STRUCTURING INFORMATION FROM UNSTRUCTURED TEXTUAL DOCUMENT

Ling Chei Huang

Bachelor of Computer Science with Honours (Network Computing) 2005
STRUCTURING INFORMATION
FROM UNSTRUCTURED TEXTUAL DOCUMENT

LING CHEI HUANG
(NETWORK COMPUTING)

This project is submitted in partial fulfillment of the requirements for the degree of Bachelor of Computer Science with Honours

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2005
DECLARATION

No portion of the work referred to in this report has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher learning.

LING CHEI HUANG

APRIL 8, 2005

8515
ACKNOWLEDGEMENTS

First of all, I dedicated this work to my God for giving me the confidence and wisdom to complete this project.

I would like to express my deepest gratitude to my supervisor, Mr. Bong Chih How for his guidance and motivation in making this project a success. I wish to thank him for being very patient and helpful to me throughout the development of the application.

My family also had been a great support. Here, I specially thanks to them. They are the inspiration for me to complete my project. All encouragements, sacrifices and prayers from them really help a lot.

Finally, heartiest thanks to my friends who have given patience and support to me towards the completion of this project. Here I would like to express my special thanks to Mr. Teng Cheng Hong for his advice and help.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>x</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xi</td>
</tr>
<tr>
<td><strong>CHAPTER 1 INTRODUCTION</strong></td>
<td></td>
</tr>
<tr>
<td>1.0 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Structuring Information from Unstructured Textual Documents</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Objectives</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Proposed Solution</td>
<td>4</td>
</tr>
<tr>
<td>1.5 Scope of Study</td>
<td>5</td>
</tr>
<tr>
<td>1.6 Expected Outcome</td>
<td>5</td>
</tr>
<tr>
<td>1.7 Significant of Research</td>
<td>5</td>
</tr>
<tr>
<td>1.8 Chapter Review</td>
<td>6</td>
</tr>
</tbody>
</table>
CHAPTER 2 LITERATURE REVIEW

