ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my supervisor, Dr. Albaraa Abuobieda for attention, encouragement, guidance and support throughout the length of this research; he has greatly helped me in a lot of ways I needed to go through this research, and he helped me to know about so many new things.

ABSTRACT

Automatic text summarization is the process of creating a small version from the original text. Extraction approach is one of way of extracting the most important sentences in document, this approach is used to select sentences after calculating the score for each sentence, and based on user defined summary ratio the top n sentences are selected as summary. The selection of the informative sentence is a challenge for extraction based automatic text summarization researchers. This research applied extraction based automatic single document text summarization method using the particle swarm optimization algorithm to find the best feature weight score to differentiate between important and non important feature. The Recall-Oriented Understanding for Gisting Evaluation (ROUGE) toolkit was used for measuring performance. DUC 2002 data sets provided by the Document Understanding Conference 2002 were used in the evaluation process. The summary that generated by PSO algorithm was compared with other algorithm (GA,ACO) and used DE algorithm as benchmark. Experimental results showed that the summaries produced by the DE algorithm are better than another algorithm.

المستخلص

التلخيص الآلى للنص هو عملية إنشاء نسخة مصغرة من النص الأصلى. طريقة الاستخراج هي أحد طرق استخراج الجمل الأكثر أهمية في المستند. هذه الطريقة تستخدم لاختيار الجمل الأكثر اهمية في المستند بعد حساب النتيجة لكل جملة، واعتماداً على نسبة الاختصار المحدد بواسطة المستخدم يتم اختيار أعلى "ن" جملة كاختصار. اختيار الجملة الغنية بالمعلومات يمثل تحدى للباحثين الذين يعتمدون على منهجية الاستخراج. هذا البحث طبق طريقة التلخيص الآلى لمستند واحد اعتماداً على منهجية الاستخراج عن طريق خوارزمية أمثلية سرب العناصر(PSO) لايجاد أوزان السمات للتفريق بين السمات الهامة وغير الهامة. لقد تم استخدام أدوات (ROUGE) لقياس الأداء واستخدمت مجموعة من البيانات تسمى(2002 DUC) لعملية التقييم. ولقد تم مقارنة التلخيص المولد بواسطة خوارزمية أمثلية سرب العناصر مع الخوارزمية الجينية (GA) ، وخوارزمية أمثلية مستعمرات النمل(ACO)،واستخدمت الخوارزمية التطورية ((Differential Evolution(DE)) كمعيار. اظهرت النتائج التجريبية أن الملخصات التى تنتجها خوارزمية التطورالتفاضلي ((Differential Evolution(DE))هي أفضل من الخوارزميات الأخرى.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IR	-	Information Retrieval
PSO	-	Particle Swarm Optimization
GA	-	Genetic Algorithm
ACO	-	Ant colony optimization
ROUGE	-	Recall-Oriented Understudy for Gisting Evaluation
DUC	-	Document Understanding Conference
AVG-P	-	Average Precision
AVG-R	-	Average Recall
AVG-F	-	Average F-measure

LIST OF SYMBOLS

Σ	-	Sum
$\in$	-	Element of
$\cup$	-	Union
$\cap$	-	Intersection

LIST OF APPENDICES

CHAPTER 1
INTRODUCTION

1-1 Introduction

At the present time, internet is widely used to find information through information retrieval (IR) tools as search engines. However the big growth of information on the internet makes the information abstraction of retrieved results has become a necessity for users. A Process of producing summary for document keeps the main content that helps users to understand and interpret large volume of information available in the document summarization. Summary that create by human is called manual summarization. Summarization that done by humans involves reading and understanding an article, web site, document, etc. summary that create by machine is called automatic summarization. The needs for automated summaries is becoming more and more apparent to automatically generating the summary and get the rich information of long textual data. Nowadays, information overload, text summarization has become an important for user to quickly understand the large volume of information. Text summary is a shorter version of the original document that keeps the main content of information in the document. This task is performed by human after deep reading and selecting the most important information and paraphrasing them into shorter version. There are several areas to take advantage of automatic text summarization such as email summary, news articles summary, short message news on mobile, information summary for businessman, government officials, and research, etc.; online search engines and so on.

The earlier effort on automatic text summarization system that more developed in the late 1950s formed of selecting important sentences from original document and concatenating them into shorter form. Automatic text summarization techniques are classified into different approaches. Some of these techniques are classified based on the input document used for the summary. Single document summarization uses only one document to produce a single summary while multi-document summarization uses many document that are related to the some topic to create a single summarization. The summarization methods can be classified into two approaches: extraction and abstraction (Lin, 1997). An extractive summarization consists of selecting important sentences form the original document and concatenating them into shorter form. An abstractive summarization is a summary at least some of whose material is not present in the input (Mani, 2001). The task of evaluation of the quality of summary is very important. The evaluation can be assessed manually and/or automatically. Manual evaluation is done by human or automatic by special tools. Two categories of methods used in text summarization are extrinsic and intrinsic (Jing et al., 1998; Mani and Maybury, 1999; Afantenos et al., 2005). Extrinsic evaluation measures the efficiency acceptability of summaries in some task based on the idea of how useful the summaries are for a given task, for example reading comprehension or relevance assessment. On the other hand, intrinsic evaluation measures a summary quality of the system of itself by comparison to some gold standard such as human generated summaries. The evaluation can be assessed manually and/or automatically. Manual evaluation is done by human. The Recall-Oriented Understanding for Gisting Evaluation (ROUGE) (Lin, 2004), for example, is an automatic intrinsic evaluator of summary systems for the Document Understanding

Conference (DUC). ROUGE is said to correlate highly with the results of human judgments of content and quality (Lin, 2004).

1-2 Problem Background

The need to automatic summarization becomes more important because the large volume of text document information overload. The first work in automatic text summarization was introduced by Luhn (1958). Automatic text summarization is using extraction and abstraction method. Most of the researchers are focus on extraction method. In extractive summarization, important feature such as the words in the sentence have high frequencies, sentence length, key word in the sentence, are concatenated to makes the summary. The summary quality is sensitive to feature scores. These features are differentiated according to their importance. In this study the researcher used supervised machine learning to produce feature weights.

1-3 Problem Statement

The features are the main entries in text summarization. Treating all features equally causes poor summary generation. Building an optimal feature weighting mechanism for high quality summary generating is considered a complex task.PSO was proposed before to solve this problem but when we surveyed this research area, especially text summarization based on evolutionary and swarm intelligence algorithms, we found that previous works established an unfair comparison criterion to test which evolutionary and swarm algorithm is better. In this work, we unified the criteria to judge which algorithm is better to be realized for implementing text summarizer application.

1-4 Research Objectives

The goals of this research are:

- To apply PSO algorithm as machine learning to learn feature-weights with our new identified criteria.
- To compare the quality of summary that adjusted its weights of the features by PSO algorithm with other evolutionary and swarm algorithms.

1-5 Research Questions

The main questions which must be answered in dealing with such a problem are as follows:

1- Can feature selection and feature weight adjustment based on Particle Swarm Optimization algorithm (PSO) produce a good summary?

2- Is reimplementation of PSO algorithm as found by (Binwahlan *et., al* 2009a) based on our new criteria may make it perform well in contest was found by (Albaraa *et al*., 2013)?

1.6 Research Scope

In this thesis, the text will be summarized by using the PSO algorithm. The PSO algorithm is trained using DUC 2002 dataset to learn the weight of each feature. This research aims to establish a good bench mark criteria for comparing different algorithms in single application.

1.7 Research Significant

The feature is an important component in text summarization. This study selected five features to apply PSO algorithm to learn feature-weights, and compare it with another algorithms that used the same features to determine which algorithm is better performance.

1.8 Thesis Structures

Chapter 1 gives an introduction that includes problem background, problem statement, research objectives, research questions, research scope, and research significance. Chapter 2 presents a background and related work for previous study in the research area. Chapter 3 describes the design and implementation PSO algorithm for text summarization. Chapter 4 describes the results and Discussion. And finally, Chapter 5 describes the Conclusions.

CHAPTER 2
BACKGROUND AND RELATED WORK

2-1 Background

2-1-1 Introduction

This chapter aims to present an overview on the basics of text summarization, types of the summarization, some areas in which the summarization has been applied and a number of significant efforts, which have been done in the automatic text summarization field. A theoretical explanation on the fundamental methods on which the current study is expected that depend on them was presented. The most important evaluation measures of automatic text summarization are also presented.

2-1-2 Text Summarization

Text summarization is a process of rewriting text into a shorter compressed form to represent the original text. This task is accomplished by humans after deep reading and well understanding of the document content, selecting the most important points and paraphrasing them to short version. In daily life, the people deal with different kinds of summaries such as news headlines, abstract of scientific publication, search results retrieved by a search engine, reviews of movies, overview of books, and so on (Mani , 2001). Newspaper headlines are a natural example of human summarization.

Automatic text summarization is a summary generation by machine. The aim of automatic text summarization is to condense the source text by extracting its most important content that meets a user's or application's needs (Mani, 2001). Summarization is a challenging problem because the characteristics of informativeness, readability, robustness, and length reduction. Those factors must be taken into account when dealing with this problem (Melander, 1993).

