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

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

1 Introduction **1**

1.1 Project Context 1

1.2 Purpose and Description 2

1.3 Objectives 3

1.4 Scope and Limitations 4

2 Review of Related Systems and Related Literature **5**

2.1 FreeCite and ParsCit 5

2.2 Identity Uncertainty 6

2.3 Data Catching and Citation Parsing 7

3 Technical Background **9**

4 Methodology **10**

4.1 Systems Analysis 10

4.2 Systems Design 10

4.3 Requirements Specification 10

4.4 Development and Testing 10

5 Contributions and Recommendations **11**

5.1 Summary of Contributions 11

5.2 Implementation Plan 11

A Code Listing **12**

B Evaluation Tool **13**

C Sample Input/Output **14**

D Sample Reports **15**

E User's Guide **16**

LIST OF FIGURES

2.1 Relational Probability Model 7

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.

The primary method used by the community on ranking journal articles is through the amount of citations it has received from other articles. Citations also serve as auxiliary support for informational purposes, like indexing and ranking [14], and quality assessment [11].

However, the flexibility of the citation system used by article authors give references of supposedly standardized citations massive potentiality for variation. 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 [5].

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. [19] 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. [12]

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. It is to be noted that the users themselves shall input the text taken from

the available PDF documentation of the articles into the system to be developed. 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 generated by 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:

- To study the various formats used by journal publications, both on articles and on citations
- To study and formulate possible variations of the aforementioned formats, and the particular formats and identifying characteristics of such variations;
- To create a front-end input system where users may input texts of information required taken from journal PDFs;
- To design and create a back-end parser that can extract the necessary data from articles, utilizing the plans and formats studied;
- To identify erroneous data from the parsed input, and remove said errors if possible;
- To categorize the parsed input to the information category used by the PJCID;
- To create an output file based on the above that shall be compatible to the PJCID system;

- To create a back-end system that shall upload the output into the PJCID system;
- To study any errors that could not be removed from the citation parser, and compare it to error rates of the existing acquisition process:

1.4 Scope and Limitations

The thesis shall only focus on the data acquisition process of the Philippine Journal Citation Index Database. Thusly, the parsing system to be designed and implemented by the proponents shall be designed with the PJCID system in mind only. The citation parser to be designed and implemented by the proponents shall only focus on the information necessary for ranking. Journals to be used as test inputs will be ones based within the Philippines. These journals shall be in PDF format, with text transfer to the system to be done by the users themselves. The comparison of error rates shall be between the designed citation parser and the existing manual data acquisition process.

The system designed shall in no way focus on the process of ranking and display of information to the PJCID website, nor will it perform image-processing functions for PDF journal data extraction. The system shall only take in textual input taken by the users from PDF documentation, and add the processed data into the PJCID database.

Chapter 2

Review of Related Systems and Related Literature

Machine-assisted parsing of journal citations encounters problems apparent in all textual retrieval tasks in mass applied media. The variations in official citation styles, the deviations within an official citation styles and same authors and journals written in different ways are but the first difficulties in the road for complete machine-assisted citation parsing.

However, there have been previous works on citation parsing, complete with online, public citation parsers with API. Much like this study, these parsers merely parse through the data inputted. They do not process the rank, nor perform image-processing data extraction. Two of the most widely used citation parsers found online are FreeCite and ParsCit.

2.1 FreeCite and ParsCit

FreeCite is an open-sourced application that parses through the citations of various documentations into fielded data. It can be freely used as an application or service, under MIT license. FreeCite was developed by the Brown University in partnership with Public Display, a Providence-based start-up company, in inspiration to ParsCit [3].

ParsCit primarily performs two tasks, the reference of string parsing, may it be citation parsing and citation extraction, and logical structure parsing of documents. Its architecture is that of a

supervised machine-learning system that utilizes Conditional Random Fields as its learning mechanism. [7]. Like FreeCite, the software is licensed in such a way (Lesser GNU Public License) that the public is free to utilize it in their own projects.

These parsers function to extract the data to make citation based ranking easier. In measuring citation impact, different measures may be employed although a naive way to do this is through its citation count [9]. Often times, the more citations a journal has, it can then be said that it is more "central" [20]. However, merely counting citations of a particular article is ineffective. In the identification of citations for proper parsing, certain conditions must be analyzed and accounted for.

