بسم الله الرحمن الرحيم

Text Summarization by using Genetic Algorithm Method

تلخيص النصوص باستخدام طريقة الخوارزمية الجينية

A Thesis Presented to the

Sudan University of Science & Technology

In Partial Fulfillment

Of the Requirements for the Degree Master

In Computer of science

By

Asem Abdullah Mohammed

Supervisor

Dr. AlbaraaAbuobieda

April 2015

﴿ نرفع درجات من نشاء و فوق كل ذي علم عليم ﴾

ACKNOWLEDGEMENT

I would like to express my gratitude and appreciation to all those who gave me the possibility to complete this search. A special thanks to supervisor, Dr. Albaraa Abuobieda, whose help, stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this research.

My completion of this project could not have been accomplished without the support of my friends, Omer Fisal, Abd-Elrhman Yousif.

All thanks and appreciation to my parents, Mr. Abdullah Mohammed and Mrs. Eklaas Hamed and my brother and my sister.

ABSTRACT

Automatic text summarization is a process of rewriting text into a shorter compressed version from the original text. Extraction 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. The selection of the informative sentence is a challenge for extraction based automatic text summarization researchers. This research applied an extraction based automatic single document text summarization method help differentiate using the genetic algorithm (GA) to find out the best feature weight score to difference between important and non-important features. The Recall-Oriented Understanding for Gusting Evaluation (ROUGE) toolkit was used for measuring the performance. DUC 2002 data sets provided by the Document Understanding Conference 2002 were used in the evaluation process. The summary that generated by GA algorithm were compared with other evolutionary algorithm (PSO,ACO) and used DE algorithm as benchmark. Experimental results showed that the summaries produced by the DE algorithm are better than other algorithms. In the meantime, recently propped algorithms such as (ACO) could out performance GA.

المستخلص

التلخيص الآلي للنصوص هو عملية إعادة كتابة النص في إصدار مضغوط أقل من النص الأصلي.ترتكزطريقة الاستخلاص على اختيارأجزاءمعينة من النص حيث تعتبرالجمل و /أو العبارات ذات الدرجات العالية هي الجمل البارزة والتي يتم اختيارها لتشكيل الملخص.اختيارالجمل عن طريق فهم المعلومة يمثل تحديا للباحثين الذين يستندون إليها في التلخيص الآلي للنص. هذا البحث يطبق الاستخلاص استناداً على التلخيص الآلي لمستند نصي واحد باستخدام الخوارزمية الجينية(GA) للعثور على أفضل سمة مستخدمة في التلخيص و التفريق بين السمة المهمة و الغير مهمة. تم استخدام(ROUGE) كأدوات لقياس الأداء. واستخدمت وثائقDUC إصدار عام 2002في عملية التقييم.وتمت مقارنة المستخلص الذي تم توليده باستخدام الخوارزمية الجينية مع خوارزميات تطوير أخرى(PSO، ACO) ويستخدم خوارزميةDEكمقياس رئيسي. وقد أظهرت النتائج التجريبية أن الملخصات التي تنتجها خوارزميةDEهي أفضل من خوارزميات أخرى.وفي ذات الوقت إن خوارزمية ACO المقترحة حديثاً قد فاقت الخوارزمية الجينية (GA)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DE	-	Differential Evolution
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization
ACO	-	Ant colony optimization
IR	-	Information Retrieval
NIST	-	National Institute of Standards and Technology
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

It is very difficult for human beings summarize large documents of text manually. There is a large amount of text files are available in the world in various fields, which offer more information than is needed. Therefore, we have a problem in two ways: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. For this reasons we need a mechanism to summarize useful information from the all documents.

Text summarization defined as an operation of summarizing texts into a summarized form. A summary is a new shorter text generated from one or more text sources**.**

The features are the basic element in the generation process of the text summary. The summary quality is sensitive for those features in terms of how the sentences scored based on the used features. Therefore, the determination of the effectiveness of each feature could lead to mechanism to differentiate between the features having high importance and those having low importance.

In order to tackle complex real-world optimization problems، scientists have been looking into techniques inspired by natural processes such as Darwinian evolution and social group behavior. Accordingly, there has been a remarkable growth in the field of nature-inspired search and optimization algorithms over the past few decades. These algorithms categorized mainly on either evolutionary computation or swarm intelligence. Evolutionary computation makes use of a population selected in a guided random search to achieve the optimization process.

The evaluation made by any Element in population this process call fitness. Element have high fitness become parents for the next population. Swarm intelligence characterized by the collective behavior of decentralized, self-organized systems typically made up of a population of simple agents interacting locally with one another and with their environment. This research-applied one of evolutionary algorithms is Genetic Algorithm. Which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, crossover, selection, and mutation.

1.2 Problem Background

Interest in automatic text summarization, arose as early as the fifties. An important paper of these days is the one in 1958, suggested to weight the sentences of a document as a function of high frequency words, disregarding the very high frequency common words. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights).

The Trainable Document Summarizer in 1995 .The ANES text extraction system in 1995 is a system that performs automatic, domain-independent condensation of news data.

The research emerge in the last years tried to enclose feature weighting to adjust feature scores in summarization problems (Fattah and Ren, 2009,Binwahlan et al., 2009a, Suanmaliet al., 2011b).

Recently (Albaraa Abuobieda and Naomie Salim,2013)used Differential Evolution to introduce text summarization methods designed to solely use a functional approximation (randomized search) approach attached to different learning techniques. In this work we noted the following. The authors have compared evolutionary computing based text Summarization methods with this DE method

1.3 Problem Statement

Building an optimal feature weighting mechanism for high quality summary generating considered a complex task. Several evolutionary computing algorithms (Genetic, Swarm and DE) are propped. As we stated in the problem background, recent works presented by (Albaraa and Salim,2013) used to compare this target algorithm (DE) with other similar existing algorithms (GA and PSO). By taking a lack to the comparison factors established, we found the following. The Number of features, which were used are not in equal and the structure of those features are different. In addition, the data set used at all works is in equivalent in terms of documents number and topics. For these reasons, this research aims to redesign a work of (Suanmali et.al,2011b) whom used Genetic algorithm for improving text summarize performance in order to establish a fair compassion study. Our improved work will be compared against several evolutionary algorithms adjusted similarly to ours.

1.4 Research Objective

The main objective of this research is to compare the performance of genetic algorithm with three optimization–based algorithms (ACO, DE& PSO) with new defined parameters, whichare number of features, structure of features, number of documents and their structure.

1.5 Research Hypothesis

Redesigning the targeted genetic algorithm (GA) with our new adjusted factors may give it a good chance to concept fairly and highly compared with its similar evolutionary algorithms.

1.6 Research Scope

This research designed by using Genetic algorithm, in order to examine it is ability compared to other summarization applications. The following aspects are the scope of this research:

1. The method designed to use is Genetic algorithm.
2. For the evaluation of data, the DUC 2002 selected as the test data. DUC 2002 chosen because it is the last dataset designed for single document summarization.
3. The Recall-Oriented Understanding for Gusting Evaluation (ROUGE) toolkit. It is selected to measure and evaluate the system's generated summaries with reference summaries
4. The genetic algorithm evaluated and compared with similar selected research (DE, PSO and ACO).

1.7 Research Significant

Mainly, the important goal of researches such as one with your hand is to design a summarizer generated high quality summaries. The importance of this researched is to re-evaluate the performance of the GA in terms of a new view of compression.

1.8 Thesis Structures

First, chapter one presents introduction. Chapter two presents Background and Literature Review. Third, Chapter three presents Materials and Methods used in research. Forth, Chapter fourintroduces results and discussion. Fifth Chapter five views Conclusion and recommendation. Finally, Chapter six contain the References have been to take advantage of them.