2-1-3 Text Summarization Basic Concepts

This section introduces the basic concepts used in the field of automatic text summarization (Lamkhede, 2005).

Coherence: A summary is said to be coherent if all its sentences or text units form an integrated whole and the sequence of ideas progressed logically.

Compression Rate: It is a ratio of summary length to source length expressing the degree of summarization required. It is calculated as:

$$\frac{\text{Summary Length}}{\text{Source Length}} \qquad (2.1)$$

Salience or Relevance: It is the information score expressing both the information relevance to the user’s or application’s need and the content of the document.

Compaction of text: It is a process of removing less salient phrases or words from sentences.

A generic summary: It presents the main topics or the most important content of the document.

A query or topic specific summary: It contains the document information that is relevant to the user’s need.

Critical summary: It contains the abstractor's opinions towards the quality of the source for evaluation purpose.
A summarizer: It is a system that creates the summary.

Monolingual Summarizer: It uses just one language for input and output.

Multilingual Summarizer: It has the ability to use many languages with output in the same language as the input.

Crosslingual Summarizer: It has the ability to use many languages with output in different language from the input.

Single Document Summarizer: It summarizes one document and produces a single summary.

Multi-Document Summarizer: It summarizes many documents and produces a single summary.

2-1-4 Text Summarization Approaches

There are two approaches for text summarization, abstraction and extraction. Extraction approach focuses on the selection of particular pieces of text from a document where the sentences and/or phrases with the highest score are considered as salient sentences and are chosen to form the summary. Abstraction approach is a more complicated task than extraction, It needs to deep understanding of the main concepts in a document by using linguistic methods in natural languages and generating a new shorter text may different from the original text document. The complexity of abstraction makes extraction more widely used in automatic text summarization.

2-1-5 Summary Types

There are four types of summaries: indicative summary, informative summary, critical summary and extract. The two top are most important. An Informative summary is to replace the original document that contains all important contents of document. An indicative summary is a condensed version of the article contents avoiding the presentation of the content details to attract the user into getting the whole document, (e.g. movie trailers, Book jackets, Headline and scientific abstract) (Mani and Maybury, 1999).

2-1-6 Automatic Text Summarization System

Automatic text summarization system is represented in Figure 2.1. The automatic text summarization process consists of three stages (Mani, 2001):

- Analyzing stage utilizes linguistic and semantic information to determine facts about the input text. This requires some level of understanding of the words and their context (discourse analysis, part of speech tagging, etc.)
- Transformation stage uses statistical data and semantic models to generalize the input text and transform it into a summary representation.
- Synthesizing stage depends on the information created from the previous two stages to synthesize an appropriate output form.

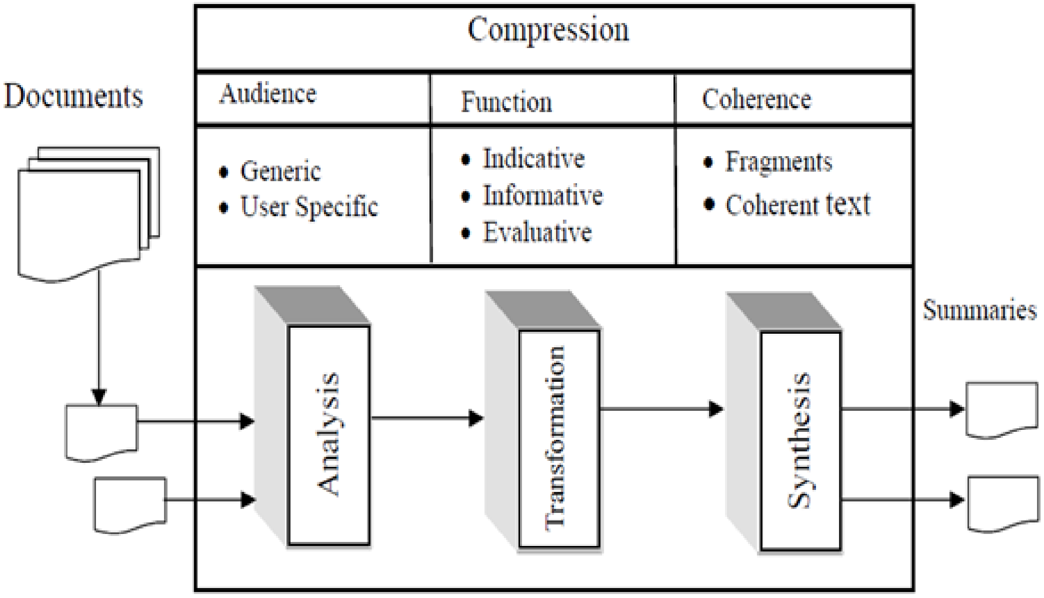


Figure 2.1 A typing automatic text summarization system (Mani and Maybury, 1999)

2-1-7 Summarization Applications

The summarization was used in several areas, including:

- Voice mails. In Koumpis and Renals's system (2005), the summary words are identified through a set of classifiers. The generated text summaries are appropriate for the applications of mobile messaging.
- Multi-party dialogs. Zechner (2002) presented a dialogue summarization system for automatically creating extract summaries for open-domain spoken dialogues in multiparty conversations.
- Newsgroups. Newman and Blitzer (2003) described an approach to condense the threads of archived discussion lists; they clustered messages into topic groups, and then extract summaries for each messages group.

- Blogs. Zhou and Hovy (2006) described computational approaches to summarize two types of data, which are blogs and online discussions.

2-1-8 Text Summarization Techniques

There are different categories of approaches for summarization techniques. These techniques are classified into two main categories: single document summarization techniques and multi-document summarization techniques. In the following subsection reviewed these techniques.

2-1-8-1 Single-Document Summarization

The work on summarization began as early as fifties when the first summarization research was presented (Luhn, 1958). It is considered the cornerstone for all works which followed it. Single document summarization is producing a single summary from only one document.

2-1-8-1-1 Machine learning-based approaches

The appearance of machine learning methods in Natural Language Processing (NLP) in 1990s encouraged many researchers to conduct many studies in the summarization to generate summaries. This kind of methods falls into two categories: supervised learning and unsupervised learning. For supervised methods, a prior knowledge is needed for derivation or estimation purposes. Mostly human generated summaries are used as a prior knowledge for such purpose. Unsupervised methods have the ability to estimate the required features and coefficients without making use of a prior knowledge. Kupiec *et al.* (1995) developed a system called “A Trainable Document Summarizer” based on Bayesian classifier algorithm where five weighting heuristics are employed in the system, as follows.

- Sentence length cut-off feature. The sentence consists of a number of words less than a predefined threshold to be excluded from the summary.
- Uppercase word feature. Proper names are considered as an uppercase thematic word under some conditions.
- Paragraph feature. This represents a sentence position in the paragraph (initial, final or middle).
- Fixed-phrase feature. Sentence including any of certain cue words or appearing directly after a section header comprising a keyword is included in the summary.
- Thematic word feature. The thematic words are the most frequent words and their function of frequency is the sentence score.

Based on those features, the score of each sentence is calculated using Bayesian classifier algorithm where the classification algorithm computes the probability of each sentence. The decision to include the sentence in the summary or excluding it is made based on its probability. If the sentence probability is equal to 1, it means the inclusion decision is taken. If it is 0, the exclusion decision is taken. The features paragraph feature, fixed-phrase feature and thematic word feature have been used previously by Edmundson (1969). Lin and Hovy (1997) built their method on the idea that the most important sentences tend to appear in fixed locations like title. The position method works through determining the sentence score by its position in the text. The manual topic words were used to calculate the yield of each sentence position. Lin (1999) examined decision tree as a machine learning method, the goal of his study was to investigate the influence of the topic importance and the query type on the performance of the heuristics, like Title and TF scores. Conroy and O'leary (2001) proposed a method to produce the generic extracts using a hidden Markov model that decide the likelihood of the including sentence in/excluding the sentence from the summary. Neto et al. (2002) used two machine learning approaches: Bayes and C4.5 which is a decision-tree algorithm to produce a trainable text summarizer. The set of the extracted features extracted from the original text is used to classify the sentences into summary sentences and un-summary sentences. The results showed that Naive Bayes classifier-based method outperforms the C4.5 classifier-based method. Fattah and Ren (2009), as part of their work, trained GA for producing weights of the features, where the average precision was used as fitness function, the method was proposed for single document summarization.

2-1-8-2 Multi-Document Summarization

Multi-document summarization is producing a single summary from multi-document. Generating a summary for multi-documents was more interesting by the mid 1990s. Generating such summaries must take into account the following things: keeping the important ideas in each document, reducing the size of each document and comparing ideas across documents (Mani and Maybury, 1999).

2-1-9 Swarm Intelligence

Computational Intelligence (CI) is a human made system for borrowing some essential properties of life being. The Computational Intelligence techniques helped in solving many computational problems. The CI was divided into many parts such as evolutionary computation and swarm intelligence. Swarm Intelligence (SI) is the collective intelligence resulting in the collective behaviors of individuals interacting locally and with their environment causing coherent functional global patterns to emerge (Ahmed, 2004). Figure 2.2 shows the swarm intelligence in nature. Particle swarm optimization (PSO) is inspired from the social behavior of bird flocking or fish schooling and Ant colony optimization (ACO) is inspired from behavior of ants.

