

University of Mississippi

eGrove

---

Electronic Theses and Dissertations

Graduate School

---

1-1-2011

## Extraction of ontology and semantic web information from online business reports

Lakisha L. Simmons  
*University of Mississippi*

Follow this and additional works at: <https://egrove.olemiss.edu/etd>



Part of the [Management Information Systems Commons](#)

---

### Recommended Citation

Simmons, Lakisha L., "Extraction of ontology and semantic web information from online business reports" (2011). *Electronic Theses and Dissertations*. 1360.  
<https://egrove.olemiss.edu/etd/1360>

This Dissertation is brought to you for free and open access by the Graduate School at eGrove. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of eGrove. For more information, please contact [egrove@olemiss.edu](mailto:egrove@olemiss.edu).

EXTRACTION OF ONTOLOGY AND SEMANTIC WEB INFORMATION FROM ONLINE  
BUSINESS REPORTS

A Dissertation  
presented in partial fulfillment of requirements  
for the degree of Doctor of Philosophy  
in Business Administration  
The University of Mississippi

by

LAKISHA L. SIMMONS

May 2011

Copyright Lakisha L. Simmons 2011  
ALL RIGHTS RESERVED

## ABSTRACT

CAINES, Content Analysis and INformation Extraction System, employs an information extraction (IE) methodology to extract unstructured text from the Web. It can create an ontology and a Semantic Web. This research is different from traditional IE systems in that CAINES examines the syntactic and semantic relationships within unstructured text of online business reports. Using CAINES provides more relevant results than manual searching or standard keyword searching. Over most extraction systems, CAINES extensively uses information extraction from natural language, Key Words in Context (KWIC), and semantic analysis.

A total of 21 online business reports, averaging about 100 pages long, were used in this study. Based on financial expert opinions, extraction rules were created to extract information, an ontology, and a Semantic Web of data from financial reports. Using CAINES, one can extract information about *global and domestic market conditions*, *market condition impacts*, and information about the *business outlook*. A Semantic Web was created from Merrill Lynch reports, 107,533 rows of data, and displays information regarding *mergers, acquisitions, and business segment news between 2007 and 2009*. User testing of CAINES resulted in recall of 85.91%, precision of 87.16%, and an F-measure of 86.46%. Speed with CAINES was also greater than manually extracting information. Users agree that CAINES quickly and easily extracts unstructured information from financial reports on the EDGAR database.

**Keywords:** Semantic web; information extraction; ontology; EDGAR

## DEDICATION

This work is dedicated to my husband, Christopher B. Simmons.

Without his encouragement, I would not have had the courage to embark on this journey.

## ACKNOWLEDGMENTS

I am deeply grateful to Dr. Sumali Colon for her compassion as an advisor and mentor. Her strength and caliber as a researcher has given me much hope. I would also like to thank Dr. Milam Aiken for keeping me focused on prioritizing the many obligations of my Ph.D. program. I offer my sincere thanks to Dr. Tony Ammeter for stressing the need to conduct quality, theory-driven research. High regards to Dr. Walter Davis for his advice and support during my coursework and dissertation.

I would also like to thank my parents David Stewart and Tanya Thomas for whom I am grateful to have in my life; my God parents, Ms. Trena Taylor and Mr. Frank Carter for their enduring love and confidence. I extend precious thanks to my Aunt Chanine and Uncle Ricky for believing in me when I didn't. I thank my rock, my grandmother Margaret Johnson, my aunts and uncles, sisters, brothers, parents-in-law, church family, and my prayer group for all of their love and support. To my son Christopher Jacob, mommy thanks you for being patient as I finished my dissertation.

## TABLE OF CONTENTS

ABSTRACT.....	ii
DEDICATION.....	iii
ACKNOWLEDGMENTS.....	iv
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x
 1. INTRODUCTION _____	 10
1.1 The Study _____	10
1.2 Motivation & Research Questions _____	11
1.3 Goal and Contributions _____	12
1.4 Methodology _____	14
2. INFORMATION EXTRACTION _____	16
2.1 Seminal IE systems and Conferences _____	16
2.1.1 Message Understanding Conferences (MUC) _____	17
2.1.2 TIPSTER _____	18
2.1.3 Text REtrieval Conference (TREC) _____	18
2.1.4 SCISOR _____	19
2.2 Types of IE systems _____	20
2.2.1 Name Entity Systems _____	20
2.2.2 Co-reference Systems _____	20
2.2.3 Template Elements, Relations and Productions _____	21
2.3 Performance of IE systems _____	22
2.4 Acceptance of Systems: Technology Acceptance Model _____	24
2.5 Information Extractions Systems and the World Wide Web _____	25
2.5.1 KnowItAll _____	25
2.5.2 TextRunner _____	26
2.5.3 CAINES _____	27
2.5.4 FIRST _____	27
2.6 Approaches to IE Systems _____	28

2.7	Lexicon	29
3.	SEMANTIC RELATIONSHIPS	31
3.1	Overview	31
3.2	Semantic Web Theory	31
3.2.1	Semantic Web Issues	33
3.3	Ontology	33
3.4	Resource Description Framework (RDF)	34
3.4.1	RDFS and OWL for Web pages	37
3.5	Current Web Ontology Extraction Systems	37
3.5.1	Dynamic Semantic Engine (DySE)	37
3.5.2	RelExt	38
3.5.3	DB2OWL	39
3.5.4	TopBraid Composer	39
3.6	The EDGAR Database	39
3.7	Current Systems that Extract Financial Information from Edgar	40
3.7.1	Edgar2xml	41
3.7.2	EDGAR-Analyzer	41
3.7.3	Edgar Extraction System (EES)	42
4.	RESEARCH METHODOLOGY	43
4.1	Semantic Based Information Extraction from EDGAR using CAINES	43
4.1.1	Corpus Development	45
4.1.2	Perl Processing	46
4.1.3	IE and Semantic Web Rule Creation	47
4.1.4	N-gram processing	50
4.1.5	KWIC and SQL	53
4.1.6	Semantic and lexical analysis	56
4.1.7	Stemming	56
4.1.8	CAINES Extraction Interface	57
4.2	Financial Markets Ontology Extraction Methodology	58
4.2.1	Specification	59
4.2.2	Knowledge Acquisition	60
4.2.3	Conceptualization	61
4.2.4	Ontology Extraction Implementation	64
4.2.5	Ontology Evaluation and Documentation	67
5.	HYPOTHESIS DEVELOPMENT & TESTING	68
5.1	IE Performance	69



5.2	Technology acceptance measurement _____	70
5.3	Pilot Testing _____	71
5.4	Main Study Participants and Task _____	72
6.	DATA ANALYSIS AND RESULTS _____	79
6.1	Overview _____	79
6.2	Main Study Analysis and Results _____	79
6.2.1	Performance Results _____	80
6.2.2	Performance Hypotheses Results _____	86
6.2.3	Survey Measurement Analysis _____	88
6.3	Discussion _____	90
7.	CONCLUSION _____	94
7.1	Implications _____	94
7.2	Limitations and Future Work _____	96
	BIBLIOGRAPHY _____	96
	APPENDIX _____	11704
	VITA _____	117

## LIST OF TABLES

Table 1. Financial Report Extraction Rule Sets.....	49
Table 2. Relevant n-gram Output for Extraction Rules .....	52
Table 3. Example Data output from The KWIC Index System.....	55
Table 4. The Extracted Company Ontology .....	65
Table 5. Finance terms ontology A-B (see the Appendix II for the full ontology) .....	66
Table 6. Demographics of Respondents (n=44) .....	73
Table 7. Speed, Recall, Precision, and F-measure results .....	86
Table 8. Results of Paired T-Test .....	87
Table 9. Performance Hypotheses Results Summary .....	88
Table 10. Survey Response Summary .....	88
Table 11. Correlations between Constructs .....	90
Table 12. Satisfaction Hypotheses Results .....	88
Table 13. Summary of Results.....	884
Table 14. Comparison of CAINES to Similar Extraction Systems .....	885

## LIST OF FIGURES

Figure 1: The RDP triple conceptualized visually .....	35
Figure 2: Example RDF triple graph .....	36
Figure 3: System Architecture of CAINES .....	44
Figure 4: Example text of Bank of America 3-31-10 10-Q.....	54
Figure 5. CAINES Extraction Interface.....	57
Figure 6. Financial Markets Ontology: Specification Document .....	60
Figure 7. Financial Markets Ontology Diagram .....	64
Figure 8. Conceptual model of the impact of CAINES .....	70
Figure 9. Qualtrics Web Screen - Manual Extraction Instructions and Questions .....	74
Figure 10. Qualtrics Web Screen - CAINES Extraction Instructions and Questions .....	75
Figure 11. Qualtrics Web Screen – User Survey .....	77
Figure 12. Manual Data Collection Questions and Answers .....	81
Figure 13. Extraction Questions and Answers as output by CAINES.....	82
Figure 14. Information Extracted dealing with the “Business Outlook in 2008” .....	83
Figure 15. Semantic Web Questions and Answers as output by CAINES .....	84
Figure 16. CAINES screen of the Semantic Web .....	85

