Text Summarization:
Taking Legal Document Summarization as an Example

By

Adilah Abu Bakar

Project dissertation submitted in partial
fulfillment of the requirements for the
Bachelor of Technology (Hons)
(Information System)

DECEMBER 2004

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

TEXT SUMMARIZATION: TAKING LEGAL DOCUMENT SUMMARIZATION AS AN EXAMPLE

By

Adilah Abu Bakar

A project dissertation submitted to the
Information System Programme
Universiti Teknologi PETRONAS
In partial fulfillment of the requirement for the
BACHELOR OF TECHNOLOGY (Hons)
(INFORMATION SYSTEM)

Approved by,

Ahmad Izuddin

(Mr. Ahmad Izuddin Zainal Abidin)

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK
DECEMBER 2004

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified source or persons.

ADILAH BT ABU BAKAR

ABSTRACT

Legal Document Summarization is an automated text summarization system which is generated by a computer program. This projects aim to generate a relevant summary from a legal tender specifically documents on tender. With the development of this module, it is hoped that this will decrease the time required for handling tender process, eliminating the need on using manual summarizing and providing an easy viewing for user. This program will be developed by incorporating Artificial Intelligence field of Natural Language Processing (NLP) techniques and also finding the most suitable methodology to handle a project development that deals on text summarization processes. Therefore a custom-made methodology are implemented which are based on SDLC methodology and a summarization process. In incorporating NLP technique, based on the existing summarization system technique on word counting and clue phrases for topic identification and word clustering are used for better interpretation of information. Apart from using NLP, other techniques such as theme extraction are also taken into consideration for better generation of the summary based on the relevant requirement for the document. With this, extraction of texts based on the results from word counting and theme extraction can be generated. The technology that are being generated here are for a single document summarization in English language.

ACKNOWLEDGEMENT

In doing this final year project, my efforts alone would not have brought me forward in completing this project. There are many individuals and groups that had spurred me to finish my project. Hence, here I would like to express my heartiest thanks to all the parties involved in any ways in assisting me throughout my project.

First of all I would like to express my gratitude to my final year project supervisor, Mr. Ahmad Izuddin Zainal Abidin in assisting and guiding me throughout this project. His advice had been helpful for me in finding my course in doing this project. I am also grateful for Mrs. Yong Suet Peng for her ideas to get me started this project.

I would also like to give my thanks to Tuan Haji Mohd Rodzi b Mansor from Bahagian Kontrak, JKR Ipoh and Mr. Nizar from Bahagian Kontrak, Universiti Teknologi Petronas in their cooperation in helping me gathered the requirements needed for this project.

Not forgetting my heartiest thanks and appreciation to my friends, family and housemates on their constant support and reassurances, helps and lending hand when assistance is greatly needed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND OF STUDY

Electronic documents are becoming more important nowadays as more businesses and organizations are opting for this kind of media for easier access and transferability. The importance of electronic documents are especially true for organizations or businesses handling large volume of tenders where the bids of formal proposal to buy at a specified price are being given by many companies, a document summarization would be useful in presenting this large volume of data. Therefore by having legal document summarized, we can extract the gist of this document and provide easy viewing.

Text summarization is an automatically generated summary of a document where it summarizes only its essential content. Here the legal document summarizing will be using natural language processing and also information retrieval and information extraction which are a part of artificial intelligence to summarize the legal tender according to its most important information and its relevance. In order to generate summary, important pieces of information from the document need to be identified, irrelevant information omitted and details minimized, and assemble them into a compact coherent report. This is how natural language processing comes in place.

Natural Language Processing is a field in Artificial Intelligence involving anything that processes natural language. It generally refers to the complete linguistic and conceptual processing of a text.

1.2 PROBLEM STATEMENT

1.2.1 Problem Identification

Large Amount of Time

Handling a considerable volume of documents can be overwhelming. Important documents such as tenders will take a *large amount of times* to be carefully read through this large amount of documents. This is especially unmanageable for large organizations.

Human Resources

In generating summaries manually, *more human resources are needed* and it is more time-consuming. Some companies approach this problem of providing abstracts or summaries by hiring additional personnel to manually scan entire documents and writes summaries. This process is incredibly expensive and time consuming. An automatically created summary can be beneficial in order to address this problem

Variation

In addition, by using many personnel for manual summarizing, *variation* among different extracts can be seen. This can be confusing to a person who handles the evaluation of the tender in figuring out the style and interpretation of each document. This is because different humans may interpret different data needed for summarization. This will resulted in a variation of styles and interpretation from the human perception of the document. Therefore in using automatic summarization, a standard can also be set. Thus later presented, it will provide easier understanding and evaluation on tender by the committee later on.

1.2.2 Significance of the Project

By producing automated summarizing for tender, important information for the bids can be viewed first which saved time before the user will proceed to read the whole documents. By generating the legal document summarizer, users can view the substance of the documents easily. This is especially useful for user who is interested in getting an overview of the document before proceeding to open the document or in searching for a certain criteria or points of a document. This enables the user to determine what should be read in more depth, and what just needs a cursory glance.

Time and resources are needed to generate manually summarized documents can be lessened. This means less workload that can be time-consuming are needed for personnel who will be handling this type of task.

Awarding tenders may be done in a faster and more convenient ways by looking first at the main criteria needed of the tenders before reading the whole document. The user will get the gist of the criteria's that they are searching for in a tender before document. This will help them understanding the importance of the tender's criteria.

1.3 OBJECTIVES

In developing this legal document summarization, objectives that are been set below need to be met and had provided for the author as guideline in doing this project.

1.3.1 Objectives of the legal document summarizer:

- To decrease the time needed for personnel to generate the documents manually.
- To ease the process of searches on important information of the legal tender.
- To improve time management and resource management hence resulting in fewer resources needed by using automated legal summarizing.

1.3.2 Objectives of the project:

- Developing a legal text summarization that can generate a summarized content of a document automatically.
- Incorporating AI field of NLP technique in developing a suitable method needed to create the module for the legal document summarizer.
- Search for a suitable methodology for developing a legal document summarizer.

1.4 SCOPE OF STUDY

For this project, a legal document summarizer has been developed as a module for the user of a company to automatically generate the document. Here, the legal document that the project is focusing on is the tender such as the form of tender and its details.

The analysis and research process for this project had been specifically to include an AI field which is NLP for text processing in order to develop this legal document summarizer.

