Genetic Based Optimization Models for Enhancing Multi-Document Text Summarization

Dr. Hilal H. Saleh
Computer Science Department, University of Technology/ Baghdad.
Email: hhsrq888@yahoo.com
Nasreen J. Kadhim
Science College, University of Baghdad/Baghdad.
Email: nasreen_jawad@yahoo.com

Received on:19/2/2015 & Accepted on:30/8/2015

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 two models for extractive multi-document summarization based on genetic algorithm (GA). First, the problem is described and modeled as a discrete optimization problem with two candidate expressions and a specific fitness function is designed to effectively cope with each candidate. Then, a binary-encoded representation together with a heuristic mutation and a local repair operator are proposed to characterize the adopted GA. The semantic roles of similarity of sentence to sentence, sentence to center of document collection and center of summary to center of document collection are exploited in the proposed model formulations. Experiments are applied to ten clusters from DUC2002 datasets (d061j through d070f) and compared with another state-of-the-art model. Results clarify the effectiveness of the proposed models. Moreover, the injection of several levels of text similarity in the model formulation shows a positive impact on enhancing the overall performance of the proposed GA.

التلخيص الاقتطاعي للمستندات النصية المتعددة باستخدام نماذج أمثلية مستندة على الخوارزمية الجينية

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 نموذجين للتلخيص الاقتطاعي مستند على الخوارزمية الجينية. حيث تم اولا وصف ونمذجة المشكلة كمشكلة افضلية متقطعة عن طريق نموذجين مع تصميم دالة ملائمة محددة لكل نموذج مقترح. والثاني هو استخدام تمثيل ثنائي مع موجه طفرة ومصحح محلي لمساعدة الخوارزمية الجينية المتبناة. تم تبني دور درجة التشابه بين كل جملة مع باقي الجمل في مجموعة المستندات النصية والتشابه بين كل جملة ومركز مجموعة المستندات النصية والتشابه بين مركز المختصر ومركز مجموعة المستندات النصية في النماذج المقترحة. التجارب طبقت على عشرة محاور من مجموعة البيانات العالمية DUC2002 وقد اظهرت النتائج فعالية النماذج المقترحة عندما

https://doi.org/10.30684/etj.2015.116733
2412-0758/University of Technology-Iraq, Baghdad, Iraq

تمت مقارنتها مع أحد النماذج الحديثة. أظهرت عملية حقن مستويات متعددة من مقياس التشابه النصي عند صياغة النموذج تأثير ايجابي على تحسين الأداء الكلي للخوارزمية الجينية المقترحة.

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 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. Text summarization problem belongs to several disciplines like computer science, multimedia, statistics, and cognitive psychology. Thus 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]. Also, 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 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 (MDS) 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 reviews optimization based works which are most related to the approach proposed here. Section 3 and 4 introduce the details of the proposed mathematical formulation and modeling. The numerical experiments and results are presented in Section 5. Finally, conclusions and some possible extensions to the current work are given in Section 6.

Related Work

In literature, multi-document summarization approaches vary in their essence. 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.

Text summarization can be categorized as a combinatorial optimization problem. The optimization based text summarization algorithms reported in literature are mainly classified as heuristic algorithms. Heuristic methods do not guarantee finding optimal solution within finite amount of time but rather they can provide acceptable and near-optimal solutions with a fraction of computation time. 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, 10, 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 the 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, namely information coverage, significance, redundancy and text coherence are involved. 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. The work [22] formulated text summarization as a modified p-median problem taking in consideration four objectives: relevance, content coverage, redundancy minimization, and bounded length that are of great necessity to generate good summaries. A self-adaptive differential evolution algorithm is created to solve the proposed model. Multiple document summarizations are modeled in [23] as a Quadratic Boolean Programming problem which is a weighted combination of two objectives that are important to generate a good summary: content coverage and redundancy reduction. The optimization problem is solved using a modified differential evolution algorithm. In [24], Text summarization is formulated as linear and nonlinear optimization models which aims to balance between content coverage and redundancy reduction in the target summary simultaneously. A novel particle swarm optimization algorithm is developed to solve the optimization problem. Work in [25] proposes a constraint-driven multi-document summarization models enforcing diversity and maximum coverage which are modeled as a quadratic integer programming problem. The optimization problem is solved by using a discrete Particle Swarm Optimization algorithm. Paper [26] proposes a model which is an optimization-based for generic multi-document summarization. The proposed model describes content coverage and redundancy minimization in the target summary as relations in sentence to document, summary to document and relations between each pair of sentences in the document collection. An adaptive crossover that makes adjustment to the crossover rate according to the fitness of individuals is used to improve the differential evolution algorithm used to solve the optimization problem.