Figure 2.2 Swarm Intelligence in nature

2-1-9-1 Particle Swarm Optimization

Particle swarm optimization was proposed by James Kennedy & Russell Eberhart (1995) .It relates both Computational Intelligence in general and bird flocking, fish schooling and swarming theory in particular. It keeps also a relation with evolutionary computation, genetic algorithm and evolutionary programming. Therefore, the particle swarm optimization can be defined as a stochastic, population-based evolutionary algorithm for problem solving. In general the idea of PSO method is to simulate the shared behavior happening among the birds flocks or fish school. PSO applies the concept of social interaction to problem solving.

2-1-9-1-1 Particle Swarm Optimization Mechanism

Particle swarm optimization depends on methodology of a population of individual to discover favorable regions of the search space. Every member in the population is called particle and the group of all particles is called a swarm. Each particle flies in the search space with a velocity that is dynamically adjusted according to its own flying experience and its companions' flying experience and retains the best position it ever encountered in memory. The best position ever encountered by all particles of the swarm is also announced to all particles, Figure 2.4 illustrates this process for each particle. In the local variant topology, each particle can be assigned to its neighbors group, which comprises a predefined number of particles, (Ahmed, 2004; Shi and Eberhart, 1998). The work of PSO starts by initially randomizing a group of solutions (particles), the swarm will update its best value in every cycle based on the equations (2.2 and 2.3) and then after several iterations finds the optimized solution. The general description of the PSO algorithm work is shown in Figure 2.3.The PSO has many versions such as continuous particle swarm optimization and binary particle swarm optimization. In following subsections discussed both two versions.

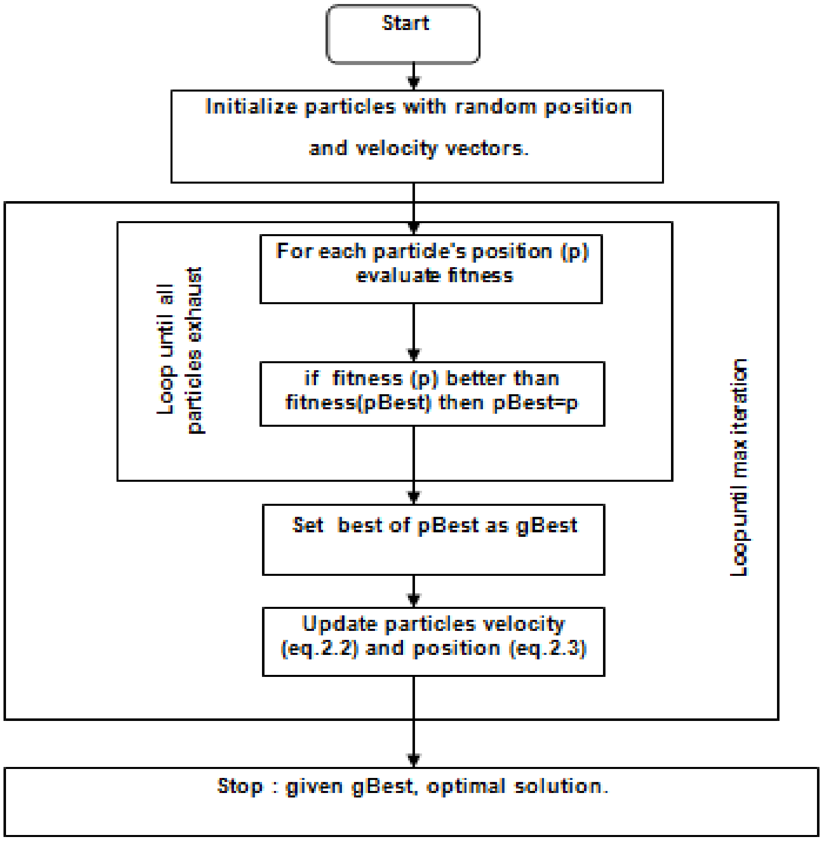

Start
Initialize particles with random position and velocity vectors.
For each particle's position (p) evaluate fitness
if fitness (p) better than fitness(pBest) then pBest=p
Loop until all particles exhaust
Set best of pBest as gBest
Update particles velocity (eq.2.2) and position (eq.2.3)
Loop until max iteration
Stop : given gBest, optimal solution.

Figure 2.3 Show flow chart of general PSO Algorithm

2-1-9-1-2 Continuous Particle Swarm Optimization

The Continuous Particle Swarm Optimization is the standard version of PSO is to optimize continuous nonlinear problems, which was presented by Kennedy and Eberhart (1995), where it consists of two equations: the velocity of the particle (Eq. 2.2) and the position of the particle in the D-dimension search space (Eq. 2.3).

$$\mathrm{V}_{id}(t+1) \leftarrow \mathrm{V}_{id}(\mathrm{t}) + \mathrm{c}_1\mathrm{r}_1(\mathrm{p}_{id}(\mathrm{t}) - \mathrm{x}_{id}(\mathrm{t}) + \mathrm{c}_2\mathrm{r}_2\left(\mathrm{p}_{gd}(\mathrm{t}) - \mathrm{x}_{id}(\mathrm{t})\right) \qquad (2.2)$$

$$\mathrm{X}_{id}(t+1) \leftarrow \mathrm{X}_{id}(\mathrm{t}) + \mathrm{V}_{id}(t+1) \qquad (2.3)$$

The velocity of the particle (Eq. 2.2) consists three parts: The velocity $V_{id}(t)$ of the particle i in the time point t in the D-dimension search space. The cognitive part $c_1r_1(p_{id}(t)-x_{id}(t))$: the cognitive part concerns the influential elements on the velocity of the particle resulting in its own behavior, where $p_{id}(t)$ is the best position in which the particle previously got high fitness value, it is called p*best*, $x_{id}(t)$ is the current position of the particle i in the search space, c_1 is a parameter used as weight to determine how much the particle is influenced by p*best* and r_1 is random generated number in the range [0,1]. The social part $c_2r_2(p_{gd}(t)- x_{id}(t))$: the velocity of the particle receives a different influence from the social part controlled by the overall best position $p_{gd}(t)$ in which a particle got best fitness value, it is called the g*best*, c_2 is a parameter used as weight to determine how much the particle is influenced by g*best*, r_2 and $x_{id}(t)$ are the same case as in the cognitive part. The parameters c_1 and c_2 are called acceleration parameters. The position of the particle (Eq. 2.3): it is the new position $x_{id}(t+1)$ which the particle must move to, where $x_{id}(t)$ is the current position of the particle and $V_{id}(t+1)$ is the new velocity of the particle resulting in the calculation in (Eq. 2.2) which mainly determines the new position of the particle. Figure 2.3 shows the applying of the two equations (Eq. 2.2 and Eq. 2.3) in the practice by the particle.

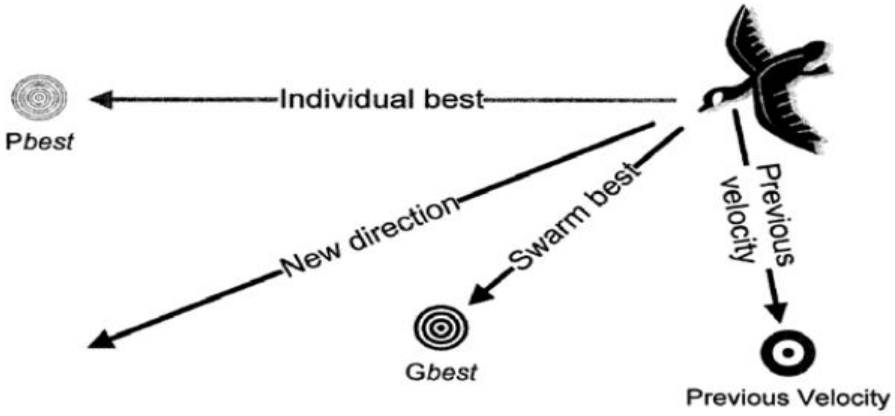


Figure 2.4 The behavior of a particle in the search space to find the optimal solution (Alrashidi, 2007).

2-1-9-1-3 Binary Particle Swarm Optimization

The binary Particle Swarm Optimization (Kennedy and Eberhart, 1997) is an extension of continuous PSO, in which the particle position is represented as bit string, the update of the position in continuous PSO is done directly by adding the velocity to the previous position but in binary PSO, the velocity is used sigmoid function to calculate the probability of the bit value to be changed to 1 or 0, where the value retrieved from the sigmoid function is compared with random generated value in the range between zero and one. Equation (2.4) shows how to calculate a particle position change.

$$\mathrm{x}_{\mathrm{ij(t+1)=}} \begin{cases} 0 \quad \text{if } \mathrm{p_{ij}(t)} \geq \dfrac{1}{1+\exp\left(-\mathrm{v_{ij}(t)}\right)} \\ 1 \quad \mathit{otherwise} \end{cases} \tag{2.4}$$

2-1-10 Evaluation Measure

The task of evaluation quality of summary is very important. The evaluation can be assessed manually and/or automatically. Evaluation measures in text summarization can be categorized into two types: extrinsic and intrinsic. Extrinsic evaluation measures the efficiency acceptability of summaries in some task based on the idea of how useful the summaries are for a given task. Intrinsic evaluation measures a summary quality of the system of itself by comparison to some gold standard such as human generated summaries.