In this prototype, the module will only admit authorized users into the page since this task involves highly confidential information. These users will then easily generate the summaries after uploading the documents required.

Detailed framework of this project will be discussed further on.

1.5 RELEVANCE OF THE PROJECT

The rationale of this project is to find the best method incorporating natural language processing technique in order to generate a summary with the necessary relevant contents from a legal document. A research and analysis on the current existing system are done to get the main idea on how to go on to this project. Therefore by doing this, it would be helpful to develop a summarizing method that will emphasize on a method to handle on a specific type of documents such as legal document.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

The volume of electronic text and documents is unmanageable without tools for retrieving and filtering. World Wide Web and search engines do a good job of locating possible relevant texts, articles, documents and many others. However, the thoughts of handling large amount of texts and data in a short time in a fast paced business world manually can be impossible to imagine since it is beyond the company's capacity to handle it. Therefore, companies need to make use of tools that can help them decrease the time needed for a process.

Now with text summarization, electronic documents can be fully utilized effectively as a tool in getting the gist of information before proceeding to read the whole document or texts.

2.2 TEXT SUMMARIZATION

Text summarization as presented in [1] by Moens, Angheluta and Dumortier strives for summarizing the most important and relevant content of the document and at the same time assisting in filtering and selecting information.

It aims at making a content made up of only essential information of the document so as to provide an easy reading on the gist of the document. Humans interpret data [2] basically by three steps which by understanding the document content, identifying essential information in the document and then writing the information in making the concepts more compact.

2.3 AUTOMATIC TEXT SUMMARIZATION

Automatic Summarization is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text.

Access to coherent and correctly-developed text summaries can be of great use, especially in our time of information overload, in which the amount of information electronically available to us, grows every day. Technologies that can make a coherent summary, of any kind of text, need to take into account several variables such as length, writing-style and syntax to make a useful summary.

As been laid out in [3] by Mani, there are several approaches on automatic text summarization which are:

1. Detail: Indicative/informative
2. Granularity: specific events/overview
3. Technique: Extraction/Abstraction
4. Content: Generalized/Query-based
5. Approach: Domain/Genre specific/independent

For this project, it had been deemed most suitable in selecting the technique approach. This type of approach makes use in generating summaries which is most suitable for a legal document summarization. Here therefore in generating these summaries, technique in extraction and abstraction will thus be used in getting the text. Extractive summarizing in [4] explained; aims at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary. While in Summaries by abstraction, [5] i.e. abstracts, make use of rather more complex linguistic technology techniques, where the output for abstracts is not a simple number of sentences retrieved as found in the input text but instead a brand new document which has been generated by processing the information contained there in providing the most important information within a document.

2.4 NATURAL LANGUAGE PROCESSING IN TEXT SUMMARIZATION

Natural language understanding is sometimes referred to as an AI-complete problem, because natural language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. Natural language processing, also called computational linguistics or natural language understanding, attempts to use automated means to process text and deduce its syntactic and semantic structure. This is done for many purposes such as to understand the nature of language, to extract specific information from text, machine translation, and to produce automated summaries such as being developed for this project.

The definition of "understanding" is one of the major problems in natural language processing. Natural Language Processing would be used for this particular project to develop text summarization as compared to other approaches as from the earlier works such as using lexical cues, linguistic approach and cohesion streamlining to today's summarization methods such as knowledge-rich methods, use of linguistic representation, statistical models and many others.

Specifically, NLP techniques will be used to extract from text documents by using semantic relationships between the words and its related meanings; in other words the relevance of the words to the document. Generally speaking, in developing the automated generated summarization as being discussed in [6] automated summarization encompassed 4 distinct processing which are:

1. Analysis of the source text
2. Identification of important source elements (Selection)
3. Condensation of information
4. Generation of the resulting summary representation. (Presentation)

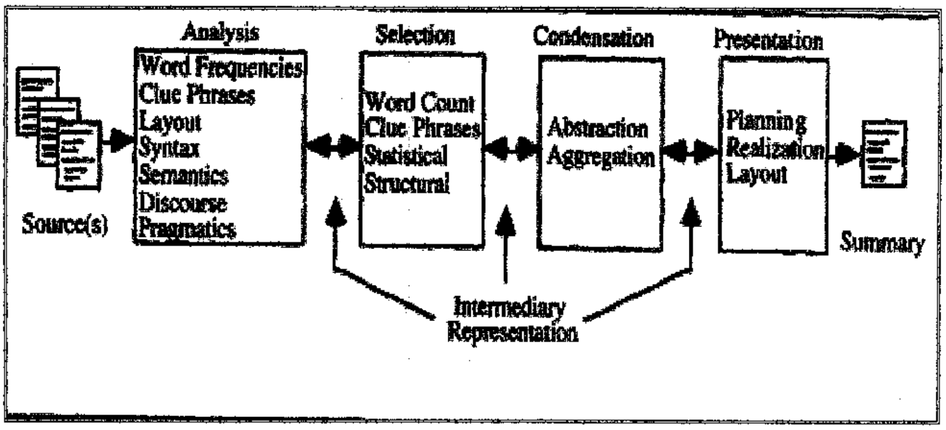


Figure 2.1: Mark T. Maybury Summarization Process

This process (Figure 2.1) is almost similar to how humans interpret and summarize texts or data as mentioned above where humans understands first the content of the document, its meaning, the style, where here it can be identified to the *Analysis* stage. Then human identify the most important pieces of information which this can be seen as similar on the selection of the *Selection* stage where word count and clue phrases gives indication on the importance of that particular words in the documents. While in *Condensation,* it can be seen that this stage is for the computer program to process the information in order for abstraction or aggregation. Next in *Presentation* of the summarized content, human processed it as writing the information.

Automated summarization also uses branches of AI such as machine translation and semantic networks to generate automatic text summarization. We can see from [7] TextAnalyst (an existing summarization system) which incorporates the use of semantic network for one of its function which is summarization where semantic network is utilized to score individual sentences in the investigated text. Here we could also see on the SUMMARIST system presented in [8] by Hovy and Chin that also use Natural Language Processing by using information retrieval and statistical techniques by incorporating technique for topic identification, topic interpretation and generation. Therefore we can see there are many methods in NLP alone for text summarization.

