ISSN: 0067-2904
GIF: 0.851

A Genetic Based Optimization Model for Extractive Multi-Document Text Summarization

Hilal H. Saleh[1]*, Nasreen J. Kadhim[2], Bara'a A. Attea[2].
[1]Computer Science Department, University of Technology, Baghdad, Iraq.
[2]Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.

Abstract
Extractive multi-document text summarization – a summarization with the aim of removing redundant information in a document collection while preserving its salient sentences – has recently enjoyed a large interest in proposing automatic models. This paper proposes an extractive multi-document text summarization model based on genetic algorithm (GA). First, the problem is modeled as a discrete optimization problem and a specific fitness function is designed to effectively cope with the proposed model. Then, a binary-encoded representation together with a heuristic mutation and a local repair operators are proposed to characterize the adopted GA. Experiments are applied to ten topics from Document Understanding Conference DUC2002 datasets (d061j through d070f). Results clarify the effectiveness of the proposed model when compared with another state-of-the-art model.

Keywords: Text summarization, genetic algorithm, local repair, content coverage.

نموذج أمثلية مستند على الخوارزمية الجينية للتلخيص الاقتطاعي للمستندات النصية المتعددة

هلال هادي صالح[1]*، نسرين جواد كاظم[2]، براء علي عطية[2].
[1]قسم علوم الحاسبات، الجامعة التكنلوجية، بغداد، العراق.
[2]قسم علوم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق.

الخلاصة
التلخيص الاقتطاعي للمستندات النصية المتعددة – تلخيص يهدف الى ازالة البيانات المتكررة بمجموعة مستندات مع الحفاظ على الجمل المهمة التي تبرز المحور الرئيسي الذي تدور حوله هذه المستندات – حصل مؤخرا على اهتمام واسع من خلال اقتراح نماذج رياضية اوتوماتيكية لصياغة هذه المشكلة. هذا البحث يقوم باقتراح نموذج تلخيص اقتطاعي للمستندات النصية المتعددة مستند على الخوارزمية الجينية. حيث تم اولا نمذجة المشكلة كمشكلة افضلية متقطعة مع تصميم دالة ملائمة محددة للنموذج المقترح. والثاني هو استخدام تمثيل ثنائي مع موجه طفرة ومصحح محلي لمساعدة الخوارزمية الجينية المتبناة. التجارب طبقت على عشرة محاور من مجموعة البيانات العالمية DUC2002 وقد اظهرت النتائج فعالية النموذج المقترح عندما تمت مقارنته مع أحدى النماذج الحديثة.

Introduction

Identification of relevant information that meets user needs becomes very difficult as a result of exponential growth of Internet and availability of huge amount of online information. This has triggered a race for developing automatic document summarization tools. This race is not necessary

*Email: hhsrq888@yahoo.com

just for professionals who aim to find the information in a short time but also for large search engines like Google, Yahoo, AltaVista, and others.

The main goal of any text summarization technique is the presentation of the common and most important information in a shorter version of the original text while preserving its main content and overall meaning to help the user to quickly understand the large volume of information. Different dimensions can be used to classify document summarization. A summary can be either generic summary or query-relevant summary [1-4]. In a generic summary, an overall sense of the document content is presented without any prior knowledge, on the other hand, the information presented in a query-relevant summary should have some relevance with a given query or topic [5].

Text summarization methods can also be either extractive or abstractive. Extractive methods tend to select a subset of existing words, phrases, or sentences found in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then create a summary that is closer to what a human might generate using natural language generation techniques. Such a summary might contain novel words that are not explicitly present in the original text.

Moreover, the summary can be created either from a single-document or from a multi-document collection depending on the number of documents to be summarized [3,6]. Single-document summarization can only produce a shorter representation of one document, whereas multi-document summarization can produce a summary of a set of documents.

The main contribution of this paper is to model the *multi-document text summarization* task as an optimization problem. The proposed model emphasizes the discovery of essential sentences that cover the main topic of the document collection while transcending the occurrence of redundant sentences. A binary-encoded genetic algorithm together with heuristic mutation and local repair operators is proposed to handle the modeled optimization problem. The organization of this paper is as follows. Section 2 presents related works on extractive summarization. Section 3 introduces the details of the proposed mathematical formulation and modeling. The numerical experiments and results are presented in Section 4. Finally, conclusions and some possible extensions to the current work are given in Section 5.

