Intelligent Control and Automation, **2010, 1, 105-111**
doi:10.4236/ica.2010.12012 Published Online November 2010 (http://www.SciRP.org/journal/ica)

Multi-Document Summarization Model Based on Integer Linear Programming

Rasim Alguliev, Ramiz Aliguliyev, Makrufa Hajirahimova
Institute of Information Technology of National Academy of Sciences of Azerbaijan
E-mail: *a.ramiz@science.az*
Received August 28, 2010; *revised October* 1, 2010; *accepted October* 3, 2010

Abstract

This paper proposes an extractive generic text summarization model that generates summaries by selecting sentences according to their scores. Sentence scores are calculated using their extensive coverage of the main content of the text, and summaries are created by extracting the highest scored sentences from the original document. The model formalized as a multiobjective integer programming problem. An advantage of this model is that it can cover the main content of source (s) and provide less redundancy in the generated summaries. To extract sentences which form a summary with an extensive coverage of the main content of the text and less redundancy, have been used the similarity of sentences to the original document and the similarity between sentences. Performance evaluation is conducted by comparing summarization outputs with manual summaries of DUC2004 dataset. Experiments showed that the proposed approach outperforms the related methods.

Keywords: Multi-Document Summarization, Content Coverage, Less Redundancy, Integer Linear Programming

1. Introduction

With the rapid growth of the Internet and information explosion automatic document summarization has drawn increasing attention in the past. The explosion of electronic documents has made it difficult for users to extract useful information from them, and a lot of relevant and interesting documents are not read by the user due to the large amount of information [1].

The information overload problem can be reduced by text summarization. Automatic document summarization aims to condense the original text into essential content and to assist in filtering and selection of necessary information. Present search engines usually provide a short summary for each retrieved document in order that users can quickly skim through the main content of the page. Therefore it saves users time and improves the search engine's service quality [2]. That is why the necessity of tools that automatically generate summaries arises. These tools are not just for professionals who need to find the information in a short time but also for large searching engines such as Google, Yahoo!, AltaVista, and others, which could obtain a lot of benefits in its results if they use automatic generated summaries. After that, the user only will require the interesting documents, reducing the flow information [1,3].

Depending on the number of documents to be summarized, the summary can be a single-document or a multi-document [4-6]. Single-document summarization can only condense one document into a shorter representation, whereas multi-document summarization can condense a set of documents into a summary. Multidocument summarization can be considered as an extension of single-document summarization and used for precisely describing the information contained in a cluster of documents and facilitate users to understand the document cluster. Since it combines and integrates the information across documents, it performs knowledge synthesis and knowledge discovery, and can be used for knowledge acquisition [5,7].

This paper focuses on the multi-document summarization. It models text summarization task as an optimization problem. This model directly discovers key sentences in the given collection and covers the main content of the original source(s). The model implemented on multi-document summarization task. Experiments on DUC2004 datasets showed that the proposed approach

outperforms the other methods.

The rest of this paper is organized as follows. Section 2 introduces the brief review of the summarization methods. The proposed text summarization model is presented in Section 3. The numerical experiments and results are given in Section 4. Section 5 concludes the paper.

2. Related Work

Document summarization methods can be divided into two categories: abstractive and extractive. In fact majority of researches have been focused on summary extraction, which selects the pieces such as keywords, sentences or even paragraph from the source to generate a summary. A human summarizer typically does not create a summary by extracting textual units verbatim from a source into the summary. Abstraction can be described as reading and understanding the text to recognize its content which is then compiled in a concise text. In general, an abstract can be described as a summary comprising concepts/ideas taken from the source which are then 'reinterpreted' and presented in a different form, whilst an extract is a summary consisting of units of text taken from the source and presented verbatim [4,5,7].

The extractive method proposed in [8] decomposes a document in a set of sentences, using the cosine measure computes the similarity between sentences and they represent the strength of the link between two sentences, and sentences extracted according to different strategies. The centroid method [9] applies MEAD algorithm to extract sentences according to the following three parameters: centroid value, positional value, and first-sentence overlap.

