Machine Assisted Data Capturing Process

John Michael C. Mariquit
Bachelor of Science in Computer Science

Redan Benedict S. Alcaide
Bachelor of Science in Computer Science

Kier Sostenes N. Guevara
Bachelor of Science in Computer Science

Senior project submitted to the faculty of the
Department of Computer Science
College of Computer Studies, Ateneo de Naga University
in partial fulfillment of the requirements for their respective
Bachelor of Science degrees

Project Advisor: Allan A. Sioson, PhD
Joshua C. Martinez, MIT
Jenilyn L. Agapito
Rey Herman R. Vidallo, MCS

June 23, 2014
Naga City, Philippines

Keywords: web service, android

Copyright 2014, John Michael C. Mariquit, Redan Benedict S. Alcaide, and Kier Sostenes N. Guevara

The Senior Project entitled

Machine Assisted Data Capturing Process

developed by

John Michael C. Mariquit
Bachelor of Science in Computer Science

Redan Benedict S. Alcaide
Bachelor of Science in Computer Science

Kier Sostenes N. Guevara
Bachelor of Science in Computer Science

and submitted in partial fulfillment of the requirements of their respective Bachelor of Science degrees has been rigorously examined and recommended for approval and acceptance.

Joshua C. Martinez, MIT
Panel Member
Date signed: _____

Jenilyn L. Agapito
Panel Member
Date signed: _____

Rey Herman R. Vidallo, MCS
Panel Member
Date signed: _____

Allan A. Sioson, PhD
Project Advisor
Date signed: _____

The Senior Project entitled

Machine Assisted Data Capturing Process

developed by

John Michael C. Mariquit
Bachelor of Science in Computer Science

Redan Benedict S. Alcaide
Bachelor of Science in Computer Science

Kier Sostenes N. Guevara
Bachelor of Science in Computer Science

and submitted in partial fulfillment of the requirements of their respective Bachelor of Science degrees is hereby approved and accepted by the Department of Computer Science, College of Computer Studies, Ateneo de Naga University.

Jenilyn L. Agapito, MS
Chair, Department of Computer Science
Date signed: _____

Allan A. Sioson, PhD
Dean, College of Computer Studies
Date signed: _____

Declaration of Original Work

We declare that the Senior Project entitled

Machine Assisted Data Capturing Process

which we submitted to the faculty of the

Department of Computer Science, Ateneo de Naga University

is our own work. To the best of our knowledge, it does not contain materials published or written by another person, except where due citation and acknowledgement is made in our senior project documentation. The contributions of other people whom we worked with to complete this senior project are explicitly cited and acknowledged in our senior project documentation.

We also declare that the intellectual content of this senior project is the product of our own work. We conceptualized, designed, encoded, and debugged the source code of the core programs in our senior project. The source code of third party APIs and library functions used in my program are explicitly cited and acknowledged in our senior project documentation. Also duly acknowledged are the assistance of others in minor details of editing and reproduction of the documentation.

In our honor, we declare that we did not pass off as our own the work done by another person. We are the only persons who encoded the source code of our software. We understand that we may get a failing mark if the source code of our program is in fact the work of another person.

John Michael C. Mariquit
4 - Bachelor of Science in Computer Science

Redan Benedict S. Alcaide
4 - Bachelor of Science in Computer Science

Kier Sostenes N. Guevara
4 - Bachelor of Science in Computer Science

This declaration is witnessed by:

Allan A. Sioson, PhD
Project Advisor

Machine Assisted Data Capturing Process

by

John Michael C. Mariquit, Redan Benedict S. Alcaide, and Kier Sostenes N. Guevara

Project Advisor: Allan A. Sioson, PhD

Department of Computer Science

(ABSTRACT)

Our thesis focuses on the theoretical design and potential implementation of a data capturing system to be utilized on journal documents. The system shall identify pertinent information from the manuscript in question through parsing, and collate the identified data to specified fields pertaining to the required data for ranking by the pre-existing system of the Philippine Journal Citation Index Database, or PJCID. This is to be done in such a way as to avoid erroneous overlap, repetition or omission of data which shall be caused by widely varying formats of journals and citations.

The ranking system utilized by the PJCID has been fully documented and proved. The problem of the system, however, lies in the data acquisition and entry done for every new journal published. The data entry for the journal ranking system requires the users to manually encode every citation for every journal entry. The sheer amount of citations makes it tedious, time-consuming, and has the potentiality for errors and overlap.

Our thesis aims to develop a system that is able to extract journal author information from documents of journal bibliographies and citations by parsing, then collate and relate the gathered information in such a way as to remove erroneous entries of journal information due to usage of varied citation formats. The system input shall rely upon documents from the users, which will be then parsed to extract relevant information. The system output shall be a fully organized data of citations per entry, with accompanying data on authors, title and associated information. Any errors and its potential rate of generation found in the output of the system shall be studied in order to find the feasibility of the system.

I dedicate this research work to all of humanity.