CHAPTER 2
BACKGROUND AND RELATED WORK

2.1Introduction

This chapter preview Basic concepts conceder the keys of research, In addition to related work inthe same filed. This chapter organized into six sections. Section 2.2 explain text summarization (Definition, Application, concepts related to summarization, approaches), while Section 2.3 explains the Genetic algorithm. Section 2.4 presents an evaluation measure for text summarization. Section 2.5 preview related work based in text summarization.

2.2Text Summarization

Text summarization is a process of rewriting text into a shorter compressed form to represent the original text. Humans finish this task after deep reading and well understanding of the document content, selecting the most important points and Reformulate 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 (trailer), overview of books, and so on.

2.2.1 Summery Definition

They are many definition of summery:

- Sparck Jones (1999) defines summary as "a reductive transformation of source text into summary text through content condensation by selection and/or generalization on what is important in the source."
- Mani (2001) defines summary, as "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."

- Hovy (2005), “a summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s).”
- Fattah and Ren (2008) said, “text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user.”

Generally, a summary should be much shorter than the source text. This characteristic is defined by the compression rate, which measures the ratio of length of summary to length of the original text in word or sentences.

2.2.2 Summarization Applications

The summarization used in several areas, including:

- **Voice mails**. In Koumpis and Renals’s system (2005), the summary words 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.

2.2.3 Text Summarization Basic Concepts

This section introduces the basic concepts used in the field of automatic text summarization.

1- **Type of document :**
   a. Single Document Summarizer: It summarizes one document and produces a single summary.
   b. Multi-document Summarizer: It summarizes more than onedocument and produces a single summary.

2- **Type of language :**
   a. Monolingual Summarizer: It uses just one language for input and output.
   b. Multilingual Summarizer: It has the ability to use many languages with output in the same language as the input.
   c. Cross lingual Summarizer: It has the ability to use many languages with output in different language from the input.

3- **Type of media :**
   The type of medium represented as text summarization or multimedia summarization such as images, speech, and video apart from textual content.

4- **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.

5- **Compression Rate**: It is a ratio of summary length to source length expressing the degree of summarization required.

6- **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.

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

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

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

10- **Critical summary:** It contains the abstractor's opinions towards the quality of the source for evaluation purpose.

11- **A summarizer:** A system creates the summary.

2.2.4 Text Summarization approaches

There are two approaches for text summarization

a. Extraction approach: is 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.

b. 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.

2.2.5Automatic Text Summarization System

Automatic text summarization process consists of three stages represented in Figure 2.1(Mani, Maybury.1999).

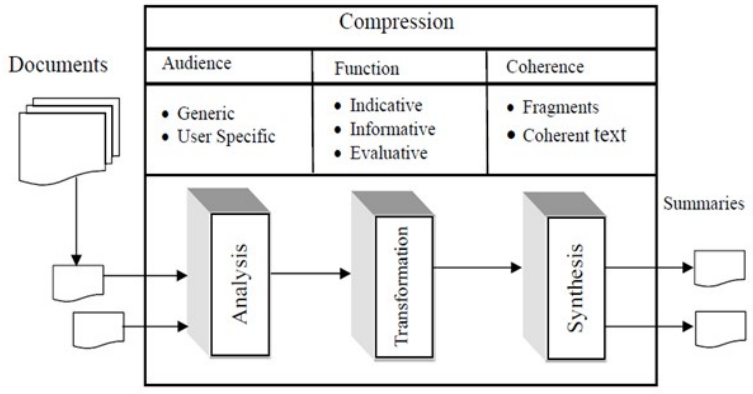

Figure2.1 Automatic text summarization system (Mani and Maybury, 1999)

- 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.

2.3Text Summarization Data set

The data set 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 (Sggion*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.

2-3-1 DUC 2002

DUC 2002 (document understanding conference 2002) data set were used in evaluation process of automatic summarization. DUC 2002 produce by (NISI) 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.

2.4Evaluation Measure

It is important to find out the performance of the various tools and techniques used to evaluate summarized text such as information coverage, grammatical and discursive coherence, readability, etc. Existing evaluation techniques for text summarization can be classified into two categories (Jing et al., 1998; Mani and Maybury, 1999; Afantenos et al., 2005); intrinsic and extrinsic. The first, an intrinsic evaluation method, measures the quality of system-generated summaries using criteria such as the summary readability, the integrity of its sentences, and the accuracy of the summary compared to the source text. Intrinsic evaluations have assesses mainly the coherence and in formativeness of summaries. An extrinsic evaluation judges the generated summaries in terms of a specific task. Thus, such an evaluation can greatly vary from system to system. Extrinsic evaluations have tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc. There is several evaluation measures used in automatic text summarization. This section mentions two measures described as follows.

2.4.1 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 (2.20). 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; Equation (2.21). The F−score measure is used to balance system performance on both “precision” and “recall” measures.

2.4.2 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.

2.4.2.1 ROUGE-N

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

$$\text{ROUGE\_N} = \frac{\sum_{\text{gram n} \in S} \text{Count match (gram n)}}{\sum_{\text{gram n} \in S} \text{Count match (gram n)}}$$

Where S is the reference sentence, n is the length of the n-gram, Count (*gram n*) the number of n-grams co-occurring in the candidate and the reference sentence. Since the denominator is the total number of n-grams occurring in the reference sentence, ROUGE-N measures the n-gram recall.

2.5Related Work

The history of automatic computerized summarization began 50 years ago. The oldest publication describing the implementation of an automatic summarizer is often cited. Luhn's method uses term frequencies to appraise the eligibility of sentences for the summary. In order to determine which sentences need to be included in the summery thisis done by determining the "significant" of those sentences. Two measurements as a significant factor have been proposed: word occurrences and sentence relative position. In addition, preprocessing steps are also applied which include: stop words removal and words stemming. The system then specifies sentences with high scores to be included in the abstract.

Same work at the same year was proposed by. Baxandale proposed a sentence selection measurement by its location on the text. Each sentence that located at the beginning or at the end of the paragraph is considered important candidate and is included in the summary. For evaluation, Baxendale tested his methodology on 200 paragraphs: 85% of the paragraphs hold the sentence topic, while 7% ends with a topic sentence. A positioning feature for sentence extraction and importance has become an important measure in text summarization researches.

The next remarkable step was taken ten years later. Edmondson's work introduced a hypothesis concerning several heuristics (e.g. positional heuristic - a high informational value of sentences at the beginning of an article).Edmundson used two features to score sentences, and added two other features which are pragmatic words: cue words, title and heading words. Cue words such as "significant", "key idea", and "hardly". Baxendale compared his work against manual extracts; a score of 44% was the result of his experiment.

The following years brought further results, but the renaissance in this field and remarkable progress came in the 1990's. It was the time of broader use of artificial intelligence methods in this area and the combining of various methods in hybrid systems. The new millennium, due to WWW expansion, shifted the interest of researchers to the summarization of groups of documents, multimedia documents and the application of a new algebraic method for data reduction.

Josef Steinberger and Karel Jezek they built SWEeT system (Summarizer of WEb Topics). A user enters a query in the system. That query should describe the topic. The system passes the query to a search engine. It answers with a set of relevant documents sorted by relevance to the query. Top n documents, where n is a parameter of the system, are then passed to our summarizer, the core of the system. The created summary is returned to the user, together with references to the searched documents that can help him to get more details about the topic. They can easily change the search engine or the summarizer or any of its modules. Summarization modules, e.g. a sentence compression module, can be easily plugged, unplugged, or changed. And thus the system output will improve with improvements in the modules.