These techniques that were mentioned above had been successful in generating the required summaries, but as had been stated in all of the papers there are not one single solution for the developing text summarization. Various other theories and technique also are needed to be incorporated apart from NLP which had made them successful in developing text summarization.

Therefore, for this project a research and analysis will be done to find out on the best methods and techniques that are available in NLP to be incorporated for legal document summarizing.

CHAPTER 3
METHODOLOGY

3.1 PROCEDURE IDENTIFICATION

Since there is limited time for this project, therefore a "scrum" methodology will be incorporated. In developing the methodology, the main procedures are being developed in adherence and with references to the process of text summarization of Mark T. Maybury as mentioned before in Chapter 2 and SDLC methodology model. From here the author had chosen to combine this process into the methodology for this particular project.

This methodology had been chosen by taking into accounts the tasks that are needed to be done and the time needed to develop and deliver this project. The basic 5 phases that are being used here are:

1. Preliminary Analysis
    a. Analysis of Company and Procedures
2. Design Analysis
    a. Identification of Important Source Elements
    b. Designing User Interface
3. Design and Development
    a. Condensation of Information
4. Generation and Testing
    a. Generate summary

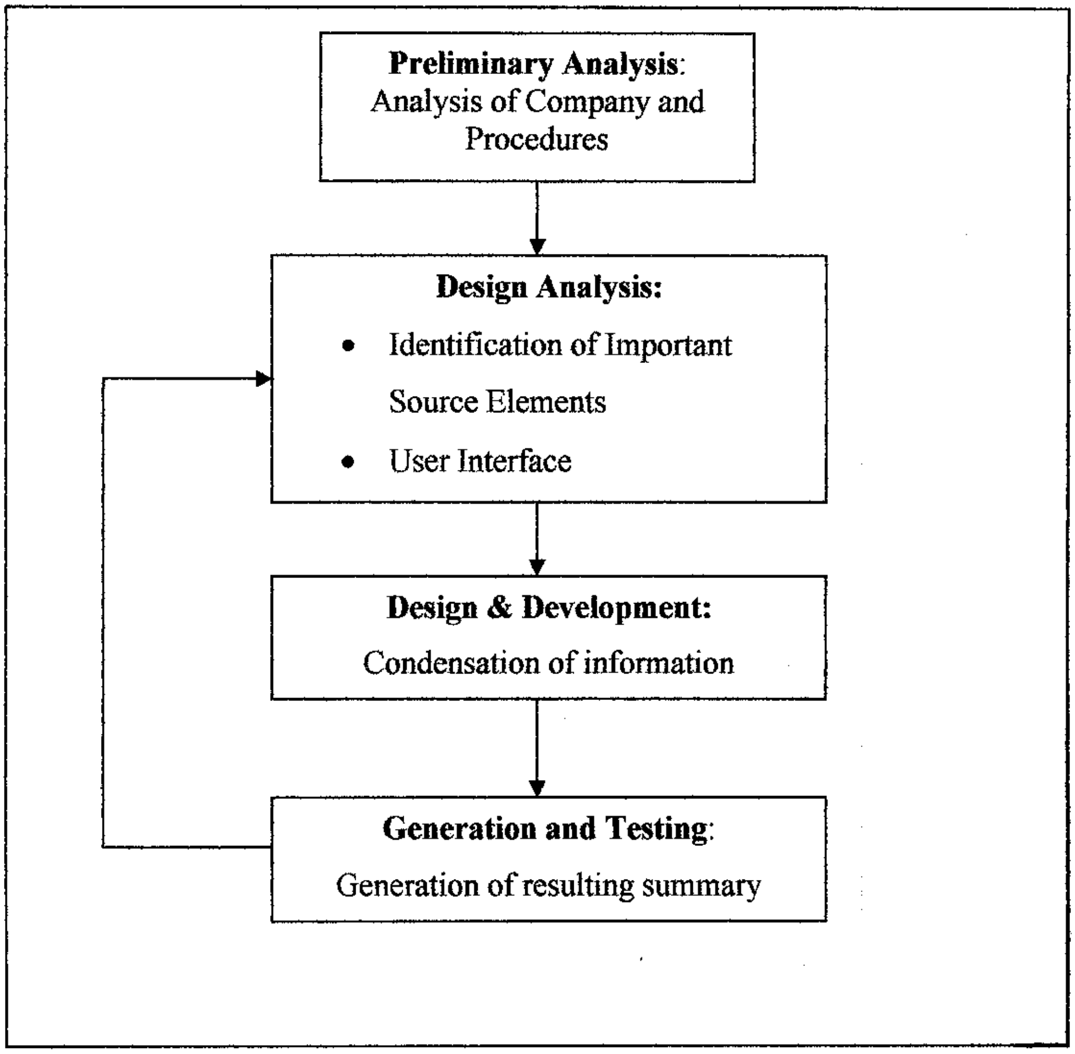

Preliminary Analysis:
Analysis of Company and Procedures
Design Analysis:
• Identification of Important Source Elements
• User Interface
Design & Development:
Condensation of information
Generation and Testing:
Generation of resulting summary

Figure 3.1: Diagram depicting methodology to be done throughout this project

3.1.1 Preliminary Analysis: Analysis of Company and Procedures

Analysis on the work or research that had been done previously on works on text summarization and various example of text or document summarization system was seen. The analysis and research focused mostly on the method that had been used to develop the text summarization system in order to get an idea and improvise it on developing this project. In doing the analysis, the methods used are mostly by studying on existing summarization system, their methods, and the white papers and journals written on text summarization and Natural Language Processing.

Concurrently, much of the work will also be done on researching and gathering the company requirements regarding the legal document summarizer. An important criteria that need to be looked on is on the specifications of what the companies are looking for in a contract. Understanding on the process of how the document is being reviewed is also vital in order to understand what type of information vital to be displayed on the summarization page.

The method that has been used in this stage is by interviewing the personnel involved in the tender process of the company on their flow, the processing of the information, requirements needed in the project and existing systems that are used. Also we will try to get on the user input regarding the legal document summarizer. Here, the criteria that are mostly taken into account in awarding a tender are being stressed on since this is the essential content needed for the summary. Example of the legal documents will need to be obtained, which then analysis of the legal tenders will be performed. A specification on the type of files used is also taken into consideration since the assumption in this project is that the company using this type of system is using an electronic form of documentation in storing their tenders or contracts.