2-1-10-1 Precision, Recall and F-measure

In text summarization systems, extraction approaches are commonly use. These approaches depend on selecting the most important sentences in the source text into summary without change the original sentences. In such setting, the commonly used information retrieval metrics of precision, recall, and F-Score. The summary that generated by human is a best choose for evaluation. Therefore, the generated summaries in this study evaluated and compared with the human generated summaries. (Nenkova ,2006) defined “precision” and “recall” for automatic text summarization as follows. Precision (P) is the number of sentences intersected between the system summary and human summary divided by the number of sentences in the system summary; see Equation (2.5). Recall (R) is the number of sentences intersected between the system summary and human summary divided by the number of sentences in the model summary; see Equation (2.6). The F–score measure is used to balance system performance on both “precision” and “recall” measures; see Equation (2.7).

$$\text{Precision} = \frac{|\text{system summary}| \cap |\text{human summary}|}{|\text{system summary}|} \qquad (2.5)$$

$$\text{Recall} = \frac{|\text{system summary}| \cap |\text{human summary}|}{|\text{human summary}|} \qquad (2.6)$$

$$\text{F} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2.7)$$

2-1-10-2 ROUGE: methodology of evaluation

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a system for measuring the quality of summaries by comparing it to summaries are created by humans, ROUGE is proposed by (Lin, 2004), the ROUGE tool depends on counting n-grams co-occurrences in the system summary and in the reference summary. ROUGE provides four different measures, namely ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. The following subsection discussed ROUGE-N.

2-1-10-2-1 ROUGE-N

ROUGE-N measures co-occurrences of n-grams. The ROUGE-N score can be calculated as:

$$\text{ROUGE} - \text{N} = \frac{\sum_{gram_n \epsilon s} Count_{match}(gram_n)}{\sum_{gram_n \epsilon s} Count(gram_n)} \qquad (2.8)$$

Where S is the reference sentence, n is the length of the n-gram, Count*match* (gramn) is the shared ngrams between a set of reference summaries and a system generated summary and Count (gramn) is the number of n-grams occurred in the system summary, ROUGE-N measures the n-gram recall. Let's look at one short example of a reference compression R and a candidate compression C:

R: The cat is on the wall.

C: The cat on the wall.

In this example, the ROUGE-1 score is 5/5; the ROUGE-2 score is 3/5.

2-1-11 Text Summarization Data set

The data set or corpus is a very important component in soft computing techniques method. In supervise machine learning the data set is use as pervious knowledge. There are many data set were proposed and presented in a number of conferences and workshops such as "SummBank" data set is multi-document and multi-language data set used for summarization documents written in English and Chinese (Saggion *et al*. ,2002). The "CAST" data set is a supervised summarization (Hasler *et al*., 2003). The "Ziff-Davis" data set is presented for a summarization of sentence reduction (Harman and Liberman, 1993). The "DUC" data set is one of the data set that has been widely used in automatic text summarization. In this research we used one of DUC data set called DUC2002. The following subsection is describes the duc2002 data set.

2-1-12 DUC 2002

DUC 2002 (document understanding conference 2002) data set were used in evaluation process of automatic summarization. DUC 2002 produce by (NIST)National Institute of standards and technology of U.S, Its contains a large set of documents with human created summaries for comparison, each document is supplied with a set of human generation summarization provided by two different experts. The data in any document related to four different categories: single natural disaster event, single event in any domain, multiple distinct events of a single type, and biographical information.

2-2 Related Work

2-2-1 PSO-Based Text Summarization

The particle swarm optimization (PSO) (Kennedy and Eberhart, 1997) is used in the current study as machine learning for features selection problem in order to study the feature structure effect on the feature selection, with the main result obtained is the learned features weights. The features scores will be combined with the features weights produced by PSO in a proposed model for automatic text summarization problem. PSO was successfully applied in some related problems like text classification. The particle swarm optimization was also applied successfully in the feature selection problem. Liu et al. (2004) used particle swarm optimization to select a subset of features for classification and training of neural network. Tu et al. (2006) used particle swarm optimization (PSO) for feature selection in the classification problem. Lee et al. (2007) adapted PSO for feature selection to enhance the performance of support vector machines and neural networks to classify the power transformer faults. Lin et al. (2008) employed PSO with support vector machine for parameter determination and features selection for improving the classification .According to successes of PSO in above studies, Binwahlan et al. (2009a) employed PSO method to investigate the effect of feature structure on the feature selection process in text summarization area. The features used are divided into two types: "complex" features and "simple" features. Complex features are "sentence centrality", "title feature", and "word sentence score"; simple features are "keyword" and "first sentence similarity". After computing each feature score, the PSO was used to identify which features are more effective. Score of ROUGE-1 was used to calculate the value of the fitness function. The dataset used for training the system comprised one-hundred articles from DUC-2002. The PSO parameters were initialized and "g*bests*" values were computed to extract the weight of each feature. Results showed that complex features received higher weights than did simple features, which indicates that feature structure plays an important role in the feature selection process. Furthermore, to calculate the features weights, Binwahlan et al. (2009b) divided the dataset into training and testing phases. They assigned 99 documents to train their PSO algorithm, while the 30th document was assigned to test the model. Consequently, scored sentences are ranked in a descending manner with top "n" sentences selected for summary where "n" is equal to summary length. To assess the results, the authors installed one human model summary as a reference and a second as a benchmark. The Microsoft Word-Summarizer and the first human summary were compared. The result showed that PSO outperformed the MS-Word Summarizer and achieved outcomes closest to the human model. The maximal margin relevance (MMR) is a method proposed by (Carbonell and Goldstein, 1998) to enhance summary diversity. Binwahlan et al. (2009c) continued optimizing the summarization problem using the PSO combined with the Maximal Margin Importance (MMI) technique. The MMI technique is a method derived from the maximal margin relevance (MMR) that enhances summary diversity. The general idea of MMI is to select a sentence that is both highly relevant to the document topic but with low relevance to selected sentences in the summary (redundancy).

2-2-2 GA-Based Text Summarization

Genetic algorithm is an optimization algorithm, which based on Darwinian principle of natural selection. (Pooya Kkhostaviyan Dehkordi,

2009) presented a genetic extractive based multi-document summarization. The genetic was also used to extract the weights of features. (Fattah and Ren, 2009) are used GA to extract features weights. (Yeh *et al.*, 2005) are used GA to extract features weights. (Suanmali *et al.*, 2011) proposed a GA for extract features weights.

2-2-3 DE-Based Text Summarization

Differential Evolution algorithm is one of an evolutionary algorithm. DE was originally presented by Storn and Price (1997). (Alguliev and Aliguliyev, 2009) presented a DE-Based text summarization for extractive-Based in multi-Document summarization. (Alguliev *et al.*, 2011) proposed a self-adaptive optimization based method for multi-Document summarization problems. (Alguliev *et al.*, 2012) published a multi-Document Summarization method. (Albaraa Abuobieda *et al.*, 2013) proposed DE algorithm as method for extractive features weights from single-document summarization, this work was compared with other algorithms and the results were good, the features that used in this work is the same features that used in this research.

2-2-4 ACO-Based Text Summarization

Ant Colony optimization is a method of heuristic search using in general artificial intelligence (swam intelligence), it simulate the behavior of the aggregate food for ants to find new solution for optimization problems. Inherently the Ant is able to find the shortest path from the nest to food source. The basis of the mathematical model for ant colony is the natural behavior of ants. The ant puts aromatic substance (pheromone) on the ground to determine paths between the source of food and their colony should be followed by the rest of the members of the colony. With passage of time, evaporate this substance aromatic, but this substance remain high proportion of these roads with the shortest distance it takes for the ant to go back again to colony. Thus, the ants follow the shorter paths that contain a higher amount of aromatic substance. This natural pheromone was the basis for the construction of the ACO algorithm. Several different aspects of the behavior of ant colonies have inspired different kinds of ant algorithms. ACO algorithm Used to solve a lot of issues that need to be the optimal solution. The first algorithm called the Ant System was initially proposed by (Marco Dorigo, 1991). To the best of our knowledge, the ACO algorithm is never used to solve the problem of ATS. As we showed in the literature review, many of evolutionary computational algorithms were presented enhance the performance of text summarization methods. The problem of these methods (PSO, DE,GA) are as fellows, Since these methods were designed in good way, but they compared unfair. We noted that, the number of features are differ, the structure of features are also not similar are differ, the number of documents are not similar and list structure of documents are also different. Due to this note an evolutionary computation research group was established. In next section, we describe this group.