The new research arise in the automatic text summarization explain in the following sub sections

2.5.1 DE-Based Text Summarization

Differential Evolution algorithm is one of an evolutionary algorithm.Storn and Price (1997) originally presented DE. (Alguliev and Aliguliyev, 2009) presented a DE-Based text summarization for extractive-Based in multi-Document summarization. (Algulie*vet al.*, 2011) proposed a self-adaptive optimization based method for multi-Document summarization problems. (Algulie*vet al.*, 2012) published a multi-Document Summarization method. (AlbaraaAbuobieda*et al.*, 2013)

proposed DE algorithm as method for extractive features weights from single-document summarization. In his research he was use five text feature in apply text summarization. In the evaluation stage he use ROUGE_N & ROUGE_L to evaluate the result of algorithm and compare it with benchmark and another algorithms to determine the quality of summary and efficiency of algorithm.

2.5.2 PSO Text Summarization

Practical swarm optimization algorithm is one of a swarm intelligence algorithm.(Algulievet al, 2011a) proposed a multi-document summarization using the PSO algorithm. The DUC 2001 and DUC 2002 were used to evaluate the performance of the method and compared against other proposed cluster-based text summarization methods.(Binwahlan et al, 2009a) presented a PSO method to investigate the effect of feature structure on the feature selection process in the domain of textsummarization. The dataset usedfor training the system comprised one-hundred articles from DUC-2002. To calculate the features weights, he divided the dataset into training and testing phases. He continued optimizing the summarization problem using the PSO combined with the Maximal Margin Importance (MMI) technique.

2.5.3 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 ants in searching food. 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 the rest of the members of the colony should follow paths between the source of food and their 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. 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). The previous work ACO was never used as method for extractive features weights.

2.6Research Group

This group consists of four members; Dr. AlbaraaAbuobieda is a group leader. (Asem Abdulla, AbdurrahmanYousif and Omer Fisal) are members of the group. In previous work has been the comparison between GA (Suanmali*et al*., 2011), (AlbaraaAbuobieda*et al*., 2013) and (Binwahlan et al., 2009), this comparison was unfair due to un-unified parameters we mentioned at our research objectivity (section 1.4). One of our group members (AbdelrahmanYousif) has re-impalement PSO algorithm based text summarization. The second member in the group (Omer Fisal) is used binary ACO algorithm based text summarization. Whereas this research aims to re-impalement genetic algorithm based text summarization. These works will becompared with DE algorithm (AlbaraaAbuobieda*et al*., 2013). This group is used the same number and features and data set which were described at latest work.

2.7Genetic Algorithm

Genetic algorithm is an adaptive algorithm for finding the global optimum solution for an optimization problem. GA often used as function optimizers. The usual applications of GA are the solution of optimization problems, where reliable and efficient results have been presented. This study use algorithm in automatic text summarization to gain better result in this field.

The concept of genetic algorithms introduced by Holland (1975). His purpose was to construct computers to do what nature do. He was concerned

with algorithms that use strings of binary digits representation of individual solutions, simple problem-independent crossover and mutation operators, and a proportional selection rule. Each artificial "chromosome" consists of a number of "genes" and each gene is represented by 0s or 1s as shown as Figure 2.1

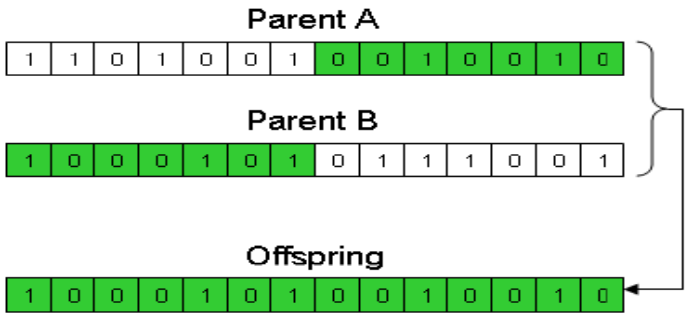

Figure 2.1: Gene crossover operator

The simple methodology in the genetic algorithm is the GA transforms a population of individual objects or chromosomes into new generation of population related with fitness value using the principle of reproduction operators such as crossover and mutation. Figure 2.2 shows the main steps in producing the GA given by Koza (1995).

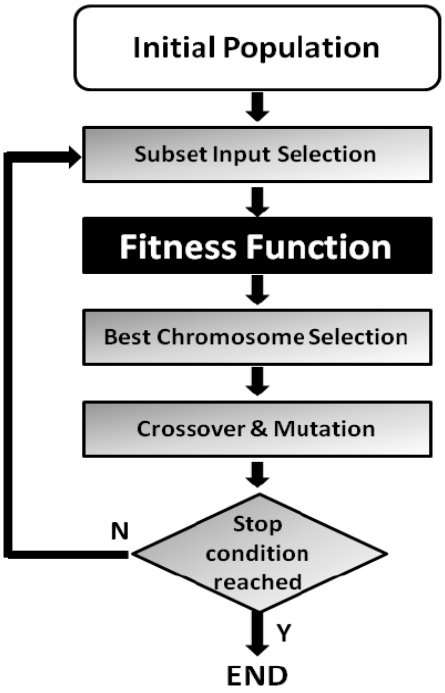

Figure 2.2: The main steps on producing the GA (Koza, 1995).

Traditionally, binary strings of 0s and 1s are used to represent an individual of the population, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. The reproduction operators select chromosomes from the population to be parents for a new chromosome. In each generation, the fitness of every individual in the population is evaluated. Selection of a chromosome for parenthood can range from a totally random process to one that is biased by the chromosome's fitness. The fitness evaluation returns the value to the fitness of an individual in GA. This evaluation function judges quality of chromosomes or the fitness to solve a problem. The fitness evaluation function acts as an interface between the GA and the optimization problem. First, decode the chromosome, and then use the fitness function to evaluate. The fitness function returns a value indicating the fitness of chromosomes to solve the problem. The results of the fitness evaluation function determine the probability that a possible solution is selected to produce new solutions in the next generation.

In the reproduction operator, two chromosomes selected from the population based on its fitness. Then, copy the individual into the next generation of the population without change. Reproduction operator established into the population in two main ways: crossover and mutation. Crossover is an interchange of parts of two individuals. The crossover operator creates two offspring chromosomes, which contains some genetic material of its parents. The concept of the crossover operator is applied to a new chromosome with hope that when it takes the best characteristics from each of the parents, it can produce a better offspring than both parents. A variety of types of crossover operator initiated in the literature (Sivanandam and Deepa, 2008) and used in

binary genes representation are single point, two point, uniform, and arithmetic.

Single Point:

The single point of crossover chosen by randomly function then interchanges bit strings between two parents from begin until the random point to produce two new offspring.

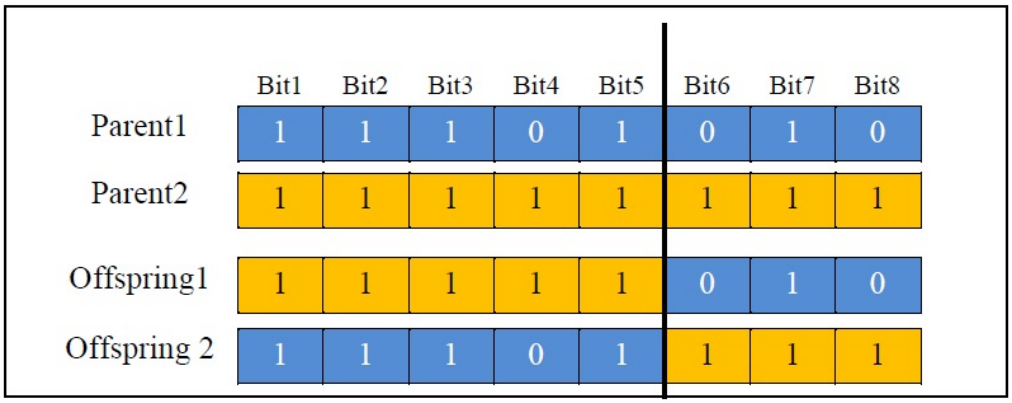

Figure 2.3: Generated next generation using single point crossover at 5th

Two Points:

The two point of crossover chosen two randomly points then the bit string is interchanged between two parents from the first random point until the second random point to produce two new offspring. However, an advantage of having more crossover points is that the problem space may be searched more thoroughly.

