Wiki Saga: An Approach for the Digitisation, Processing and Visualisation of Historical Documents

Daniel Tan Yong Wen
Master of Science
(Computer Science)
2015
Wiki Saga: An Approach for the Digitisation, Processing and Visualisation of Historical Documents

DANIEL TAN YONG WEN

A thesis submitted
in fulfilment of the requirements for the degree of Master of Science

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2015
ABSTRACT

A historical document contains information about past events which can be a source of reference. In this research, the selected historical document is the Sarawak Gazette, a monthly newspaper that reported on what happened in Sarawak. With one hundred and forty four years of reports since its first publication on Friday, August 26, 1870, the Sarawak Gazette is one of the most important historical document for information on the history of Sarawak. The task of gleaning for information by laboriously going through pages of printed pages is an arduous task in terms of time and effort. This research focuses on enabling a semantic search on the Sarawak Gazette, as a case study, for visualising a summary of what actually happened in Sarawak during a certain period. This research proposes a pipeline process that involves digitising the Sarawak Gazette, a natural language process that extracts named entities and a timeline generator to display events as reported. Due to the difficulties of the task, the current state-of-the-art approach makes use of human power as part of a mass digitisation projects by Google. A prototype system, Wiki SaGa, visualises the digitised documents in conjunction with the generated timeline. Through Wiki Saga, researchers who use the Sarawak Gazette can search for specific information on an event that happened in Sarawak during a certain timeframe by using the timeline display. By extracting named entities and displaying them within events in a timeline, researchers can have a summary of the event. By visualising events in a timeline, semantic patterns are recognised and related events can be identified. Through this research, Wiki Saga, a new archival and retrieval system, has been produced. In the process a semi-automated approach for digitising all the documents is also now available to researchers.
ABSTRAK

ACKNOWLEDGMENT

First and foremost, I thank God for giving me the inspiration to conduct this research. It has motivated me to be involved in doing research.

My heartfelt gratitude to my supervisors, Professor Dr. Narayanan Kulathuramaiyer, Associate Professor Dr. Bali Ranaivo-Malançon. Without your wisdom and guidance, my progress would not be so smooth. Thank you for the time spent on our discussions and the experience you have shared with me.

I would like to thank my parents as well, for your support all these while. To the staff of International East Asian Study Institute, thank you for sharing with me the knowledge I need regarding the Sarawak Gazette.
LIST OF TABLES

Table 2-1: Comparison of digitisation methods ................................................................. 16
Table 2-2: Recommended scanned document settings ......................................................... 17
Table 2-3: Comparison of Wiki systems .............................................................................. 21
Table 2-4: Comparison of NER tools .................................................................................. 24
Table 2-5: Comparison of timeline generators .................................................................... 27
Table 4-1: SaGa used for the research ................................................................................. 52
Table 4-2: Errors and cause of errors ................................................................................ 60
Table 4-3: Comparison of words and characters in The New School using ABBYY OCR and Tesseract OCR ......................................................................................... 62
Table 4-4: Comparison of words and characters in the collection ...................................... 62
Table 4-5: Results of Annotation Diff Tool ......................................................................... 65
Table 4-6: Evaluators list .................................................................................................... 69
Table 4-7: Wiki SaGa evaluation result for forty random users ........................................... 70
Table 4-8: Wiki SaGa evaluation result for ten selected users ............................................ 71
LIST OF FIGURES

Figure 1-1: Research Methodology ................................................................. 8
Figure 2-1: Digitisation methods ................................................................. 15
Figure 3-1: Framework Design ..................................................................... 30
Figure 3-2: Digitisation process ................................................................. 32
Figure 3-3: Errors / recommendations during OCR process ..................... 33
Figure 3-4: OCR result .................................................................................. 34
Figure 3-5: NER stage .................................................................................. 36
Figure 3-6: Annie system .............................................................................. 37
Figure 3-7: Result of Annie NER document .............................................. 38
Figure 3-8: Human corrected document ..................................................... 38
Figure 3-9: Excerpt of Exported XML document ......................................... 39
Figure 3-10: Wiki SaGa ............................................................................... 41
Figure 3-11: Sample Timeline ...................................................................... 42
Figure 3-12: Implementation of Filter and Highlight Function .................. 45
Figure 3-13: Interface of timeline with filter and highlight function .......... 45
Figure 3-14: The Batang Lupar Troubles as shown in Wiki SaGa ............... 47
Figure 3-15: Wiki SaGa structure ................................................................. 48
Figure 4-1: Scanned pages of SaGa: January 1903 .................................... 53
Figure 4-2: Scanned pages of SaGa: February 1903, page 30 and 31 ......... 54
Figure 4-3: SaGa header .............................................................................. 56
Figure 4-4: OCR processed texts ................................................................. 57
Figure 4-5: Extraction of event report .......................................................... 58
Figure 4-6: Corrected event report ............................................................... 59
Figure 4-7: Comparison of different versions of same document ............... 61
Figure 4-8: Details of an event ................................................................................................. 67