## CHAPTER 1

### INTRODUCTION

#### ***1.1 The Study***

Useful information drives decision making, business solutions, and even competitive advantages. The increased adoption and diffusion of the Internet allows dissemination of all types of information easy and widespread. For example, the United States Securities and Exchange Commission (SEC) requires major financial businesses in the United States to submit financial reports to their electronic data gathering, analysis, and retrieval (EDGAR) online database. The reports dispersed throughout the EDGAR database are very useful for businesses searching for data for benchmarking and recent competitor moves. However, many reports on EDGAR are over 100 pages long and keyword searches are prone to irrelevant results. The future of the Web, Web 3.0 or the Semantic Web, is one solution to this problem. The basic Semantic Web technologies have been defined and components of the architecture are becoming more standardized. However, there is still little practice oriented research in demonstrating how it truly enables the connections of people and computers (Hendler & Berners-Lee, 2010).

Content Analyzer and INformation Extraction System (CAINES) was developed to analyze content, lexicons, and relationships after performing extraction of text. This paper presents the idea of using an information extraction (IE) technique to extract unstructured data from online text documents and make the semantic relationships within the text available for computer search agents. Information extraction is automatic extraction of structured information

such as entities, relationships between entities and attributes describing entities from unstructured sources (Sarawagi, 2008). In this dissertation, the research focus is on how relevant information can be extracted from unstructured text. This dissertation extracts an ontology of descriptions and a Semantic Web of relationships using IE.

## ***1.2 Motivation & Research Questions***

There are two problems in the current Web. First, it is easy for humans to visit a Web page and understand the content, but computers cannot. A search engine scans and indexes pages for words that are represented many times on a page, but it cannot understand how those keywords are used in the context of the page, which leads to less accurate searches (Strickland, 2008). With the Semantic Web, computers will interpret information on Web pages through collections of information called ontologies that will be stored in the metadata code of Web pages. Second, Web personalization and browser profiles cannot reach their full potential in today's Web 2.0. Berners-Lee's original vision of the Web was that as humans search the Web, the browser learns what that person is interested in (more personalized than iGoogle). The more you use the Web, the more your browser learns about you and the less specific you'll need to be with your searches. CAINES can answer user questions such as *what is the business outlook for equity markets?* Or, *what acquisitions occurred?*

This dissertation plans to further the research in the development of a solution to the first problem: reaching more accurate search goals on the Semantic Web by extracting information, not just keywords. Studies have been able to demonstrate the basic extraction of unstructured and semi-structured data from Web pages (e.g. KnowItAll) and financial databases (e.g. FIRST). No known study has used IE to extract data from a semi-structured financial reports database to create a Semantic Web of data readable and searchable by computer agents. The purpose of this

study is to demonstrate that a Semantic Web can be developed using CAINES that will be useful for researchers and practitioners by having their computer agents easily search EDGAR for more exact search results. Those who will benefit most from this study are those who are interested in reviewing business information in SEC reports. The future of the Web depends on semantic relationships to increase search efficiency. This research asks three questions:

- (1) Can an IE system (CAINES) be designed to assist users in extracting semantic based information from online financial reports?
- (2) Can using CAINES result in performance greater than manually extracting Semantic Web information from financial reports?
- (3) Will users be satisfied with using CAINES as a semantic based search system?

Analysis of these research questions can further the development of the Semantic Web and semantic based information extraction.

### ***1.3 Goal and Contributions***

In this paper, IE is used to extract unstructured text semantically and to create a Semantic Web of data. CAINES also extracts ontology information which describes the financial reports. This dissertation uses the United States SEC EDGAR database to build our corpus since it houses required financial reports with written descriptions and explanations submitted by major businesses in the United States. Since each report is lengthy, around 100 pages long, there is an ample supply of unstructured text to build a Semantic Web. The EDGAR database is also a useful source for our study due to the amount of important information within each text report. But due to the length of the reports it can be time consuming to find and process large amounts of information for humans.

CAINES will contribute to practice and research in a number of ways over current systems. Current ontology extraction systems such as DySE and RelExt use limited extraction

techniques. Rinaldi's (2009) DySE uses a query structure formed by a list of terms (subject\_keyword) and a domain of interest (domain\_keyword). However, results show that when recall is low, for example, 15%, precision is high at 90% and when recall is high at 90%, precision is merely 40%. However, CAINES will use several techniques, including KWIC, n-gram processing, and semantic analysis to improve recall and precision. RelExt uses predicate-argument-pairs to construct semantic triples that are specifically used in a football "ticker" corpus (Schutz & Buitelaar, 2005). RelExt is concerned with the extraction of domain specific verbal relations rather than traditional is-a, has-a type ontology triples. Unlike DySE, RelExt is directed towards ontology extension, relying on an already existing ontology in football, in order to map concepts. Results showed that recall and precision were only 36% and 23.9% respectfully. Contrary to RelExt, CAINES uses the traditional triple as the standard set forth by the W3C. CAINES plans for more successful results because it uses natural language versus just short "ticker" based language.

FIRST is similar to CAINES in that both systems utilize IE but for different outcomes. CAINES is more advanced than FIRST in that it will create ontology out of unstructured text for search purposes. FIRST is for exchanging information across financial applications via XML. Furthermore, FIRST extracted short articles that were about half a page long that consisted of structured financial data and converted it into XML for use in business applications. For example, the extraction template used in FIRST for each article consisted of slots to be filled such as:

Company Name: NEXTEL

Financial item: earnings

Financial status: fell

Percent change: 20%

Over most systems, the advantage of CAINES is the extensive use of information extraction from natural language to build the Semantic Web of data. CAINES will deal with lengthy documents with a wide range of information and will extract more semantic facts than FIRST. CAINES will output semantic relationship phrases from the system to a user interface. This study makes several other contributions to IE research. It builds upon the use of IE techniques in the financial markets (e.g. Conlon, Hale, Lukose, & Strong, 2008) to advance more efficient and effective searches. This study makes extensive use of information extraction from natural language for training to analyze content in text. The RDF format: subject/predicate/object, semantics and synonyms of a sentence are important, and Patterns in the text are important to effectiveness of the extraction.

#### **1.4 Methodology**

The purpose of CAINES is to extract pure unstructured text (i.e. excluding tables) from quarterly financial statements on the SEC's EDGAR database and create a Semantic Web. The first step in developing CAINES is to create a corpus from 10-Q financial statements. The corpus consists of the *Management's Discussion and Analysis of Financial Condition and Results of Operations* section of 10-Q's. Knowledge engineering techniques are applied to the corpus to determine the format of information and to detect patterns in the text. The next step is to apply subsystems to the corpus for semantic extraction through a Web based user interface.

Chapter 2 begins with a discussion of Information Extraction, notable IE systems, and acceptance of systems. Chapter 3 of this dissertation discusses Semantic Web, technology acceptance and ontology theories. The semantic and ontological extraction methodologies are detailed in Chapter 4. The hypotheses and testing procedures are outlined in Chapter 5. Chapter



6 explains the results. The final chapter concludes with a summary of our research question results, a discussion of the limitations, and suggestions for future research.

## CHAPTER 2

### INFORMATION EXTRACTION

Information extraction (IE) is automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources such as text corpus or text documents (Sarawagi, 2008). IE, like text summarization, machine translation, and natural language interface, is a sub-field of natural language processing. This chapter defines IE, compares and contrasts types of IE systems, as well as discusses acceptance and performance of systems.

#### ***2.1 Seminal IE systems and Conferences***

IE is an extension of artificial intelligence (AI) and statistical natural language processing (NLP). IE is typically achieved by scanning a set of documents written in a natural language and populating a database with the extracted information. The origins of IE date back to the Cold War era of the 1960s. Several prominent IE systems were developed in the 1960s and 1970s. The ability to incorporate syntactic and semantic information in NLP techniques brought about new ideas in IE (Wilks, 1997). Naomi Sager and her colleagues first successfully applied IE technology to extract hospital discharge information from patient records in 1970. Dictionary lookup and pattern matching were used to extract relevant medical information in a limited context (Sager, et al. 1987).