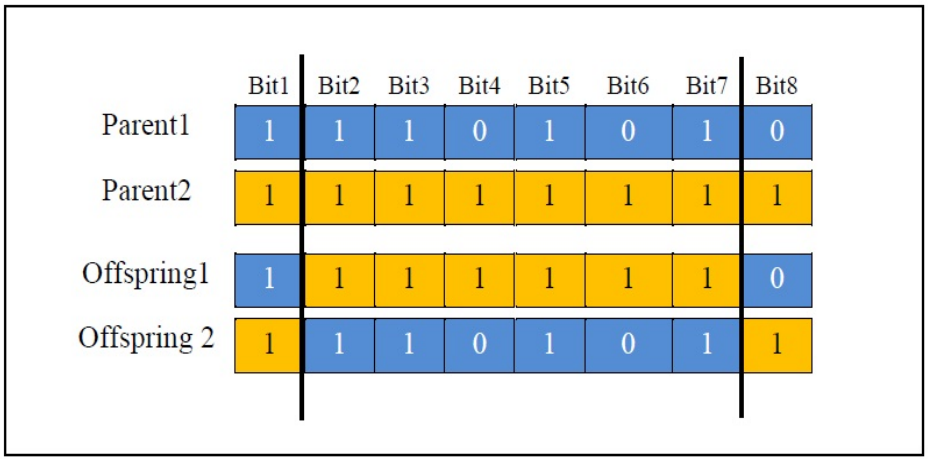

Figure 2.4: Generated next generation using two points crossover at 2ndand 7th

Mutation

Mutation operator is viewed as a background operator to maintain different genetics in the population. The basic mutation operator generated is the random position of one of bits in a bit string then the bit corresponding to that position is changed. Finally, copy the generated individual into the new generation of the population. There are many different types of mutation operator used in binary representation such as flip bit and interchange.

Flip Bit: The flip bit mutation operator crates the new offspring chromosome based on randomly generated mutation chromosome by changing 0 to 1 and 1 to 0. Figure 2.7 explains flip bit mutation operator. Offspring chromosomes are produced by flipping a bit (0 to 1 and 1 to 0), if a mutation chromosome is 1then in parent chromosome, the corresponding bit is flipped.

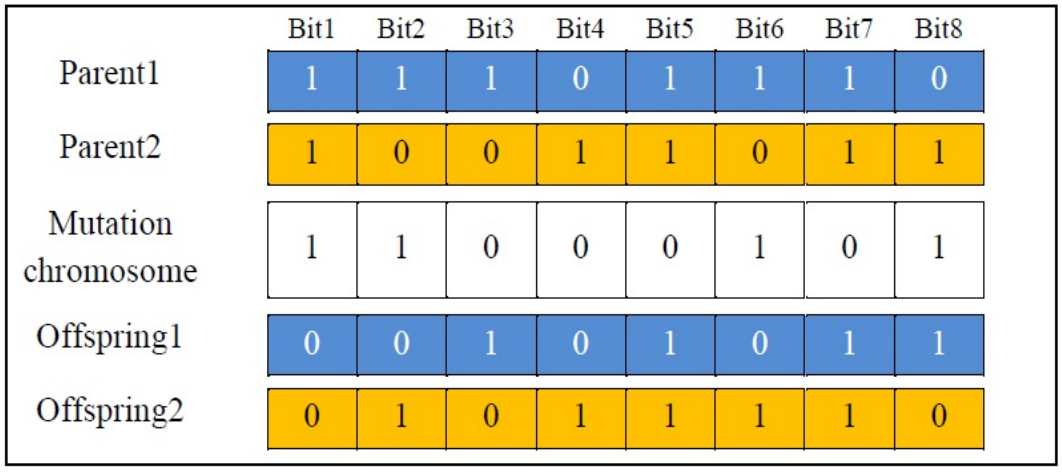

Figure 2.7: Generated next generation using flip bit mutation

2.8Genetic Algorithm Based Text Summarization

Besides statistical approaches, artificial intelligence models present an attractive paradigm to improve the quality of text summarization systems, and the GA represents one of them. How to improve feature selection using GA? This study considers answering this question. GA is

frequently observed as an optimization function, while the range of the problem to which GA has been applied is quite broad. For this reason, the implementation of genetic algorithms in IR has increased recently such as automatic document indexing (Song and Park, 2009), documents categorization (Uguz, 2011), query learning (Bueno, et al., 2007), text summarization using GA based attribute selection. Fist, multi-objective genetic algorithm (MOGA) is used select attribute subsets for classification. Second, the classification algorithm used Naïve Bayes and C4.5. The method extracts the individual sentence which is associated with a vector of features and one of the following two classes: the sentence is selected to the summary (Summary) or the sentence is not selected to the summary (Not-Summary).

2.9Summary

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. It discussed two techniques of text summarization, Single-Document Summarization and multi-document summarization and reviewed the related work to this research. This chapter reviewed evaluation measures that used in the summarization. The next point chapter explain the Genetic algorithm as the general and the operations followed in implementation it. Finally, overview about Genetic algorithm based text summarization.

CHAPTER 3
RESEARCH METHODOLOGY

3.1Introduction

This chapter presents the methodology used in this research. It describes the implementation of the chosen methods in achieving the goal and objectives of the research. One of the objectives of this study is to rebuild automatic text summarizer using genetic algorithm and compare it with another summarizer designed by another algorithms.

This chapter discusses the steps taken to carry out this research. There are five sections in this chapter where Section 3.1 is for the introduction. Section3.2 presents the phases followed in implementation. Overview of the operational framework shown in Figure 3.1. This framework separated into four phases. Section 3.3 is about the evaluation and reporting of the result. Section 3.4 explain genetic Algorithm Based Text Summarization. Finally, section 3.5 summarized all that were presented in this chapter.

3.2Phases followed in implementation

There are three phases fowled in implementation. Phase 1: Initial Study and Data Preparation, Phase 2: Feature-Based General Statistical Method (GSM), Phase 3: Feature-based Genetic Algorithm (GA) Method.