In order to enhance the performance of summarization, recently cluster-based approaches were explored in the literature [1,10-14]. The approaches proposed in [10-14] consist of two steps. First sentences are clustered, and then on each cluster representative sentences are defined. To optimize the objective functions in these works developed the evolutionary algorithms. A reinforcement approach [1] integrates ranking and clustering together by mutually and simultaneously updating each other. In other words, clustering and ranking are regarded as two independent processes in this approach although the cluster-level information was incorporated into the sentence ranking process. MDS [15] uses the minimum dominating set to formalize the sentence extraction for document summarization. Weighted consensus summarization (WCS) method [16] combines the results from single summarization systems. Multi-document summarization by maximizing informative content-words (MICW) [17] first assign a score to each term in the document cluster, using only frequency and position information, and then finds the set of sentences in the document cluster that maximizes the sum of these scores, subject to length constraints. LexPageRank [18] first constructs a sentence connectivity graph based on cosine similarity and then selects important sentences based on the concept of eigenvector centrality.

Filatova and Hatzivassiloglou [19] modeled extractive text summarization as a maximum coverage problem that aims at covering as many conceptual units as possible by selecting some sentences. McDonald [20] formalized text summarization as a knapsack problem and obtained the global solution and its approximate solutions. Takamura and Okamura [21] represented text summarization as maximum coverage problem with knapsack constraint (MCKP). Shen et al. [22] represented document summarization as a sequential labeling task and it solved with conditional random fields. Although this task is globally optimized in terms likelihood, the coverage of concepts is not taken into account. In [23], text summarization formalized as a budgeted median problem. This model covers the whole document cluster through sentence assignment, since in this model every sentence is represented by one of the selected sentences as much as possible. An advantage of this method is that it can incorporate asymmetric relations between sentences in a natural manner. Wang et al. [24] propose a new Bayesian sentence-based topic model for multi-document summarization by making use of both the term-document and term sentence associations. This proposal explicitly models the probability distributions of selecting sentences given topics and provides a principled way for the summarization task.

Position information has been frequently used in document summarization. Paper [25] defines several word position features based on the ordinal positions of word appearances and develops a word-based summarization system to evaluate the effectiveness of the proposed word position features on a series of summarization data sets.

3. Modelling Document Summarization

In general, the goal of text summarization is to find the subset of sentences in text which in some way represents main content of source text. In other words, generate such summary that similarity between a document collection and a summary is maximized. As input given a document collection $\mathbf{D} = \{d_1, d_2, ..., d_{|D|}\}$, where $|\mathbf{D}|$ is the number of documents. For simplicity, the document collection represented as the set of all sentences from all the documents in the collection, *i.e.* $\mathbf{D} = \{s_1, s_2, ..., s_n\}$, where s_i denotes i th sentence in $\mathbf{D}$, n is the number of sentences in the document collection.

3.1. Sentence Representation

Let $\mathbf{T}=\{t_1,t_2,...,t_m\}$ represents all the terms occurred in the document collection $\mathbf{D}$. Each sentence is represented using the vector space model. According to this model each sentence s_i is located as a point in a m dimensional vector space, $s_i=\{w_{i1},w_{i2},...,w_{im}\}$, $i=1,...,n$. Each component of such a vector reflects a term connected with the given sentence. The value of each component depends on the degree of relationship between its associated term and the respective sentence:

$$w_{ik}=n_{ik}\times\log\left(n/n_k\right) \tag{1}$$

where n_{ik} is the number of occurrences of term t_k in sentence s_i, n_k is the number of sentences containing term t_k.

Inverse sentence frequency $isf=\log(n/n_k)$ accounts for the global weighting of term t_k. Indeed, when a term appears in all sentences in the collection, $n_k=n$ and thus the balanced term weight is 0, indicating that the term is useless as a sentence discriminator. The *isf* factor has been introduced to improve the discriminating power of terms in the traditional information retrieval.

3.2. Similarity Measure

The cosine measure has been one of the most popular similarity measures due to its sensitivity to text vector pattern. The cosine measure computes the cosine of the angle between two feature vectors and is used frequently in text mining where vectors are very large but sparse. The cosine similarity between two sentences $s_i=\{w_{i1},w_{i2},...,w_{im}\}$ and $s_j=\{w_{j1},w_{j2},...,w_{jm}\}$ calculate as:

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

3.3. Multiobjective Integer Linear Programming Model

The proposed approach attempts to find a subset of the sentences $\mathbf{D}=\{s_1,s_2,...,s_n\}$ that covers the main content of the document collection and to reduce redundancy in the generated summaries.