2-2-5 Research Group

This group consists of four members; Dr. Albaraa Abuobieda[1] is the leader of the group. (Asem Abdulla, Abdelrahman Yousif and Omer Fisal)[2] are members of the group. In previous work has been the comparison between

[1] Dr.Albaraa Abuobieda Mohammed Ali he received his PhD from Univeristi Teknologi Malaysia in the area of Text Summarization. He received his B.Sc in Computer Science from the International University of Africa, Sudan, in 2004. He earned M.Sc in Computer Science from Sudan University of Science and Technology in 2008. His current areas of research include text summarization, plagiarism detection, Ontology, network and network security. Currently he is a dean of the Faculty of Computer Studies - International University of Africa.

[2] Asem Abdulla, Abdelrahman Yousif and Omer Fisal are master students.

GA (Suanmali *et al.*, 2011), (Albaraa Abuobieda *et al.*, 2013) and (Binwahlan et al., 2009), this comparison was unfair because the use different features. (Asem Abdullah and Albaraa Abuobieda, 2014) are applied GA to extract features weights. (Omer Fisal and Albaraa Abuobieda, 2014) used binary ACO algorithm as new method to extractive features weights. This research applied PSO algorithm to extract features weights. These works are compared with DE algorithm (Albaraa Abuobieda *et al.*, 2013). These group is used the same features.

2-3 Summary

This chapter reviewed text summarization concepts, approaches, types and some details of automatic text summarization system. This chapter gave a brief about automatic text summarization. This chapter discussed two techniques of text summarization, Single-Document Summarization and multi-document summarization and reviewed machine learning approach as one of the approach that used in those techniques. Also, the particle swarm algorithm has been discussed above in some detail followed by its characteristics and continuous, binary versions. This chapter reviewed evaluation measures that used in the summarization. Finally, reviewed the work related to this research.

CHAPTER 3
RESEARCH METHODOLOGY

3-1 Introduction

This chapter present the methodology used in this research. It describes the implementation of the PSO algorithm to achieve the objectives of the research. There are three sections in this chapter, where section 3.1 is for introduction, section 3.2 describes the research design, and section 3.3 provides the operational framework of this research.

3-2 Research Design

This research is based on the functional approximation (random research) approach using an intelligent swarm algorithm named "particle swarm optimization". The PSO algorithm is provided with learning approach (feature weighting). The PSO used to obtain an appropriate feature weights. We have set the particle swarm optimization parameters as follows: number of particles=30, maximum number of iteration=500. In the previous works, it was found that those values of PSO parameters (except the number of particles, which can take any value) are suitable (Shi and Eberhart, 1998; Eberhart and Shi, 2001; Wang et al., 2007).

3-3 Operational Framework

The operational framework of this research consists of four phases, phase1: Basic Elements, Phase 2: Binary Particle Swarm Based Text Summarization, Phase 3: Training Procedure and Phase 4: Testing Procedure.

3-3-1 Phase 1: Basic Elements

This phase is responsible of collecting the standard data set for evaluate method. The document understanding conference (DUC 2002) has become the main standard data set used for evaluating automatic text summarization research. The DUC data set are collections of newswire articles. Subsection 3-3-1-2 describes the pre-processing steps that are required for configuring where the articles are processing. The feature scoring process is cornerstone in processing a document for automatic text summarization. Subsection 3-3-1-3 describes how to calculate a set of selected features over all preprocessed document in the data set.

3-3-1-1 The DUC2002 Data set

DUC2002 data set are created by National Institute of Standard and Technology of the U.S (NIST) which consists of 60 data sets. The following ten documents D075b, D077b, D078, D082, D087, D089, D090, D092c, D095c, and D096c comprising of one hundred documents are used see Appendix A for more details. Each document has two human-experts in field to produce two model summaries (see appendix B, C and d). The first is assigned as a reference summary and called (H1); the second is assigned as a reference method and called (H2). The goal of produce these models are comparison with summary that generated from the document.

3-3-1-2 Text Data Preprocessing

The text preprocessing is an important process in text summarization because the qualities of the generated summary depend on the efficient of the text representation. In this stage there are four main steps performed: sentence segmentation, tokenization, stop word removal, and stemming.

- Sentence Segmentation

Sentence segmentation is a task of separating source text into sentences after deleting sentence boundary. There are several notation marks that share the characteristic of sentence end point such as ".", "?", "!". Appendix E illustrates example of this process.

- Tokenization

Tokenization is a task of separating sentences into words. There are several notation marks that share the characteristic of word end point such as the tab, white space, colon, semi colon, comma, and so on.

- Stop Word Removal

Stop words are words that appear frequently in document but less effect in identifying the important content in a document such as "a", "the", "in", "and" etc. see Appendix F.

- Word Stemming

Stemming process is returning each word to its base or root. For example to stem the terms "fishing", "fished" and "fisher" into root from which is "fish".

3-3-1-3 The Selection Features

The features are used to extract salient sentences from the text. The features scoring process is cornerstone of the summary sentence selection approach. In this research, five features are selected to score each sentence in document. The features are: Title Feature "TF"(Edmundson, 1969), Sentence Length "SL"(Nobata *et al.*, 2001), Sentence Position

"SP"(Edmundson, 1969), Numerical Data "ND"(Fattah and Ren, 2009) and Thematic Word "TW" (Luhn, 1958, Edmundson, 1969, Luo et al., 2010).

1- Title Feature (TF)

The sentence that shared words with title gives a high score. We calculate the score for this feature using Equation 3.1. Where CountWord ( ) is a function used to count words of the input parameter such as i^{th} sentence in the document S_i that are intersected with the Title words, Count Length ( ) is a function computes the length of title.

$$\mathrm{TF(Si)} = \frac{\mathrm{CountWord(Si)} \cap \mathrm{CountWord(Title)}}{\mathrm{CountLength(Title)}} \qquad (3.1)$$

2- Sentence Length (SL)

The short sentences are not usually belonging to the summary. We use normalized length of the sentence by using Equation 3.2. Where S_i refers to number of words in i^{th} sentence in document and Sj refers to number of words in longest sentence in document, and CountLength ( ) is a function that computes the length of each input sentence.

$$\mathrm{SL(Si)} = \frac{\mathrm{CountLength(Si)}}{\mathrm{CountLength(Sj)}} \qquad (3.2)$$

3- Sentence Position (SP)

The position of sentence in paragraph can effect on generated summary. The first sentence in paragraph is considered an important sentence and highest ranking for generating summary. Equation 3.3 is used to calculate the SP feature, where S_i refers to i^{th} sentence in document, and CountTotal( ) is a function that retrieves the total number of the sentence in document d and CurrentPostion( ) is a function that retrieves the current order of sentence S_i in document d.

$$\mathrm{SP(Si)} = \frac{\mathrm{CountTotal(d)} - \mathrm{CurrentPosition(Si)}}{\mathrm{CountTotal(d)}} \qquad (3.3)$$

4- Numerical Data (ND)

Numerical Data are refers to the number of numerical data in a sentence such as a date, money transaction, and etc. a sentence that contains numerical data is important. Equation 3.4 is used to calculate this feature, where CountND( ) is a function that computes the Numerical Data in i^{th} sentence S in the document, and CountLength( ) is a function used to compute the sentence length of S_i.

$$\mathrm{ND(Si)} = \frac{\mathrm{CountND(Si)}}{\mathrm{CountLength(Si)}} \qquad (3.4)$$

5- Thematic Word (TW)

Thematic word is the words that have most frequencies in a document. We used the top ten words most frequency as thematic. Equation 3.5 is used to calculate this feature, where CountThematic ( ) is a function used to compute the number of thematic words in sentence s_i.

$$\mathrm{TW(Si)} = \frac{\mathrm{CountThematic(Si)}}{\mathrm{max(TW)}} \qquad (3.5)$$

This research depends on these five features. Table 3.1 shows feature score vector for a document after preprocessing process, this document contains ten sentences. The ratio of summary is 20%; according to equation 2.1 so the Top two sentences will be selected.

Sentence / feature	TF	SL	SP	ND	TW	Total
S1	0.83	0.47	0.36	0.94	0.52	3.12
S2	0.61	0.45	0.64	0.58	0.33	2.61
S3	0.73	0.61	0.81	0.57	0.63	3.35
...	...	...	...	...	...	...
S10	0.89	0.71	0.44	0.89	0.65	3.58

Table 3.1 Example for Feature Score vectors

3-3-2 Phase 2: Binary Particle Swarm Based text summarization

The feature scoring is considered the base of the text summarization process. The particle swarm optimization (PSO) is used as a machine learning method to learn the feature weight from the training data. The extracted weights are used to adjust the feature scores.

3-3-2-1 Particle Position Representation and Configuration

The binary PSO (kennedy and eberhart, 1997) is used which the particle position is represented as a bit string. Each bit takes the value of one or zero for represents the case of one feature. If the bit contain the value 1, that means the feature is selected otherwise the feature unselected. The first bit represents the first feature; second bit represents the second feature and so on. The particle position was represented as Figure 3.1.