3.2.1 Phase 1: Initial Study and Data Preparation

This phase consists of five main elements: problem formulation, literature review, identifying existing technique, obtaining data set, and data preprocessing. Problem formulation involves the process of identifying the issues that exist in the automatic text summarization

system. It is done by doing literature review to analyze the existing text summarization technique. This phase has the following five activities:

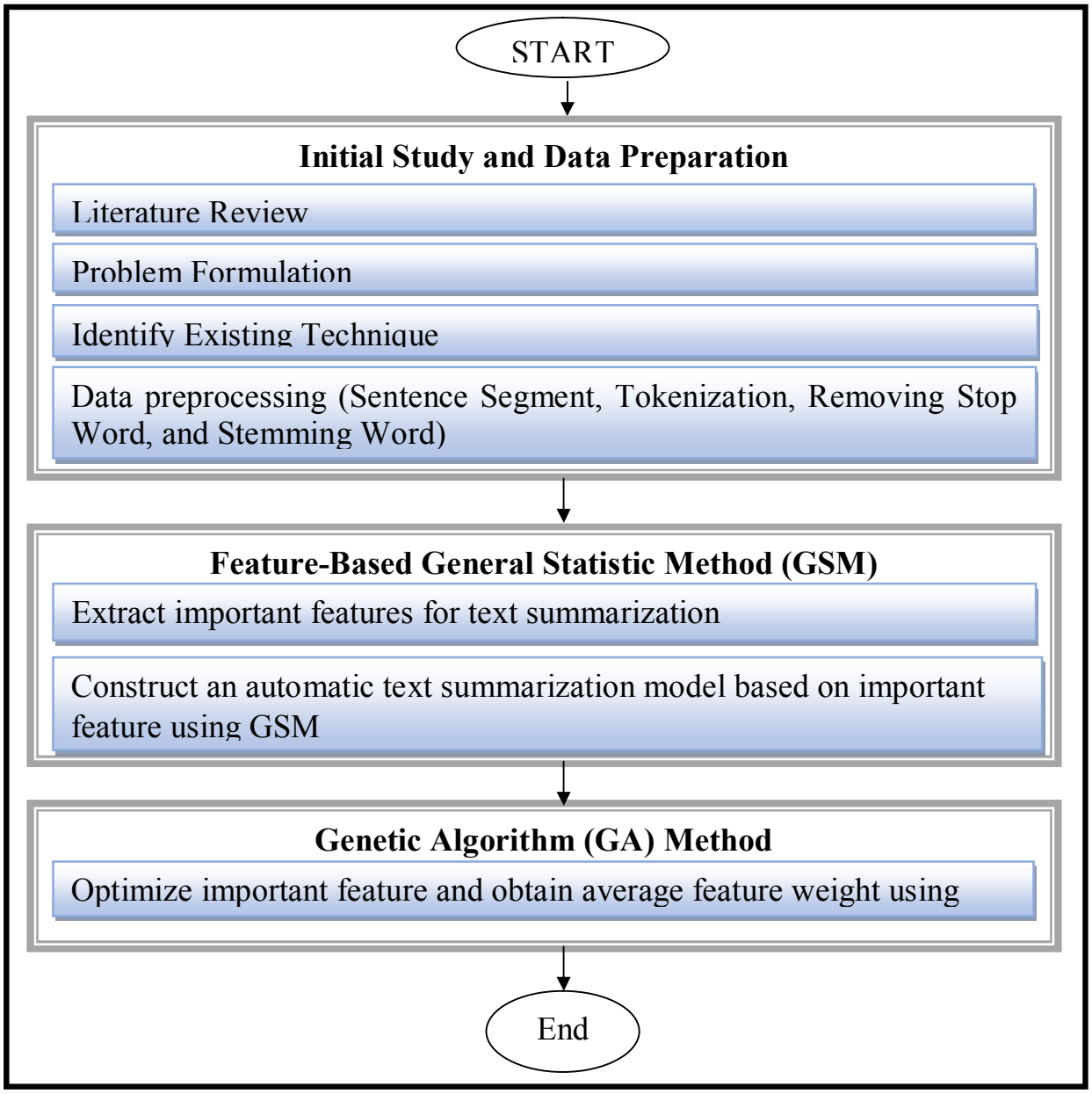

Figure (3.1)

3.2.1.1Literature Review:

This phase reviews and studies the research works related to automatic text summarization approaches and genetic algorithm. The reviews of previous works have been done with a related research topic. Literature review was continuously performed until the research is finished. It is important to make sure the novelty of the research and to identify

useful information related to the research. Throughout the literature review, related information will be recorded. This phase is already achieved and presented as shown in chapter 2.

3.2.1.2Problem formulation:

An overview of the problem is first formulated from automatic text summarization and many additional related areas. Identify Existing Technique: Various existing text summarization have been discovered during the literature review h same gaps and problem statement have been formulated.

3.2.1.3Analyze data sets:

The data set of initial data analysis was obtained from The Document Understanding Conference (DUC), which became a standard data set for testing and evaluation any summarization method. The DUC data set is collected from famous newswires used by most researchers in automatic text summarization. The DUC data collection is prepared and peer-reviewed by language professional people. Most researchers in text summarization used DUC data set. The data collection in DUC is famous newswires. The evaluation data of 100 data sets, which were used in (DUC, 2002), was created by the National Institute of Standards and Technology of the U.S. (NIST). Each set contained documents, single-document abstracts, and multi-document abstracts/extracts, with sets defined by different types of criteria: single natural disaster event, single event in any domain, multiple distinct events of a single type, and biographical information. Each document in DUC2002 collection is supplied with a set of human generation summaries provided by two different experts. Each document in DUC2002 collection is provided with a set of human-generation summaries provided by two different experts. DUC2002 data sets are

used in this study because it is used in the similar automatic text summarizer

3.2.1.4Data Preprocessing:

Preprocessing is an important process in text summarization since the quality of the generated summary depends on the efficiency of the text representation. In this step, input documents are of plain text format as shown in Table 3.1 in APPENDIX A.

There are five main activities performed in this stage: sentence segmentation, tokenization, stop word removal, lower case letter and word stemming as described as follows. The results after data preprocessing is shown in Table 3.3 in APPENDIX A.

a. Sentence segmentation

Sentence segmentation is an important task in text processing approaches such as information extraction, plagiarism detection, machine translation, syntactic parsing and text summarization. Sentence segmentation is performed by boundary detection and separating source text into sentences. In most cases, a simple matter; a period (.), an exclamation mark (!) or a question mark (?) usually signals in the sentence boundary (Mikheev, 2000). An example is show in Table 3.2 in APPENDIX A.

b. Tokenization

The stream of characters in a natural language text must be broken up into distinct meaningful units or tokens (Ronald, 2005). Tokenization is separating the input document into individual words. In this task, we use method impeded in java program language to separate text into words and punctuation tokens. For example show Table 3.3 in APPENDIX A

c. Stop word removal

Next step in preprocessing is stop words removal, where words which rarely contribute to useful information in terms of document relevance and appear frequently in document but provide less meaning in identifying the important content of the document are removed. There are various approaches used for determination of such stop words list. Currently, several English stop words list is commonly used to assist information retrieval. Those words including articles, prepositions, conjunctions and some other high-frequency words, such as *'a', 'an', 'the', 'in', 'and', 'I', etc.*.Despite its redundancy and having no influence to the meaning, these words contribute a significant percentage on the overall documents. Elimination of such words can gradually increase the effectiveness and efficiency of information retrieval process. The size of the documents can be minimized without affecting its contents and less time or memory is consumed during the retrieval process. APPENDIX B shows some of wordlist used in this research.

d. Word Stemming

The last step for preprocessing is word stemming. Word stemming is the process of reducing inflected or derived words to their stems, base or root form to represent the same concept. This research performed words stemming by removing suffixes proposed by Porter's stemming algorithm (Porter, 1980). The Porter's algorithm is probably the stemmer most widely used in IR research. The algorithm applies a set of rules to iteratively remove suffixes from a word until none of the rules apply. The algorithm implements a non-dictionary-based stemmer, which works well enough for information retrieval. The removal of suffixes by automatic mean

is an operation which is especially useful in the field of information retrieval. For example, a stemming algorithm for English should stem the words ‘compute’, ‘computed’, ‘computer’, ‘computable’, and ‘computation’. This may be done by removal of the various suffixes –e, –ed, –er, –able, –ation to leave the single term ‘comput’. Example of output from this process is show in Table 3.4 in APPENDIX

3.2.2Phase 2: Feature-Based General Statistical Method (GSM)

The main task of this phase is to identify the difference important features for text summarization, to construct the automatic text summarization model based on the most important features and to report the results. The details are shows as follows.