In the summarization process redundancy control is necessary. To choose summary sentences an approach similar to the maximal marginal relevance [26] is proposed. After each selection, the current candidate sentence is compared against the already-included sentences. The sentence is added to the summary only if it is not significantly similar to any already-selected sentence, which is judged by the condition that the cosine similarity between the selected sentences is minimal.

Assuming that each sentence is a candidate-summary sentence to be selected for inclusion in the summary, then text summarization task can be formalized as follows:

$$f(\mathrm{X})=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}sim(\mathbf{O},s_i)\cdot sim\left(\mathbf{O},s_j\right)x_{ij} -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}sim\left(s_i,s_j\right)x_{ij}\rightarrow\max, \tag{3}$$

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[len\left(s_i\right)+len\left(s_j\right)\right]x_{ij}\le L, \tag{4}$$

$$x_{ij}\in\{0,1\},\ \forall i,j, \tag{5}$$

where x_{ij} denotes a variable which is 1 if pair of sentences s_i and s_j are selected, for inclusion to the summary, otherwise 0, L is length of summary, and $len(s_i)$ denotes the length of sentence s_i.

In Equation (3), $\mathbf{O}$ denotes the centre of the collection $\mathbf{D}=\{s_1,s_2,...,s_n\}$ where k th coordinate o_k define as follows:

$$o_k=\frac{1}{n}\sum_{i=1}^{n}w_{ik}\ . \tag{6}$$

In Equation (3) the term $sim(\mathbf{O},s_i)$ measures the content coverage degree of the document by sentence s_i. (3)-(5) is an integer linear programming (ILP) problem, where both the multiobjective function (3) and the constraint (4) are linear in the set of integer variables (5). The first term in Equation (3) guarantees that the main content of the source (s) will be covered by the summary. The second term provides a high diversity (*i.e.*, minimum redundancy) in the summaries. The selection process is repeated until the length of the sentences in the summary reaches the length limitation (4). The integrality constraint on x_{ij} (6) is automatically satisfied in the problem above. Now the objective is to find the binary assignment $\mathrm{X}=[x_{ij}]$ (6) with the best coverage (3) and high diversity such that the summary length is at most L (4).

4. Experiments

4.1. Dataset and Experimental Setting

The DUC2004 dataset from DUC [27] was tested to examine the effectiveness of the proposed summarization method. This dataset consists of 50 document clusters. Each cluster contains 10 newswire articles. For each

group, four NIST assessors were each asked to read all the documents and to create a brief summary. The manually-generated summaries are treated as gold-standard summaries to evaluate the qualities of machine-generated summaries. Following the most relevant previous methods the target length is limited to 665 bytes.

In the dataset used in the experiment, the original documents are all pre-processed by sentence segmentation, stop-word removal and word stemming. For removing the stopwords we used the stoplist from [28]. In our experiments, stopwords were stemmed by Porter's stemmer [29].

Solving arbitrary ILPs is an NP-hard problem. However, ILPs are a well studied optimization problem with efficient branch-and-bound algorithms for finding the optimal solution. Since our model is an NP-hard problem, it cannot generally be solved in polynomial time. However, if the size of the problem is limited, sometimes we can obtain the exact solution within a practical time by means of the branch-and-bound method. Modern commercial ILP solvers can typically solve moderately large optimizations in a matter of seconds. To solve the optimization problem (3)-(5) the GNU Linear Programming kit [30] is used, which is a free optimization package.

4.2. Evaluation Metrics

Machine-generated summaries are evaluated using the ROUGE-1.5.5 (Recall-Oriented Understudy for Gisting Evaluation) package [31]. ROUGE is adopted by DUC as the official evaluation metric for text summarization. It includes measures, ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU, which automatically determine the quality of machine summary by comparing it to ideal summaries created by humans. These measures evaluate the quality of the summarization by counting the number of overlapping units, such N-grams, between the generated summary by a method and a set of reference summaries.

Basically, the ROUGE-N measure compares N-grams of two summaries, and counts the number of matches. This measure is computed as [31]:

$$\text{ROUGE-N} = \frac{\sum_{S \in Summ_{ref}} \sum_{\text{N-gram} \in S} Count_{match}(\text{N-gram})}{\sum_{S \in Summ_{ref}} \sum_{\text{N-gram} \in S} Count(\text{N-gram})}, \quad (7)$$

where N stands for the length of the N-gram, $Count_{match}(\text{N-gram})$ is the maximum number of N-grams co-occurring in candidate summary and the set of reference-summaries. $Count(\text{N-gram})$ is the number of N-grams in the reference summaries.