3.1.2 Design Analysis:

3.1.2.1 Identification of Important Source Elements

In the first stage we analyze the legal documents which are the form of tenders and the contracts. Therefore on the completion of the first stage, identification on important source elements of the legal documents will be done. The relevance of the information to users will also be taken into account for creating both the abstract and the important details.

Methods on the generation of the text summarization will be done by using Natural Language Processing which contains techniques of using word clustering and word counting. First the usage on word counting will be used. In using word counting on the form of tender, it will be highlighted on what type of word that can be counted as important. Here comparison will be seen on either the lowest or the most words test can be deemed important with exception on articles, prepositions and conjunctions. Analysis on the words that had been counted will be done to see the relevance of the word in the document.

Then word clustering technique can be used under these most important words whereas a group of words relating to this particular word will be set up.

Also a manual version of a summarized legal document will be done in order to set a guideline by *specifying* characteristics expected in a legal document summarizing as in the process elicited by Edmundson.

3.1.2.2. User Interface

The storyboard for the module will be initially created first, depicting the interface of the system to give an idea on the flow of the system and user navigational flow. This will then be a base for the author in designing the user interface in Visual Basic. The user interface will then be designed first based on the storyboard developed. This user interface will just have the main functions design to begin with. Background color and images added in the module are being developed and edited at Adobe Photoshop.

Design flow of the user interface is also being done as will be presented in Chapter 4 which is the activity diagram. This diagram will depict the navigational flow of the system.

3.1.3 Design & Development: Condensation of information

In the design and development stage; this is where most of the technical work in this project will be done to develop the details of the working product of this project. Various software tools needed to be worked into here such as using Microsoft Excel and Microsoft Words in generating analysis for words generation and also programming tools such as Visual Basic will be needed in the practical development of making it into an automatic summarizer. A suitable coding will be built based on the technique to generate the summaries generation.

Also, other functions will be done in here apart from the developing a suitable code for the summarizer. Other functions such as uploading the file, editing, searching and help will also be included in order to create a more user friendly module. Also another function in displaying the company profile is being included in the text summarization module for easier viewing of company information. This is being done by using Microsoft Access.

3.1.4 Generation and Testing: Generation of resulting summary representation

The summarized output of the documents will be generated. An analysis on the output of the summarized data will be done and compared with the relevancy of the documents. This are also done as to test the system and to see if it meets the relevance of the documents and requirements. If there was an error then it will go back to the design analysis stages, through the design and development stage and down till this stage. Comparison with a manually summarized content is also being done with in mind that automated generated summaries is hardly perfect, therefore a comparison on relevance on the manually generated summary will be done.

When the summarized report of the tender had been met successfully with the relevancy of the requirements, then it will goes to the implementation stage.

3.2 TOOLS

In doing this project, some tools are needed to make this project a successful one. This includes both software and hardware that is listed below had been used throughout this project. The list below stated on the software's name and its functions, the browsers used for search on research materials, database used for this project and the hardware utilized for the development of this project.

Software:

- Microsoft Excel
  - Analysis on keyword occurrences
- Microsoft Words
  - Reporting
  - Designing storyboard
  - Generating analysis and graphs
- Microsoft Visio
  - Analysis on module workflow
  - Designing module workflow
- Macromedia Dreamweaver MX
  - Create electronic documents sample to be made into HTML files
- Adobe Photoshop
  - Generation for images and background for user interface
  - Image editing
- Visual Basic Studio
  - Development of module
  - Testing and Debugging of module

- Browsers:
    - Internet Explorer (IE), Version 6
    - Opera, Version 7.23

- Database:
    - Microsoft Access

- Hardware:
    - AMD Athlon 2500++
    - 1.83GHz
    - 256MBs Samsung DDR
    - 120GB Western Digital HDD

CHAPTER 4
RESULTS AND DISCUSSION

4.1 FINDINGS

4.1.1 Analysis on Technique for Summarizing

During the first month of this project, the initial analysis of the project was completed. From this stage it was seen that there were various techniques being used by Natural Language Processing for text summarization which included word counting, word clustering and more complex word clustering which are called concept fusion.

NLP technique was needed in developing automatic summarization along with various other techniques such as in the implementation of the SUMMARIST system. First, in identifying the topic of the document, word counting would be used along with theme extraction in comparing the words that had been counted if it was the most relevant. This would be done for the abstract of the summarized version.

As for highlighting the important details, it is a little more difficult since the data were widely dispersed into several documents. Therefore a template with combination in using word clustering can be for this type of extracting information.

4.1.2 Analysis on Criteria Needed for Document Summarization

Interviews had been conducted with 2 different companies in finding out on the criteria's most looked at in awarding a tender. While both of these companies had some similar criteria's in finding out the most suitable criteria's for the project that they are offering, huge dissimilarity can be found in a more technical details. In general the companies offering for tenders to bidders are mostly looking for these particular criteria in the bidder:

- Bidder profile
- Price
- Contractors involved

Adjustments are needed to be made if these systems were being made for both of these companies since the users have different criteria that they were supposed to look at. Therefore, this type of system would be suitable if it were made custom-tailored to the need of a particular company in highlighting important details; exception for generating abstracts for the document summarizing.

4.1.3 Word Counting and Sentence Extraction

Topic identification by using word counting for Information Retrieval may be used for abstracts in finding out the least important or most important word. This step would also be done by comparing the relevance of the counted words for the form of tender. Here, a combination for this step on using theme extraction can also be done. In theme extraction, the process works through the document to extract the most relevant facts and concepts.

In developing the extraction for the abstracts, sentence extraction or word extraction for summarization can be done if we identify on what is the most suitable candidate to the part of the summary. Some features that had been identified were based on Goldstein and Edmundson works:

- *Keyword-occurrence* – words most often used in document usually represent theme of document
- *Title* – The title and the following sentences are indicative of the themes of the document
- *Location heuristic* – Location on the document on the most important information lies where genres put important sentences in fixed positions, whereas here, the data needed are mostly included are at the beginning of the document e.g. for Form of Tender.
- *Upper-Case word Feature* - Sentences containing acronyms or proper names are included, example "Power Trainer".
- *First sentence*: First sentence of each paragraphs are the most important sentences

While the above features increase the score of a sentence to be included in the summary, there are exceptions are made in using these technique which are by not including:

- *Pronouns*: e.g. "she, they, it"
- *Articles*: e.g. "The, a, an"
- *Conjunction*: e.g. "and, if, or, but"
- *Preposition*: e.g. "to, in, on, near"

4.1.3.1 Word Counting on Form of Tender

For the document which is the legal tender for form of tender, we had counted the words and the frequency of the words occurring within the document. There are 220 different words in this document. As mentioned above, pronouns, articles, conjunctions and preposition were omitted which amounted to 34 words. And other words that are omitted are names that appear at the top of the form. At the next page, are table and graph depicting the data on the words that had been counted. Most of the words only appeared once whereas the most frequent appeared words (10 times) have been counted to only 2 words.