Problem Statement and Formulation

Preliminaries

Several methodologies have been explored for text similarity; however, they are centered on four major categories. These are word co-occurrence/vector-based methods, corpus-based methods, hybrid methods, and descriptive feature-based methods [27].

In text summarization, vector-based methods are commonly used [28]. Let $T = \{t_1, t_2, t_3, \dots, t_m\}$ represents m distinct terms in a document collection. *Cosine similarity* is the most popular measure that evaluates text similarity between any pair of sentences being represented as vectors of terms. For a set of m different terms composing n sentences of a document collection$\mathbb{D}$, cosine similarity associates weight w_{ik} to term t_k according to its magnitude in sentence s_i. Cosine similarity metric can be formulated, according to *term-frequency inverse-sentence-frequency* scheme (tf_isf), as [28]:

$$w_{ik} = tf_{ik} \times isf, \qquad \dots(1)$$

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 sentence s_i, w_{ik} should be zero. 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 $Eq. (2)$:

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

Quantitatively, the *main* content of a document collection $\mathbb{D}$ 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 $\mathbb{O} = [o_1, o_1, \dots, o_m]$ can be computed. The k^{th} coordinate o_k of the mean vector $\mathbb{O}$ can be calculated as [6]:

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

Problem statement and Formulation

The proposed text summarization problem is expressed here while considering three challenges:

- *Content Coverage*: the main topic of the document collection $\mathbb{D}$ *should be covered* by the generated summary.
- *Redundancy Reduction*: similar sentences in the document collection $\mathbb{D}$ *should not be duplicated* in the generated summary.
- *Length*: summary should be of a bounded length

Let $\mathbb{D}$ be a document collection of N documents, i.e. $\mathbb{D} = \{d_1, \dots, d_N\}$. By the language of sentences, $\mathbb{D}$ can be noted by $\mathbb{D} = \{s_i | 1 \le i \le n\}$, where n is the number of distinct sentences from the documents in $\mathbb{D}$. the aim of this paper is to generate a summary $\overline{\mathbb{D}} \subset \mathbb{D}$ that can satisfy the above three criteria. The first attempt, here, to model $\overline{\mathbb{D}}$ and to formulate text summarization problem is given in the following two definitions.

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

Definition 2 (*text summarization problem* Φ_1). Let $x_i \in \{0,1\}$ be a binary decision variable denoting the existence (1) or absence (0) of the sentence s_i in $\overline{\mathbb{D}}$ (see Eq. 4). Also, let $x_{ij} \in \{0,1\}$ be another binary decision variable relating to the existence of both sentences s_i and s_j in $\overline{\mathbb{D}}$ (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)

$$x_i = \begin{cases} 1\ if\ s_i \in \overline{\mathbb{D}} \\ 0\ otherwise \end{cases}, \qquad \dots(4)$$

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

$$Maximize \quad \Phi_1(x) = \frac{\sum_{i=1}^{n} sim(s_i,\mathbb{O}) x_i}{((\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(s_i,s_j) x_{ij}) * \sum_{i=1}^{n} x_i)} \qquad \ldots(6)$$