For evaluation, are used the ROUGE-1 and ROUGE-2 metrics. ROUGE-1 and ROUGE-2 compares the unigram and bigram overlap between the system summary and the manual summaries created by human, respecttively.

4.3. Performance Evaluation

The following methods as the baseline systems are used to compare with the proposed method (denoted as ILPS): 1) Centroid [9]; 2) LexPageRank [18]; 3) BSTM [24]; 4) MICW [17]; 5) MCKP [21]; 6) WCS [16]; 7) Re- inforcement [1]; 8) MDS [15].

The experimental results are shown in **Table 1**. The ILPS method outperforms the simple Centroid method and another graph-based LexPageRank, and its performance is close to the results of the Bayesian sentence-based topic model and those of the best team in the DUC competition. Note however that, like clustering or topic based methods, BSTM needs the topic number as the input, which usually varies by different summarization tasks and is hard to estimate.

With comparison to the average ROUGE values for other methods, the proposed method can achieve signifycant improvement. Results of comparison reported in **Table 2**. Improvement refers to the difference between the ROUGE scores and the relative improvement in the parentheses when ILPS is compared to other methods. The relative improvement is calculated as $(b-a)*100/a$ when b is compared to a. Results of comparison reported in **Table 2**.

For more evident representation the comparisons of the methods are reported on **Figures 1** and **2**.

Table 1. ROUGE values of the methods.

Methods	ROUGE-1	ROUGE-2
MCKP	0.30908	0.08042
MICW	0.32714	0.08609
Centroid	0.36728	0.07379
Reinforcement	0.37082	0.08351
LexPageRank	0.37842	0.08572
MDS	0.37934	0.08934
DUC Best	0.38224	0.09216
BSTM	0.39065	0.09010
ILPS	0.39121	0.09113
WCS	0.39825	0.09641