2.2 Identity Uncertainty

Identity uncertainty is a prevalent and unavoidable problem in citation matching. This phenomenon occurs if identifiers are absent in token sequences which may result in ambiguous observations for certain objects. This problem is addressed by constructing Relational Probability Models (RPMs) which consist of attributes that a citation may contain. RPMs semantics assume that unique names exist for different papers (although they may be the same) unless proven otherwise. Uniqueness is decided by the mapping which has the least co-referring terms. The RPMs can then be expressed to an equivalent Bayesian network where attributes of the objects are constructed as nodes. If citation C1 and C2 correspond to the paper P1 and P2 where P1 and P2 are the same object, the models will share one set of the same attributes [16]. A Markov Chain Monte Carlo based algorithm is then used to match the citations. [18]

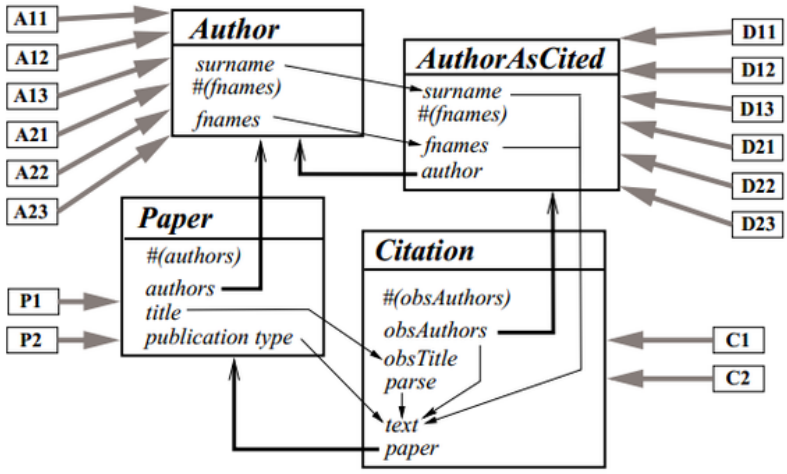

Figure 2.1: Relational Probability Model

The image above serves as an example of Citeseer. The large rectangles represent classes: the dark arrows indicate the ranges of their complex attributes, and the light arrows lay out all the probabilistic dependencies of their basic attributes. The small rectangles represent instances, linked to their classes with thick grey arrows. We omit the instance statements which set many of the complex attributes.

2.3 Data Catching and Citation Parsing

In order to extract the necessary information, a parser is required to sift through the text and separate the necessary from the unnecessary. The problem of citation extraction is the segmentation of oftentimes unstructured or ill-structured citations into proper segments, where data and its classification can be properly identified. Several methods have been proposed in the recent years for the extraction of data from textual documents. The methods can be generally defined into two categories: knowledge-based approach, and learning -based approach. [4]

The knowledge-based approach derives ontology that describes the data of interest using domain knowledge, where the knowledge includes relationships, lexical appearances and context keywords. From the parsing the ontology, several rules and an extractor can be generated, which is then used to extract the data. This particular method is more widely used in real-world applications, with a

well-known example on CiteSeer. CiteSeer is a popular search engine and digital library that extracts metadata from citations using heuristics, with an identification accuracy of titles and authors at 80

Other applications of knowledge-based approach are CRAM [2], FLUX-CiM [6], and INFOMAP [8]. CRAM develops an automatic segmentation system for its inputs by mining tables in relational databases and data warehouses. FLUX-CiM can automatically create ontology for a given area, constructed from an existing set of sample metadata records. The FLUX-SiM dataset, its own set of reference data to be used for testing and benchmarking, contains citations from two domains, Computer Science and Health Science. INFOMAP relies on a tree-based representation scheme that organizes reference concepts in a hierarchical manner. For the six major citation formats, it has an overall average of 92.39

The learning-based approach focuses on the classification of the citation data in question, using machine learning in order to solve it. This requires training data in order to function properly, slightly similar to the FLUX-CiM stated above. Currently, there are three major machine learning techniques, the Support Vector Machines [13], Hidden Markov Model [18], and Conditional Random Fields [17]. They are used to extract information from research papers and journal articles, and has shown great performance to the Cora dataset. The Cora dataset is the most widely used benchmarking dataset for machine learning techniques [1].