2.0 Introduction 8

2.1 Review on existing system 8

2.1.1 FASTUS (Finite State Automaton Text Understanding System) 8

2.1.2 Smart Ideas 11

2.1.3 KMap 16

2.1.3.1 Generating Concept Maps by Browsing in Netscape 19

2.2 Comparison for system features 20

2.3 Proposed System 20

2.4 Conclusion 23

CHAPTER 3 RESEARCH METHODS AND ANALYSIS

3.0 Introduction 24

3.1 GATE Overview 26

3.1.1 GATE - General Architecture for Language Engineering 26

3.1.2 Resources in GATE 26

3.1.3 Applications and Datastores 27

3.1.4 Annotations 27

3.2 ANNIE (A Nearly-New Information Extraction System) 29

3.2.1 Tokeniser - gate.creole.tokeniser.DefaultTokeniser 30

3.2.2 Gazetteer - gate.creole.gazetteer.DefaultGazetter 30

3.2.3 Sentence Splitter - gate.creole.splitter.SentenceSplitter 31

3.2.4 Part of Speech Tagger - gate.creole.POSTagger 31

3.2.5 Semantic Tagger 31
CHAPTER 5 TESTING AND EVALUATION

5.0 Introduction

5.1 Testing Methodology

5.2 Effectiveness Measurement

5.3 Evaluation

5.3.1 Result

5.3.2 Discussion

5.4 Conclusion

CHAPTER 6 CONCLUSION AND FUTURE WORKS

6.0 Introduction

6.1 Achievements

6.2 Constraints and Limitations

6.2.1 Process document in text format

6.2.2 Process document in English language

6.2.3 Point Extraction

6.4 Conclusion

REFERENCES

APPENDIX A Research Questionnaires

APPENDIX B Documents

APPENDIX C Outputs
LIST OF FIGURES

Figure 2.1 FASTUS System 9
Figure 2.2 A hyperspace (large thick oval) and sub views of the hyperspace 12
Figure 2.3 Hyperspace in Smart Ideas 13
Figure 2.4 An example of a live concept map embedded in an HTML document 15
Figure 2.5 Example of Concept map indexing 16
Figure 2.6 Popup menu for node types 17
Figure 2.7 Node type editing and creation 18
Figure 2.8 The ANNIE pipeline in GATE 22
Figure 3.1 The Information Structuring Process 25
Figure 3.2: Components of ANNIE 29
Figure 4.1: Sample of paragraph annotation 42
Figure 4.2: Sample of sentence annotation 43
Figure 4.3: Sample of token annotation 44
Figure 4.4: Sample of point generated 44
Figure 4.5 The sample mind map with three points 45
Figure 4.6 Sample output for the actual implementation 46
Figure 5.1: Formula for calculation of the percentage of error 49
Figure 5.2: Formula of calculation of the percentage of effectiveness 49
Figure 5.3: Calculation of the percentage of effectiveness 52
Figure 5.4 Accuracy Responses from Users 53
LIST OF TABLES

Table 2.1: Comparison for system features 20
Table 5.1: Percentage of Effectiveness 51
Table 5.2: The accuracy of the point marked by user 51
ABSTRAK

Kertas ini membentangkan satu cara untuk mempersembahkan informasi satu dokumen dalam bentuk grafik. Ini bermakna satu teks dokumen akan ditukarkan kepada informasi yang senang dan mudah difahami dalam bentuk grafik seperti peta minda. Disebabkan semakin banyak dokumen dihasilkan setiap hari, oleh itu, informasi dalam bentuk grafik adalah sangat berguna kepada pengguna untuk mendapatkan idea dalam dokumen dengan cepat tanpa membaca dokumen tersebut perkataan demi perkataan.

Aplikasi ini dihasilkan dengan menggunakan "General Architecture for Text Engineering (GATE)". GATE dihasilkan oleh Universiti Sheffield, United Kingdom.

ABSTRACT

This paper presents an approach for structuring information. Structuring information means to convert a textual document into a simple structured information representation such as graphical map. As more and more documents are generated everyday, the structured information is useful to enable users easily grasp the ideas of a document without reading the document word by word.

The application for structuring information from the unstructured textual document is developed by using the General Architecture for Text Engineering (GATE) tool, which is developed by University of Sheffield, United Kingdom.

The implementation of this application enables the users to get the structured information of a document quickly. The application first read an unstructured textual document, such as a paragraph of text. Then the document’s main points will be extracted. Finally, the main points will be illustrated in graphical map, which is simple and easy to understand. The graphical map represents the structured information from the document.
1.0 Introduction

Structuring information means to convert a textual document into a simple structured information representation such as graphical map and diagram, but this undertaken project will concentrate on the graphical map. Structuring information is done by extracting the main points from an unstructured textual document and then rearranges the significant points in a graphical map.

Unstructured textual documents are the writing that contains text information, which is interpreted subjectively. An unstructured document is data rich. On the other hand, structured information is a simple representation of information of a certain text that is interpreted in graphical methods. It is a summary of a text document that consists of the main ideas and its points.

It is very time consuming if we have to read through the whole document to get the main idea in that particular text. Instead of skim through the whole document, it would be preferable analyzing and extracting the document to get the structured information in the form of graphical map.

Graphical map is a graph that consists of nodes with connecting arcs. It expresses the relationship between nodes. It tends to be easier to understand by users compared to the full text documents. It is useful as it is easier to get the main ideas by looking at the graphical map and to get the points that are showed in each node in the graphical map.
1.1 Structuring Information from Unstructured Textual Documents

We can extract and structure information from documents if we can match attributes with document data values and associate these matched attribute-value pairs as tuples in relations (David W. Embley & others). A textual document can be extracted to reformulate as a structured document. But not all documents can be structured with 100% of accuracy ratios of results. However, it is expected to provide the result with up to 80% of accuracy.