Frequency of words	Words counted
1	111
2	18
3	11
4	6
5	2
6	1
7	2
8	0
9	0
10	2

Table 4.1: Frequency of words appearing in the document

As can be seen below (Figure 4.1), words appearing only once at the document accounted for are 111 words altogether, while only 2 words appearing most of the times (10 times). Looking at the words counted, it can be deduced that words that appear 4 times (as per shown in the red line) and more can be deemed as important to the document.

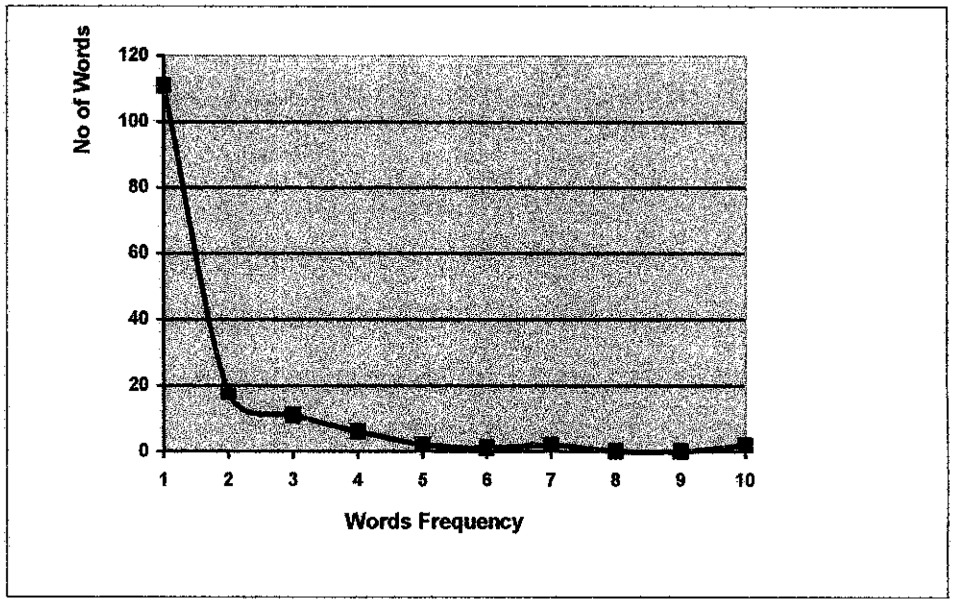


Figure 4.1: Frequency of words over number of words

As identified by Goldstein and Edmundson on the most suitable candidate to be the part of the summary, which are keyword occurrences, title, location heuristic, upper-case word and first sentences in paragraph. The keyword occurrence is often the theme of the documents. As being analyzed, the most often used words which are "tender" and "undersigned" represent the theme of the document that involves a tender and the undersigned (bidder). Also, most of the words that appear frequently are in the title of the form. Location heuristic find out where the most important information lies is being determined here by keyword occurrences within paragraphs which have the frequency more than 4 times.

Location	Keyword Occurrences
Paragraph 1	17
Paragraph 2	13
Paragraph 3	1
Paragraph 4	8
Paragraph 5	5
Paragraph 6	2
Paragraph 7	5
Paragraph 8	8
Paragraph 9	3
Paragraph 10	3

Table 4.2: Keyword occurrences within a paragraph

As stated before, important words that have upper-case word feature is also important where in this document "Power System Trainer" is the main item for this tender. It is also shown through keyword occurrences that "Power System Trainer" occurred more than once in this document. As for the first sentences is regarded as the most important in the documents, is yet to be seen due to the need of more research and documents to prove for this particular module. This is because in this document, some paragraphs have only one sentence in it.

Therefore from here we could see on the most suitable part of the document that can be brought in generating a summary by seeing on the title, paragraph containing the most frequently word occurrences and upper-case word features.

4.1.4 Word Clustering and Theme Extraction

Word clustering may be done on the important details of the tender such for example here, the word 'undersigned' in the document Form of Tender, are referred to as the bidder. Therefore, it would be useful in setting a word group under the word 'undersigned' e.g. bidder or company for a more coherent summary. This can also be incorporated together with several other techniques such as using template for generating a standardized version of important details and theme extraction.

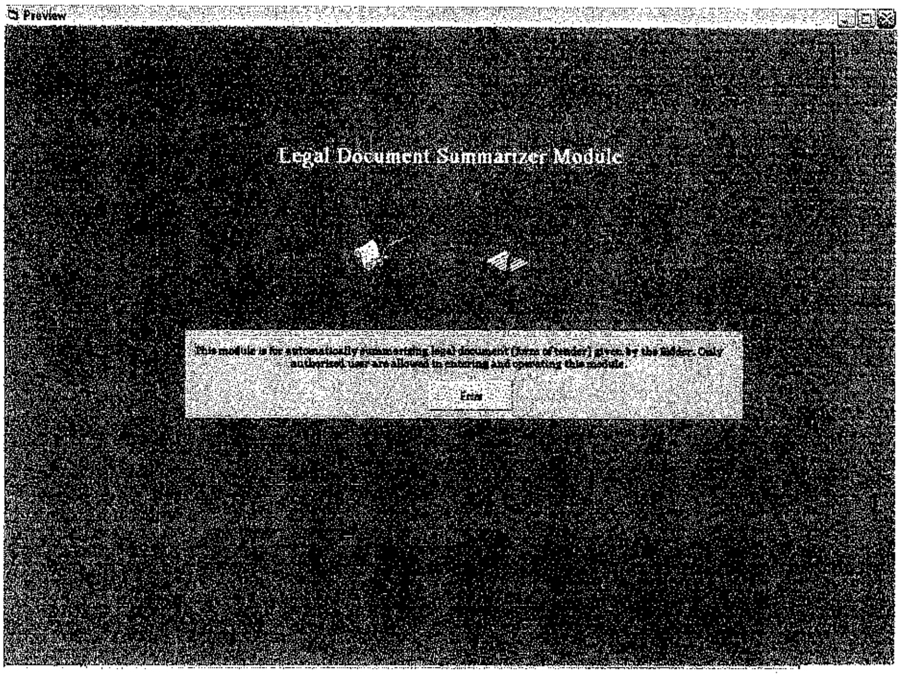
In incorporating the use of a template for summarization of technical writing; this can be similar to the Active Navigation solution in theme extraction which is by extracting and making the summaries in ways most relevant to the users.