ACKNOWLEDGEMENTS

I thank everyone who helped me finish this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Journals publications, their entries and the authors of these entries, serve as one of important foundations of modern academic studies, serving as a convenient source of new discoveries and citation sources. However, finding the best of or even the reputable authors among the mass of entries and journals can be difficult to any reader. This is the reason for the several journal citation index databases that has been implemented in many academic circles, a close example of which is the Philippine Journal Citation Index Database. Authors of studies, texts and documents published in the many academic journals around the world are ranked in these databases, giving users quick access to authors whose works have been referenced and trusted by many other authors to their own published papers.

1.1 Project Context

The Philippine Journal Citation Index Database, or PJCID, is a web-based citation index database funded by the Commission on Higher Education (CHED), Republic of the Philippines, in cooperation with Ateneo de Naga University, to track publications of Philippine Journals accredited by the Journal Accreditation Service (JAS) of CHED. The database currently has records of up to 44 Philippine-based journals, more than 1000 articles in total.[1]

The PJCID system records data on journal citations and all its associated information, like article authors and co-authors, article title, journals, publishers, along with year of publication and even kind of citations. The data collected will then be processed and resulting conclusions shall be

displayed on the PJCID website as reports pertaining to the author, publisher or journal in question. The PJCID only focuses on direct citations - articles citing earlier documents - and not bibliographic coupling - two or more articles sharing one or more references. [2]

Data gathering for the PJCID system is relies on the manual input of data from the PJCID administrative team. Identification of necessary information from the myriad citation formats and forms is done by the people involved, and they will be the ones to input the identified data into the database. Properly identifying the information, however, has proven to be rife with difficulties, like unclear nomenclature, synonyms, and publication volume, which has been recorded with a yearly increase of 3.7

In practice, searching articles for necessary information starts with title and author acquisition and continues with extraction of authors, titles, publishers and journals of entries of the reference section of the article. In reality, however, even advanced solutions for identifying related literature, like co-word analysis, collaborative filtering, Subject-Action-Object (SAO) structures or citation analysis do often not deliver satisfying results. [4]

This forces the current implementation of the PJCID data acquisition process to a manual approach. This makes data acquisition time-consuming and tedious to the people involved, especially on journal articles with substantial citations used. This thesis aims to create a machine-assisted process for the data acquisition of PJCID, specifically the utilization of an automated parsing system for the extraction of necessary information from journal articles.

1.2 Purpose and Description

The purpose of this thesis is the creation of a machine-assisted process for the data acquisition of PJCID. It shall focus on the parsing of journal articles derived from pdf format documents of several Philippine-based journals, deriving information necessary for report and data generation within the PJCID ranking system, like authors, titles, and citations used, along with information regarding said citations. This information shall be extracted from the input text by way of parsing, primarily using general citation formats commonly used by the many academic journals and certain keywords, an example of which is journal names, common among citations.

The thesis shall also study any errors that occur in the output of the machine-assisted process, and compare its rate to that of the existing manual data acquisition process of the PJCID. A

comparative study on the potential error rate for the machine-assisted process will serve as evidence for the potential feasibility of machine-assisted data acquisition process.

Implementation of the process in question shall be performed by way of a citation parser to be designed and implemented by the proponents. With the main difficulty of data acquisition from citations coming from the varied and uncommon nomenclature, particular focus for the design will be on coverage of as much potential variations of citation formats as possible. This includes, but is not limited to, lack of authors, shortened names, interchange of article title and journal name, lack of date, or lack of either journal or publisher name.

1.3 Objectives

The objective of this thesis is the development of a machine-assisted data acquisition process to be utilized specifically by the PJCID system. This end objective is further expounded by the following:

1.4 Scope and Limitations

Chapter 2

Review of Related Systems and Related Literature

Blah, blah, and blah.

Chapter 3

Technical Background

Blah, blah, and blah.

Chapter 4

Methodology

Methodology stuff will appear here.

4.1 Systems Analysis

4.2 Systems Design

4.3 Requirements Specification

4.4 Development and Testing

Chapter 5

Contributions and Recommendations

5.1 Summary of Contributions

Blah, blah, and blah.

5.2 Implementation Plan

Implementation plan in terms of Infrastructure and Deployment.

Appendix A

Code Listing

The following is a source code listing of programs developed in this research project.

Appendix B

Evaluation Tool

Evaluation tool used goes here.

Appendix C

Sample Input/Output

Describe and discuss the details of sample I/O here.

Appendix D

Sample Reports

Describe and discuss the details of sample reports here.

Appendix E

User's Guide

REFERENCES

[1] R. Albert, H. Jeong, and A. Barabasi, *Internet diameter of the world-wide web*, Nature, 401 (1999), pp. 130–131.

[2] G. Ausiello, G. F. Italiano, and U. Nanni, *Hypergraph traversal revisited: cost measures and dynamic algorithms*, in Proceedings of the Mathematical Foundations of Computer Science, 23rd International Symposium, August 1998.

[3] The Ateneo de Naga University website. `http://www.adnu.edu.ph/`.

[4] The SWI-Prolog website. `http://www.swi-prolog.org/`.

VITA

JR is BS Information Technology student of the Department of Computer Science at the Ateneo de Naga University.