Related Work

Various extraction-based techniques have been proposed for generic text summarization. One of the popular extractive summarization methods is the centroid-based method [7]. This paper briefly reviews only optimization based works which are most related to the approach proposed here.

In [8], a method using latent semantic analysis is proposed to identify semantically important sentences for generation of a summary and selection of highly ranked sentences and different from each other for summarization. Other methods include Non-negative Matrix Factorization (NMF-based) topic specification [9-11] and Conditional Random Fields based (CRF-based) summarization [1]. In [9], a multi-document summarization framework based on sentence-level semantic analysis and symmetric Non-negative Matrix Factorization is proposed. The relationships between sentences can be captured by sentence-level semantic analysis in a semantic manner and the similarity matrix can be factorized by symmetric Non-negative Matrix Factorization to obtain sentences groups that are meaningful for extraction. In [12], text summarization is modeled as a maximum coverage problem that aims at covering as many conceptual units as possible and avoiding redundancy in summarization and question-answering. The problem is formalized by positing a textual unit space, a conceptual unit space, and a mapping between them. McDonald [13] models text summarization as a knapsack problem. Text summarization is represented as a maximum coverage problem with the knapsack constraint in [14]. In this work three algorithms are studied for global inference in the summarization of multi-document. It is found that an algorithm of dynamic programming that is based upon solutions to the knapsack problem satisfies optimality in accuracy and scaling characteristics corresponding to both an exact algorithm and greedy algorithms. In addition to this, the compatibility of the knapsack and the greedy algorithms with arbitrary scoring functions that can be of great benefit to the performance is noticed. Shen *et. al.* [1] presents a framework based on Conditional Random Fields for generic document summarization to keep the merits of supervised and unsupervised approaches taking in consideration avoiding disadvantages of them. This approach treats the text summarization task as a sequence labeling problem. A feature that is common for all these works is that they all rank sentences based on classification models. Multi-document generic summarization is modeled in [15] as a budgeted median problem. This model covers the entire relevant part of the document cluster through

sentence assignment and incorporates asymmetric relations between sentences in a natural manner. The work [10] proposes a Bayesian sentence-based topic model (BSTM) for multi-document summarization by making use of both the term-document and term-sentence associations. It models the probability distributions of selecting sentences given topics and provides a principled way for the summarization task. In [16], document summarization is formalized as a multi objective optimization problem. In particular, four objective functions are involved, namely information coverage, significance, redundancy and text coherence. These four objective functions measure the generated summaries according to the cluster of semantically or statistically related core terms. In [17], an optimization-based method for opinion summarization based on the p-median clustering problem from facility location theory is proposed, in which content selection is viewed as selection of clusters of related information. A formulation for the widely used greedy maximum marginal relevance (MMR) algorithm as an integer linear programming is introduced in [18]. In [19], text summarization of multi-document based on sentence-extraction is formalized as a discrete optimization problem and solved using an adaptive differential evolution algorithm. The approach is presented toward all of the three aspects of summarization: content coverage, redundancy and length. In [20], text summarization is modeled as an integer linear programming problem. The proposed model demonstrates that the summarization result depends on the similarity measure. A combination of the NGD-based and cosine similarity measures conducts to better result than their use separately. In [21], document summarization is modeled as a nonlinear 0-1 programming problem where an objective function is defined as Heronian mean of the objective functions defining content coverage and redundancy minimization. The optimization problem is solved using discrete particle swarm algorithm, which is based on estimation of distribution algorithm.

Problem Statement and Formulation

Preliminaries

There are four major categories to measure the similarity between texts: word co-occurrence/vector-based methods, corpus-based methods, hybrid methods, and descriptive feature-based methods [22].

In text summarization, vector-based methods are commonly used [23].