The structuring information from unstructured textual documents can be done manually or automatically by the help of computer processing power. The graphical map with the higher accuracy can be produced by using the manual structuring approach. The manual structuring approach required the users to read through the unstructured textual documents, which can be in either hardcopy or softcopy. After understanding the content of the document, the main points in the document to be illustrated in the graphical map. This approach relied on the understanding of a user to develop the graphical map. Different user may provide with the different graphical map.

The automatic approach involved the utilization of the processing power of a computer. The application for structuring information is to be used to help in structured the information from the unstructured documents. The application will first read the textual document in softcopy, and then follow by the extraction of the points. From the points, a graphical map will be developed. This approach provides faster manner in extracting a document, it also depends on the processing power of the computer and the performance
of the application. The powerful application will provide higher accuracy of graphical map, while others may not.

Structuring information from unstructured textual documents is very useful as there are more and more documents for user to read before they can get the right information. By extracting the documents into structured information, it is easier for the users to acknowledge the main ideas of the documents. Thus, it is no need for users to read every single document.

1.2 Problem Statement

The problems encounter especially when reading documents are:

- Difficulties in getting main ideas

  Usually we have to read through whole document word by word, sentence by sentence to understand the main text in the document and ideas of the text. It is difficult and laborious if we have to read through every document to get the main ideas. Furthermore, some documents making turns before getting into main ideas.

- Time constraint

  It is very time consuming to digest the contents of a document. Further, as the resources from Internet increase, it is not economic in term of our time to read every single document.
• Problem of misleading title
   As we always scan the main title to select the related document to be read, we might end up by reading document which is unrelated or not in our interest due to its misleading title.

1.3 Objectives
The objectives of this project are:

i) To provide a solution based on Natural- Language processing (NLP) to perform summarization for documents

ii) To construct a graphical map from the summarization to help readers to grasp the document quickly before reading.

1.4 Proposed Solution
As stated in the problem statement, we need a system to analyze text based on the content structure of text in order to extract the main points, and then represent it in the simple form, which make it easier to be interpreted and comprehend. This will also speed up the document reading. In addition, the relevancy of the document can be justified.

It is proposed to develop a prototype of application to extract and structure the information from unstructured textual documents. The application should be able to read an unstructured document, such as a paragraph of text, and then extract and represent the document’s main ideas in structured information, such as in graphical map, which is simple and easy to understand.
1.5 Scope of Study

To complete this project, we are going to study on the Natural-Language processing (NLP) approach in document contents extraction. The General Architecture for Text Engineering (GATE) tool will be studied for its architecture and to be applied in this research project as necessary. GATE tool is developed by University of Sheffield, United Kingdom.

1.6 Expected Outcome

At the end of this research project, it is expecting to come out with an application prototype for structuring information from unstructured textual documents. The proposed application should be able to illustrate the contents of unstructured document, and then extract and represent the document’s main ideas in structured information, such as in graphical map.

1.7 Significant of Research

This research project is mainly to help the user to digest the textual document easier in sense that user no need to read through every single document to get the main ideas. This shortens the time for reading lots of irrelevant documents daily.
1.8 Chapter Review

Each chapter documented in this report covers an aspect of the entire research done.

Chapter 1: Introduction

This chapter gives a short tour on the entire project. The introduction of the structuring information is highlighted. This chapter gives a summary of this research project.

Chapter 2: Literature Review

This chapter outlines the review on the existing system in the market. The technique and the method used in each system were studied and it is to be used in the development of the project when necessary.

Chapter 3: Research Methods and Analysis

This chapter highlights the analysis of the methods in completing this research project. The modules will be identified and studied in details. Each module will be outlined in details.

Chapter 4: Design and Implementation

This chapter outlines the design and implementation of the application to be carried out to complete this research project. It will be illustrated in details of the flow of the application to be developed.
Chapter 5: Testing and Evaluation

This chapter includes the report for the testing on the prototype that had been developed. The observations are made on the reports. Also included is the evaluation of the research project.