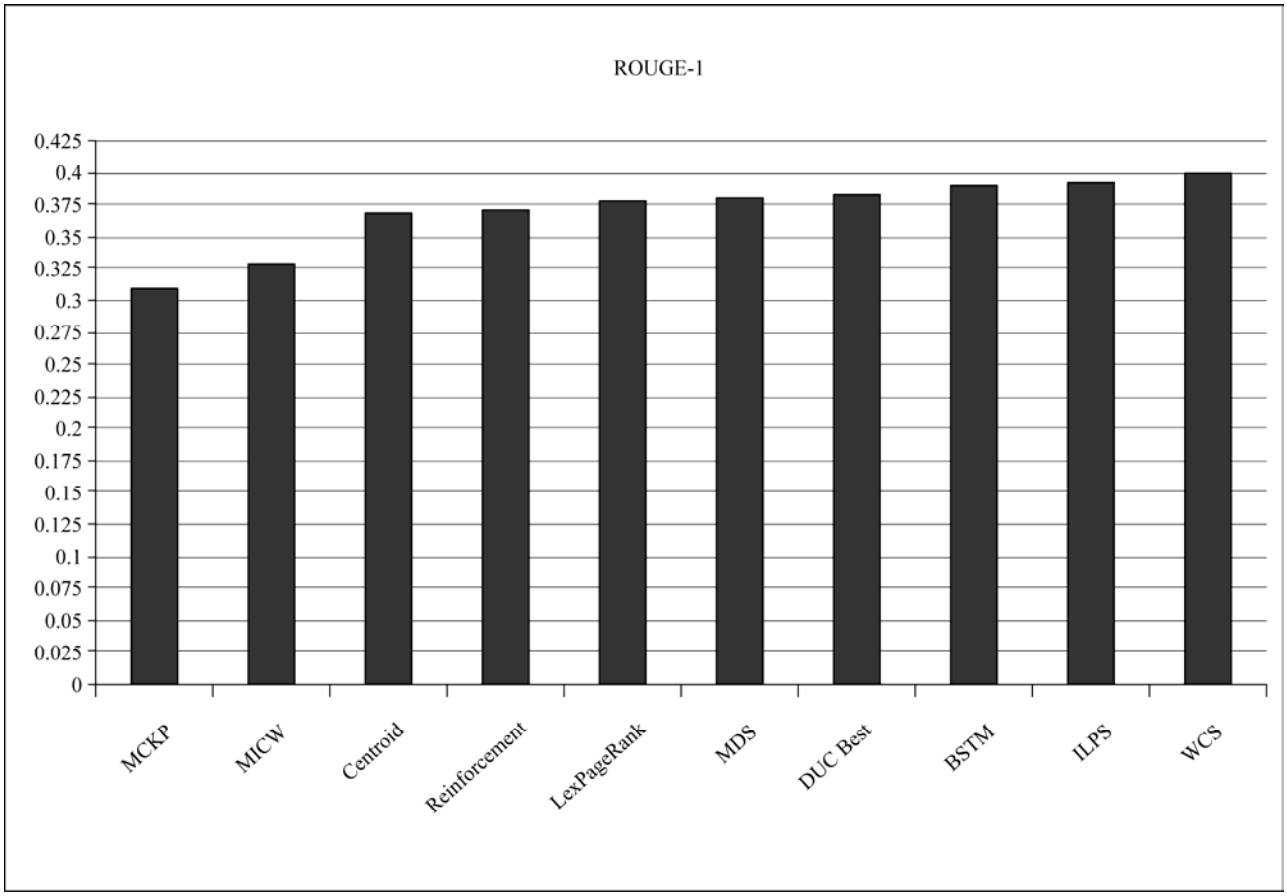

Figure 1. Comparison of the methods: ROUGE-1 score.

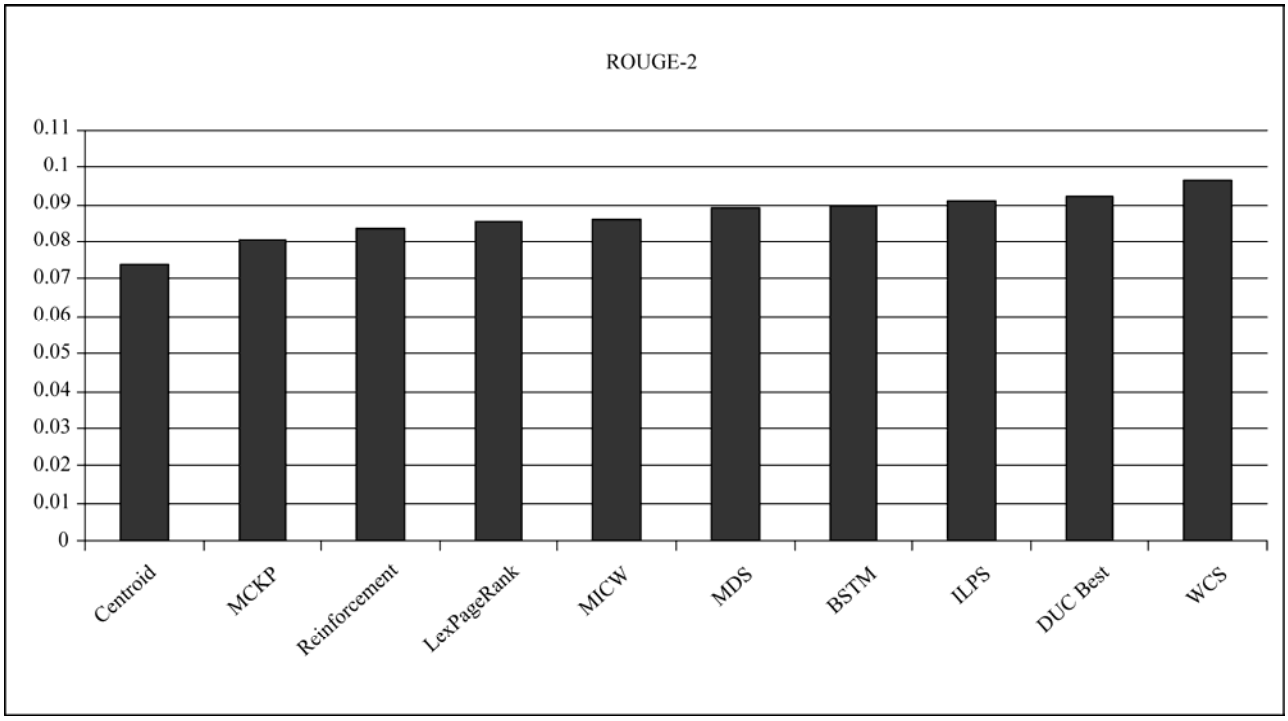

Figure 2. Comparison of the methods: ROUGE-2 score.