Let $T = \{t_1, t_2, t_3, \dots, t_m\}$ represents m distinct terms in a document collection$D$. *Cosine similarity* is the most popular measure that evaluates text similarity between any pair of sentences being represented as term vectors. Cosine similarity measure needs to calculate m different weights for all m terms composing the sentences of the document collection *D* [24, 25]. The weight w_{ik} associated with term t_k in sentence s_i can be calculated using *term-frequency inverse-sentence-frequency* scheme (tf_isf) [23]:

$$w_{ik} = tf_{ik} \times isf \quad (1)$$

Where:

tf_{ik}: is the measure of how *frequently* a term t_k occurs in a sentence s_i, and

$isf = \log(n/n_k)$ is the measure of how *few* sentences $n_k$contain the term t_k .

Intuitively, if a term t_k does not exist in weight sentence s_i, w_{ik} should be zero. Sentences in vector-based methods are represented as vectors of term weights. Now, given two sentences $s_i = [w_{i1}, w_{i2}, \dots, w_{im}]$ and $s_j = [w_{j1}, w_{j2}, \dots, w_{jm}]$, the cosine similarity between these two sentences can be calculated as in $Eq.(2)$:

$$sim(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \sum_{k=1}^{m} w_{jk}^2}} \qquad i, j = 1,2,3, \dots, n \quad (2)$$

Quantitatively, the *main* content of a document collection *D* being represented in $T = \{t_1, t_2, t_3, \dots, t_m\}$ space, can be reflected by the *mean* weights of the m terms in T. Thus, for $T = \{t_1, t_2, t_3, \dots, t_m\}$ vector, a mean vector $O = [o_1, o_1, \dots, o_m]$ can be computed. The kth coordinate o_k of the mean vector O can be calculated as [6]:

$$o_k = \frac{1}{n} \sum_{i=1}^{n} w_{ik} \qquad k = 1,2,3, \dots, m \quad (3)$$

Problem statement and Formulation

In order to model the proposed summarization problem, three issues are to be considered. These three issues jointly make the global summarization problem as one of challenging tasks:

- *Content coverage*: the inclusion of sentences in the summary should take in consideration that they cover the main topic of document collection.

- *Information redundancy*: sentences that carry the same information should not be included in the summary.
- *Length*: summary should be of a bounded short length.

Then, to formulate the problem, let us consider the following assumption. Let D be a document collection of N documents, i.e. $D = \{d_1, \dots, d_N\}$. D also can be described by the set of all n distinct sentences from all the documents in the collection, i.e. $D = \{s_i | 1 \le i \le n\}$. Our aim is to find a summary being represented as subset $\overline{D}$ of sentences in D, i.e., $\overline{D} \subset D$ that satisfies both content coverage and redundancy reduction. Thus, we seek to encapsulate the characterization of the generated summary $\overline{D}$ by the following definition.

Definition 1 (*Summary* $\overline{D}$). Let $s_i \in D$ be a sentence included in the summary $\overline{D}$, then the *content coverage*, or quantitatively, the similarity $sim(O, s_i)$ between the set of sentences in the document collection D (be represented by its mean vector O) and s_i should be *maximized*. On the other hand, the *redundancy reduction*, or quantitatively, the similarity $sim(s_i, s_j)$ between any two sentences belong to $\overline{D}$ should be *minimized*. Now, to formalize our suggestion, the *text summarization problem* will be modeled using the following definition:

Definition 2 (*text summarization problem*). Let $x_i \in \{0,1\}$ be a binary decision variable denoting the existence (1) or absence (0) of the sentence s_i in $\overline{D}$ (see Eq. 4) and $x_{ij} \in \{0,1\}$ be another binary decision variable relating to the existence of both sentences s_i and s_j in $\overline{D}$ (see Eq. 5). Now, let $X = \{x_i | 1 \le i \le n\}$ be a vector of n such decision variables corresponding to n sentences. Then for the vector X, text summarization problem (see Eq. 6 & Eq. 7) is a constrained maximization problem taking a combination of maximizing the content coverage (numerator) and minimizing information redundancy (denominator)

$$x_i = \begin{cases} 1 \; if \; s_i \in \overline{D} \\ 0 \; otherwise \end{cases} \quad (4)$$

$$x_{ij} = \begin{cases} 1 \; if \; s_i and \; s_j \in \overline{D} \\ 0 \; otherwise \end{cases} \quad (5)$$

$$Maximize \;\; f(X) = \frac{\sum_{i=1}^{n} sim(s_i, O) x_i}{((\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(s_i, s_j) x_{ij}) * \sum_{i=1}^{n} x_i)} \quad (6)$$

$$subject \; to \;\; L - \varepsilon \le \sum_{i=1}^{n} l_i x_i \le L + \varepsilon \quad (7)$$

where:

L: Summary length constraint,

l_i: Length of sentence s_i,

O: Center of the document collection $D = \{s_1, s_2, \dots, s_n\}$.