In the early 1970s, Gerald DeJong developed an IE system named FRUMP. FRUMP used newswire articles to determine the relevant information using keywords and sentence analysis (Cowie & Lehnert, 1996). Even though it had weaknesses in performance, FRUMP became the basis of a later commercial IE system named ATRANS, the earliest IE system to be used for commercial purposes. ATRANS was able to extract bank money transfer information from

telexes by using sentence analysis as FRUMP. The assumption of ATRANS was that the sentence structures of money transferring telex messages are predictable (Lytinen, 1993). A system named JASPER (Journalist's Assistant for Preparing Earnings Report), developed by Carnegie Group of Reuters Ltd., was designed to extract corporate news stories related to earnings, dividends, or income. JASPER requires manual evaluation using a set of test documents collected from PR newswire, which is a drawback (Andersen et al., 1992). JASPER employs knowledge representation, syntactic and semantic knowledge of sentences and domain dependent regularities of patterns, with higher precision of extraction than the previously developed systems. Its performance has been used as a baseline by TIPSTER and MUC (Message Understanding Conferences) systems.

### **2.1.1 Message Understanding Conferences (MUC)**

Message Understanding Conferences (MUC), Text REtrieval Conferences (TREC), Information Summarization, Organization and Retrieval system (SCISOR), and the TIPISTER text program helped define and promote research in IE. MUC's were initiated and financed by the Defense Advanced Research Project Agency (DARPA). DARPA is the research and development arm of the U.S. Department of Defense and often supports research and technology (DARPA, 2004). The purpose of MUC was to encourage development of high performance IE systems through competition of research teams (Grishman & Sundheim, 1996). MUC organized 7 different conferences (MUC-1 in 1987 to MUC-7 in 1997). Prior to each conference, MUC invited different research teams to develop IE systems based on the same data set. In order to develop unbiased systems, MUC provided the test data set one month before conference.

MUC-1 (1987) and MUC-2 (1989), focused on extracting information from short naval messages. Many of the first systems that analyzed natural language text-based information came

from MUC-1 and MUC-2 (Jacobs & Ra, 1990). MUC-3 (1991) and MUC-4 (1992) centered on systems that extracted data about terrorists in Latin America from newspaper and newswire articles. The conferences continued in 1993, 1995 and 1997 (MUC-5, MUC-6, and MUC-7) using news articles to extract information about joint ventures, space vehicles, and missile launches (Appelt & Israel, 1999).

### **2.1.2 TIPSTER**

In 1990, TIPSTER was jointly sponsored by DARPA and the Central Intelligence Agency (CIA) and partially managed by the National Institute of Standards and Technology (NIST). The goal of the program was to improve document processing efficiency, document detection, information extraction, and summarization (NIST, 2001).

TIPSTER also produced major advancements in algorithms used in information retrieval (IR) systems. IR is sometimes called ‘search’ and is finding unstructured text or documents within a large collection of data or documents (Manning et al., 2008). While IE is a type of IR, IR is finding text in a collection of unstructured text. IE extracts structured information from unstructured sources and imports it into a database. Although IR and IE differ in their objective, they are complementary. IR is a first step in IE systems that extract information from Web documents (Adams, 2001). In summary, IR is document retrieval and IE is fact retrieval (Gaizauskas & Wilks, 1998). TIPSTER ended in 1998.

### **2.1.3 Text REtrieval Conference (TREC)**

Another very effective conference, TREC (Text REtrieval Conference), started in 1992 and is still held each year. It supports the data mining research community by providing a huge collection of English and non-English documents for developing and experimenting with large

scale IR and IE systems. TREC is co-sponsored by the United States Department of Defense and NIST, managed by members of government, academia, and industry to further promote IR and IE research. Their goal was to speed the use of IR and IE products for commercial use. In 1999, 16 countries were represented at the TREC-8 conference. Between 1992 and 1999, TREC research succeeded in doubling the efficiency of IR systems (NIST, 2000). TREC has had steady participation over the years. In 2003, 22 countries participated in TREC.

#### **2.1.4 SCISOR**

SCISOR (Information Summarization, Organization and Retrieval System) is designed to extract topics from news stories of various sources. SCISOR was designed by Paul S. Jacobs and Lisa F. Rau at the General Electric (GE) Artificial Intelligence Lab in the late 1980s (Jacobs & Rau, 1990b). It uses lexical analysis, bottom-up linguistic analysis, word-based text search, and top-down conceptual interpretation (Jacobs & Rau, 1990b). SCISOR is an IE system that performs text analysis in financial news articles. SCISOR is based on the GE NLToolset that uses two text-processes: stories about corporate mergers and acquisitions in real time as they come across newswires and then presents the extracted output in a template format. It is an interactive environment where the user can browse a topic of interest. If the user-defined topic arrives to the newswire, it is extracted. SCISOR can process about six stories per minute with precision and recall of 80-90%, accuracy at 90% in extracting correct stories and 80% accuracy in extracting correct values (Jacobs & Rau, 1990a; Jacobs & Rau, 1990c).

## **2.2 *Types of IE systems***

Cunningham's IE user guide (1999) outlined three major types of IE systems that have been researched by MUC: name entity recognition systems, co-reference resolutions, and template based systems.

### **2.2.1 Name Entity Systems**

A name entity system finds and classifies names, places, and organizations, performing at over 90% accuracy when compared to human extraction (Cunningham, 1999). Names are widely used for extraction in these systems, other entities, such as dates, times, numbers, and addresses, can also be incorporated (Chen, 2003). Due to the emphasis placed on name entity systems by MUC, they are the most common systems developed and the most widely studied.

One downfall of name entity systems is that there may be a large number of name entities (e.g. company names) of the same class that may cause confusion. Therefore, pattern matching based on linguistic knowledge, machine learning, and user relevance feedback are the most prominent approaches to extracting named entities (Chen, 2003).

### **2.2.2 Co-reference Systems**

Co-references are words in a text that refer to the same thing. For example in the way pronouns refer to a proper noun. Co-references identify relations between entities in texts. This method was introduced by MUC-6 in 1995 and usually performs at accuracy levels in the 60% to 70% range (Cunningham, 1999).

A team at The University of Massachusetts developed a co-reference resolution system that uses manually coded rules. The system depends on statistical probability to determine the co-reference of a noun-phrase (Lehnert, 1993). The system is considered conservative in that it

requires overwhelming evidence of support. This limitation has been addressed by McCarthy and Lehnert (1995) who developed RESOLVE. RESOLVE was developed to resolve co-reference issues. It uses orders and relative weights of different pieces of evidence found in the given documents. A decision tree is the underlying data structure of the system that has been used to determine whether a pair of phrases is co-referent or not.

### **2.2.3 Template Elements, Relations and Productions**

The remaining systems researched by MUC are template element construction, template relation construction, and scenario template production. Template element construction builds on the name entity and co-reference systems by associating descriptive information within the entities. A template consisting of several slots that are related to each other is seen as a representational device (Appelt & Israel, 1999). The users or clients determine the number and types of slots needed to construct a template based on their information needs. Template element is very sensitive to domain and topic changes and requires major changes in the design and template filling methods (Cunningham, 1999). Cunningham (1999) reported that matching template element construction is a very difficult task and the current reports on known system performance are around 50%.

To attempt to increase system performance, Sheikh (2009) used a Symbolic Learning Model (SLM) to extract individual slot-fills and the “Max-Strength” method to determine the exact template to be filled by determining the relevance of the extracted entities. Two different SLM models were built by using Greedy Search and Tabu Search methods. The “Max-Strength” method was applied on top of the results of the SLM model built by Tabu Search. The combined results of SLM model by Tabu Search and Max-Strength method showed significant

improvement over the baseline result obtained by using the SLM model built by Greedy Search. Recall performance averaged at 77.9, precision at 87.3, and f-measure at 82.2.

Template relation construction finds relations between template entities, similar to the entity relationship that is found in a relational database. Elements of relational template exhibit relationships among objects that exist in nature. EMPLOYEE\_OF, PRODUCT\_OF are examples of relationships that MUC-7 attempted to discover. EMPLOYEE\_OF relationship requires an organization object and an employee object where the employee object has to work for the organization object.

Scenario template production fits template entities and template relations into specific event scenarios. When the targeted data matches the instructions associated with the template, the data is extracted and displayed in template format. The template positions are then filled with the extracted data and are referred to as slots. Matching the data with the program instructions is the most difficult part of template mining and often performs at less than 50% accuracy (Cunningham, 1999). MUC-6 attempted to produce scenario templates in order to reduce development time of IE systems. Scenario template production is a difficult task and the overall precision level of the best MUC based systems was around 60% compared to human precision level of 80% (Grishman & Sundheim, 1996).