4.2 MODULE DESIGN

4.2.1 Module Workflow

From the interview that had been done, a flow on how the user will access the data and summarized the information had been done. This flow which is an activity diagram depicting from the basic understanding on how what does the user needed to do in order to generate the summaries.

In starting the module the user will enter the page, then user needs to login in order to generate those summaries. Here a selected few users will be given the authority to enter this page.



Legal Document Summarizer Module

This module is for automatically summarizing legal document (form of tender) given by the bidder. Only authorised user are allowed in entering and operating this module.

Figure 4.2: The Enter page for Legal Document Summarizer Module

Then the user will then open the file containing the document of tenders or contacts by clicking on the "Open" button available and it will be uploaded into the text box provided. When uploading HTML file or documents into the text box, the text box will also displayed the tags that usually come along in HTML files. For that reason, the "Extract Tag" button will need to be click to clear away the tag of HTML file and it will displayed the content of the document. Then user will only need to click on summarizing button "Click" to generate the summary. If user chose to directly save the summary, then they can directly save the summary by clicking on the "Save" button.

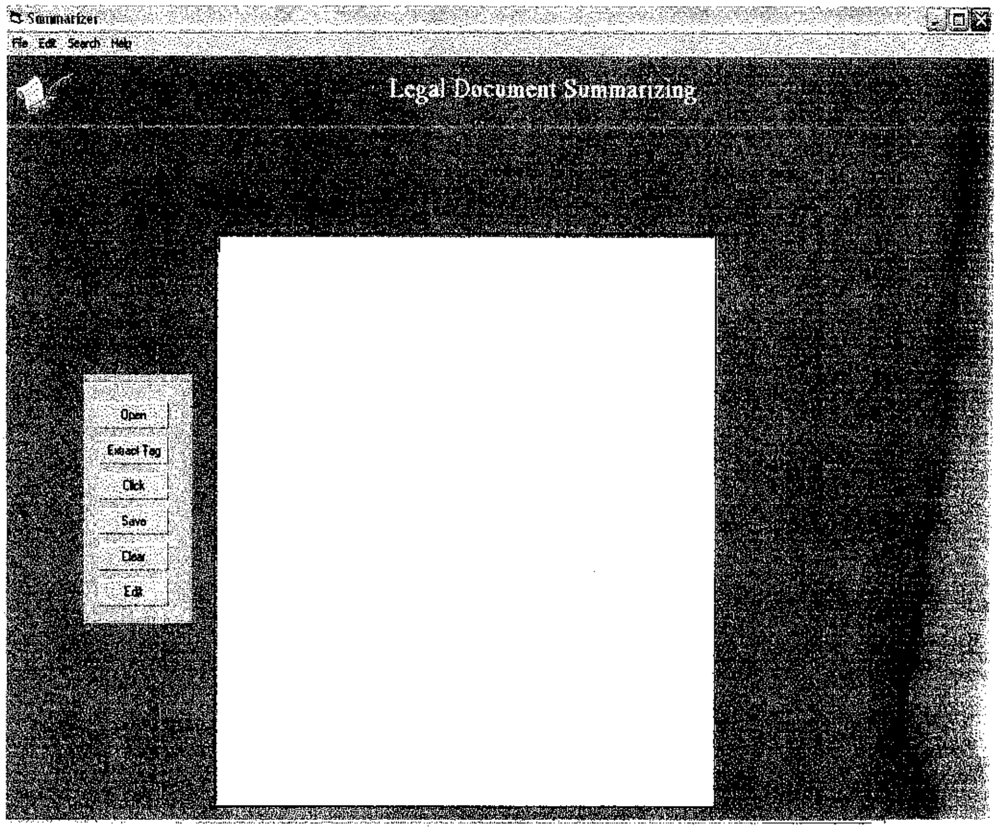


Figure 4.3: Main page for Legal Document Summarizer Module

User may also have the options to edit the summary. User might want to edit the summary by making it clearer, more understandable and easy to read for readers later. Therefore from the main page, the user may navigate to the "Edit Document" page by clicking on the button "Edit" and the Edit Document page will be displayed as shown on the next page for Figure 4.4. Here some of the functions that are available on Notepad and Words are available for the user editing purposes such as bold, italic, underline, align summary and changing the font size and type. Subsequently user will then save changes that were made by clicking "Save" under the "File" menu options.