ε: A tolerance introduced in this model and defined as:

$$\varepsilon = max_{i=1,\dots,n}(l_i) - min_{i=1,\dots,n}(l_i) \quad (8)$$

Notice that the first two issues of the global summarization problem, i.e. content coverage and information redundancy have contradictory meaning in objectives. To attain the best coverage, i.e. to maximize the numerator of Eq. 6, may result in increasing the number of sentences, s_i, to be included in the summary $\overline{D}$. This may imply increasing redundancy of the obtained information (or in other words, increasing denominator). To this end, the third issue, i.e. summary length is formalized as a constraint using Eq. 7.

The proposed Genetic Algorithm

Genetic algorithm (GA) is a population-based optimization algorithm with the aim of how to evolve a population of initial solutions toward better and better ones through a sequence of generations. In the design of the proposed algorithm, each genotype solution is represented as a fixed-length vector of size n, where each gene value indicates the presence or absence of the corresponding sentence. Then, the whole search space δ for the proposed GA can be computed by the Cartesian product of presence/absence of all n sentences, i.e.:

$$\delta = \prod_{i=1}^{n}(\{0,1\}) = 2^n \quad (9)$$

Let us consider a population ρ of $K \ll \delta$ genotype solutions, $\mathbb{P}_{1 \le k \le K} \in \rho$. Then, $\forall k \in \{1, \dots, K\} \; and \; \forall j \in [1, n]: \mathbb{P}_k = (\mathbb{P}_{k1}, \mathbb{P}_{k2}, \dots, \mathbb{P}_{kn}) \; s.t. \; \mathbb{P}_{kj} \in \{0,1\}$. The proposed GA can be described as a process formulated in an iterative function $\Psi: \rho \rightarrow \rho'$ with $\Psi(\rho_{iter}) = \rho_{iter+1}$, where $iter$ is the iteration index and ρ_{iter} is the population at iteration $iter$. The population starts with an initial random population ρ_0 and continues until a maximum number of iterations $iter_{max}$ is reached.

The evolution function Ψ in each iteration $iter$ will be composed of three main operators: selection, crossover, and heuristic mutation, each of which is controlled by its control parameter. Formally speaking:

$$\Psi = sel_{\Theta_s} \circ c_{\Theta_c} \circ m_{\Theta_m} \quad (10)$$

By applying selection operator, sel_{Θ_s}, bad chromosomes are eliminated whereas good quality chromosomes that are fittest are copied for the next generation to improve the average quality of the population Tournament selection has been adopted in our work. In tournament selection, only one individual from several randomly selected individuals is selected for the next generation if it is fittest. The number of randomly selected individuals, i.e. *tournament size* is determined by the control parameter Θ_s.

Uniform Crossover has been adopted in the proposed work. According to this type of crossover, each gene of each chromosome is created by randomly selecting respective gene from one of both parents. An equal chance is given to both parents to contribute in the chromosomes that are created from them [26]. Crossover rate is determined by the control parameter Θ_c.