Figure 4-9: Timeline search and highlight result ..................................................................... 74

Figure 4-10: Knowledge model constructed with evaluator by using Wiki SaGa ................. 75

Figure 5-1: Wiki Saga as a semi-automatic knowledge base .................................................. 79
# LIST OF ABBREVIATIONS USED IN THE THESIS

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>Named Entity</td>
</tr>
<tr>
<td>NER</td>
<td>Named Entity Recognition</td>
</tr>
<tr>
<td>SaGa</td>
<td>Sarawak Gazette</td>
</tr>
<tr>
<td>OCR</td>
<td>Optical Character Recognition</td>
</tr>
<tr>
<td>TXT</td>
<td>Text File</td>
</tr>
<tr>
<td>DOC</td>
<td>Microsoft Word document file</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

ABSTRACT........................................................................................................................................... III

ABSTRAK.............................................................................................................................................. IV

ACKNOWLEDGMENT.................................................................................................................. V

LIST OF TABLES.......................................................................................................................... VI

LIST OF FIGURES........................................................................................................................ VII

LIST OF ABBREVIATIONS USED IN THE THESIS ................................................................. IX

TABLE OF CONTENTS.............................................................................................................. X

Chapter 1. Introduction....................................................................................................................... 1
  1.1. Research Motivation and Problem Statement ...................................................................... 1
  1.2. Related Works ..................................................................................................................... 3
  1.3. Research Questions .......................................................................................................... 4
  1.4. Research Objectives ....................................................................................................... 5
  1.5. Research Methodology .................................................................................................. 6
  1.6. Research Scope .............................................................................................................. 7
  1.7. Thesis Organisation......................................................................................................... 9

Chapter 2. Literature Review.......................................................................................................... 10
  2.1. Current Work ................................................................................................................. 10
  2.2. Historical Documents: Digitisation and Storage .......................................................... 11
    2.2.1. Digitisation .......................................................................................................... 12
2.2.2. Different Methods for Digitising a Document

2.2.3. Best Practice in Digitisation

2.2.4. OCR as a software for digitisation

2.2.5. Wiki as a repository

2.3. Named Entity Recognition (NER) and Extraction

2.3.1. What is a Named Entity (NE)?

2.3.2. What is a NER?

2.3.3. Existing NER tools

2.4. Timeline Display

2.4.1. What is a Timeline?

2.4.2. What is a Timeline Generator?

2.4.3. What are existing timeline generators?

2.5. Summary

Chapter 3. Pipeline Framework Design and Implementation

3.1. Pipeline Framework Design

3.2. Situational Study

3.3. Digitisation

3.4. SaGa Named Entity Recognition and Extraction

3.4.1. NER Implementation

3.4.2. Post NER correction

3.5. Archival System
Chapter 3. Timeline Generation and Display with SIMILE Timeline Widget ...............42

3.5.1. Timeline Generation and Display with SIMILE Timeline Widget ...............42

3.5.2. Wiki SaGa ...........................................................................................................46

3.6. NER, Timeline and Wiki.........................................................................................49

3.7. Summary ..................................................................................................................49

Chapter 4. Wiki SaGa: Experimental Results and Analyses ......................................52

4.1. Initial Data ................................................................................................................52

4.2. OCR Results and Analysis .....................................................................................55

4.2.1. OCR Results .......................................................................................................55

4.2.2. OCR Error Analysis ...........................................................................................60

4.3. NER Results and Analysis .....................................................................................63

4.3.1. NER Results .......................................................................................................63

4.3.2. NER Error Analysis ...........................................................................................64

4.4. Wiki SaGa Evaluation Results and Analysis .........................................................67

4.4.1. Wiki SaGa Evaluation Results ..........................................................................68

4.4.2. Wiki SaGa Analysis ...........................................................................................72

4.4.3. Semantic Model Construction ..........................................................................74

4.5. Summary ..................................................................................................................75