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

where:
L: Summary length constraint,
l_i: Length of sentence s_i,
$\mathbb{O}$: Center of the document collection $\mathbb{D} = \{s_1, s_2, \ldots, s_n\}$.
ε: A length tolerance introduced in this model as:

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

In the second attempt, the text summarization problem is re-defined again by projecting the first criterion, i.e. content coverage in the light of text similarity. The proposed model hypothesizes a possible decomposition of text similarity into three different levels of optimization formula. First, aspire to global optimization; the candidate summary should cover the summary of the document collection. Then, to attain, more or less global optimization, the sentences of the candidate summary should cover the summary of the document collection. The third level of optimization is content with local optimization, where the difference between the magnitude of terms covered by the candidate summary and those of the document collection should be small. The summary $\overline{\mathbb{D}}$ and text summarization problem Φ_1 can then be formulated as in definition 3 and 4, respectively.

Definition 3 (*Summary* $\overline{\mathbb{D}}$). Let $s_i \in \mathbb{D}$ be a sentence to be included in the generated summary $\overline{\mathbb{D}}$, then three different semantics of coverage (*summary level, sentence level, and term level*) can be cooperated together to define *content coverage* criterion. *Summary level* is to be expressed by the degree of similarity $sim(\,O, \mathbb{O})$ between the mean vector of a candidate summary O and the center $\mathbb{O}$ of the document collection$\mathbb{D}$. *Sentence level* is to be defined by the degree of similarity $sim(s_i, \mathbb{O})$ between sentence s_i and the mean vector $\mathbb{O}$ of the document collection$\mathbb{D}$. *Term level* to be defined by the degree of similarity $sim(O_k, \mathbb{O}_k)$ between the mean vector of term k in a candidate summary O and its correspondence term in the center $\mathbb{O}$ of the document collection $\mathbb{D}$. On the other hand, the *redundancy reduction*, or quantitatively, the similarity $sim(s_i, s_j)$ between any two sentences belongs to $\overline{\mathbb{D}}$ should be *minimized*.

Definition 4 *text summarization problem* Φ_2 can be expressed as a constrained optimization problem taking a combination of maximizing the content coverage (numerator) and minimizing information redundancy (denominator)**.** Content coverage is expressed by maximizing both $sim(\,O, \mathbb{O})$ and $sim(s_i, \mathbb{O})$ while simultaneously minimizing $sim(O_k, \mathbb{O}_k)$.

$$Maximize \quad \Phi_2 = \frac{sim(O,\mathbb{O}) * \mathbf{10}}{\sum_{i=1}^{n} x_i} + \frac{\sum_{i=1}^{n} sim(s_i,\mathbb{O}) x_i - \sum_{k=1}^{m} |O_k - \mathbb{O}_k|}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(s_i,s_j) x_{ij} * \sum_{i=1}^{n} x_i} \qquad \ldots(9)$$

$$subject \; to \quad L - \varepsilon \leq \sum_{i=1}^{n} l_i x_i \leq L + \varepsilon,$$

As can be seen in Eq. 9, the magnitude O_k of term k in the candidate summary $\overline{\mathbb{D}}$ can be expressed by its impact, i.e. average of total weights of k occurring in the sentences of $\overline{\mathbb{D}}$. Likewise, the magnitude $\mathbb{O}_k$ of term k in $\mathbb{D}$ can be computed by the average of total weights of k occurring in the sentences of $\mathbb{D}$. Intuitively, the difference between these two magnitudes should be small over all terms of $\mathbb{D}$ and $\overline{\mathbb{D}}$. Moreover, the similarity in the summary level, i.e. $sim(O,\mathbb{O})$ is multiplied by 10 to unify its scale with values of the other two similarity levels expressed at the numerator.

