-sentence overlap) in the cluster and determines more important sentence using linear combination of the features. Selection of the sentences is restricted to checking the proximity of the length and new selected sentences and the cosine [11]. Huang and et.al. applied fuzzy hybrid scheme to determine the relevance of the sentences [12].

The process of the selection of various ideas of the document in automatic text summarization is called differently. Diversity is very important for redundancy monitoring and for the development of better text summarization. In [13] diversity based summarization model, i.e. MMR (Maximal Marginal Relevance) is presented. In this approach "greedy" algorithm chooses more relevant sentences and delete the ones that are closer to pre-selected sentences, thus, redundancy shall be relatively prevented. The most important problem of MMR is being non-optimal. In [14] general document summarization model without training is presented with the use of Non-negative Matrix Factorization (NMF). NMF chooses the with more informative sentences, uses better interpreted semantic features and identifies the natural structure of the document rather than Latent Semantic Analysis (LSA) methods. LSA methods present the sentences with the linear combination of the semantic features [15]. In [16] Semantic Sentence-Level Analysis (SLSS) and Symmetric Non-Negative Matrix Factorization (SNMF) based model is proposed based on the. SLSS builds relationships between the sentences using semantic analysis and creates the similarity matrix first. Then, SNMF is used to group the sentences in clusters. Finally, the most informative sentences are selected from each cluster to generate summarization.

[17-20] offer variety of methods based on the summarization graphs theory. These methods present the documents in the form of graphs. LexRank first creates cosine proximity based sentence combination graphs and then, it selects important sentences basing on the eigenvector centralization concept [18]. [21] uses Text Relationship Map (TRM) methodology to remove the paragraphs. Top sentences of the graphs, the weight of the edge shows the level of their proximity. Proposed extractive methods split the documents into paragraph (sentence) sets, and calculate the proximity between them using cosine. Clustering methods based approaches are widely used to improve the summarization quality [4, 22, 23]. These approaches consist of two steps: clustering and ranking. First, the sentences are grouped to identify thematic units, and centroid value of the sentence based on average cosine proximity between the sentence in the

cluster and the other ones is defined. The sentences are ranked by the values, and the top ranked sentences in each cluster are selected as the candidate sentences to summarization. Clustering is the most effective tool to identify the diversity between the sentences [22]. Optimal summarization can be viewed as an optimization problem in extractive document summarization. Because, identification of informative sentences is a matter of optimization essentially. In last decade, optimization based approaches in summarization are studied more intensively [2,3,7,24-30]. The idea of creating optimal summarization was first put forward by Filatova and Hatzivassiloglou in 2004 [26]. They presented the documents in two-dimensional space as text and conceptual units. At the same time, they proposed a formal model to select the key text units and to minimize data replication [26]. Okamura and Takamura presented text summarization as a matter of maximum coverage knapsack problem (MCKP) [27]. Huang and et.al. included four purpose functions (information coverage, importance, redundancy, and text sequences), viewed the summarization as a multi-criteria optimization problem [3]. They used spectral clustering to eliminate redundancy, and classified each sentence into group of sentences with related semantics. Importance of sentences within the document is determined by Markov model. Studies conducted by R.M.Alguliyev and R.M.Aliguliyev in the field of optimal summarization are more attractive. In [2,7,24,29-31] the authors formalized selection of the sentences as a matter of optimization and achieved solution with the use of evolution algorithms. In [30] summarization is modeled as a matter of p-median method-based multi-criteria (relevance, coverage and difference) optimization and achieved solution with the use of adaptive ant algorithms. In [2], the summarization model, which provides maximum coverage and minimum redundancy, is formulated as a quadratic Bull programming problem. The function of the model is presented as a combination of content coverage and weighted redundancy, and binary differential evolution algorithm is developed to solve optimization. The advantage of the model is to provide high diversity of the summary, i.e. to minimize redundancy by deleting sentence overlap while selecting. In [31] summarization is modeled as an integer quadratic linear programming problem, and resolved through herd intelligence based discrete algorithm. The authors propose a mathematical model for the multi-document summarization in [29]. In this approach, they use sentence-document collection, summary-document collection, and sentence-sentence relationship to extract important sentences and to reduce redundancy. Modified differential evolution algorithm is developed to solve optimization problem.

Proposed summarization model

The studies show that the coverage and the quality of the summary are the key criteria. To this end, we propose an optimization model to control redundancy for automatic text summarization.