Chapter 5. Conclusion and Future Works ....................................................................77

5.1. Main Contributions .................................................................................................77

5.1.1. Design a framework to visualise historical documents’ content ....................77

5.1.2. Propose an interactive access to historical documents ....................................77
5.1.3. Allow searching named entities within historical documents.................78
5.1.4. Implement a new archival and retrieval system for historical documents....78
5.2. Research Limitations.............................................................................79
5.2.1. Limited Number of Processed Documents..........................................79
5.2.2. Low performance of OCR and NER Tools...........................................79
5.3. Future Work ............................................................................................80
5.3.1. Increasing the Amount of Historical Documents....................................80
5.3.2. Increasing the Accuracy of OCR Results and NER Results....................80
5.5. Concluding Remarks .............................................................................81
References ....................................................................................................82
Appendix A: Sample of OCR Results by ABBYY Fine Reader........................88
Appendix C: Wiki SaGa Evaluation Form .......................................................96
Appendix D: Sample Exported XML Document.............................................100
Appendix E: Integrating SIMILE .................................................................104
Chapter 1. Introduction

1.1. Research Motivation and Problem Statement

A research Centre, Institute of East Asian Studies (IEAS), at the Universiti Malaysia Sarawak (UNIMAS) has acquired a subset of Sarawak Gazette, which henceforth will be referred as SaGa. It is the oldest newspaper published in Sarawak. Its first publication was on Friday, August 26, 1870, during the reign of the first White Rajah, James Brooke. The Rajah’s Civil Service edited the monthly gazette. The published articles reflected the official thinking of the Rajah on major issues (Sarawak Gazette Delima Edition, 2006). The documents acquired were all scanned images stored as portable document file (pdf) in compact disks (cd).

According to a former IEAS director, Prof. Datuk Dr. Abdul Rashid Abdullah (personal communication, August 15, 2011), IEAS requested the pdf files to be stored in a digital library. The purpose is to provide an access to the files from a computer. SaGa keeps a record of what happened in Sarawak since the reign of James Brooke until recent years. The bi-annual report was later changed into the annual report of each district. These reports contained useful information on trade and economic activities, law and order, and conditions of life in the various residencies or districts. Various articles on important events, ethnic relations, and general comments on the country, the landscape, and social life were also published in the Sarawak Gazette. IEAS rely heavily on SaGa to conduct their research into Sarawak’s history.

SaGa is used to obtain specific information the researchers are interested. For example, there was a court case on a native land dispute between a Penan tribe and the Sarawak government. The tribe claimed that a particular land belonged to their ancestors but there was no hard evidence to back their claim. IEAS helped by scouring SaGa, which was still stored in microfiche form and some in printed form. After a long year of laborious search, involving
several researchers, they found an article which stated that the tribe indeed lived on the claimed land. That was the evidence that the tribe needed to claim their ancestral land.

According to one of the researchers, Jayl Langub (personal communication, August 15, 2011), the biggest obstacle is lack of access in their research centre. Even with access to the documents, they had to browse through all the documents, page by page tediously, so that no information was overlooked. Searching for specific information such as the tribe’s name or the location’s name took a lot of time because of the lack of specific date as reference. During their research into SaGa, they also found out that in order to understand a major event, such as Asun rebellion, several reports were needed to piece together the story. These reports were difficult to search without again browsing through the whole collection. A method to help the researches to “connect” events together will greatly facilitate their research.

A portion of the documents had been scanned and their images published online through *E-Sarawak Gazette* (Pustaka Negeri Sarawak, 2013). The researchers still could not search for specific information by using the word search because words scanned, in the form of an image, cannot be identified without an indexing setup.

Based on the situation, three problems had been identified and listed as follows:

**Problem 1.** Inexistence of SaGa repository for content access

Although researchers can access documents by physical means, it is not efficient because of location differences. *E-Sarawak Gazette* may solve this problem but the contents are all scanned images. Searching for information from printed documents will still require the researchers to transfer the information in written or typed format. This is because the image files are for display of text as a picture, not machine-readable text. Any attempt to manipulate these texts by computer access cannot be done. This leads to the second problem.
Problem 2. Searching for specific information

As the text is currently in image form, the computer system only recognises it as a single file of image, nothing more. As such, the advantage of using a computer system for text mining such as searching by using a keyword is not possible. The researchers will still need to go through pages of documents painstakingly to obtain the specific information that they seek. This is a sheer waste of time and effort.