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 by means of some evolutionary operators. In the proposed GA, each genotype solution is represented by 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.:

$$\delta = \prod_{i=1}^{n}(\{0,1\}) = 2^n \qquad ...\ (10)$$

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 $\Psi: \rho \to \rho'$ with $\Psi(\rho_{iter}) = \rho_{iter+1}$, where 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 noted as:

$$\Psi = sel_{\Theta_s} \circ c_{\Theta_c} \circ m_{\Theta_m} \qquad ...(11)$$

By applying selection operator, sel_{Θ_s}, bad chromosomes are eliminated whereas good quality chromosomes that are fittest are copied to the next generation to improve the average quality of the population. Tournament selection has been adopted in this 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. 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 [29]. Crossover rate is determined by the control parameter Θ_c.

A heuristic mutation operator is proposed in this 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:

$$sim(s_i,\mathbb{O}) \ge \frac{1}{n}\sum_{j=1}^{n} sim(s_j,\mathbb{O}) \qquad ...(12)$$

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

$$\forall i \in \{1, \dots, n\} \wedge r_i \le p_m \qquad \dots(13)$$

$$x_i' = \begin{cases} 1 \; iff \; sim(s_i, \mathbb{O}) \ge \frac{1}{n}\sum_{j=1}^{n} sim(s_j, \mathbb{O}) \\ 0 \; otherwise \end{cases} \qquad \dots (14)$$

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

$$\mathbb{P}^*: \Leftrightarrow \nexists \mathbb{P} \in \rho_{iter_{max}} | \, \Phi(X_{\mathbb{P}}) > \Phi(X_{\mathbb{P}_{best}}) \qquad \dots(15)$$

However, the phenotype of the best solution may still suffer from violating the length constraint. i.e.:

$$\sum_{i=1}^{n} x_i \in \mathbb{P}^* > L \qquad \dots(16)$$

To this end, a *local repair* operator is proposed to handle the existence of more than constraint needs. Firstly, this repair operator removes from $\mathbb{P}^*$ those redundant sentences which have a high degree of similarity between them. Considering a *similarity threshold* $\delta = 0.9$ and two sentences x_i and x_j in $\mathbb{P}^*$, one of them will be excluded from the final generated summary if their similarity is more than or equal to δ (see Eq. 17). Secondly, this operator will only handle the selection of high importance sentences in $\mathbb{P}^*$. Each sentence belongs to $\mathbb{P}^*$ is ranked according to the formula in Eq. 18 to gain a corresponding score:

$$\forall i, j \in \{1, \dots, n\} \wedge x_i, x_j \in \mathbb{P}^* = 1$$

$$sim(s_i, s_j) \le \delta \qquad \dots(17)$$

$$\forall i \in \{1, \dots, n\} \wedge x_i \in \mathbb{P}^* = 1$$

$$score_{s_i} = sim(s_i, \mathbb{O}) + \left(\left(sim(O^{sum}, \mathbb{O}) - sim(O^{sum-s_i}, \mathbb{O}) \right) * 10 \right) \qquad \dots (18)$$

Where

$sim(O^{sum}, \mathbb{O})$ refers to the similarity of the centre of the generated summary (including sentence s_i) and the centre of document collection $\mathbb{O}$. On the other hand, $sim(O^{sum-s_i}, \mathbb{O})$ denotes the similarity between the generated summary (excluding sentence s_i) and the centre of document collection $\mathbb{O}$. 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

Qualitative evaluations of the proposed two models were made quantitatively based on the multi-document summarization datasets provided by Document

Understanding Conference (DUC), particularly using DUC2002 dataset [30]. A brief statistics of the dataset is given in Table-1. Like all other related works, the documents in DUC2002 dataset are, first, preprocessed as follows:

- segmentation of the documents into individual sentences,
- sentences are tokenized,
- stop words are removed and
- finally, the remaining words are stemmed using Porter stemming algorithm [31].

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 set to $p_c = 0.7$ and $p_m = 0.1$, respectively.