A heuristic mutation operator is proposed in our work. Here, the mutation operator is controlled by two parameters. The first parameter is the well-known mutation probability, p_m, controlling the probability of mutation on each gene. The second parameter is *mutation action*, which controls the role of mutation on each *mutated* gene. Mutation action can be projected by the following similarity condition:

$$sim(s_i, O) \geq \frac{1}{n}\sum_{j=1}^{n} sim(s_j, O) \quad (11)$$

For a given gene i and for a random uniform variable r_i~[0,1], if p_m is satisfied (i.e., $r_i \leq p_m$) then the similarity condition should be checked. The condition checks whether the similarity between the i^{th} sentence and mean vector, O is more or less than the average similarity of document collection sentences. If it is satisfied, then the corresponding sentence, s_i can be selected in the solution's summary. Otherwise, it can be removed from the summary. Formally speaking,

$$\forall i \in \{1, \dots, n\} \wedge r_i \leq p_m \quad (12)$$

$$x_i^{'} = \begin{cases} 1 \; iff \; sim(s_i, O) \geq \frac{1}{n}\sum_{j=1}^{n} sim(s_j, O) \\ 0 \; otherwise \end{cases} \quad (13)$$

The best solution, $\mathbb{P}_{best}$, of the final generation of GA can be selected as the result to the maximization problem.

$$\mathbb{P}_{best} : \Leftrightarrow \nexists \mathbb{P} \in \rho_{iter_{max}} | \, f(X_{\mathbb{P}}) > f\left(X_{\mathbb{P}_{best}}\right) \quad (14)$$

The phenotype of the best solution may still suffer from violating the length constraint. Formally speaking:

$$\sum_{i=1}^{n} \mathbb{P}_{best_i} > L \quad (15)$$

To this end, a *local repair* operator is proposed to handle the existence of, more than constraint need, sentences. First, this repair operator removes from $\mathbb{P}_{best}$ those redundant sentences which have a high degree of similarity. Considering a *similarity threshold* $\delta = 0.9$ and two sentences x_i and x_j in $\mathbb{P}_{best}$, one of them will be excluded from the final generated summary if their similarity is more than or equal to δ. Second, the proposed local repair operator will only handle the selection of high importance sentences. Each sentence exists in $\mathbb{P}_{best}$ is selected according to the following formula in order to gain a corresponding score:

$$\forall i \in \{1, \dots, n\} \wedge x_i = 1 \quad (16)$$

$$score_{s_i} = sim(s_i, O) + \left(\left(sim(O^{sum}, O) - sim(O^{sum-s_i}, O)\right) * 10\right) \quad (17)$$

Where $sim(O^{sum}, O)$ refers to the similarity of the centre of the generated summary (including sentence s_i) and the centre of document collection O. On the other hand, $sim(O^{sum-s_i}, O)$ denotes the similarity between the generated summary (excluding sentence s_i). The right term of the proposed formula is multiplied by 10 in order to unify the scale of the two terms. The basic idea behind the right term of the formula is to measure the impact of each of the sentences exist in the best phenotype summary. The sentence with the highest score has a great impact on the summary and it is of high importance whereas the sentence with the lowest score has a little impact on the final summary. The sentences are sorted in descending order and the high scored sentences are selected to be included in the final summary until the required length L is reached.

Experiments

Evaluations to the quality of the proposed model were made based on the multi-document summarization datasets provided by Document Understanding Conference (DUC) [27]. We have evaluated the quality of our model according to DUC2002 dataset. A brief description of the dataset is given in table 1. First, documents in DUC2002 dataset are preprocessed as follows:

- Segmentation of documents into individual sentences,
- Sentences are tokenized,
- Stop word removal operation is implemented,
- Finally, the remaining words are stemmed using Porter stemming algorithm [28].

The proposed algorithm is coded in Visual Basic and the experiments were executed on a THINK-PC Lenovo with Intel(R) Core(TM) i5-2410M CPU @2.30 GHz and a Memory of 4 GB RAM. GA's parameters are set as follows: a population of $pop_{size} = 50$ individuals is used and evolved over a sequence of $iter_{max} = 1000$. For the tournament selection, a tournament size equals to 2 has been chosen. Crossover probability and mutation probability are $p_c = 0.7$ and $p_m = 0.1$, respectively.

Table1-Description of the DUC2002 dataset.

Description	DUC2002 dataset
Number of topics	60 (d061j through d120i)
Number of documents in each topic	~ 10
Total number of documents	567
Data source	TREC
Summary length	200 words

Evaluation metrics

The proposed method is measured using the Recall-Oriented Understudy for Gisting Evaluation (denoted by $ROUGE$ evaluation metric) [29]. $ROUGE$ is considered as the official evaluation metric for text summarization by *DUC*. It includes measures that automatically determine the quality of a summary generated by computer through comparison made between it and human generated summaries. This comparison satisfied by counting the number of overlapping units, such as $N-grams$, word sequences, and word pairs between the summary generated by a machine and a set of reference summaries generated by humans.