### ***2.3 Performance of IE systems***

Performance measures for IE systems were developed by MUC and refined with each conference. Precision and recall performance metrics were developed for MUC-3 and MUC-4 to set a standard performance measure for IR and IE systems. Precision measures what percentage of the information extracted are correct information, thus measuring reliability or accuracy. Precision is calculated by dividing the total number of correctly extracted items by the total



number of extracted items. Recall measures what percentage of the available correct information is extracted, thus measuring the ability of the system to extract relevant information (Adams, 2001). Recall is the number of correct answers produced divided by the total possible correct answers. For example, if system extracts 8 correct slot values and the total possible correct slot values is 10, the recall of that system is 80% (8/10).

$$\text{Precision} = \frac{\text{Total correct extracted}}{\text{Total extracted}}$$

$$\text{Recall} = \frac{\text{Total correct extracted}}{\text{Total possible to be correctly extracted}}$$

A combined weighted measure of precision and recall, the F-measure, was used at MUC-5 and the final TIPSTER evaluation. Per Manning and Schutse (2002), a higher F-measure indicates greater performance. An equal weight for precision (P) and recall (R) is commonly used along with the simplified formula of (Appelt & Israel, 1999):

$$F = \frac{2RP}{R + P}$$

By the mid 1990s TIPSTER and MUC systems showed average recall performance of 40%, with precision performance somewhat better at 50%. Some simple systems can reach performance levels in the 90% range (Cowie & Lehnert, 1996). Various sub tasks of IE had attained the performance level in the range of 60-80% till late 1990s (Appelt and Israel, 1998). Human analysis has shown that professional analysts earn performance in the range of 60% to 80% for overall IE on Tipster based systems (Cowie & Lehnert, 1996).

## **2.4 Acceptance of Systems: Technology Acceptance Model**

For years researchers have studied why users accept certain systems and reject others (Goodwin, 1987; Venkatesh & Davis, 1996). The technology acceptance model (TAM) is one of the most widely used behavioral prediction models in the information systems (IS) field (Davis, 1986; Davis, 1989; Davis, Bagozzi, & Warshaw, 1989). The TAM is an information systems model to predict a behavioral outcome, the adoption and use of an information system. TAM consists of two technology acceptance measures that model how and when a user will accept a new technology. The two antecedents, perceived usefulness (PU) and perceived ease of use (PEOU), are related to behavior intentions (BI) in the TAM model. PU was first operationalized to refer to a user's subjective probability that using a specific system will increase job performance, and PEOU refers to the perception degree that the system will be free of effort (Davis, 1989). TAM is an application of Theory of Reasoned Action (TRA), a behavioral model of prediction of behavioral intention, developed by Fishbein & Ajzen (1975).

Perceived usefulness is the perception that using a certain information system will improve the user's performance or productivity on a task. Adams et al. (1992) and Straub et al. (1997) report that user acceptance of an IT system is driven to a large extent by perceived usefulness. Goodwin (1987) argues that the effective functionality of a system, that is perceived usefulness, depends on its usability. The degree of perceived usefulness leads to a higher degree of behavioral intention and actual system usage (Lee et al., 2003).

Perceived ease of use is a form of Web site quality assessment by a user before any experience with the site. Perceived ease of use is essentially a user's individual perceptual regardless of many external factors. Studies have also found that the availability of training and support for the use of an IT artifact had no impact on perceptions of ease of use (Karahanna &

Straub, 1999). Gefen et al. (2003) found strong evidence that ease of use leads to higher levels of trust. Perceived ease of use is a well replicated construct and useful in assessing information systems. In summary, perceived usefulness and perceived ease of use may impact actual system usage.

## ***2.5 Information Extractions Systems and the World Wide Web***

The Web is an important means of transmitting and communicating privately and publicly and providing a wealth of information. As of February 2010, it is estimated that the Indexed Web (numbers of pages indexed by Google, Bing, Yahoo Search and Ask) contains at least 19.8 billion pages (de Kunder, 2010). There is no wonder there are recent developments to extract unstructured Web data in a format that is easy to use and search.

Extracting useful information from Web documents is an intricate task (Jain & Ipeirotis (2009). Web documents differ dramatically in their structure, format, and quality of information. One of the main challenges in IE is extracting structured data from unstructured Web documents (Chen, 2003). The following systems are applications of IE on the World Wide Web.

### **2.5.1 KnowItAll**

The KnowItAll Web IE system is known for learning to label its own training examples in what is considered a self-supervised method, using only a small set of domain-independent extraction patterns (Etzioni, 2008). KnowItAll was the first published system to extract data from Web pages that was unsupervised, domain-independent, and on a large-scale (Etzioni et al. 2004). The authors demonstrated IE from a targeted method, appropriate for finding instances of a particular relationship in text, to an open-ended method (Open IE) that scales to the entire Web and can support a broad range of unanticipated questions over arbitrary relations. Open IE

supports aggregating information across a large number of Web pages in order to provide comprehensive answers to questions rather than just a search string.

In an experiment, KnowItAll ran for four days on a single machine and 54,753 facts were extracted through its ontology and extraction rules (Etzioni et al. 2004). The rules were applied to Web pages identified via search-engine queries, and the resulting extractions were assigned a probability using information-theoretic measures derived from search engine hit counts. The researchers foresee opportunities to unify Open IE with information provided by ontologies such as Word-Net to improve the quality of extracted information and facilitate reasoning.

### **2.5.2 TextRunner**

Etzioni, Soderland, and Weld (2008) created an Open IE called TextRunner. TextRunner extracts information from sentences in a scalable and general manner. It learns the relations, classes, and entities from its corpus using its relation-independent extraction model. Specifically, it trains a graphical model called a conditional random field (CRF) to maximize the conditional probability of a finite set of labels, given a set of input observations. Using a CRF, the extractor learns to assign labels to each of the words in a sentence. TextRunner's extractor then scans sentences linearly and rapidly extracts one or more textual triples that aim to capture (some of) the relationships in each sentence.

TextRunner uses the triple as described in Semantic Web notation and consists of three strings, in which the first and third are meant to denote entities and the second to denote the relationship between them. TextRunner has run on a collection of over 500 million Web pages on its own and on over one billion Web pages in conjunction with Google (the system can be found at <http://www.cs.washington.edu/research/textrunner/>). The system works by having a user enter two arguments and a predicate, or a phrase in the form of a question, then answers are

returned. For example, “what has the FDA banned” returns answers such as ‘FDA banned ephedra (53)’ with the number in parenthesis corresponding to the number of Web pages found. It has extracted over 500 million tuples from Web pages. Results show the precision of the extraction process exceeds 75% on average.

### **2.5.3 CAINES**

CAINES (Content Analyzer and INformation Extraction System) was built using a knowledge engineering approach. It analyzes texts using syntactic and semantic techniques. Syntactic analysis identifies whether a word functions as a subject, verb, or object. The system can analyze word frequency, co-occurrence information, and extract unstructured text to display explicit information. The extracted information can be inserted to data warehouses, processed by data mining systems, and used in business intelligent tasks. CAINES was built using Perl and MySQL database management system.

CAINES can analyze documents from several online sources including web blogs, customer reviews, business and government reports, and online news articles. In 2009, CAINES was used to analyze eWOM reviews from a sample of 18 action and adventure movies consisting of 20,677 individual reviews (Simmons, et al., 2009). It analyzes texts using syntactic and semantic techniques. Results from the system showed storyline is most important and consumers tend to leave more positive than negative reviews. Findings also reveal key sentiments of consumers’ evaluations towards movies, something not found in many other studies.

### **2.5.4 FIRST**

Conlon, Hale, Lukose, and Strong (2008) created FIRST (Flexible Information extRaction SysTem) that is able to extract financial information from The Wall Street Journal (WSJ). Using

a training set of documents from the WSJ, FIRST builds a knowledge base. The FIRST system relies on a service-oriented framework with IR and IE components. The IR component retrieves source documents and the IE component analyzes the documents and converts news articles from The Wall Street Journal into a data template. It extracts specific information such as “sales rose 5%”. FIRST demonstrated that it can be employed to extract information from unstructured Web documents and translate it into extensible markup language (XML). Recall was 85%, precision 90%, and F-measure at 87%.

## **2.6 *Approaches to IE Systems***

Two general approaches, knowledge engineering and automated training, have been used to address various IE problems (Appelt & Israel, 1999). Both approaches rely heavily on the use of a domain specific corpus, a set of documents that is annotated and used to train the system. Annotations could include parts-of-speech and semantic tagging.