Chapter 6: Conclusion

This chapter concludes the whole research project. This concludes the efficiency and the effectiveness of the prototype that had been developed.
CHAPTER 2 LITERATURE REVIEW

2.0 Introduction

To have sufficient knowledge to develop a new application, strong literature reviews are required. Studies on the problems and limitation of existing systems will help in enhancing the new developed application with more useful new features.

This chapter reviews all the existing systems that are currently available in the market. There are various systems in the market today with the different features. However, only the systems that structuring the unstructured textual document are focused for this research paper.

The reviews are done on the systems that have similarity to the proposed system. The tools, techniques, and features of each system are identified and compared.

2.1 Review on existing system

Reviewed below is some of the existing that are found to be related to the proposed application. The highlighted features will be the information extraction and the graphical concept mapping tools.

2.1.1 FASTUS (Finite State Automaton Text Understanding System)

FASTUS is an acronym for Finite State Automaton Text Understanding System. It is a system for information extraction. It works with free text in English, and other languages as well, as the English and Japanese versions of system are currently
available. FASTUS was developed in response to the needs of the intelligence community for scanning and processing huge volumes of written texts.

FASTUS is one of the top technologies in the information extraction field. It can analyze an average length of news or article in under 10 seconds. This helps to translate 10,000 articles per day, on a single processor.

![Diagram of FASTUS System](image)

Figure 2.1: FASTUS System (Jerry Hobbs and others, 2002)
The operation of FASTUS is comprised of three steps: (Jerry R. Hobbs and David Israel, 2002)

1. Recognizing Phrases: Sentences are segmented into groups such as noun, verb, and other phrases.
2. Recognizing Patterns: The sequence of phrases is scanned for patterns of interest, and when they are found, corresponding "incident structures" are built.
3. Merging Incidents: Incident structures from different parts of the text are merged if they provide information about the same incident.

The current version of FASTUS may be thought of as using five levels of processing:

1. Complex Words: This includes the multi words and proper names recognition.
2. Basic Phrases: Sentences are segmented into groups such as noun, verb, and other phrases.
3. Complex Phrases: Complex noun and verb groups are identified.
4. Domain Events: The sequence of phrases that are produced at (3) is scanned for interested event to the application, and the structures information are built for entities and events contained in the pattern.
5. Merging Structures: Structures arising from different parts of the text with same entity or event are merged.

The advantages of the FASTUS system are as follows: (Jerry R. Hobbs and David Israel, 2002)

- It is conceptually simple. FASTUS is a set of cascaded finite-state automata.
- The basic system has a relatively small footprint.
• It is effective.
• It has very fast run time. The average time for analyzing a message is less than 10 seconds.
• In part because of the fast run time, it has a very fast development time. This is also true because the system provides a very direct link between the texts being analyzed and the data being extracted.

The FASTUS system has been extended in the following ways:
• The development of a convenient interface that will allow grammar writers to define patterns more easily.
• The implementation in Japanese language version of FASTUS.
• The adoption of FASTUS for the use in high-precision Information/Document Retrieval engine.

FASTUS is a mature, robust, effective, efficient information extraction system. It is not a text understanding system, but an information extraction system. It is perhaps the most convenient and most effective system that has been developed for extracting the information.

2.1.2 Smart Ideas
Smart Ideas is a concept mapping tools and hypermedia program which allows users to draw concept maps with a graphical direct-manipulation interface. Most people enthusiastically adapt to reading and understanding informal concept maps rather than to read the full text documents. Kremer (1996) has introduced computer-supported
concept mapping to individuals who could then independently draws concept maps, and shared concept map workspace. *(Rob Kremer and Brian R. Gaines, 1996)*

Users can draw the concept maps using an editor, every node and link drawn in the map will also be represented in an underlying hyperspace *(Figure 2.2)*, which can be shared with other users and other concept maps. The hyperspace contains the hypernodes that have hyperlinks between them. The window labelled “The Complete Protocol” contains the entire hyperspace view.

*Figure 2.2: A hyperspace (large thick oval) and sub views of the hyperspace (Rob Kremer and Brian R. Gaines, 1996)*