$ROUGE-N$ is an $N-gram$ Recall between a computer generated summary and a set of human generated summaries. It counts the number of $N-grams$ matches of two summaries, and it is calculated as follows [29]:

$$ROUGE-N = \frac{\sum_{S\in\{reference\ Summaries\}}\sum_{N-gram\in S} Count_{match}(N-gram)}{\sum_{S\in\{reference\ Summaries\}}\sum_{N-gram\in S} Count(N-gram)} \quad (18)$$

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

The similarity between reference summary sentence X of length m and candidate summary sentence Y of length n is calculated using $ROUGE-L$ measure (also called $F-Measure$ which is denoted by F_{lcs}) which is defined as the ratio between the length of the longest common subsequence of the two summaries $LCS(X,Y)$ and the length of the reference summary as follows [28]:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (19)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (20)$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad (21)$$

Where recall and precision of the $LCS(X,Y)$ is denoted by R_{lcs} and P_{lcs}, respectively and $\beta = \frac{P_{lcs}}{R_{lcs}}$.

If the definition of $ROUGE-L$ is applied to summary-level, the union LCS matches between a reference summary sentence, r_i, and sentences of the candidate summary, C which is denoted by

$LCS_{\cup}(r_i, C)$ is taken. Given a reference summary of u sentences containing a total of m words and a candidate summary of v sentences containing a total of n words, then summary-level $ROUGE-L$ is calculated as follows [29]:

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m} \quad (22)$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{n} \quad (23)$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (24)$$

Results and Discussion

To evaluate our proposed model, comparison with some other related models should be performed. In this paper, the model proposed in [19] is used for comparison. This model formulates content coverage and redundancy reduction issues in a different meaning. For comparison fairness, model in [19] has been solved using GA algorithm proposed in this paper. A comparison between the two models is made using $ROUGE-2$ and $ROUGE-L$ evaluation metrics. These evaluation metrics were calculated by comparing computer-generated summaries against summaries generated by human. The machine-generated summary is evaluated by comparing it with multiple reference summaries generated by experts and supported by DUC2002 dataset (each topic in DUC2002 dataset is supplied with a two human reference summaries generated by two different experts).

The proposed model and the model introduced in [19] have been run on ten topics from DUC2002 dataset [d061j, d062j, d063j, d064j, d065j, d066j, d067f, d068f, d069f, d070f]. Table 2 presents some statistics that describe documents of these topics in order to give an identification of the search space size for the problem.

Table 2-Some Statistics Describing Documents of Topics Mentioned Below.

Topic number	No. of Words before Preprocessing	No. of Words after Preprocessing and Removing Multiple Occurrences	Final No. of Sentences
d061j	3679	675	184
d062j	2669	626	118
d063j	4760	841	242
d064j	4038	921	181
d065j	5449	1071	280
d066j	3863	916	189
d067f	2796	634	121
d068f	2550	528	126
d069f	7609	1300	325
d070f	3160	628	151

Table 3 and 4 present detailed average $ROUGE$ scores in addition to the best and worst values for the 20 run $ROUGE-2$ scores of each topic.

Table 3-Detailed $ROUGE-2$ Score.

Topic number	Model Proposed in [19]			Proposed Model		
	$ROUGE-2$	$\boldsymbol{ROUGE_{best}}$	$\boldsymbol{ROUGE_{worst}}$	$\boldsymbol{ROUGE-2}$	$\boldsymbol{ROUGE_{best}}$	$\boldsymbol{ROUGE_{worst}}$
d061j	0.266	0.418	0.128	**0.306**	0.411	0.148
d062j	0.188	0.336	0.061	**0.200**	0.468	0.046
d063j	0.245	0.366	0.158	**0.275**	0.388	0.109
d064j	0.194	0.336	0.056	**0.233**	0.418	0.062
d065j	0.144	0.278	0.069	**0.182**	0.290	0.082
d066j	**0.201**	0.313	0.056	0.181	0.319	0.074
d067f	0.239	0.387	0.152	**0.260**	0.407	0.109
d068f	0.491	0.711	0.327	**0.496**	0.647	0.366
d069f	0.184	0.274	0.108	**0.232**	0.368	0.129
d070f	0.224	0.396	0.136	**0.262**	0.363	0.148

Table 4-Detailed $ROUGE-L$ Scores For Proposed Model.