Table(1). Description of the DUC2002 dataset

Description	DUC2002 dataset
Number of topics	59 (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 work is quantitatively measured using Recall-Oriented Understudy for Gisting Evaluation $ROUGE$evaluation metric [32]. $ROUGE$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. The comparison is 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 counting the number of $N-grams$ matches of two summaries, and it is calculated as follows [32]:

$$ROUGE-N = \frac{\sum_{S\in\{reference\ Summaries\}}\sum_{N-gram\ \in\ S} Count_{match}(N-gram)}{\sum_{S\in\{reference\ Summaries\}}\sum_{N-gram\ \in\ S} Count(N-gram)} \quad \ldots (19)$$

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}). $ROUGE-L$ evaluates 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 [32]:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad ...(20)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad ...(21)$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad ...(22)$$

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 [32]:

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

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i,C)}{n}, \quad ... (24)$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}, \quad ...(25)$$

Results and Discussion

To evaluate the proposed models, a comparison with another related model should be performed. In this paper, the model proposed in [19] is used for comparison. This model formulates content coverage and redundancy reduction issues as in Eq. 26. For comparison fairness, model in [19] has been solved using GA algorithm proposed in this paper. A comparison between the three models is made using $ROUGE - 2$ and $ROUGE - L$ evaluation metrics. These evaluation metrics were calculated by comparing the summary generated by the three GA-based models against summaries generated by human. The reference summaries generated by experts are supported by DUC2002 dataset (each topic in DUC2002 dataset is supplied with a two human reference summaries generated by two different experts).

$$Maximize \;\; f(x) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[sim(s_i,\mathbb{O})+sim(s_j,\mathbb{O})]x_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sim(s_i,s_j)x_{ij}}, \quad ..(26)$$

$$subject\ to \;\; L - \varepsilon \leq \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (l_i, l_j)x_{ij} \leq L + \varepsilon,$$

The proposed models and the model introduced in [19] have been run on ten clusters 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 Table-4 present detailed average ROUGE scores in addition to the best and worst values for the 20 runs. In these tables, the best results obtained are shaded.

Table(3) Rouge-2 Scores

Topic #	Model in [19]			Proposed Model 1 (Φ_1)			Proposed Model 2 (Φ_2)		
	ROUGE − 2	Best Value	Worst Value	ROUGE −2	Best Value	Worst Value	ROUGE − 2	Best Value	Worst Value
d061j	**0.266**	0.418	0.128	**0.306**	0.411	0.148	**0.320**	0.464	0.184
d062j	**0.188**	0.336	0.061	**0.200**	0.468	0.046	**0.278**	0.422	0.161
d063j	**0.245**	0.366	0.158	**0.275**	0.388	0.109	**0.296**	0.470	0.161
d064j	**0.194**	0.336	0.056	**0.233**	0.418	0.062	**0.245**	0.372	0.138
d065j	**0.144**	0.278	0.069	**0.182**	0.290	0.082	**0.194**	0.314	0.111
d066j	**0.201**	0.313	0.056	**0.181**	0.319	0.074	**0.206**	0.381	0.085
d067f	**0.239**	0.387	0.152	**0.260**	0.407	0.109	**0.272**	0.504	0.140
d068f	**0.491**	0.711	0.327	**0.496**	0.647	0.366	**0.498**	0.680	0.211
d069f	**0.184**	0.274	0.108	**0.232**	0.368	0.129	**0.221**	0.303	0.147
d070f	**0.224**	0.396	0.136	**0.262**	0.363	0.148	**0.300**	0.418	0.172