In knowledge engineering, a human expert in the domain and an expert in rule construction have to manually determine relevant extraction patterns. This approach is labor intensive and may take several iterations to produce a high performance system (Appelt & Israel, 1999). He or she also needs to have a clear understanding of the structure of the extracted information to satisfy user requirements. Knowledge engineering deals primarily with producing rules rather than training data. A major disadvantage is the dependence on the knowledge and skill of the engineer and the reliance on the test, re-test, and de-bug cycle. Although the automated system is catching up, the human expertise and intuition of the knowledge engineer have given the handcrafted approach an advantage thus far (Appelt & Israel, 1999).

Conversely, in the automated training or machine learning approach, a person needs enough knowledge about the domain to annotate the corpus of text used since a subset of

documents of the target domain is used to automatically train an extraction model by using a learning algorithm. The trained model learns extraction rules for extracting information from unseen documents of the same domain. Afterwards, the trained model is tested on the unseen test document set to determine the real performance of the model. Benefits include significantly less dependency on human labor, more scientific methods, and robustness of the developed systems to handle huge volume of data.

Systems developed by automated training require the training documents to be annotated and a number of quality training documents must be available. Unfortunately, performance is not as good as that of a system developed by the knowledge engineering approach (Appelt & Israel, 1999). Another drawback of this approach is that whenever the users specify new requirements, all of the training documents need to be annotated again. Knowledge engineering and machine learning both have benefits and drawbacks. It's up to the researcher to decide which method is more beneficial to the study at hand.

## **2.7    *Lexicon***

In English grammar, "patterns" are sentences, the "lexicon" consists of words, and the "constraints" are the rules of English syntax (Pentland, 1995). In information retrieval systems, a lexicon consists of all words recognized by the system, their grammatical categories, synonyms, and any associations with database objects and forms (Adam, Gangopadhyay, & Clifford, 1994). Natural language provides reliable and valid indicators of personality, cognitive processes, and social processes (Pennebaker & King, 1999). An array of words can be defined in terms of its lexical co-occurrence, a lexicon or semantic analysis (Bardi, Calogero, & Mullen, 2008). Therefore, for a construct, a lexicon of words indicative of a construct can be developed. Bardi, Calogero, and Mullen (2008) showed in their study where a value lexicon was developed on the

basis of the Schwartz (1992) value theory to extract lexical indicators of values (e.g. power, achievement) from texts (i.e. newspapers). Their basic process was to come up with three words for each value/construct to create a value lexicon then search 'pages' (Internet or newspaper) for all three words (more reliable than one word) to show evidence of that value/construct. Evidence of convergent and discriminant validity of their value lexicon was demonstrated by using American newspaper content from 1900 to 2000.

Lexicons have been created in management to expose some of the pitfalls and traps that plague post-modern managers (Cooksey, Gates, & Pollock, 1998). Cady and Hardalupas (1999) created a lexicon to attempt to identify a single term to identify the phenomena of major organizational change. They searched 2168 issues out of 15 publications and gathered an extensive list of terms. The terms were weighted to devise the top 5 mentioned terms. The terms appeared in 82% of the relevant articles.

Aggarwal, Vaidyanathan, and Venkatesh, (2009) proposed, a method based on lexical semantic analysis of discovering brand-descriptor associations in online search engines such as Google. By examining associations between brands and carefully selected adjectives/descriptors, one can go beyond merely counting text content to uncover the meaning of the content.



## CHAPTER 3

### SEMANTIC RELATIONSHIPS

#### **3.1 *Overview***

The current World Wide Web is fundamentally designed for human use. The Semantic Web aims to achieve a greater degree of communication, coordination and collaboration between computer systems for the benefit of people and organizations. The World Wide Web contains billions of documents that contain information. Search engines currently use keyword searches within these documents to find text that a user is looking for. The problem is that the document must then be read and interpreted by the human before any useful information can be deemed the right data and extrapolated. In other words, the computer can present the text but cannot understand it well enough to display the most relevant data in a given circumstance (What is the Semantic Web?, 2009).

#### **3.2 *Semantic Web Theory***

Berners-Lee et al. (2001) define the future Web as the Semantic Web, where information is given well-defined meanings, better enabling computers and people to work in cooperation. The Semantic Web provides a mechanism that is useful for formatting data into machine readable form, linking individual data properties to globally accessible schemas, matching local references to entities against various kinds of standard names, and providing a range of inferences over data in scalable ways (Hendler & Berners-Lee, 2010). In this dissertation, it is hypothesized that CAINES can extract semantic information more efficiently than humans.

The Semantic Web is an initiative of the World Wide Web Consortium (W3C) whose mission is to lead the Web to its full potential (W3C Mission, 2009). An example of a Semantic Web application is a service agent that automatically, given constraints and preferences, gives the user custom travel suggestions (Horrocks, 2008). For example the agent searches for the cheapest flight leaving at or around 10am on July 30, 2010 to Montego Bay and books it using the users stored American Express credit card number and personal information. Such a software agent would not simply exploit a predetermined set of sources but searches the Web for relevant information in much the same way a human user might without caring about size, color, and font of a webpage. The current Web uses keyword searches which can produce less accurate results. In other words, the computer can present the information but cannot understand it well enough to display the most relevant data in a given circumstance (What is the Semantic Web?, 2009).

That is where the Semantic Web can add value to the Web. Much of the world business news is published online in government reports, financial reports (e.g. EDGAR), and Web articles, but the data could be more useful if it was more accessible and understandable by computers. Basically, online sources contain a lot of information but in forms that computers cannot use directly and must be processed by people before the facts can be put into a database (Conlon et al., 2007). The Semantic Web is able to process data, such as dates, titles, and names, not just documents on the Web, independently of application, platform, or domain. Software programs called intelligent agents will be built to navigate the Semantic Web, searching not only for keywords or phrases, but also for concepts semantically encoded into Web documents (Berners-Lee, et al., 2001).

To make the Web of data a reality, two things are important according to W3C. First, the data that is currently on the Web must be made into a standard format that Semantic Web tools

can reach and manage and that intelligent agents can navigate. Second, the Semantic Web needs access to the relationships among the data, to create a Web of Data, not a collection of datasets. This collection of interrelated data on the Web can also be referred to as linked data. The current collection of Semantic Web technologies available from W3C provides query, inference drawing, and vocabulary creation functionalities. RDF is the common format and official W3C recommendation to design linked data and make it available for either conversion or on-the-fly access to existing databases (relational, XML, HTML, etc).

### **3.2.1 Semantic Web Issues**

Besides the issue of html tags and text information in a format unfriendly for semantic agents, integrating different ontologies may prove to be as difficult as integrating the resources they describe (Horrocks, 2008). There must be consistent standards for ontology designs. Emerging problems include how to create suitable and consistent annotations and ontologies. Progress is being made in the infrastructure needed to support the semantic Web, particularly in the development of languages and tools for content annotation and the design and deployment of ontologies which will be discussed next.

### **3.3 *Ontology***

Ontologies are the foundation of the Semantic Web, enabling automated tasks such as searching, merging, sharing, maintaining, and customizing (Corby, et al. 2006). The term “ontology” has a long-standing tradition in philosophy, meaning “the study of being or existence” (Cahn, 2002). In computer science, ontologies are a means to share and reuse knowledge. Neches (1991) notes that an ontology not only comprises the vocabulary of a topic area, but also the rules for combining terms and relations to define extensions to the vocabulary,

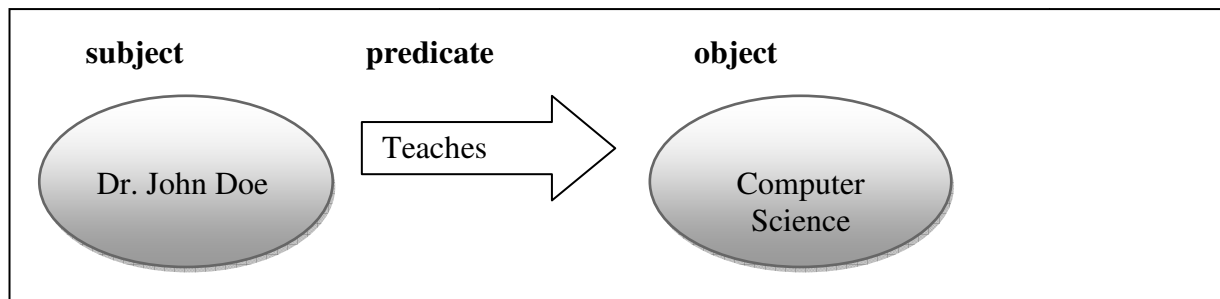
meaning an ontology can include terms that can be derived using the defined rules and properties. They are central to the development of knowledge-based systems and for modeling some domain of the world in terms of labeled concepts, attributes, relationships and classifications. Ontologies consist of vocabulary, rules for combining terms, and are reusable structures.