Topic number	Model in [19]			Proposed Model		
	$ROUGE-L$	$ROUGE_{best}$	$ROUGE_{worst}$	$ROUGE-L$	$ROUGE_{best}$	$ROUGE_{worst}$
d061j	0.542	0.649	0.441	**0.554**	0.635	0.430
d062j	0.473	0.603	0.364	**0.481**	0.679	0.373
d063j	0.493	0.578	0.422	**0.528**	0.616	0.445
d064j	0.462	0.588	0.353	**0.488**	0.626	0.339
d065j	0.431	0.516	0.375	**0.457**	0.554	0.380
d066j	**0.455**	0.553	0.357	0.441	0.506	0.357
d067f	0.509	0.649	0.417	**0.529**	0.636	0.392
d068f	**0.666**	0.796	0.570	0.626	0.728	0.502
d069f	0.454	0.549	0.414	**0.476**	0.583	0.392
d070f	0.496	0.606	0.433	**0.513**	0.587	0.429

Table 5 provides the average $ROUGE$ scores and standard deviation of $ROUG$ results of the two models on the determined topics from DUC2002 datasets. All the results reported in Tables-5 are averaged for each topic over 20 runs with the same parameter setting. Results for $Rouge-2$ and $ROUGE-L$ show that the proposed model has better performance than the model proposed in [19].

Table 5-Average $ROUG$values obtained from implementing models below on DUC2002 dataset

Method	$ROUGE-2_{avg}$	$ROUGE-2_{\sigma}$	$ROUGE-L_{avg}$	$ROUGE-L_{\sigma}$
Proposed model	**0.263**	**0.087**	**0.509**	**0.051**
Model in [19]	0.238	0.091	0.498	0.064

Conclusion

The need for effective multi-document summarization techniques to extract the important information from a document collection becomes of necessity. A good summary should have the ability to keep the key sentences representing the main topic of the document collection while simultaneously reducing irrelevant and redundant ones from the whole collection. An optimization model is introduced in this paper to satisfy content coverage and diversity in the document collection. A genetic algorithm together with a heuristic mutation and a local repair operators have been proposed to solve the modeled problem. The performance of the proposed model shows improvement over the model proposed in [19].

In a future work, we plan to enhance the results of the proposed model through designing another optimization model for the problem and to extend our experiments to the rest of DUC2002 topics and compare the results with results obtained from other state-of-the-art models and to solve the optimization problem with different algorithms.

References

1. Shen, D., Sun, J.-T., Li, H., Yang, Q. and Chen, Z. **2007**. Document summarization using conditional random fields, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6–12, pp. 2862–2867.
2. Tao, Y., Zhou, S., Lam, W. and Guan, J. **2008**. Towards more text summarization based on textual association networks, in: Proceedings of the 2008 4th International Conference on Semantics, Knowledge and Grid, Beijing, China, December 03–05, pp. 235–240.
3. Fattah, M.A. and Ren, F. **2009**. GA, MR, FFNN, PNN and GMM based models for automatic text summarization, *Computer Speech and Language* 23 (1) 126–144.

4. Dong, H., Yu, S. and Jiang, Y. **2009**. Text mining on semi-structured e-government digital archives of China, in: Proceedings of the 2009 Second Pacific-Asia Conference on Web Mining and Web-Based Application, Wuhan, China, June 06–07, pp.11–14.

5. Aliguliyev, R.M. **2009**. A new sentence similarity measure and sentence based extractive technique for automatic text summarization, *Expert Systems with Applications* 36 (4) 7764–7772.

6. Zajic, D.M., Dorr, B.J. and Lin, J. **2008**. Single-document and multi-document summarization techniques for email threads using sentence compression, *Information Processing & Management* 44 (4) 1600–1610.