Extract important features for text summarization*:* The first step in summarization by extraction is the identification of important features such assentence length (Lin and Hovy, 2003), sentence position (Fattah and Ren, 2008), title feature (Salton and Buckley, 1997), thematic word or key word (Edmundson, 1969; Kupiec et al, 1995), and number of numerical data (Lin, 1999). In order to use a statistical method, it is necessary to represent the sentences as vectors of features. These features are attributes that attempt to represent the data used for the task. This method used five features for each sentence. Each document is converted into a list of sentences. Each sentence is represented as a vector [S_F1, S_F2,S_F3S_F4,S_F5]. Each feature is given a value between ‘0’ and ‘1’as described as the following.

3.2.2.1 S_F1: Title feature

The word in a sentence that also occurs in title gives a high score. This is determined by counting the number of matches between the content words in a sentence and the words in the title. We calculate

the score for this feature, which is the ratio of the number of words in the sentence that occur in the document title over the number of words in document title.

$$S\_F1(S) = \frac{No.Title\ word\ in\ S}{N0.Word\ in\ Title}$$

3.2.2.2 S_F2: Sentence Length

This feature is useful to filter out short sentences such as datelines and author names commonly found in news articles. The short sentences are not expected to belong to the summary. We use normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

$$S_{F2(S)} = \frac{\text{No.Word in Sentence}}{\text{N0 Word in the longest sentence}} \quad (3.2)$$

3.2.2.3 S_F3: Sentence Position (SP)

The first sentence in the paragraph is considered an important sentence and a good candidate for inclusion in the summary. Equation (3.3) is used to compute the SP feature, where Si refers to the i^{th}sentence in the document wanted to extract its position score, and CountTotal() is a function that retrieves the total number of the sentences in the input parameter document d and Current Position () is a function that retrieves the current order of sentence Si in document d.

$$SP(Si) = \frac{\text{CountTotal(d)-CurrentPostion(Si)}}{\text{CountToyal(d)}} \quad (3.3)$$

3.2.2.4 S_F4:Numerical Data (ND)

A sentence that contains numerical data often have important information such as a date of event, money transaction, damage

percentage, etc. Equation(3.4) shows how to compute this feature where Count ND() is a function computes the Numerical Data (ND) found in the ith sentence S in the document, and Count Length () is a function used to compute the sentence length of Si.

$$\text{ND(Si)}=\frac{\text{CountND(Si)}}{\text{CountLength Si(word)}} \tag{3.4}$$

3.2.2.5 S_F5:Thematic Words (TW)

Thematic words are a list of top n selected terms with the highest frequencies. To compute the thematic words, first the frequencies of all terms in the document are computed. Then, a threshold is specified in order to assign terms that should be selected as thematic words. In this case, the top ten frequent-terms max(TW) would be assigned as a threshold. To compute the ratio of TW found in thei^{th} sentence S in the document Equation (3.5) is used where Count Thematic () is a function used to compute the number of the thematic words found in Sentence Si.

$$\text{TW (Si)} =\frac{\text{Count Thematic (Si)}}{\text{Max(N0 of thematic word)}} \tag{3.5}$$

Text summarization base on general statistic method was exploited to integrate all the five feature scores as the sentence weight. A set of highest score sentences are extracted as the document summary based on the compression rate. After features were extracted by the system, the sentence scores can be calculated. First, a weighted score function for a sentence *S* is exploited to integrate all the five features, as calculated using equation (3.6).

$$\text{Score(S)} =\sum_{k=1}^{5} Score\ s\_fk(S) \tag{3.6}$$

Where Score (S) is the score of the sentence S and Score S_Fk (S) is the score of the feature k.

3.2.3 Phase 3: Genetic Algorithm (GA) Method

Optimize important feature and obtain average feature weight using GA: In this phase, we use GA for training and testing to optimize weight of features. The approach can easily be applied to feature weighting. 100 documents from DUC2002 data collection are used in this experiment; the 100 documents data set used as training and testing. We divided the data into 70 documents for training and 30 documents for testing**.** In the training section, the fitness measure used by the GA for feature selection is the average recall measure generated by ROUGE-1 (Lin, 2004) in equation (3.7).

$$\frac{\sum s \in ref summaries \sum_{gram_n \in S} count\ match^{gram_n}}{\sum s \in ref summaries \sum_{gram_n \in S} count\ gram_n} \quad (3.7)$$

Where n is the length of the n-gram and $Count_{Match}$ is the most possible number of n-grams shared between a systems generated summary and a set of reference summaries.

3.3 Evaluation and Reporting Results

The evaluation of this research uses a set of metrics called Recall-Oriented Understudy for Gusting Evaluation (ROUGE), evaluation toolkit (Lin, 2004) that has become standards of automatic evaluation of summaries in the DUC text summarization which compares a system generated summaries against human generated to measure the quality as reported in this study. For comparison, it uses n-gram statistics. This study chooses ROUGE-1 as the measurement for the experimental results using the average precision, recall and f-measure The reason for using this evaluation toolkit is based on the report by a similar study (Lin, 2004) that showed those measures work well for single document summarization.

- **Precision:**

  It is the number of sentences intersected between the system summary and human summary divided by the number of sentences in the system summary

$$\text{Precision} = \frac{(\text{S\_Ssystem Summaries} \cap \text{S\_Human Summaries})}{\text{S\_Ssystem Summaries}} \quad (3.8)$$

- **Recall:**

  It is the number of sentences intersected between the system summary and human summary divided by the number of sentences in the model summary

$$\text{Recall} = \frac{(\text{S\_Ssystem Summaries} \cap \text{S\_Human Summaries})}{\text{S\_Human Summaries}} \quad (3.9)$$

- **The F −score measure:**is used to balance system performance on both “precision” and “recall” measures. In other words, the F-score counts all co-occurring words regardless of their orders.

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.10)$$

3.4 Genetic Algorithm Based Text Summarization

The simple methodology in genetic algorithm follow static steps to solve any problem. Firstly generate random population of individual objects or chromosomes (or the genotype of the genome), each with an associated fitness value, Secondly select best chromosomes to generate new generation of population using the principle of reproduction operators such as crossover and mutation. In the population of each generation, the fitness of every individual is evaluated. The fitness evaluation function in GA returns the value of the fitness of each individual. This evaluation function judges the quality of chromosomes in terms of the fitness to solve the

optimization problem. First, it decodes the chromosome, and then uses a function to evaluate its fitness. The results of the fitness evaluation function determine the probability that a possible solution is selected to produce new solutions in the next generation.

Figure 3.2 and Algorithm 3.1 illustrate the proposed model. Firstly, the document is preprocessed using: sentence segmentation, stop words removing and words stemming. Next, the sentence features are extracted. Then, sets of document are used for training and testing in GA method phase.