In the Semantic Web context, ontologies are formalized using Web-suitable semantically unambiguous representation languages and are meant to be shared and reused across the Web. Ontology reuse can be defined as the process in which existing ontological knowledge is used as input to generate new ontologies (Frakes, 1996). Ontologies are intended to provide knowledge engineers with reusable pieces of declarative knowledge for problem-solving methods and reasoning services (Neches, 1991). The ability to efficiently and effectively perform reuse is acknowledged as a crucial role in the reality of the Semantic Web. The sharing and reuse of ontologies increases the quality of the applications using them by becoming interoperable and a deeper, machine-processable understanding of the domain of interest (Frakes, 1996). Even though ontologies are expected to play a significant role in many application domains on the emerging Semantic Web (Staab & Studer, 2004), emerging ontologies seldom reflect a consensual or application-independent view of the modeled domain. Without any claim of being formal, application-independent or built to be shared or reused the reality of the Semantic Web will be continue to be an unreachable goal (Simperl, 2009).

### ***3.4 Resource Description Framework (RDF)***

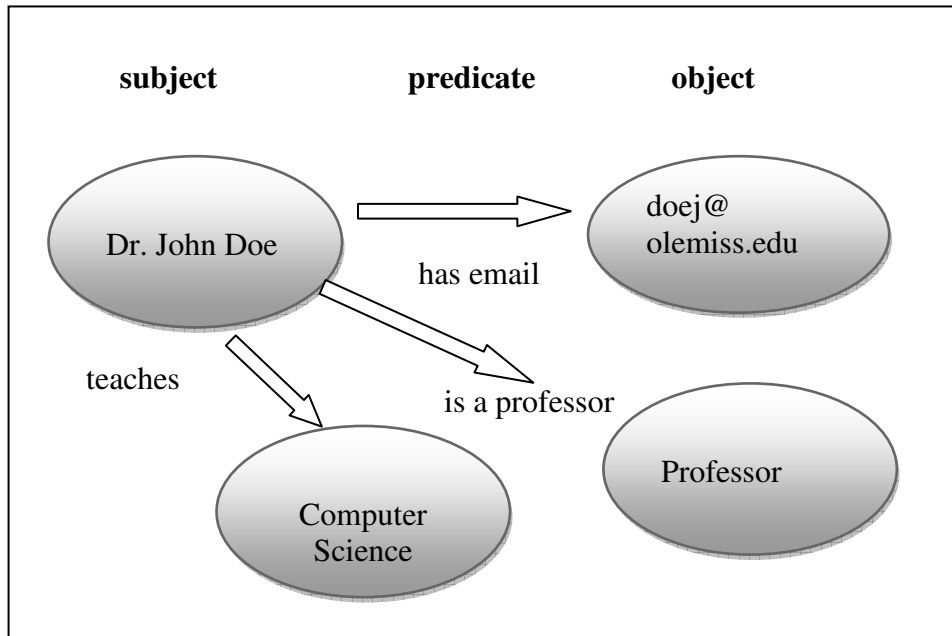
RDF is a simple ontology language with its data structure in the form of a labeled directed graph, and its only syntactic construct is the triple, which consists of three components: subject, predicate, and object (Horrocks, 2008). RDF is an XML-based standard that builds on

URI (Uniform Resource Identifier) for describing resources that exist on the Web, intranets, and extranets. RDF statements are usually first conceptualized graphically before being coded for programmatic use in XML tags using an ontology language. RDF statements contain three components: The resource (which can be identified by a URI), the resource's properties, and the values of those properties. RDF statements are often referred to as "triples" that consist of a resource (subject) a property (predicate), and a property value (object) (Figure1).



**Figure 1:** The RDP triple conceptualized visually

After creating this initial triple, additional triples can be created to associate the resources with additional objects. A set of triples is called an RDF graph (see Figure 2).



**Figure 2: Example RDF triple graph**

Triples with subjects, predicates, and objects, allow computer applications to make logical assertions based on the associations between subjects and objects. With the use of URI's each resource is tied to a unique definition available on the Web. However, while RDF provides a model and syntax for describing resources, it does not specify the semantics (the meaning) of the resources. The RDF vocabulary description language (RDF schema or RDFS) extends RDF to include the basic features needed to define ontologies. The next step is to define semantics through RDF schema (RDFS) and Web Ontology Language (OWL). RDFS and OWL are two of the technologies used for defining semantics.

### **3.4.1 RDFS and OWL for Web pages**

The RDF vocabulary description language (RDF schema) extends the RDF to include the features needed to define ontologies. This extension allows for more resources such as `rdfs:Class`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, and `rdfs:domain`, where `rdfs` is an abbreviation for the URI `www. w3.org/2000/01/rdf-schema#`. An ontology language standard is a prerequisite for the development of the Semantic Web. In 2004, the OWL ontology language became the standard set forth by W3C. OWL is to be used when the information in documents needs to be processed by applications, not just presented to humans. According to W3C, OWL goes beyond XML, RDF, and RDFS in its ability to represent machine interpretable content on the Web. This dissertation will present the base RDF style framework.

## **3.5 *Current Web Ontology Extraction Systems***

This section describes systems that perform extraction of ontology information and a modeling and development tool.

### **3.5.1 Dynamic Semantic Engine (DySE)**

Rinaldi (2009) proposed a system for information retrieval based on ontologies with the goal of designing a system capable of retrieving and ranking results, taking into account the semantics of the pages. Dynamic Semantic Engine (DySE) is composed of a dynamic semantic network and lexical chains that retrieve results based on scoring and ranking results by semantic relatedness (perceived relations between words and concepts) between words. DySE implements a context-driven approach in which the keywords are processed in the context of the information in which they are retrieved. Rinaldi's DySE (2009) uses a query structure formed by a list of terms to retrieve (`subject_keyword`) and a domain of interest (`domain_keyword`). For example,

if a user wants to get information about the Miles Davis, the musician the subject keywords:=Davis, and domain keyword:=music. DySE relies on WordNet (Miller, 1995) which is a knowledge base organized from a linguistic point of view and one of the most used linguistic resources in the research community. However, results show that when recall is low for example, 15%, precision is high at 90% and when recall is high at 90% precision is merely only 40%. Similarly to DySE, Baziz et al. (2005) described the use of ontologies for information retrieval by identifying important concepts in documents using co-occurrence and semantic relatedness, and then disambiguating them via WordNet.

### **3.5.2 RelExt**

A named entity recognition system RelExt, is capable of identifying relevant triples over concepts by extracting relevant verbs and their grammatical arguments (i.e. terms) from a German football ticker text collection (Schutz & Buitelaar, 2005). The authors say benefits of this kind of a ‘ticker symbol’ corpus compared to the more detailed text is that the sentences are concise, which significantly reduces the error rate of grammatical function assignment. RelExt exercises computing by corresponding relations through a combination of linguistic and statistical processing. RelExt uses dependency structure, grammatical function assignment, phrase structure, part-of-speech, statistical relevance ranking, and cross-referencing relevant nouns and verbs with the predicate-argument-pairs, in order to construct triples that are specifically used in the football domain.

RelExt contrasts to the majority of work carried out in ontology learning. RelExt is concerned with the extraction of domain specific verbal relations rather than traditional is-a, has-a type ontology triples. It is directed towards ontology extension, relying on an already existing ontology in football, in order to map nouns to concepts. Triples were produced and 3 annotators



were chosen to decide if the triples were appropriate for the football domain. There was only a 27% agreement rate. Recall was only 36% and precision was 23.9%.

### **3.5.3 DB2OWL**

DB2OWL was developed to generate an ontology from a database schema (Ceci et al., 2007). It starts by mapping the tables to concepts and then mapping the columns to properties. Each database component (table, column, constraint) is then converted to a corresponding ontology component (class, property, relation). The set of correspondences between database components and ontology components is conserved as the mapping result to., be used later. The mapping process starts by detecting particular cases for elements in the database and then converts database components to corresponding ontology components.

### **3.5.4 TopBraid Composer**

TopBraid Composer is a commercially available modeling and development tool for Semantic Web standards such as RDFS, OWL, and SPARQL (query language). Composer is implemented as an eclipse format plug-in and can be used to develop ontology models, test them, and customize dynamic forms and reports. TopBraid seems appropriate from ground up development, not beginning with actual data (an extractor is not part of the tool).