Table(4) $ROUGE - L$ Score

Topic #	Model in [19]			Proposed Model 1 (Φ_1)			Proposed Model 2 (Φ_2)		
	$ROUGE - L$	Best Value	Worst Value	$ROUGE - L$	Best Value	Worst Value	$ROUGE - L$	Best Value	Worst Value
d061j	**0.542**	0.649	0.441	**0.554**	0.635	0.430	**0.565**	0.660	0.448
d062j	**0.473**	0.603	0.364	**0.481**	0.679	0.373	**0.526**	0.650	0.440
d063j	**0.493**	0.578	0.422	**0.528**	0.616	0.445	**0.532**	0.672	0.455
d064j	**0.462**	0.588	0.353	**0.488**	0.626	0.339	**0.508**	0.597	0.434
d065j	**0.431**	0.516	0.375	**0.457**	0.554	0.380	**0.461**	0.553	0.416
d066j	**0.455**	0.553	0.357	**0.441**	0.506	0.357	**0.465**	0.634	0.350
d067f	**0.509**	0.649	0.417	**0.529**	0.636	0.392	**0.541**	0.692	0.420
d068f	**0.666**	0.796	0.570	**0.626**	0.728	0.502	**0.634**	0.723	0.502
d069f	**0.454**	0.549	0.414	**0.476**	0.583	0.392	**0.470**	0.528	0.403
d070f	**0.496**	0.606	0.433	**0.513**	0.587	0.429	**0.536**	0.585	0.487

Table-5 reports $\boldsymbol{ROUGE}$ scores in terms of average ($\overline{\boldsymbol{ROUGE}}$) and standard deviation (σ) over all topics. From the results reported in Tables 3-5, one can easily see that the two proposed models perform better than the model proposed in [19]. Moreover, inspecting content coverage objective into three distinct similarity sub-objectives, as suggested in Φ_2, improve the overall quality of the generated summary.

Table (5) Average and standard deviation of ROUGE

Model	$\boldsymbol{ROUGE-2}$		$\boldsymbol{ROUGE-L}$	
	$\overline{\boldsymbol{ROUGE-2}}$	σ	$\overline{\boldsymbol{ROUGE-L}}$	σ
Φ_2	**0.283**	**0.082**	**0.523**	**0.049**
Φ_1	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. Two optimization models are introduced in this paper to satisfy *content coverage* and *diversity* in the document collection. An improved performance is reported by introducing the second model where text similarity has been decoupled along three dimensions: sentence to sentence similarity, sentence to document collection similarity and summary to document collection similarity. 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 models shows improvement over the model proposed in [19]. The results reported in this paper encourage us for further investigation study. The current interest is to take a further step towards capturing the essence of text summarization problem. Taking the benefit of implicit contradictory nature of both content coverage and content diversity, designing the text summarization problem can be modeled as a *multi-objective* optimization problem. Moreover, one of multi-objective evolutionary algorithms will be adopted to handle the formulated multi-objective problem.

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.714.

[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].Alguliev, R. M., Aliguliyev, R. M. and Mehdiyev, C. A. 2011. pSum-SaDE: A Modified p-Median Problem and Self-Adaptive Differential Evolution Algorithm for Text Summarization. Applied Computational Intelligence and Soft Computing. Volume 2011, Article ID 351498, 13 pages.

[23].Alguliev R.M., Aliguliyev R.M. and Hajirahimova M.S. 2012. Quadratic Boolean Programming Model And Binary Differential Evolution Algorithm For Text Summarization. İnformation Technology Problem, No 2(6), 20-29.

[24].Alguliev, R. M., Aliguliyev, R. M. and Mehdiyev, C. A. 2013. An Optimization Approach To Automatic Generec Document Summarization. Computational Intelligence, Volume 29, Number 1.

[25].Alguliev R.M., Aliguliyev R.M. and Isazade, N. R. 2013. CDDS: Constraint-driven document summarization models. Expert Systems with Applications 40 (2013) 458–465.

[26].Alguliev R.M., Aliguliyev R.M. and Isazade, N. R. 2013. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications 40 (2013) 1675–1689.

[27].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.

[28].Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval, *Information Processing & Management* 25 (5) 513–523.

[29].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.

[30].Document understanding conference: http://duc.nist.gov.

[31].Porter stemming algorithm: http://www.tartarus.org/martin/PorterStemmer/.

[32].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.