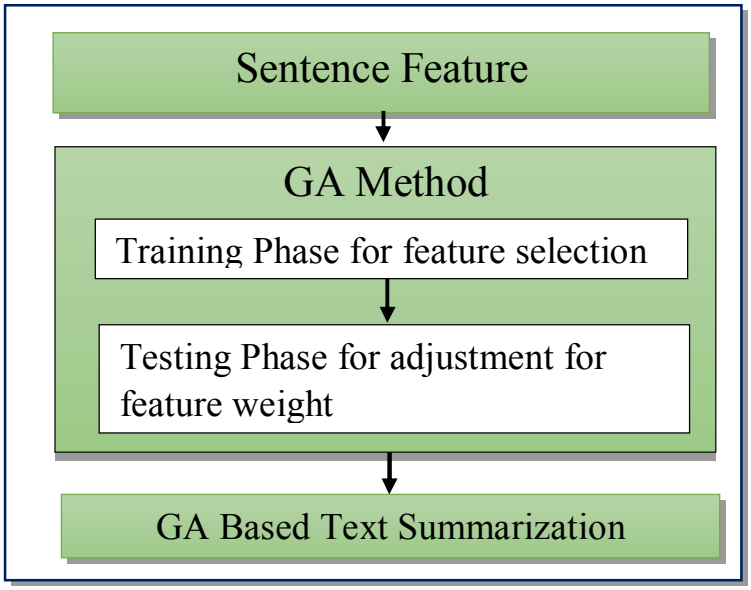

Figure 3.2 Algorithm 3.1

Step	Main process	Process detail
1	Read source document D:	Read the source document D into the system, *D* = {Title, S1, S2, S3,…, Sn}.
2	Do preprocessing:	Extract the individual sentences of the original documents. Then, separate the input document into individual words. Next, remove stop words. The last step for preprocessing is word stemming.
3	Extract feature:	Each sentence is associated with vector representing weights of five features, S =

		{S_F1, S_F2, S_F3, S_F4, S_F5}. These values are derived from the content of the sentence.
4	GA Method 4.1 feature selection:	Select the optimal features based on GA using 70 documents in training phase.
	4.2 feature weight Adjustment :	Apply the best selected features from step 4.1 to 30 documents in testing phase for adjusting feature weight.
5	Calculate sentence score:	The features are calculated to obtain the sentence score based on GA method.
6	Extract sentences:	A set of the highest scoring sentences generated by step 5 are extracted as a document summary based on the compression rate.
7	Construct document summary	After step 6, the summary sentences are arranged in the original order.

3.4.1 Initializing population

Each individual is made up of a sequence of a binary bit (0 and 1). Let N is the size of a chromosome population. The population of 30 chromosomes is randomly generated at the beginning. Random function is used to generate a random chromosome consisted of binary numbers 0s and 1s .

3.4.2 Fitness Evaluation

The fitness evaluation function in GA returns the value of the fitness of each individual. In this experiment, a set of summary documents in previous section is used as input for the fitness function, and then uses a function to evaluate its fitness which obtain the best average recall score generated by ROUGE-1 (Lin, 2004) as fitness function presented in equation (3.7).

3.4.3 Selecting best chromosome for the next generation

The selection or reproduction operator determines which of the individuals will survive and continue in the next generation. The selection operator implemented here is selecting the top two chromosomes of the population to be parents of a new generation that give the highest average recall through the fitness.

3.4.4 Performing crossover and mutation

The genetic operations applied are the crossover and mutation. The 30 new chromosomes are generated among the offspring of the previous selection chromosomes by crossover and mutation operation.

3.4.4.1 Crossover

The main task of the crossover is to create child or new chromosomes from two parent chromosomes by combining the information extracted from the parents. Each generation of this method generates 10 chromosomes using one point crossover. The one point crossover is chosen by random function then bit strings between two parents from beginning until the random point were interchanged.

3.4.4.2 Mutation

A mutation operator implies a possibility that in a chromosome series, an arbitrary bit will be changed from its original. In this work, a child or new chromosome from 10 chromosomes was generated by crossover process.

3.5 Training phase

In each training data set, we generated 500 generations, keep the highest fitness value from each generation, and then compare all highest fitness values. The best fitness value used to represent the features that are suitable for these data sets. After finish 500 generation

sum all highest chromosomes for 70 documents and divided by 70 to gain weight of feature for use in testing phase.

3.6 Testing phase

In testing phase we used 30 documents from DUC2002 data set. The testing phase is similar to training phase begin with implementing the preprocessing process (segmentation, tokenization, remove stop word and stem the word), then extracting features for each sentence. The different begin by modifying the score of each feature based on the features weights that produced in training process. After that calculate the score and complete all other steps in text summarization to get summaries of all 30 documents.

The next step calculate recall, perception and f-score for any document of that, then sum any meager for 30 document and divided by 30 to calculate the percentage value for any meager.

These values used to compare this algorithm with anther algorithms in the same filed of Automatic text summarization

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 generated summary is sensitive to the selected features. Therefore, developing a mechanism to compute feature weight is very important. The weighting approach helps identify the important 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 (AlbaraaAbuobieda 2013), These methods are used different feature to generate summary. This chapter describes the results of implementing the Genetic algorithm in problem text summarization and compares the generated summary after apply features weights with other algorithms such asPSO algorithm, and Differential Evolution algorithm as benchmark. Additionally, we added a work based on ACO algorithm which has designed recently in our new established research group(omer and albaraa,2014). As we stated in chapter 2, this group mainly focus on studying the performance of evolutionary algorithms in problem of text summarization. We also would like to inform that, A work of (omer and albaraa,2014) has not been yet published under any local or international scientific database. 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 evaluation 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) was compared with DE algorithm summary. Table 4.1, 4.2 compare the three proposed methods using ROUGE-1, ROUGE-2. These methods are calculating it the average recall (avg-R), average precision (avg-P) and average F-measure (avg-F) based on H1 reference. 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 generalization of obtained results, the performance of the GA model is 32% approximately similar to human performance (H2) using ROUGE-1and 15% approximately similar to human performance (H2) using ROUGE-2. If we refer back to the previous comparison between DE algorithm (albr'aa and Naomi, 2013) and GA (Suanmali*et al*2011) we found DE algorithm is better than GA with differentnumber of features and documents. Thatmeans, the unification of features and documents had not do affected in the results, so we can search for another reasons (crossover and mutation) may be generate duplicate chromosomes,

number of chromosomes used in testing phase or number of generation). Finally we concluded that GA could performance better than PSO algorithm.