## **3.6 *The EDGAR Database***

Congress established the Securities and Exchange Commission in 1934 after the stock market crash of 1929 caused great losses and confidence in the market to bottom out (U.S. Securities and Exchange Commission, 2010). During the peak year of the Depression, congress passed the Securities Act of 1933 then the Securities Exchange Act of 1934, which created the SEC. The SEC was designed to restore investor confidence in capital markets by establishing rules,

monitoring, and making information available on companies. Companies that offer public stock are required to register with the SEC. The Securities Acts also require full financial disclosure and periodic financial information to be filed.

As of 1996, all public domestic and foreign companies were required to file their financial statements and other required SEC forms electronically using the electronic format via the Web based EDGAR system. EDGAR's main purpose is to increase efficiency of the receipt, dissemination, and analysis of corporate information filed with the SEC.

The EDGAR database is one of the richest and timeliest sources of financial information available on the Web. The SEC allows companies to submit and access filings that include documents in Hypertext Markup Language (HTML), Portable Document Format (PDF), and eXtensible Business Reporting Language (XBRL). These technology breakthroughs in business reporting have enhanced financial information on the EDGAR system. Therefore, EDGAR is attractive to show how a Semantic Web can become a reality.

Although the SEC requires various types of financial information, the business annual report (10-K) and quarterly reports (10-Q) includes the most comprehensive collection of financial information. These reports include sections for executive management to discuss their financial condition in their own words. This study concentrates on analyzing what managers are reporting in their *Management's Discussion and Analysis of Financial Condition and Results of Operations* (10-Q).

### ***3.7 Current Systems that Extract Financial Information from Edgar***

This section describes systems that perform information extraction from financial statements on the EDGAR database. The first two use IE and NLP techniques in their processes.

The third, EES builds on these two extraction systems and expands on the development of financial information extraction from documents on the EDGAR database.

### **3.7.1 Edgar2xml**

Leinemann et al. (2001) introduced Edgar2xml, a software agent to extract company balance sheets from Edgar, which were in ASCII text and transformed them into XML. The result shows that it is possible to extract specific financial information and transform it into XML format. This system was designed for use on the company balance sheet, specifically the quantitative portions of the balance sheet for investors so they do not have to read the entire balance sheet. Edgar2xml fragments and extracts a balance sheet into components. This system uses an input buffer of text files and parses using regular expressions for keyword identification. Keywords are then detected by a Document Object Model element listener and written to an XML output. The authors encourage further research is necessary in order to maximize the usefulness of EDGAR's financial information.

### **3.7.2 EDGAR-Analyzer**

John Gerdes (2003) developed EDGAR-Analyzer, a tool that automatically analyzes EDGAR filings. The system was used to analyze the unstructured text sections of corporate Year 2000 (Y2K) disclosures in 18,595 10-K filings from 1997 to 1999. EDGAR-Analyzer uses index files on the EDGAR web site to identify specific files. The files are then downloaded and a key-word search is used to extract the paragraph that contains the specified information. The information is further processed to extract blocks of data pertinent to the user search. EDGAR-Analyzer reduced the amount of text to be manually processed by 96% by automatically extracting an average text block of 11.1 KB from each filing.

### **3.7.3 Edgar Extraction System (EES)**

In 2006 emerged the EDGAR Extraction System (EES) that extracts information about stock options from the disclosure notes of 10-K annual reports on the EDGAR database (Grant & Conlon, 2006). The system displays stock option information in a useful structured format. The NASDAQ-100 Index companies were used as a sample for building a non-annotated, domain specific corpus for testing the system. Machine learning techniques combined with a knowledge based approach were used to analyze the patterns in the corpus. Algorithms were then designed and incorporated into a wrapper. The EES wrapper extracts pro-forma information about net income and earnings per share, as well as the fair value of the options and the assumptions and model used to calculate the fair value.

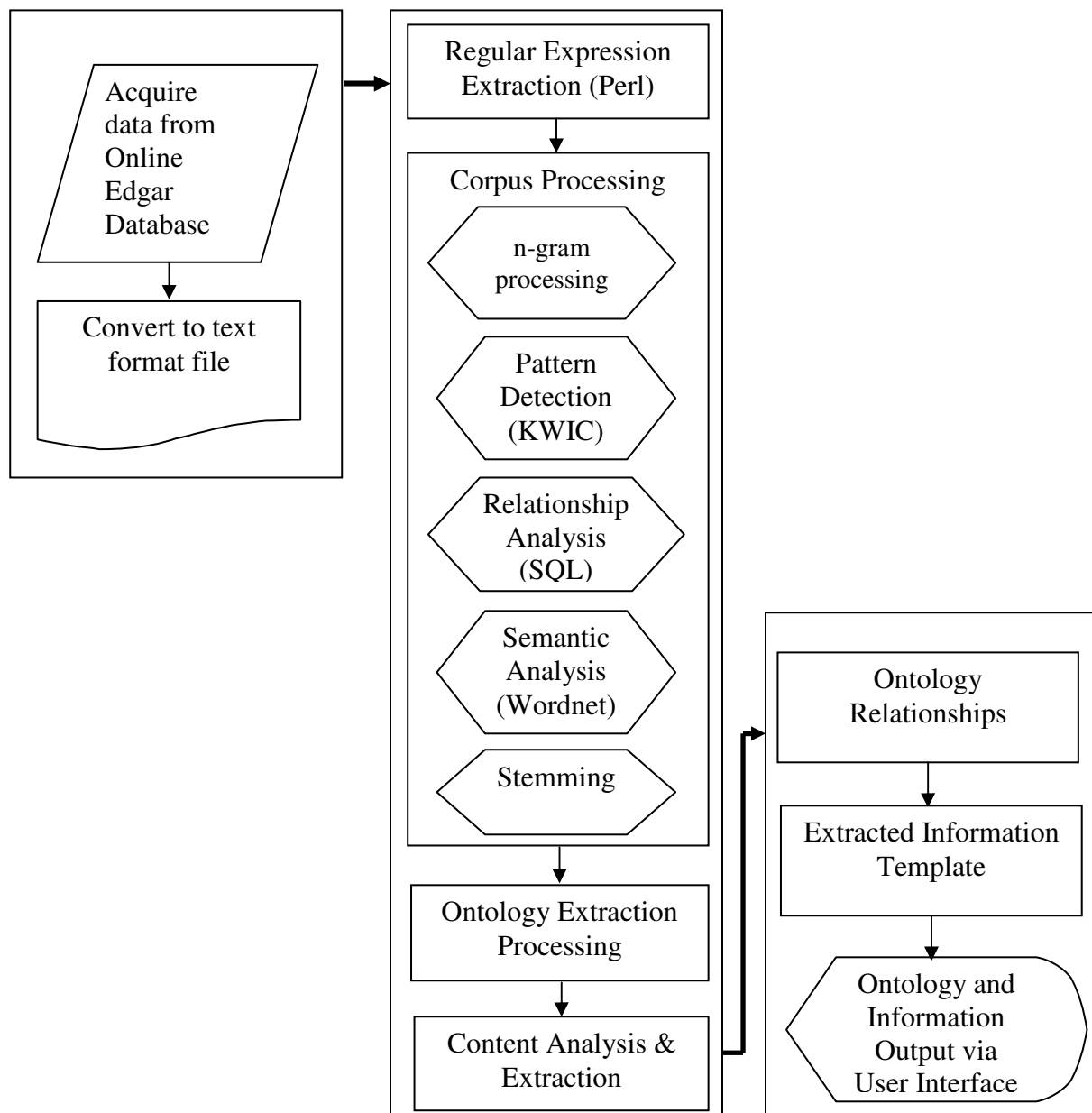
Different from other systems built to extract information from EDGAR, EES was tested and compared to human extraction of the same information. With overall recall, precision, and F-measure at 82.71%, 72.62%, and 77.34%, respectively, EES allows users to quickly and easily analyze and compare financial statements of companies that use stock options as a means of compensating employees. When compared to human extraction, there was no significant difference in EES recall ( $p = .9970$ ), precision ( $p = .7454$ ), and F-measure ( $p = .7368$ ). However, there was strong evidence that the speed of EES was significantly faster than human extraction ( $p < 0.001$ ). EES provides evidence of the usefulness and accuracy of an automated approach to extract specific financial information (stock options from disclosure notes) from files on the SEC's EDGAR Database.