This particular method has great adaptability, mainly due to its nature as a machine-learning method. However, it does possess some limitations. The quality of training data directly affects the performance of the method [4]. A faulty set of training data can render the entire process useless.

This machine-assisted process that this thesis will focus on shall be under knowledge-based approach. This method allows for the use of keywords and lexicons, without relying on the creation of proper training data.

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] *Andrew mccallum's code and data.* `http://www.cs.umass.edu/~mccallum/code-data.html`, 2005.

[2] E. Agichtein and V. Ganti, *Mining reference tables for automatic text segmentation*, in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, New York, NY, USA, 2004, ACM, pp. 20–29.

[3] *Freecite citation parser.* `http://freecite.library.brown.edu/`. Home Page.

[4] C.-C. Chen, K.-H. Yang, C.-L. Chen, and J.-M. Ho, *Bibpro: A citation parser based on sequence alignment*, in IEEE Transactions on, vol. 24, 2012.

[5] *Philippine journal citation index database.* `http://pjcid.adnu.edu.ph/about.php`, 2009. About Page.

[6] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura, *Flux-cim: Flexible unsupervised extraction of citation metadata*, in Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07, New York, NY, USA, 2007, ACM, pp. 215–224.

[7] I. G. Councill, C. L. Giles, and M. yen Kan, *Parscit: An open-source crf reference string parsing package*, in INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION, European Language Resources Association, 2008.

[8] M.-Y. Day, R. T.-H. Tsai, C.-L. Sung, C.-C. Hsieh, C.-W. Lee, S.-H. Wu, K.-P. Wu, C.-S. Ong, and W.-L. Hsu, *Reference metadata extraction using a hierarchical knowledge representation framework*, Decis. Support Syst., 43 (2007), pp. 152–167.

[9] E. Garfield, *Citation frequency as a measure of research activity and performance*, Essays of an Information Scientist, 1 (1973), pp. 406–408.

[10] C. L. Giles, K. D. Bollacker, and S. Lawrence, *Citeseer: An automatic citation indexing system*, in Proceedings of the Third ACM Conference on Digital Libraries, DL '98, New York, NY, USA, 1998, ACM, pp. 89–98.

[11] M. A. Gonçalves, B. L. Moreira, E. A. Fox, and L. T. Watson, *"what is a good digital library?" - a quality model for digital libraries*, Inf. Process. Manage., 43 (2007), pp. 1416–1437.

[12] B. Gripp and J. ran Beel, *Citation proximity analysis (cpa) a new approach for identifying related work based on co-citation analysis*, in Proceedings of the 12th International Conference on Scientometrics and Informetrics, vol. 2, 2009, pp. 571–575.

[13] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, *Automatic document metadata extraction using support vector machines*, in Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '03, Washington, DC, USA, 2003, IEEE Computer Society, pp. 37–48.

[14] S. Lawrence, C. L. Giles, and K. Bollacker, *Digital libraries and autonomous citation indexing*, IEEE COMPUTER, 32 (1999), pp. 67–71.

[15] R. M. May, The Scientific Wealth of Nations, Science, 275 (1997), pp. 793–796.

[16] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser, *Identity uncertainty and citation matching. in advances in neural information processing systems*, in Advances in Neural Information Processing Systems 15, 2002, pp. 1401–1408.

[17] F. Peng and A. McCallum, *Accurate information extraction from research papers using conditional random fields*, in HLT-NAACL04, 2004, pp. 329–336.

[18] K. Seymore, A. Mccallum, and R. Rosenfeld, *Learning hidden markov model structure for information extraction*, in In AAAI 99 Workshop on Machine Learning for Information Extraction, 1999, pp. 37–42.

[19] H. Small, *Co-citation in the scientific literature: A new measure of the relationship between two documents*, Journal of the American Society for Information Science-August-July, (1973), pp. 265–269.

[20] J. D. West, *The Eigenfactor: ranking and mapping scientific knowledge*, PhD thesis, 2008.

VITA

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