7. Radev, D., Jing, H., Stys, M. and Tam, D. **2004**. Centroid-based summarization of multiple documents, Information Processing & Management 40 (6) 919–938.

8. Gong, Y. and Liu, X. **2001**. Generic text summarization using relevance measure and latent semantic analysis, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, USA, September 9–12, pp. 19–25.

9. Wang, D., Li, T., Zhu, S. and Ding, C. **2008**. Multi-document summarization via sentence level semantic analysis and symmetric matrix factorization, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, July 20–24, pp. 307–314.

10. Wang, D., Li, T., Zhu, S. and Ding, C. **2009**. Multi-document summarization using sentence based topic models, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Singapore, August 04, pp. 297–300.

11. Lee, J.-H., Park, S., Ahn, C.-M. and Kim, D. **2009**. Automatic generic document summarization based on non-negative matrix factorization, Information Processing & Management 45 (1) 20–34.

12. Filatova, E. and Hatzivassiloglou, V. **2004**. A formal model for information selection in multi-sentence text extraction, in: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 23–27, pp. 397–403.

13. McDonald, R. 2007. A study of global inference algorithms in multi-document summarization, in: Proceedings of 29th European Conference on IR Research, Rome, Italy, April 2–5, 2007, in: LNCS, vol. 4425, *Springer-Verlag*, pp. 557–564.

14. Takamura, H. and Okumura, M. **2009**. Text summarization model based on maximum coverage problem and its variant, in: Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, March 30–April 3, pp. 781–789.

15. Takamura, H. and Okumura, M. **2009**. Text summarization model based on the budgeted median problem, in: Proceedings of the 18th ACM International Conference on Information and Knowledge Management, Hong Kong, China, November 2–6, pp. 1589–1592.

16. Huang, L., He, Y., Wei, F. and Li, W. **2010**. Modeling document summarization as multiobjective optimization, in: Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, pp. 382–386.

17. Cheung, J.C.K., Carenini, G. and Ng, R.T. **2009**. Optimization-based content selection for opinion summarization, Proceedings of the 2009 Workshop on Language Generation and Summarization (ACLIJCNLP), Singapore, 6 August, pp.7–14.

18. Riedhammer, K., Favre, B. and Hakkani-Tür, D. **2010**. Long story short – global unsupervised models for keyphrase based meeting summarization, *Speech Communication*, vol.52, no.10, pp.801–815.

19. Alguliev, R. M., Aliguliyev, R. M. and Mehdiyev, C. A. **2011**. Sentence selection for generic document summarization using an adaptive differential evolution algorithm, *Swarm and Evolutionary Computation* 1 213–222.

20. Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S. and Mehdiyev, C. A. **2011**. MCMR: Maximum coverage and minimum redundant text summarization model, *Expert Systems with Applications* 38 14514–14522.

21. Alguliev, R. M., Aliguliyev, R. M. and Mehdiyev, C. A. **2011**. An Optimization Model and DPSO-EDA for Document Summarization, *I.J. Information Technology and Computer Science*, 5, 59-68

22. Islam, A. and Inkpen, D. **2008**. Semantic text similarity using corpus-based word similarity and string similarity, *ACM Transactions on Knowledge Discovery from Data* 2 (2) Article 10, 25 p.

23. Salton, G. and Buckley, C. **1988**. Term-weighting approaches in automatic text retrieval, *Information Processing & Management* 25 (5) 513–523.
24. Singhal, A. **(2001)**. "Modern Information Retrieval: A Brief Overview". *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.
25. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., and Pinto, D. **2014**. "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". *Computación y Sistemas* 18 (3): 491–504. doi:10.13053/CyS-18-3-2043. Retrieved 7 October.
26. Shopova, E. G. and Vaklieva-Bancheva, N. G. **2006**. BASIC—A genetic algorithm for engineering problems solution, *Computers and Chemical Engineering* xxx (2006) xxx–xxx.
27. Document understanding conference: http://duc.nist.gov.
28. Porter stemming algorithm: http://www.tartarus.org/martin/PorterStemmer/.
29. Lin, C.-Y. **2004**. ROUGE: a package for automatic evaluation summaries, in: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25–26, pp. 74–81.