Table 2. Improvement of the ILPS method.

Methods	Improvement			
	ROUGE-1		ROUGE-2	
DUC Best	0.00897	(2.35)	–0.00103	(–1.12)
MICW	0.06407	(19.58)	0.00704	(8.17)
MCKP	0.08213	(26.57)	0.01271	(15.80)
Reinforcement	0.02039	(5.50)	0.00962	(11.52)
BSTM	0.00056	(0.14)	0.00303	(3.36)
LexPageRank	0.01279	(3.38)	0.00741	(8.64)
Centroid	0.02393	(6.52)	0.01934	(26.21)
WCS	–0.00704	(–1.77)	–0.02262	(–3.40)
MDS	0.01187	(3.13)	0.00379	(4.24)

From **Table 2** and the **Figures 1** and **2** the following results are obtained:

- The method ILPS concedes only to the WCS (on both ROUGE-1 and ROUGE-2 scores) and DUC Best (only on ROUGE-2 scores), and its performance is close to the results of the method BSTM. On ROUGE-1 score the method ILPS shows the best result than DUC Best.
- On ROUGE-1 metric the best result is shown by the method WCS and the worst result is obtained by the method MCKP. On ROUGE-1 metric the proposed method ILPS concedes only to the method WCS, and its result close to the result of the method BSTM.
- On ROUGE-2 metric the best result is also shown by the method WCS and the worst result is demonstrated by the method Centroid. On ROUGE-1 metric the proposed method ILPS concedes only to the method WCS, and its result close to the result of the method BSTM.
- ROUGE-1 values of the methods Centroid and Reinforcement are close.
- ROUGE-1 values of the methods LexPageRank and MDS are almost identical, and are close to the DUC Best result.
- ROUGE-2 values of the methods LexPageRank and MICW are almost identical.

5. Conclusion

In this paper, a novel text summarization model based on the assignment problem is proposed. The proposed approach covers the main content of the given document(s) through sentence assignment and reduces the redundancy in the summary. The model represented as an integer linear programming problem. When comparing the proposed method to several existing summarization methods on an open DUC2004 dataset, are found that the method can improve the summarization results significantly. The methods were evaluated using ROUGE-1 and ROUGE-2 metrics.

6. References

[1] X. Cai, W. Li and Y. Ouyang, "Simultaneous Ranking and Clustering of Sentences: A Reinforcement Approach to Multi-Document Summarization," *Proceedings of the* 23*rd International Conference on Computational Linguistics*, Beijing, China, 23-27 August 2010, pp. 134-142.

[2] C. C. Yang and F. L. Wang, "Hierarchical Summarization of Large Documents," *Journal of the American Society for Information Science and Technology*, Vol. 59, No.6, 2008, pp. 887-902.

[3] Y. Tao, S. Zhou, W. Lam and J. Guan, "Towards More Text Summarization Based on Textual Association Networks," *Proceedings of the* 2008 4*th International Conference on Semantics, Knowledge and Grid*, Beijing, China, 3-5 December 2008, pp. 235-240.

[4] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM Based Models for Automatic Text Summarization," *Computer Speech and Language*, Vol. 23, No. 1, 2009, pp. 126-144.

[5] X. Wan, "Using Only Cross-Document Relationships for Both Generic and Topic-Focused Multi-Document Summarizations," *Information Retrieval*, Vol. 11, No. 1, 2008, pp. 25-49.

[6] R. M. Aliguliyev, "Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization," *Computational Intelligence*, Vol. 26, No. 4, 2010, pp. 1-29.

[7] I. Mani and M. T. Maybury, "Advances in Automatic Text Summarization," MIT Press, Cambridge, 1999.

[8] R. M. Alguliev and R. M. Aliguliyev, "Effective Summarization Method of Text Documents," *Proceedings of the* 2005 *IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne, France, 19-22 September 2005, pp. 264-271.