Problem 3. Gaining overview of the content and events related to each other

The current method of displaying historical documents is found in the table of content. For researchers that require an overview of the documents, the table of content is not a good source of reference. It only shows the topics and sub topics, which do not provide additional information. The table of content does not show patterns of events. History has shown us that a single event could be a part of a story containing a chain of events. The table of content does not determine the relation of events by certain entities.

1.2. Related Works

Numerous efforts had been undertaken to preserve historical documents, particularly in museums. Traditional library preservation may protect the material but can be very limited in terms of access.

Efforts to digitise historical documents by web technology are an ongoing effort. Existing works include The British Newspaper Archive\(^1\), NewspaperSG\(^2\), National Library of Australia\(^3\),

\(^1\)The British Newspaper Archive, http://www.britishnewspaperarchive.co.uk/
California Digital Newspaper Collection\(^4\), and PapersPast\(^5\). Most of the documents are scanned from microfiche into pdf, gif or similar graphic formats. Many of the graphic archives have been indexed into searchable text databases utilizing Optical Character Recognition (OCR) technology. To present the content of historical documents in a different view, numerous works have been done: for examples Europeana\(^6\), World Digital Library\(^7\), and Geotopia\(^8\). They visualise their content’s information by timeline display, map and graphs.

1.3. Research Questions

Based on the identified problems, this research focuses on answering the following questions:

**Question 1. How to make SaGa easily accessible to the researchers?**

SaGa can be made accessible to the researchers by implementation of a wiki system, called Wiki SaGa. Wiki SaGa contains digitised documents from the original scanned image sources. The original image as well as the digitised text are displayed side by side for reference. Furthermore, the wiki system resides in a server, which can be accessed by the researchers via Internet and thus the document can be accessed anywhere at any time. The main reason for choosing a wiki system is that it supports collaborative work. Any mistakes on digitising the text can be pointed out or even corrected by the readers.

\(^4\)California Digital Newspaper Collection, http://cdnc.ucr.edu/cgi-bin/cdnc


\(^6\)Europeana, http://www.europeana.eu/portal/


\(^8\)Geotopia, http://www.geotopia.fr/
Question 2. How can specific information be searched in SaGa?

Specific information such as a person’s name and a location can be easily searched after digitising and annotating the SaGa documents. Digitising these documents will convert text images into machine-readable texts, which can be processed by the computer system using search function. Named entities (NEs) that occur in each report will be extracted using Named Entity Recognition (NER) process. These NEs can be used by a search function for a timeline display, which will be discussed in the next question.

Question 3. How can the events of SaGa be visualised to support researchers’ effort?

The events of SaGa can be displayed by a timeline display. Timeline display has always been used as a form to show historical events in a chronological order (Grassie, 2012). By using NER, NEs can be extracted to serve as a description of an event as reported in an article. These NEs will be displayed in the timeline display to act as an overview of the event article. Furthermore, graphical representation of events in a timeline are much more interesting to researchers.

1.4. Research Objectives

The main objective of this study is to design a pipeline process which can display a visual timeline of selected content of SaGa as well as digitised SaGa documents for knowledge acquisition process.

The more specific objectives are as follows:

- To **design** a new pipeline process from the digitising stage of historical documents until the contents visualisation display.
To **identify** suitable methods for the pipeline process by comparing existing digitisation methods and named entity recognition approaches.

To **design and develop** the Wiki SaGa environment to contextualise digitised documents with regards to events in timeline and **enable** the visualisation of data according to users’ perspective.

To **propose** an evaluation method to properly access the performance of the environment.

To **evaluate** the need of this pipeline process by observing how the researchers can benefit from Wiki SaGa.

### 1.5. Research Methodology

The first step in this research was to study the situation. Researchers from IEAS had brought forward scanned images of SaGa in PDF form and stored in a cd. The initial request was to create a digital library to store all these pdf files. An interview had been arranged to meet up with IEAS researchers to discuss on how they are using SaGa. Based on their situation, a case study is conducted to identify the problems and propose a better environment to help the researchers in their research.

Current works that are done to archive historical documents and methods to visualise the content are studied. Based on the study, suitable methods are adopted and necessary adjustments to the method are proposed so that the new process can address SaGa’s problems. This is necessary because each set of the historical documents are different from the other. A pipeline process that is able to archive and visualise SaGa to help IEAS researchers is proposed.
To ensure the design is able to solve the researchers’ problem, an experimental environment was designed. Proper method to evaluate the environment was formulated so that similar works of this research can also adopt the evaluation method. Results of the evaluation are gathered to study the response of participants. Further works to enhance the design will be proposed. Detailed explanations of all these steps are provided in Chapter 3.