Description and proximity of sentences. In general, summarization aims at identifying the sentence sets, which convey the main content of the document. In other words, we need to create a summary that would maximize the proximity between the document collection and the summary. Before presenting the model, let's show the document as a set of sentences $D = \{s_i, i = \overline{1,n}\}$. s_i denotes the i-th sentence in D, n- the number of sentences in the document. $T = \{t_1, t_2, ..., t_m\}$ denotes all the words in the document D. The sentences are presented using the known vector model. According to the model, each sentence s_i is described as a characteristic vector containing the words in m-dimensional space, $s_i = \{w_{i1}, w_{i2}, ..., w_{im}\} = \{w_{ij}, i = \overline{1,n}, j = \overline{1,m}\}$. m denotes the number of words in the document, w_{ij} is the weight of the j-th word in the i-th sentence, it is calculated using the model $tf*isf$ (term frequency-inverse sentence frequency):

$$w_{ij} = f_{ij} \times \log(n / n_j),$$

Here, f_{ij} - is the frequency of the t_j word in the sentence s_i, *isf* is the logarithm of the coefficient of the number of all sentences to the sentences where the word t_j is used. n_j denotes - the number of sentences with the terms t_j, $i = \overline{1,n}$, $j = \overline{1,m}$. The focal problem of the model is the large size of the signs space. Most commonly used approaches in space minimization include reducing stop words and defining stemmings. One of the key issues is to identify the proximity between the summarized sentences. Usually, text documents are considered to be similar when their terminological compositions are analogous. Euclidean distance, cosine, Jaccard, Pearson, and Kullback-Leibler divergence are used to define the "similarity" between the text units. Cosine is the most popular proximity measure of text vectors. Cosine measure, which is often used in text analysis, calculates the cosine of the angle between two vectors. The cosine proximity between the sentences $s_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$ and $s_j = \{w_{j1}, w_{j2}, ..., w_{jm}\}$ is calculated as follows:

$$sim(s_i, s_j) = \cos(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \; i, j = 1, ..., n$$

Mathematical statement of the problem. Three properties are usually optimized in summarization: relevance – summary should contain informative sentences that carry the main topics of the source text; redundancy – summaries should not contain multiple sentences that convey the same information; length – summary is bounded in length. Usually, the summary should contain 5 - 30% of initial text depending on demand.

Joint optimization of these three properties is a matter of difficult and global summarization. Containing relevant parts of the text is based not only on their own properties, but also on the properties of the other parts of the text in the summary [25]. Let's assume that each sentence of D set has a chance to be included in the summary. To this end, the following variables should be included

$$x_i = \begin{cases} 1, if\ s_i\ is\ included\ in\ the\ summary \\ 0, \quad otherwise \end{cases}$$

$$x_i = \begin{cases} 1, if\ s_i\ and\ s_j\ are\ included\ in\ the\ summary \\ 0, otherwise \end{cases}$$

Then, text summarization can be formalized as follows:

$$\sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (w_i + w_j) . sim_{ij} . x_{ij} \to \max \tag{1}$$

$$\sum_{i=1}^{n} l_i x_i \le L \tag{2}$$

$$x_i \in \{0, 1\} \tag{3}$$

$$x_{ij} \in \{0, 1\} \tag{4}$$

Here, $w_i -$ denotes importance (weight) of the sentence s_i, sim_{ij} –a proximity measure between the sentences s_i and s_j. The weight w_i can be defined as follows:

$$w_i = \sum_{j=1}^{n} m_j . e^{-\left(-\frac{1}{sim_{ij}}\right)^2} \quad , \; i = 1,...,n$$

here, m – is calculated as follows:

$$m_j = \frac{\sum_{k=1}^{m} w_{jk}}{\sum_{k=1}^{m} \overline{w_k}} \quad , \; j = 1,...,n$$

here $\overline{w_k} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}$, $k = 1,...,m$ – is the coordinates of the centre of sentence sets.

In (2), l_i denotes the length of sentence s_i, and L- the length of the future summary. The length can be defined as the number of words or volume (byte).

Conclusion

With the development of e-government the volume of its information to be processed is growing rapidly, while the great majority of the documents are non-structured texts. Analyzing text documents in a short period of time, and making operational decision becomes very serious problem. Automatic text summarization seems to be most advisable. The key problems of summarization include content coverage and redundancy. For this purpose, document summarization is modeled as a linear optimization problem to minimize redundancy. Proposed model can support officials in decision-making providing intellectualization of EDMS applied in government agencies. It is more appropriate to apply naturally-inspired algorithms (ants, bees, etc.) to solve complex optimization problems. The goal of our further research is to develop algorithms to solve (1-4) optimization problems.