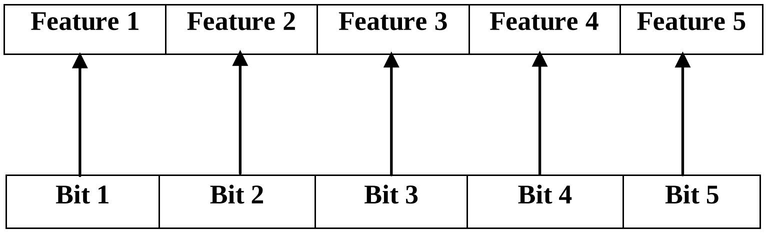


Figure 3.1 Structure of Particle (Binwahlan *et al.*, 2009)

3-3-2-2 Particle Velocity Representation and Configuration

The particle velocity is represented in the same way of particle position, where the value of each bit is retrieved from the sigmoid function. The velocity is used only in the sigmoid function as in equation 2.4(see chapter 2, subsection 2-1-9-1-3) to calculate the probability change the bit value to 1 or 0, where the value retrieved from the sigmoid function is compared with randomly generated value in range between zero and one. If the value retrieved from sigmoid function is less or equal that random number; the bit in the particle position is changed to 0 otherwise it is changed to 1.

3-3-2-3 Binary Modulation formula

Particle Swarm Optimization algorithm is one of the Swam Intelligence algorithms. The Swarm Intelligence algorithms are used real values. To enable PSO to search in a binary space, we need a modulation formula in order to modulate real value into binary values. In this research we used modulation as in equation 2.4(see chapter 2, subsection 2-1-9-1-3) to perform this task.

3-3-2-4 The Fitness Function

The fitness function is used as a measuring unit in optimization techniques. The fitness function is used to determine which particle obtains the best solution and is considered fittest value. This value will change if the new particle generated better fitness than previous particle. In this research the recall value of the generated summary is assigned as a fitness value for each particle. Recall has been successfully in many previous works. Equation 2.6 shows how to calculate the recall value for each generated summary.

3-3-3 Phase 3: Training procedure

In this research the PSO used 70 documents from DUC2002 data set in training stage. At begin each document is deal with by preprocessing process (sentence segmentation, tokenization, stop word removal and stemming), then extracting the text features. The score of each sentence features are present a vector see table 3.1. The resulting of the features scores are used as input for PSO scoring function Equation 3.6.

$$\mathrm{Score(si)} = \sum_{j=1}^{5} \mathrm{s}(fj) \times \mathrm{vopp(i)} \qquad (3.6)$$

Where Score (si) is the score of the sentence si, s(f j) is the score of the jth feature and vopp (i) is the value of ith bit in the position, after calculated the scores of document sentences by using Equation 3.6 and ranked in descending order selecting the top *n* sentence as summary, where *n* is a predefined summary length. In this research, the summary length is 20% of the total number of the document sentences. The PSO method uses Equation 3.7 to selects *n* sentences from the document to compose the summary. The generated summary is used as input for the fitness function. The ROUGE-1 is used as the fitness function (see chapter 2, subsection 2-1-10-2) for more details. depend on the summary evaluation, the p*best* is determined , that is indicate the evaluation value of the best summary generated by that particle, and also the g*best* is determined, that is indicate the evaluation value of the best summary generated by a particle in population. By the end of run, the position of the particle with the g*best* value is selected as a vector for the best selected features of each document. The final features weights are calculated over the vectors of the features weight of all documents in the data collection.

$$\text{Summary} - \text{Length} = \frac{(user\ defined\ Ratio \times document\ length}{100} \quad (3.7)$$

3-3-4 Phase 4: Testing procedure

The goal of employ the PSO is to find and optimize the corresponding weight wj of each feature fj. Equation 3.8 is calculating the features weights.

$$\text{score\_weight(si)} = \sum_{j=1}^{5} \text{w}j \times \text{score}_{\text{f}j\text{(si)}} \quad (3.8)$$

Where score_weights (si) is the score of sentence s, wj is the weight of the feature *j* that produced by PSO, *j* is the number of feature and score_fj(si) is a function that calculate the score of the feature *j*. The testing procedure used 30 documents from DUC2002 data set. The testing procedure is begin with input document, then implementing the preprocessing process (segmentation, tokenization, remove stop word and stem the word), then extracting features for each sentence, then modify the score of each feature based on the features weights that produced in training process, then calculate the score of each sentence in document by Equation 3.8, then order the sentences based on their score in descending order, then select top *n* sentence as summary sentence, where *n* is equal to predefined summary length, then order the summary sentences in the same order as in the original document. The ROUGE package (lin. 2004) is used as evaluation measure. Figure 3.2 illustrates testing procedure.

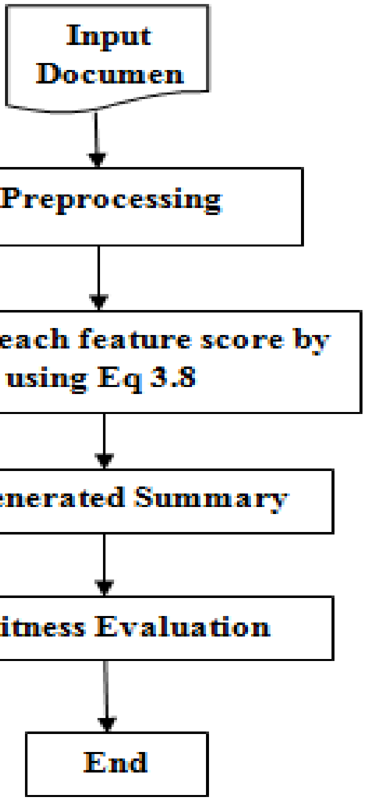


Figure 3.2 Testing Model

CHAPTER 4
RESULTS AND DISCUSSION

The previous works in extractive-based text summarization proved that designing a method with a powerful feature-weighting mechanism could generate a high quality text summary, so the quality of generate summary is sensitive to the selected features. Therefore, developing a mechanism to compute feature weight is very important. The weighting approach helps identify the importance of each feature separately in the document collection. Some researchers have proposed features weighting mechanisms using other optimization techniques such as Genetic algorithm (Fattah and Ren, 2009, Suanmali *et al* 2011), particle swarm optimization (Binwahlan *et al* 2009) and Differential Evolution algorithm (Albaraa Abuobieda *et al.*,2013), These methods are used different feature to generate summary. This chapter describes the results of the applied PSO algorithm in text summarization and compares the generated summary after apply features weights with other algorithms which are Ant-colony algorithm and Genetic algorithm; the Differential Evolution algorithm is used as benchmark. The H2-H1 Compression method is used as benchmark too; this method is produced from compare the human summary called (H2) with reference human summary called (H1). This compression is established to evaluate the summary against human performance. These algorithms are used same five statistical features (Title Feature, Sentence Length, Sentence Position, Numerical Data and Thematic Words) and same data set .ROUGE packet is used to evaluate the obtained results. When implementation of the testing process, we used ROUGE-N evaluation measure .ROUGE-N measure is counting all occurring (shared) words. The generated summary by these algorithms (PSO, GA, ACO) are compared with DE algorithm summary. Table 4.1, 4.2 compare the three methods using ROUGE-1, ROUGE-2. Average recall (avg-R), average precision (avg-P) and average F-measure (avg-F) are calculated for each method. Figures 4.1, 4.2 visualize the same results obtained.

Table 4.1: Methods comparison using ROUGE-1 result

Method	Avg-R	Avg-P	Avg-F
H2-H1	0.51642	0.51656	**0.51627**
DE	0.4561	0.52971	0.48495
ACO	0.3105	0.4508	0.3289
GA	0.3074	0.4169	0.3183
PSO	0.2871	0.4101	0.3011

Table 4.2: Methods comparison using ROUGE-2 result

Method	Avg-R	Avg-P	Avg-F
DE	0.2402	0.2841	**0.2568**
H2-H1	0.23394	0.23417	0.23395
ACO	0.1422	0.2318	0.1589
GA	0.1359	0.2028	0.1464
PSO	0.1023	0.1317	0.1017

Based on the generalization of the obtained results, the performance of the PSO model is (30%) similar to human performance (52%) using ROUGE-1 and (10%) similar to human performance (23%) using ROUGE-2. The performance of the PSO model is 30% similar to DE model performance (48%) using ROUGE-1, and 10% similar to DE model performance (26%) using ROUGE-2. Particle Swarm Optimization shows poor performance when compared with all algorithms, but the literature proved that PSO could outperformance several evolutionary computing algorithms in different domains. For instance, Binwahlan 2009a used binary PSO in text summarization for extract features weights but he used features differ from the features that used in this research.