[9] D. Radev, H. Jing, M. Stys and D. Tam, "Centroid-Based Summarization of Multiple Documents," *Information Processing and Management*, Vol. 40, No. 6, 2004, pp. 919-938.

[10] R. M. Aliguliyev, "A Novel Partitioning-Based Clustering Method and Generic Document Summarization," *Proceedings of the* 2006 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Hong Kong, China, 18-22 December 2006, pp. 626-629.

[11] R. M. Alguliev and R. M. Alyguliev, "Automatic Text Documents Summarization through Sentences Clustering," *Journal of Automation and Information Sciences*, Vol. 40, No. 9, 2008, pp.53-63.

[12] R. M. Aliguliyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization," *Expert Systems with Applications*, Vol. 36, No. 4, 2009, pp. 7764-7772.

[13] R. M. Alyguliyev, "The Two-Stage Unsupervised Approach to Multidocument Summarization," *Automatic Control and Computer Sciences*, Vol. 43, No. 5, 2009, pp. 276-284.

[14] R. M. Alguliev and R. M. Aliguliyev, "Evolutionary algorithm for Extractive Text Summarization," *Journal of Intelligent Information Management*, Vol. 1, No. 2, 2009, pp. 128-138.

[15] C. Shen and T. Li, "Multi-Document Summarization via the Minimum Dominating Set," *Proceedings of the* 23rd *International Conference on Computational Linguistics*, Beijing, China, 23-27 August 2010, pp. 761-769.

[16] D. Wang and T. Li, "Many are Better than One: Improving Multi-Document Summarization via Weighted Consensus," *Proceeding of the* 33*rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva, Switzerland, 19-23 July 2010, pp. 809-810.

[17] W.-T.Yih, J. Goodman, L. Vanderwende and H. Suzuki, "Multi-Document Summarization by Maximizing Informative Content-Words," *Proceedings of the* 20*th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6-12 January 2007, pp. 1776-1782.

[18] G. Erkan and D. R. Radev, "LexPageRank: Prestige in Multi-Document Text Summarization," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25-26 July 2004, pp. 365-371.

[19] E. Filatova and V. Hatzivassiloglou, "A Formal Model for Information Selection in Multi-Sentence Text Extraction," *Proceedings of the* 20*th International Conference on Computational Linguistics*, Geneva, Switzerland, 23-27 August 2004, pp. 397-403.

[20] R. McDonald, "A Study of Global Inference Algorithms in Multi-Document Summarization," *Proceedings of* 29*th European Conference on IR Research*, Rome, Italy, 2-5 April 2007, Springer-Verlag, LNCS, No. 4425, 2007, pp. 557-564.

[21] H. Takamura and M. Okumura, "Text Summarization Model Based on Maximum Coverage Problem and Its Variant," *Proceedings of the* 12*th Conference of the European Chapter of the ACL*, Athens, Greece, 30 March -3 April 2009, pp.781−789.

[22] D. Shen, J.-T. Sun, H. Li, *et al.*, "Document Summarization Using Conditional Random Fields," *Proceedings of the* 20*th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6-12 January 2007, pp. 2862-2867.

[23] H. Takamura and M. Okumura, "Text Summarization Model Based on the Budgeted Median Problem," *Proceedings of the* 18*th ACM International Conference on Information and Knowledge Management*, Hong Kong, China, 2-6 November 2009, pp. 1589-1592.

[24] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-Document Summarization Using Sentence-Based Topic Models," *Proceedings of the ACL-IJCNLP*, Singapore, 2-7 August 2009, pp. 297-300.

[25] Y. Ouyang, W. Li, Q. Lu and R. Zhang, "A Study on Position Information In Document Summarization," *Proceedings of the* 23*rd International Conference on Computational Linguistics*, Beijing, China, 23-27 August 2010, pp. 919-927.

[26] J. G. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents And Producing Summaries," *Proceedings of the* 21*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 24-28 August 1998, pp. 335-336.

[27] Document Understanding Conferences: http://duc.nist.gov/

[28] English stoplist: ftp://ftp.cs.cornell.edu/pub/smart/english.stop

[29] Porter Stemming Algorithm: http://www.tartarus.org/martin/PorterStemmer/

[30] GNU Linear Programming: http://www.gnu.org/software/glpk/

[31] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation Summaries," *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, 25-26 July 2004, pp. 74-81.