References

1. M.S. Hajirahimova. Actual problems and solutions of electronic document management systems // Information Society Problems, 2010, No2, pp.21-29.
2. Alguliev R.M., Aliguliyev R.M., Hajirahimova M.S. GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy // Expert Systems with Application, 2012, vol.39, no.16, pp.12460–12473.
3. Huang L., He Y., Wei F., Li W. Modeling document summarization as multi-objective optimization / Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, 2010, april 02–04, pp.382–386.
4. Aliguliyev R.M. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization // Computational Intelligence, 2010, vol.26, no.4, pp.420–448.
5. Jones K.S. Automatic summarizing: the state of the art // Information Processing and Management, 2007, vol.43, no.6, pp.1449–1481.
6. Das D., Martins A. F.T.A Survey on Automatic Text Summarization // Language, 2007, no.4, pp.1–31. http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf
7. Alguliev R.M., Aliguliyev R.M., Isazade N.R. MR&MR-SUM: maximum relevance and minimum redundancy document summarization model // İnternational Journal of İnformation Technology & Decision Making, 2013, vol.12, no.3, pp.361–393
8. Tucker R. Automatic summarizing and the CLASP system, PhD thesis, University of Cambridge, UK, 1999, 190 p.

9. Zajic D.M. Mutipe alternative sentence compressions as a tool for automatik summarization task, PhD Thesis, University of Maryland College park,. 2007, 229 p. www.umiacs.umd.edu
10. Ouyang Y., Li W., Li S., Lu Q. Applying regression models to query-focused multi-document summarization // Information Processing & Management, 2011, vol.47, no.2, pp.227–237.
11. Radev D., Jing H., Stys M., Tam D. Centroid-based summarization of multiple documents // Information Processing and Management, 2004, vol.40, no.6, pp.919–938.
12. Huang H.H., Yang H.C., Kuo Y.H. A fuzzy-rough hybrid approach to multi-document extractive summarization / Proceedings of the Ninth International Conference on Hybrid Intelligent Systems, Shenyang, China, 2009, august 12–14, pp.168–173.
13. Carbonell J.G., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries / Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998, august 24–28, pp.335–336.
14. Lee J.H., Park S., Ahn C.M., Kim D. Automatic generic document summarization based on non-negative matrix factorization // Information Processing and Management, 2009, vol.45, no.1, pp.20–34.
15. Gong Y., Liu X. Generic text summarization using relevance measure and latent semantic analysis / Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval, New Orleans, USA, 2001, september 9–12, pp.19–25.
16. Wang D., Li T., Zhu S., Ding C. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization / Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, Singapore, 2008, july 20–24, pp.307–314.
17. Wan X., Xiao J. Graph-based multi-modality learning for topic-focused multi-document summarization / Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09), Pasadena, USA, 2009, july 11–17, pp.1586–1591.
18. Erkan G., Radev D. Lexrank: graph-based centrality as salience in text summarization // Journal of Artificial Intelligence Research, 2004, vol. 22, pp.457–479.
19. Zhang J., Xu H., Cheng X. GSPSummary: a graph-based sub-topic partition algorithm for summarization / Proceedings of the Asia Information Retrieval Symposium, Harbin, China, 2008, january 15–18, pp.321–334.
20. Zhao L., Wu L., Huang X. Using query expansion in graph-based approach for query-focused multi-document summarization // Information Processing and Management, 2009, vol.45, no.1, pp.35–41.
21. Mitra M., Singhal A., Buckley C. Automatic text summarization by paragraph extraction / Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 1997, pp.39–46.
22. Binwahlan M.S., Salim N., Suanmali L. Fuzzy swarm diversity hybrid model for text summarization // Information Processing and Management, 2010, vol.46, no.5, pp.571–588.
23. Nomoto T., Matsumoto Y. The diversity-based approach to open-domain text summarization // Information Processing and Management, 2003, vol.39, no.3, pp.363–389.
24. Alguliev R., Aliguliyev R., Hajirahimova M. Multi-document summarization model based on integer linear programming // Intelligent Control and Automation, 2010, vol.1, no.1, pp.105–111.
25. McDonald R. A study of global inference algorithms in multi-document summarization / Proceedings of the 29th European Conference on IR Research, Rome, Italy, Springer-Verlag, LNCS, 2007, april 2–5, no.25, pp.557–564.
26. Filatova E., Hatzivassiloglou V. A formal model for information selection in multi-sentence text extraction / Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), Geneva, Switzerland, 2004, august 23–27, pp.397–403.

27. Takamura H., Okumura M. Text summarization model based on maximum coverage problem and its variant / Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 2009, march 30–april 3, pp.781–789.
28. Lin J., Madnani N., Dorr B. Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization / Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, USA, 2010, june 1–6, pp.305–308.
29. Alguliev R.M., Aliguliyev R.M., Isazade N.R. Multiple documents summarization based on evolutionary optimization algorithm // Expert Systems with Applications, 2013, vol.40, no.5, pp.1675–1689.
30. Alguliev R.M, Mehdiyev Ch.A. Modeling the document summarization as a modified task of p-median and adaptive ant algorithm for optimization solution // Information Technologies, 2011, No 9, pp. 9-17.
31. Alguliev R.M., Aliguliyev R.M., Isazade N.R. CDDS: Constraint-driven document summarization models // Expert Systems with Applications, 2013, vol.40, no.2, pp.458–465.