Figure 4.1: Methods comparison using ROUGE-1 result

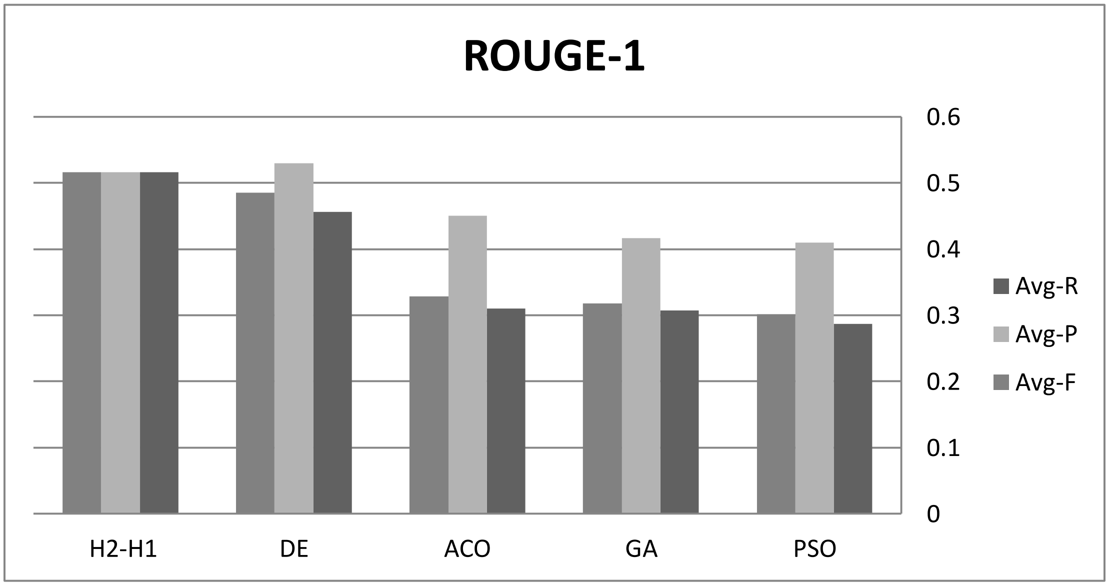
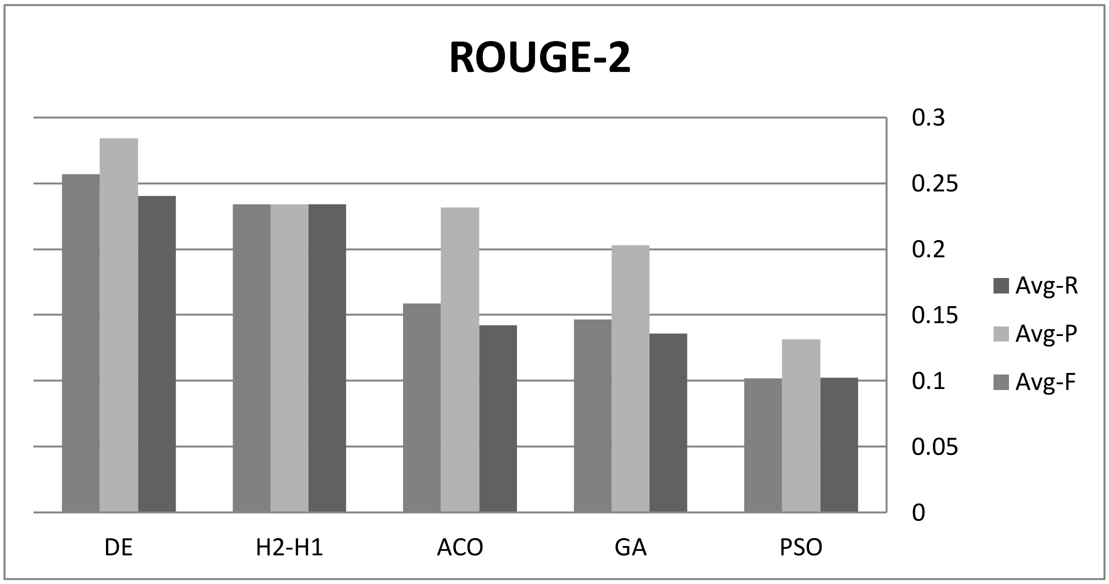


Figure 4.2: Methods comparison using ROUGE-2 result



CHAPTER 5
CONCLUSIONS

5-1 Introduction

In automatic text summarization, there are several techniques which used for selecting important sentences. The features were used to determine these sentences that should be selected in the final summary. The feature is an important component in the summary process. There are several methods proposed to study these features, and proved that unfair treatment features equally. The performance of feature-weighting in automatic text summarization has been proven to generate high quality summarization. This is presented in the related work of this thesis.

5-2 Particle Swarm Optimization Based Text Summarization

In this research, binary PSO used to obtain features weights. The standard version of PSO used real-values for position and velocity of the particles. The modulation function used to modulate the real-values into binary values to determine the inclusion of features for weighting. In this research, five effective statistical features were selected (Title Feature, Sentence Length, Sentence Position, Numerical Data and Thematic Word). This research contains three stages. The first stage is collection of data set. The second stage is training procedure, in this stage binary PSO used as machine learning to learn features. Binary PSO used to determine better particle for each document. The structure of particle includes five bits, each bit represent one feature. After end of PSO iterations, calculate the average of each feature, and that is called feature-weight. The third stage is testing procedure. In this stage used set of data set to testing process. After obtaining the feature-weight for each feature from training process, should adjusting each feature according to it feature-weight.the top *n* sentence select as summary. The generated summary in this stage compared with other algorithms (GA, DE and ACO). The summary that generated by DE algorithm is better than another algorithms. These algorithms used same features that used in this research. This research used standard data set called DUC2002, and standard evaluation tools called ROUGE.

REFERENCES

Afantenos, S.D., Karkaletsis, V. and Stamatopoulos, P. (2005). Summarization from Medical Documents: A Survey. *Artificial Intelligence in Medicine,* vol. 33, 157-177.

Ahmed, T. (2004). Adaptive Particle Swarm Optimizer for Dynamic Environments. Master Thesis. The University of Texas, Texas.

Alguliev, R. M. and Aliguliyev, R. M. (2009). Evolutionary Algorithm for Extractive Text Summarization. Intelligent Information Management. 1(2), 128–138.

Alguliev, R. M., Aliguliyev, R. M. and Isazade, N. R. (2012). DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. Knowledge-Based Systems.

Alrashidi, M. (2007). Improved Optimal Economic and Environmental Operations of Power Systems using Particle Swarm Optimization. PhD Thesis. Dalhousie University, Halifax, Nova Scotia.

Binwahlan, M., Salim, N. and Suanmali, L. (2009a). Swarm based features selection for text summarization. International Journal of Computer Science and Network Security IJCSNS. 9(1), 175–179.

Binwahlan, M., Salim, N. and Suanmali, L. (2009b). Swarm based text summarization. In Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of. IEEE, 145–150.

Binwahlan, M., Salim, N. and Suanmali, L. (2009c). Swarm Diversity Based Text Summarization. In Neural Information Processing. Springer, 216–225.

Blitzer and Newman (2003). Summarizing Archived Discussions: a Beginning. Proceedings of the 8th international conference on Intelligent user interfaces, 12-15 January. Miami, Florida, USA.

Conroy, J. M. and O'leary, D. P. (2001). Text Summarization via Hidden Markov Models. Proceedings of SIGIR '01. 9-12 September. New Orleans, Louisiana, USA, 406-407.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. Journal of the Association for Computing Machinery. 16(2), 264-285.

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.

Harman, D. and Liberman, M. (1993). Tipster complete. Corpus number LDC93T3A, Linguistic Data Consortium, Philadelphia

Hasler, L., Orasan, C. and Mitkov, R. (2003). Building better corpora for

Summarization. In Proceedings of Corpus Linguistics. 309–319.

Kennedy, J. and Eberhart, R. (1995). Particle Swarm Optimization. Proceedings of the IEEE International Conference on Neural Networks. 27 Nov - 1 Dec. Perth, Australia, 1942- 1948.

Kennedy, J., Eberhart, R. C. (1997). A discrete Binary Version of the Particle

Swarm Algorithm. Systems, Man, and Cybernetics. 'Computational Cybernetics and Simulation', IEEE International Conference on, 5. New York, 4104-4108.

Koumpis, K. and Renals, S. (2005). Automatic Summarization of Voicemail

Messages using lexical and prosodic features. *ACM* Transactions on Speech and Language Processing.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A Trainable Document Summarizer. In Proceedings of the ACM. SIGIR conference. July. New York, USA.

Lamkhede, S. (2005). Multi-Document Summarization using Concept Chain Graphs. Master Thesis. State University of New York, New York.

Lee, T., Cho, M. and Fang, F. (2007). Features Selection of SVM and ANN using Particle Swarm Optimization for Power Transformers Incipient

Fault Symptom Diagnosis. International Journal of Computational Intelligence Research. 3(1), 60-65.

Lin, C. Y. (1999). Training a Selection Function for Extraction. In Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM). 2-6 Nov. Kansas City, Kansas, 55-62.

Lin, C. Y. and Hovy (1997). Identifying topics by position. In Proceedings of the Fifth conference on Applied natural language processing, San Francisco, CA, USA, 283-290.

Lin, C. Y. and Hovy, E. (1997). Identifying Topics by Position. In Proceedings of the Fifth conference on applied natural language processing. March. San Francisco, CA, USA, 283-290.

Lin, C.Y. (2004). Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., editor, Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain. 74-81.

Lin, S., Ying, K., Chen, S. and Lee, Z. (2008). Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines. Expert Systems with Applications. 35, 1817–1824.