1.6. Research Scope

The following guidelines will be adhered to focus on the main objective of the research.

- The documents used will be historical documents, specifically SaGa.
- The digitised documents will consist of three months of SaGa publication, from January 1903 until March 1903.

The NEs to be extracted are person, location, organisation and date. This is because these are the most common NEs concerning reports of an event. This research focuses on proof of concept so the four NE are chosen.
SaGa

Study The Situation
- Interviews with IEAS staff
- Identify research problems
- Define research questions
- State research objectives

Study Current Work
- Newspaper portals
- Similar websites that visualize information other than text
- Related components
- Drawbacks if used on SaGa

Propose A Pipeline Framework Design
- Choosing component methods
- Adjust to apply on SaGa
- Digitization
- Named Entity Recognition and Extraction
- Wiki SaGa & Timeline

Prepare Experimental Design
- Develop environment to evaluate the pipeline framework
- Prototype: Wiki SaGa

Formulate Evaluation Method
- OCR Methods
- NER Methods
- Wiki SaGa Usability

Conclusion

Figure 1-1: Research Methodology
1.7. Thesis Organisation

The thesis is organised in five main chapters:

Chapter 1 serves as an introduction to the thesis: containing research motivation and problems, research questions, objectives, methodology and scope.

Chapter 2 covers the study done on current works and their limitations if used on SaGa.

Chapter 3 will illustrate how Wiki SaGa, an environment to evaluate the proposed pipeline system, is designed and implemented.

Chapter 4 contains the experimental results of three main components of Wiki Saga, which are digitisation using OCR, NER using ANNIE GATE, and displaying timeline using SIMILE Timeline.

Chapter 5 concludes this thesis and presents the main contributions, research limitations and future works.
Chapter 2. Literature Review

2.1. Current Work

In order to address the problems stated in Chapter 1, studies have been conducted to access similar preceding works that have been done. The objective is to review their methodology and determine whether their methods can be used for SaGa.

This section will discuss similar works that had been carried out to get a general idea of how news portal visualise their content. Studies were done on current work for archiving historical documents and their visualisation of content other than text and how historical documents can be made accessible online.

Historical documents record events of the past, often containing important information which leads the present scholars to have a glimpse of the past. However, with the massive amount of historical documents around, looking for a specific information is a big challenge. The usual method of browsing through page after page proves to be too much if the documents cover a span of several years. The digitisation of historical documents enables readers to search for texts by using search engine provided by the computer systems.

Even when the documents are fully digitised, sifting meaningful patterns through all these dataset takes a lot of effort as well. Researchers often found themselves exploring archives by using text search which results in texts returned. The results often do not mean anything unless some layers of visualisation are performed on the information. (Torget & Christensen, 2012) in The Mapping Text project uses timeline to visualise digitisation quality and important patterns of American historical newspapers published from 1829 to 2008.
Europeana Newspapers project (www.europeana-newspapers.eu) is one of the similar projects associated with this research. The aim is to enable over 10 million pages of digitised historical newspapers searchable. The project applies OCR and Optical Layout Recognition (OLR) to digitise the documents. NER (person, location and organisation) is also applied to materials in Dutch, German and French language to enhance the searchability of information. In addition to that, the Europeana Data Model (EDM) was created as a metadata standard (Neudecker, Wilms, Faber, & van Veen, 2014). The Europeana Project enables users to perform basic text search and refine their search by applying filters based on discipline, language, contributor and collection. However the project faces several difficulties, mainly on the results of OCR and NER process. This research will try to compare the results of different component methods in order to determine the method which is more appropriate. This project used Stanford NER, a machine learning-based system to extract the NEs because of the unavailability of domain experts. Test data are used to train the system.

Although most of these works involve digitisation of historical content, either by image or digital text, very few archive systems provide a visual representation of their content to offer different perspectives.

The subsequent sections will discuss the current works and tools used particularly in processing historical documents.

2.2. Historical Documents: Digitisation and Storage

Historical documents are documents that record events of the past. These documents are important to us to study what had happened in the past. The documents from the past are usually in a physical form such as papers, obelisks, turtle shells, bones and so on. In this particular research, SaGa is the focus as it is a historical document that records daily affairs since 1870 during the Brooke administration until present day. These documents are usually stored in