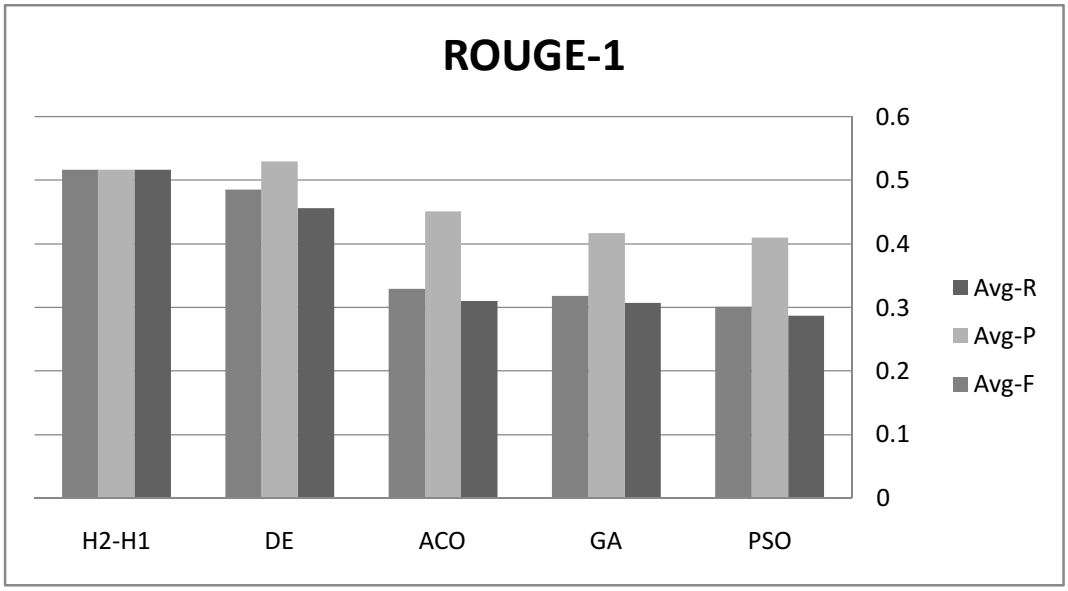

Figure 4.1: methods comparison using ROUGE-1 result

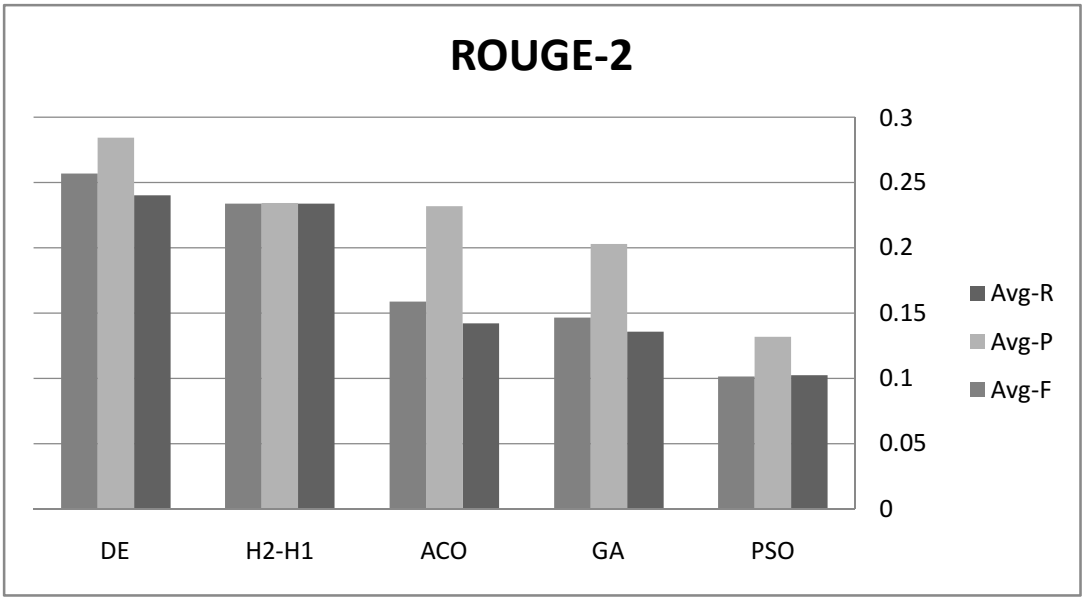

Figure 4.2: comparison using ROUGE-2 result

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 been proven to generate high quality summarization.

5.2 Genetic algorithm Based Text Summarization

In this research, Genetic algorithm used to obtain features weights. It follows three phases. The first phase is Initial study and data preparation that contain (literature review, problem formulation, analyze data sets and data preprocessing). The second phase explains features used for extract summary, five effective statistical features were selected (Title Feature, Sentence Length, Sentence Position, Numerical Data and Thematic Word). The third phase is apply genetic algorithm that divided in two stages (training stage and test stage). It is consider a machine learning to learn features. GA algorithm used to determine better chromosome for each document by using evaluate tool called ROUGE for each summery produced by this chromosome. After that we produce feature weight for all documents testing by summation for all chromosomes selected and divided by number of document testing. In testing stage was used set of data set to testing process by apply the feature weight in extract summary of document testing. The generated summary in this stage compared with other algorithms (DE, PSO and ACO). The summary that generated by DE algorithm is better than another algorithms.

REERENCES

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., 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, 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.

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.

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.

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

APPENDIX A
EXAMPLE DOCUMENT FROM DUC2002

Example Document from DUC2002

Sakharov Receives Human Rights Award.

Andrei D. Sakharov, the father of the Soviet dissident movement, finally received a 1973 human rights award Thursday night during his first trip to New York City. Sakharov received the Human Rights Award from the International League for Human Rights at a reception at the home of Ronald Lauder, the former U.S. ambassador to Austria. The 67-year-old Nobel Peace Prize winner is on his first trip to the West, less than two years after Soviet leader Mikhail Gorbachev freed him from internal exile in the city of Gorky. The human rights activist was banished in 1980 to the closed city for opposing the Soviet occupation of Afghanistan. About 200 guests, including author Elie Wiesel, Brooklyn District Attorney Elizabeth Holtzman, designer Caroline Roehm and ABC television's Barbara Walters, attended the reception. The ceremony was closed to the media. Sakharov, wearing a blue beret and plaid scarf, made no statement as he left the reception and entered a waiting limousine. Sakharov is the honorary president of the league, which has its headquarters in the United States and more than 40 affiliate groups around the world, including the Moscow Human Rights Committee, which Sakharov and his colleagues founded in 1971.Sakharov arrived in the United States on Sunday to visit relatives and receive medical treatment before going to Washington for a White House visit and a board meeting of the International Foundation for the Survival and Development of Humanity, of which he is a director.

APPENDIX B

Sentence Segmentation

Sakharov Receives Human Rights Award.

S1: Andrei D. Sakharov, the father of the Soviet dissident movement, finally received a 1973 human rights award Thursday night during his first trip to New York City.

S2: Sakharov received the Human Rights Award from the International League for Human Rights at a reception at the home of Ronald Lauder, the former U.S. ambassador to Austria.

S3: The 67-year-old Nobel Peace Prize winner is on his first trip to the West, less than two years after Soviet leader Mikhail Gorbachev freed him from internal exile in the city of Gorky.

S4: The human rights activist was banished in 1980 to the closed city for opposing the Soviet occupation of Afghanistan.

S5: About 200 guests, including author Elie Wiesel, Brooklyn District Attorney Elizabeth Holtzman, designer Caroline Roehm and ABC television's Barbara Walters, attended the reception.

S6: The ceremony was closed to the media.

S7: Sakharov, wearing a blue beret and plaid scarf, made no statement as he left the reception and entered a waiting limousine.

S8: Sakharov is the honorary president of the league, which has its headquarters in the United States and more than 40 affiliate groups around the world, including the Moscow Human Rights Committee, which Sakharov and his colleagues founded in 1971.

S9: Sakharov arrived in the United States on Sunday to visit relatives and receive medical treatment before going to Washington for a White House visit and a board meeting of the International Foundation for the Survival and Development of Humanity, of which he is a director.

APPENDIX C

Tokenization, Stop word removal, Lower case letter and word stemming

sakharov receiv human right award.

S1:andrei d sakharov father soviet dissid movement final receiv 1973 human right award thursdai night first trip new york city

S2:sakharov receiv human right award intern leagu human right recept home ronald lauder former u s ambassador austria

S3: the 67-year-old nobelpeac prize winner first trip west, less two year soviet leader mikhailgorbachev freed intern exilcitigorky

S4:the human right activist banish 1980 close citioppossoviet occupafghanistan.

S5:about 200 guests, includ author eliwieselbrooklyn district attorneielizabethholtzman, design carolinroehmabc television' barbarawalters attend reception

S6:the ceremoni close media

S7:sakharov wear blue beret plaid scarf made statement left recept enter wait limousine

S8:sakharovhonoraripresidleague headquart unit state 40 affili group around worldincludmoscow human right committee sakharovcolleagu found 1971

S9:sakharovarriv unit state sundai visit relreceiv medic treatment go washington white hous visit board meet intern foundatsurviv develop humanity

APPENDIX D

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