Liu, Y., Qin, Z., Xu, Z. and He, X. (2004). Feature Selection with Particle Swarms. In Zhang, J., He, J.-H. and Fu, Y. (Eds.). Computational and Information Science, LNCS 3314. (pp. 425–430). Heidelberg: Springer Verlag.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, vol. 2, 159-165.

Mani, I. (2001). Automatic Summarization. (1st ed.) Amsterdam: John Benjamins Publishing Company.

Mani, I. and Maybury (1999). Advances in automatic text summarization. MIT Press.

Melander, N. M. (1993). Multiple Document Summarization for Written

Argumentative Discourse. Master Thesis. Johns Hopkins University.

Nenkova, A. (2005). Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI2005). 9-13 July. Pittsburgh, USA.

Neto, J. L., Freitas, A. A. and Kaestner, C. A. A. (2002). Automatic Text Summarization using a Machine Learning Approach. In Bittencourt, G. and Ramalho, G. (Eds.). Proceedings of the 16th Brazilian Symposium on Artificial intelligence: Advances in Artificial intelligence. (pp. 386-396). London: Springer-Verlag.

Pooya Khosraviyan Dehkordi, H. K., Farshad Kumarci (2009). Text Summarization Based on Genetic Programming. Journal of Computing and ICT Research (IJCIR). Vol.3 No.1, 57–64.

Saggion, H., Radev, D., Teufel, S., Lam, W. and Strassel, S. (2002). Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. Ann Arbor. 1001, 48109–1092.

Suanmali, L., Salim, N. and Binwahlan, M. S. (2011). Genetic Algorithm Based for Sentence Extraction in Text Summarization. International Journal of Innovative Computing. 1(1).

Tu, C., Chuang, L., Chang, J. and Yang, C. (2006). Feature Selection using PSOSVM. IAENG International Journal of Computer Science. 33(1), 138-143.

Yeh, J. Y., Ke, H. R., Yang, W. P. and Meng, I. H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information processing & management. 41(1), 75–95.

Zechner, K. (2002). Automatic Summarization of Open-domain Multiparty Dialogues in Diverse Genres. Computational Linguistics.

APPENDIX A

COLLECTION THE DUC2002 DATA SET

Table A.1: The 100 documents used in all carried out methods

No.	Folder	Set of Documents
1	D075b	AP880428-0041; AP880818-0088; AP880829-0222; AP881115-0113; AP890115-0014;
		AP900322-0112; AP900705-0149; AP901003-0006; WSJ880603-0129; WSJ910418-0105
2	D077b	AP891017-0195; AP891017-0199; AP891017-0204; AP891018-0084; AP891019-0037;
		LA101889-0066; LA101889-0108; LA102089-0172; LA102089-0177; LA102389-0075
3	D078b	AP880217-0100; AP880325-0239; AP880328-0206; AP890323-0218; AP890324-0014;
		AP890330-0123; AP891110-0043; AP900220-0065; LA033089-0190; LA033189-0114
4	D082a	AP880512-0096; AP880512-0157; AP881109-0161; AP881110-0227; AP890320-0158;
		AP891216-0037; AP891217-0053; LA012189-0060; LA051589-0055; LA121589-0192
5	D087d	AP880228-0013; AP880228-0097; AP880929-0042; AP881002-0048; AP881003-0066;
		AP900328-0128; FT923-8765; LA040790-0121; LA082889-0067; WSJ881004-0111
6	D089d	AP891115-0199; AP891116-0035; AP891116-0115; AP891116-0133; AP891116-0184;
		AP891116-0191; AP891116-0198; AP891117-0002; AP891118-0136; LA111689-0160
7	D090d	AP880625-0142; AP890519-0060; AP890519-0117; AP890710-0170; AP900408-0059;
		AP900829-0044; LA052089-0075; LA101390-0087; LA120189-0122; LA120389-0170
8	D092c	AP900621-0186; AP900622-0025; AP900623-0022; AP900624-0011; AP900625-0036;
		AP900626-0010; LA062290-0134; LA062290-0169; LA062390-0068; LA062590-0096
9	D095c	AP890117-0004; AP890117-0160; AP890118-0013; AP890118-0051; AP890118-0094;
		AP890119-0221; AP890121-0050; AP890121-0123; LA011889-0131; LA012189-0073
10	D096c	AP890122-0087; AP890203-0164; AP891117-0248; AP900128-0063; AP900130-0113;
		LA013090-0161; LA020890-0197; SJMN91-06025182; SJMN91-06025282; WSJ870122-0100

APPENDIX B

ORIGINAL DOCUMENT

Table B.1: Example Document from DUC2002

THE WORLD SERIES; OAKLAND ATHLETICS VS. SAN FRANCISCO GIANTS; EXPERIENCE AT CANDLESTICK IS ONE REPORTER WILL NOT SOON FORGET.
There are events one never forgets, anyone who was alive in 1963 remembers where he was when President Kennedy was assassinated. And anyone who was at Candlestick Park Tuesday night for Game 3 of the World Series will never forget. I was in a trailer just outside the stadium, about to watch the telecast when the earthquake hit. It was shortly after the network went on the air. My immediate reaction was that a jet was flying very low overhead, but soon I knew what was happening. To me, the earthquake was not as bad as the Whittier quake in October, 1987, mainly because I was awake this time. I opened the door to the trailer and looked out. The stadium was shaking and the special concrete joint -- one of several strategically spaced around the top of the stadium to prevent earthquake damage -- was doing was it was designed to do. It was opening and closing as if the place was made of cardboard instead of concrete. A few people were running out of the stadium, but there did not seem to be great alarm, because the quake did not last long. I walked into the stadium to interview fans. And again, I sensed little panic. Most people stayed in their seats, waiting to hear whether the game would be postponed. And later, when fans started leaving en masse, the reaction of those I talked to varied. Those who were sitting in the upper deck seemed considerably more shaken than those in the lower levels. LARRY STEWAR.

APPENDIX C

Human1 Summary

Table C.1: Example of human 1 summary

Experience of a reporter at Candlestick Park for the World Series the night of the 1989 San Francisco quake: I was sitting in a trailer outside the stadium when the earthquake hit. I looked out and the stadium was shaking. A special concrete joint on top of the stadium was opening and closing like cardboard. A few people were running out. I walked into the stadium to interview fans and sensed little panic. Most people stayed, waiting to hear whether the game would be postponed. When leaving, those in the upper deck seemed more shaken that those in lower levels.

APPENDIX D

Human2 Summary

Table D.1: Example of human 2 summary

A reporter at Candlestick Park Tuesday night when the earthquake hit first thought a low-flying jet had passed overhead. The stadium shook and the special concrete joints strategically spaced around the top of the stadium to prevent earthquake damage did what they were designed to, opening and closing as if the stadium was cardboard instead of concrete. There was no great alarm or panic because the quake didn't last long. Those on upper levels seemed more shaken than those on lower. A few people ran out of the stadium, but most waited in their seats to hear if Game Three of the World Series would be postponed.

APPENDIX E

Sentence Segmentation

Table E.1: Example of Sentence Segmentation

T : THE WORLD SERIES; OAKLAND ATHLETICS VS. SAN FRANCISCO GIANTS; EXPERIENCE AT CANDLESTICK IS ONE REPORTER WILL NOT SOON FORGET.

S1 : There are events one never forgets, anyone who was alive in 1963 remembers where he was when President Kennedy was assassinated.

S2 : And anyone who was at Candlestick Park Tuesday night for Game 3 of the World Series will never forget.

S3 : I was in a trailer just outside the stadium, about to watch the telecast when the earthquake hit.

S4 : It was shortly after the network went on the air.

S5 : My immediate reaction was that a jet was flying very low overhead, but soon I knew what was happening.

S6 : To me, the earthquake was not as bad as the Whittier quake in October, 1987, mainly because I was awake this time.

S7 : I opened the door to the trailer and looked out.

S8 : The stadium was shaking and the special concrete joint -- one of several strategically spaced around the top of the stadium to prevent earthquake damage -- was doing was it was designed to do.

S9: It was opening and closing as if the place was made of cardboard instead of concrete.

S10 : A few people were running out of the stadium, but there did not seem to be great alarm, because the quake did not last long.

S11 : I walked into the stadium to interview fans.

S12 : And again, I sensed little panic.

S13 : Most people stayed in their seats, waiting to hear whether the game would be postponed.

S14 : And later, when fans started leaving en masse, the reaction of those I talked to varied.

APPENDIX F

List of Stop Words

Table F.1: Sample of List of Stop Words

a	again	although	anyone	around	against	always	anything
because	before	below	between	by	beforehand	beside	beyond
came	causes	com	considering	couldn't	can	certain	come
do	despite	different	doesn't	done	doing	down	definitely
each	else	et	everybody	exactly	elsewhere	etc	everyone
first	follows	formerly	from	for	forth	further	far
getting	go	gone	greetings	get	given	goes	got
have	hence	hereupon	his	haven't	her	hers	hither
if	indeed	instead	it'd	ignored	indicate	into	it'll
mainly	me	might	mostly	myself	mean	more	much
name	need	next	needs	nine	nor	nowhere	Namely
ones	otherwise	outside	ok	of	old	onto	our