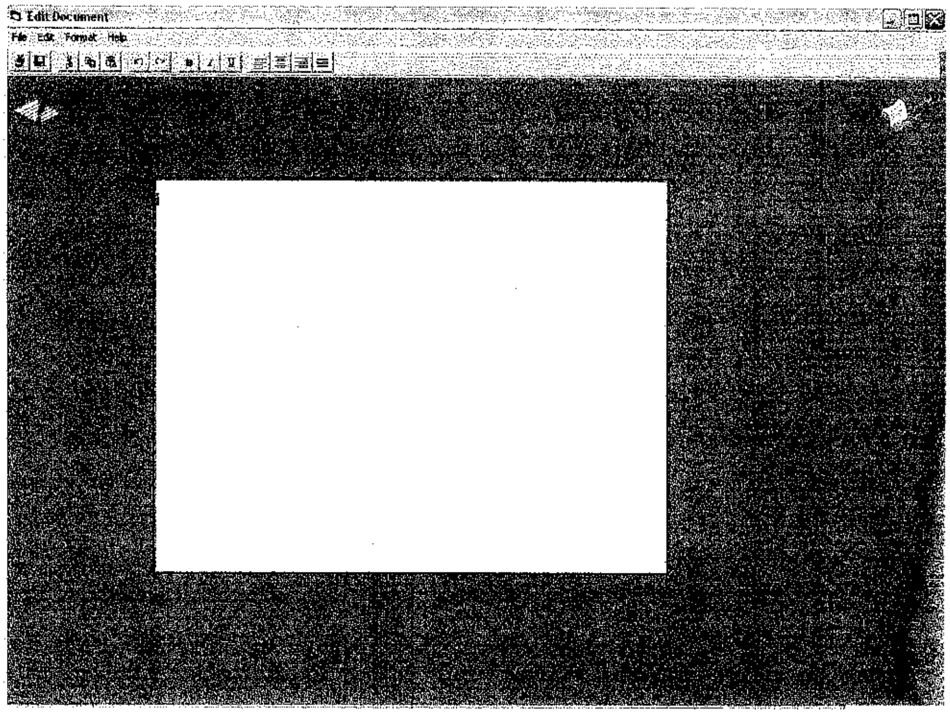


Figure 4.4: Edit page for Legal Document Summarizer Module

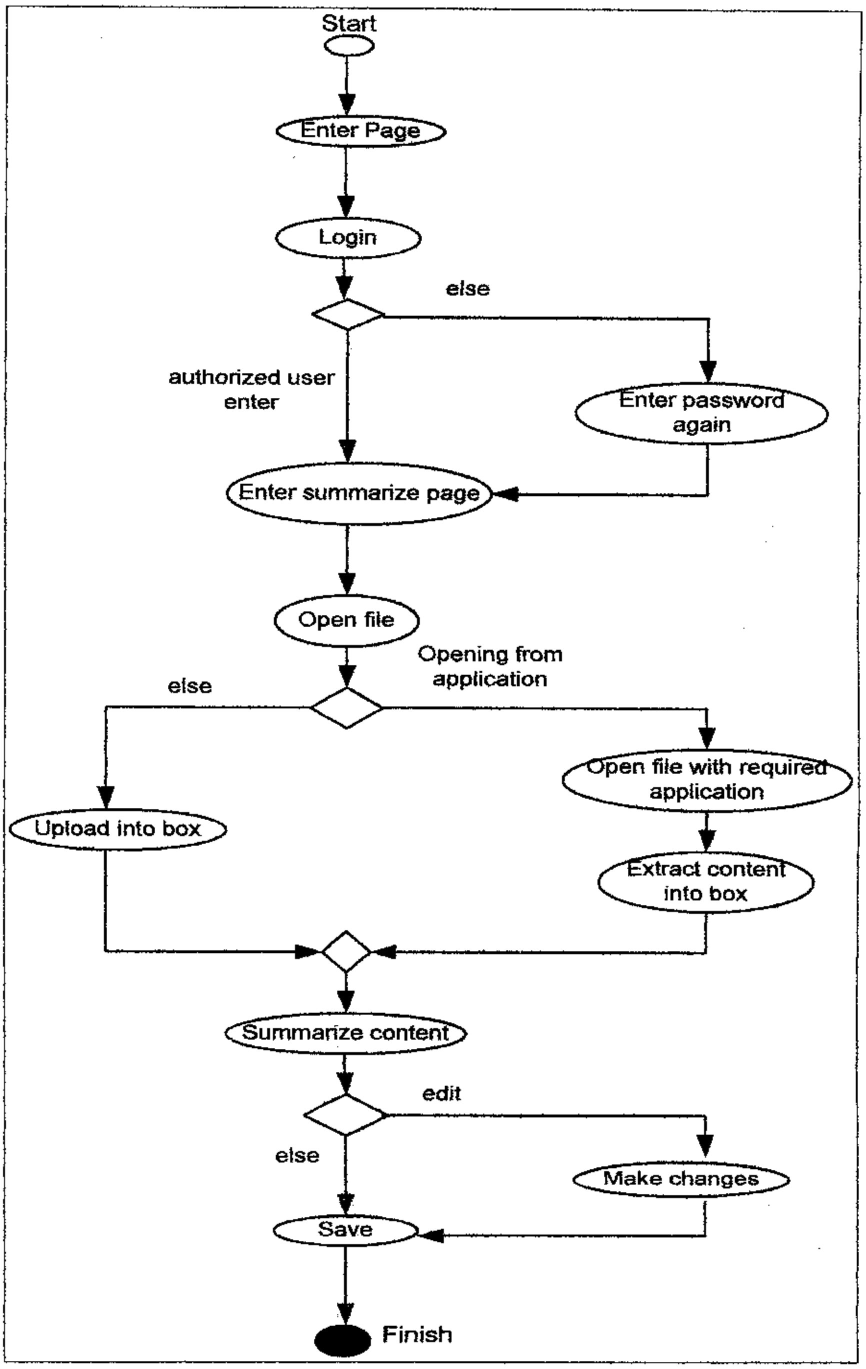


Figure 4.5: Activity Diagram for Legal Document Summarizing Module

Above is the activity diagram for the user navigational flow on this module. This diagram shows the user probable action courses that user may take in this module from when the user enters the page until they finished their task on this module.

4.3 DISCUSSION

4.3.1 Advantages

Coherent Summaries

This system was developed for developing automatic summarization of legal document for tender or contract purposes. Therefore a coherent summary will be generated for the company automatically. Here coherent summary meant that the summary would reflect the document essential content.

Flexibility in Changes

This module allowed for user to make changes to the summarized version of the document. This is to allow the user to add any important data to the content that had been left out by the system. User may also edit the summary to change it to a format that is more easily understandable and readable to the persons who may read them. This would then increase the summarized document reliability and integrity.

Alternative method from manual summarizers

This type of method has potential to outperform the manual summarization method. Decreases of time can be achieved compared to manually summarizing a large amount of documents and fewer personnel are needed for this daunting task if it were being given to them. This would also set a standardized type of summarized document for the company which will be an added advantage since confusion in understanding variation of interpretation by several users of the summaries can be avoided.

4.3.2 Limitations

Custom Made for Company

As mentioned before, it had been found that criteria needed amongst the companies for evaluation of bidder varies. Therefore this module needed to be custom made for a particular company based on the criteria's looked for by the company. Even though this can be a drawback in developing the module, this will increase the reliability on the summarized version of the document and a template can be derived based on this works.

Single Document Summary

This particular module would only be functional with a single document for summarization. Since tenders consists of several documents together, user needs to upload only one particular document from this documents of tender that are deemed necessary for abstraction and summarization. Therefore criteria looked for by each company are necessary developing this module in order to determine on the information that needed to be summarized. Therefore, this module focused on the gist of the overall documents as presented by Form of Tender.

English Summarization

Summary generated by this module are in English. Hence, this projects focusing mostly on documents that are being presented in English. Even though some of the technique can include other languages (e.g. word counting) to determine its importance but some of the proposed technique (e.g. word clustering) needs be in one language since it is related to the meaning of the words.

4.3.3 Objectives Met

Below are the objectives that had been met for this project as per comparison with the proposed objectives:

Objectives of the legal document summarizer:

- *To decrease the time needed for personnel to generate the documents manually.*
    - By successfully extracting the information in the documents, it can be said that objectives had been met in decreasing times needed from manual summarizing.

- *To ease the process of searches on important information of the legal tender.*
    - The keyword occurrences highlighted the important information in the legal tender. Therefore the objectives had been successful in easing the search process.

- *To improve time management and resource management hence resulting in fewer resources needed by using automated legal summarizing.*
    - Since by automatically summarizing the information, the summarizer module does improve time management by decreasing time needed to generate summary manually and resource management and fewer personnel are needed to handle automatic summarization. But this is yet to be proven within an organization.

Objectives of this project:

- *Developing a legal text summarization that can generate a summarized content of a document automatically.*
    - Here the scope had been changed a bit to generating a summarized content based on extraction of texts. Therefore this objective requires a longer time and research process in order for it to truly achieve its objective.

- *Incorporating AI field of NLP technique in developing a suitable method needed to create the module for the legal document summarizer.*
    - Here methods of NLP such as word counting and keyword occurrences for topic identification had been found useful in developing this particular module. Nevertheless, this objective is still not completely successful since there are various other NLP technique e.g. word clustering, that still needed more works and research to be done into it.

- *Search for a suitable methodology for developing a legal document summarizer.*
    - During the analysis phase, it had been found out on the most suitable methodology in creating legal document summarization by incorporating Mark T. Maybury's summarization process (Figure 2.1) with the traditional SDLC methodology. This in created a suitable method for developing a text summarization based module.

CHAPTER 5

CONCLUSION

By doing this project, it is expected that a suitable way of summarizing legal documents can be obtained by focusing on using Natural Language Processing techniques with combination of various technique and concept. Hence a technique that incorporates NLP such as word counting for topic identification and word clustering for abstraction on better interpretation of data had been found useful for this project. As mentioned before, summarization methods can be improved by including other technique such as theme extraction and sentence extractions.

This project focuses on the methodology and ways on how to generate a legal text summarization by extracting and getting the most relevant information to the document. Further enhancements can be done for this project such as including database, search functionality and since this project concentrates mostly on summarization.

With the development of this project, it is hoped that this will eliminate the need for manual summarizing for documents particularly in a large organization. Hence, benefits like easy viewing of the documents and decrease on time needed in handling this type of documents could be gained. This will also eliminate the need for more human resources and task that needs to be done.

A successful text summarization is based on how easy for a user to understand the summarized version of the document. To make a summarized document with the ability of abstraction as almost similar to manual summarizing by human, it is a highly complex area and much research is still being done conclusively on abstraction in the human level. Therefore more research is still required.

Nonetheless this project had achieved its objective in providing methods used in incorporating NLP for the generation of summaries.

RECOMMENDATION

Due to the limited time of the project given, there are some issues that may still need to be addressed:

1. Current prototype was set as a template in developing the legal document summarizing. More research needs to be done here such as developing a suitable algorithm and thorough research on the techniques being used. Further research word clustering and incorporating semantic technique could be done for better interpretation on meaning of the words.

2. Further extensive research could be done on more companies to generate better statistics. A large number of documents are also needed from these companies to get a more correct findings since the author have difficulty getting samples of tender documents due to the fact that this type of module deals with high confidentiality level of information.

3. More functionality added for the module. Since the module developed here is just a prototype, other functionality can be added to this system such as including a search engine for searching a specific contract and various other functionalities such as more editing functions for the user to edit the changes for the document.

4. Additional research can also be done for multi-document summarization for further expansion for this module and research,

since this type of document consisted of several documents for one particular bid of project.

5. A dual language document summarization module can also be done as enhancement later on for this module since currently this module only catered in summarizing documents that are presented in English. Hence, later this module could also summarized documents that are being presented in Bahasa Melayu.

6. The prototype can be fully integrated with existing or a new tender database system which may include all the processes involved for tender. This will promotes easier handling of the documents within the organization itself.

REFERENCES

[1] Moens, Marie-Francine., Angheluta, Roxana. and Dumortier, Jon., "*Generic technology for single and multi-document summarization*", 3rd July 2004, <www.sciencedirect.com>.

[2] Salton,Gerard., Singhal, Amit., Mitra, Mitra. and Buckley, Chris "*Automatic Text Structuring and Summarization*", 13th July 2004. <www.sciencedirect.com.>

[3] Mani, I., "*Automatic Summarization*". 2001: John Benjamin's Publishing Company.

[4] Ganapathiraju, MK., "*Relevance of Cluster size in MMR based Summarizer*", 26th November 2002, <www-2.cs.cmu.edu/~madhavi/11-742/report.pdf>

[5] Roca, S.C, "*Automatic Text Summarization*", 15th July 2004, <www.uoc.edu>

[6] T. Maybury, Mark, "*Generating summaries from event data,*" 13th July 2004, <www.sciendirect.com>.

[7] Ananyan,Sergei. and Kharlamov, Alexander., "*Automated Analysis of Natural Language Texts*", 14th July 2004, <www.megaputer.com.>

[8] Hovy, Eduard. and Chin Yew Lin., "*Automated Text Summarization in SUMMARIST*", 20th July 2004 <www.hemmingse/gslt/IA/summarist.pdf>.

[9] Norvig, Peter and Russell Stuart., "*Artificial Intelligence: A Modern Approach*", Prentice Hall, 2nd Edition 2003.

[10] Zak, Diane., "Visual Basic 6.0", Course Technology, Enhanced Edition 2001

[11] Futrelle, R.P. and Gautch S., "The role of automated word classification in the summarization of contents of sets of documents", 2nd International Conference on Information and Knowledge Management, November 1993, 18th August 2004.
<portal.acm.org>

[12] Kan, M.Y. and Klavans J.L., "Using Librarian Techniques in Automatic Text Summarization for Information Retrieval", 14th January 2002, 23rd September 2004.
<portal.acm.org>