## CHAPTER 4

### RESEARCH METHODOLOGY

#### ***4.1 Semantic Based Information Extraction from EDGAR using CAINES***

The purpose of CAINES in this dissertation is to extract information from unstructured documents, extract a Semantic Web of activities of companies, and extract an ontology of descriptions from financial reports. The first step in developing CAINES for information extraction was to create a corpus from the 10-Q files of the financial statements. The corpus was inserted into a MySQL database using Perl regular expressions. Knowledge engineering techniques were applied to the corpus to determine the format of information and to detect patterns within the text. Various tools and techniques were used to analyze the patterns in the corpus. These include n-gram processing, Key Word In Context Index System (KWIC), Structured Query Language (SQL), WordNet (as a guide), stemming, and knowledge engineering (for rule based extraction). A Web interface was designed to extract relationships and display the information. Figure 3 shows the overall structure of CAINES.



**Figure 3: System Architecture of CAINES**

#### 4.1.1 Corpus Development

The use of corpora and knowledge engineering techniques to develop a domain-specific knowledge base has become increasingly popular since the 1990s (Chen, 2003). The corpus was first developed by selecting the sample of companies to be used in the study. The second stage prepared the text documents used in the corpus for analysis. To build the corpus the downloaded 10-Qs for the companies were converted to text format. The files containing the *Management's Discussion and Analysis of Financial Condition and Results of Operations* text portion of the 10-Q were used as the basis of the corpus.

Since the US financial industry has bottomed out in 2007, it was interesting to analyze this industry to get their take on the situation through their quarterly reports. According to the *Management's Discussion and Analysis* of the U.S. SEC 2008 Performance and Accountability Report, the subprime mortgage crisis led to dramatic changes in U.S. financial markets (U.S. Securities and Exchange Commission, 2008a). The deterioration of mortgage origination standards and the rise of abusive lending practices were a real issue in 2008 causing the Enforcement Division to sweep financial institutions. The division reached the largest settlements in the SEC's history—over \$50 billion—on behalf of investors in auction rate securities (ARS) from Merrill Lynch, Wachovia, UBS, Citigroup, Bank of America and RBC Capital Markets.

In their 2009 Performance and Accountability Report, the SEC reported that the confidence of American investors was shaken by a deep financial crisis and a deterioration of the world economy. The SEC pursued significant cases of misconduct during fiscal year 2009, most notably, the SEC charged the former CEO of Countrywide Financial and two other former executives with fraud for allegedly misleading investors about the significant risks the company

was undertaking (U.S. Securities and Exchange Commission, 2008b). The SEC charged that Countrywide portrayed itself as underwriting mainly prime quality mortgages, while privately describing as “toxic” certain loans it was extending.

The crisis began to affect the financial sector in February 2007, when the world's largest (2008) bank HSBC, wrote down its holdings of subprime-related mortgage lending business with a \$10.5 billion loss, the first major subprime related loss to be reported (BBC News, 2008).

Thus the sample for this study consists of U.S. companies that were outlined in the SEC's 2008 and 2009 reports or news reports as affecting the overall US condition in some way. For the *information extraction*, six 10-Q Merrill Lynch reports were used for *training* to learn the patterns and create the rules: first and second quarter of 2007, third quarter of 2008, and first, second, and third quarter of 2009. *Information extraction testing* of CAINES was conducted using the following three files: third quarter of 2007, first and second quarter of 2008. The *Semantic Web training* corpus was comprised of Bank of America reports between 2007 and 2009. *Semantic Web Testing* occurred with Merrill Lynch 2007 to 2009 10-Q reports.

#### **4.1.2 Perl Processing**

Each company's 10-Q files were stored in a company sub-directory by year for Perl processing. Perl, a free text processing language with numerous modules, is a popular program for system designers and web developers. It is dependable at creating, managing, and extracting information from the Web using its LWP and the fact that it supports regular expressions. With a few lines of code, regular expressions assist in isolating specific passages, finding and replacing text, and performing data manipulation. Perl can extract and print text for patterns that are matched. Perl is stable open source code that can operate on Unix, Windows and Macintosh platforms (Burkes, 2002).



#### 4.1.3 IE and Semantic Web Rule Creation

In line with the knowledge engineering approach, three major sets of rules were created to extract semantic based information and one additional rule was created to extract the Semantic Web. The rules were first created in a pseudo code fashion and are based on three major categories of information that experts would be interested in. The pseudo code was then programmed into CAINES to extract many specific types of information could be extracted for each major rule. In order to ensure CAINES extracts useful information, financial experts who regularly analyze financial reports provided their opinions on what information the rules should extract. The experts consisted of an Area Manager for SunTrust Bank, a Senior Audit Officer of a Tennessee based bank, and a Senior Cost Accountant for PepsiCo Beverages and Foods (See Appendix I for the biographies of the financial experts). The three experts were shown a portion of the *Management's Discussion and Analysis of Financial Condition and Results of Operations* of a Bank of America 10-Q that was filed on 9/30/09. They were then interviewed about what they think is important in general and were asked to point out specific examples from the report shown to them.

Results of the expert interviews revealed that the experts were interested in three major categories of information (1) understanding what current market conditions were impacting the growth of balance sheets (2) management's discussion of potential risks or uncertainties moving forward (3) management's discussion of significant financial activities over the past few years and going forward. The three categories made up the three sets of information extraction rules, and one additional rule was created to extract activity information. For example, rule three regarding significant financial activities over the past few years and going forward would produce specific rules such as overall business outlook, demand forecasting, or other forward

looking information. Rule 4, the Semantic Web Rule extracts information about mergers, acquisitions, and business segment news. See Table 1 for the four sets of extraction rules.

**Table 1. Financial Report Extraction Rule Sets**

	<b>IE and Semantic Web Major Rule Sets</b>
<b>1. A rule to identify market conditions (globally and domestic)</b>	<p><i>for</i> each row in the portion of the report from which we want to extract information</p> <p><i>if</i> the key word is a candidate denoting economic and market conditions (e.g., markets, economic conditions, market conditions, credit environment, indices, credit spreads, oil prices, commodity prices)</p> <p><i>then</i> return the keyword noun phrase</p> <p><i>and</i> return verb phrase immediately following the keyword and present them as the State of US or Global market (e.g. verb phrases-continued to, increased, slowed, declined,)</p> <p><i>if</i> keyword is a verb phrase denoting cause (e.g. driven by)</p> <p><i>then</i> return the verb phrase and following noun phrases as reasons</p> <p><i>end if</i></p> <p><i>end for</i></p>
<b>2. A rule to identify specific economic and market conditions</b>	<p><b>impacting growth and earnings to specific business assets</b></p> <p><i>for</i> each row in the portion of the report from which we want to extract information</p> <p><i>if</i> the key terms are a verb phrase denoting impact (e.g. lower revenues , adversely impacted, resulted in)</p> <p><i>then</i> return the noun phrase appearing prior to the key verb phrase as the market condition</p> <p><i>and</i> return the noun phrases after the key verb phrase and present them as the business segment or asset affected</p> <p><i>end if</i></p> <p><i>end for</i></p>
<b>3. A rule to identify forward looking statements</b>	<p><i>for</i> each row in the portion of the report from which we want to extract information</p> <p><i>if</i> the key word is a candidate denoting forward looking statements (e.g., outlook, anticipate, demand)</p> <p><i>then</i> return the noun phrases immediately following the keyword and present it as the Business segment outlook</p> <p><i>and</i> return the verb phrase following the keyword and present it as the Outlook</p> <p><i>end if</i></p> <p><i>end for</i></p>
<b>4. A Semantic Web rule to identify mergers, acquisitions, and new business segments</b>	<p><i>for</i> each row in the portion of the report from which we want to extract information</p> <p><i>if</i> the verb phrase is a candidate denoting mergers, acquisitions, or new business segments</p> <p><i>then</i> return the subject noun phrase as ‘subject’</p> <p><i>and</i> return verb phrase as ‘predicate’ (e.g. verb phrases- acquired, ceased, created, became, entered)</p> <p><i>then</i> return the following noun phrase as ‘object’</p> <p><i>end if</i></p> <p><i>end for</i></p>

#### 4.1.4 N-gram processing

In 1994, Philip Clarkson and Ronald Rosenfeld at Carnegie Mellon University developed a UNIX based set of software tools called the Statistical Language Modeling (CMU-SLM) Toolkit. The purpose of the CMU-SLM Toolkit is to process larger corpora of textual data into n-grams, specifically bi-grams and tri-grams, and provide related statistical data (Clarkson & Rosendfeld, 1997). The CMU-SLM toolkit provides word frequency lists and vocabularies, word bigram and trigram counts, bigram- and trigram-related statistics, and various bigram and trigram language models. N-gram models are used to predict the probability of the sequence of words in a phrase, where n represents the number of words in the phrase. The most common n-grams are n=2 (bi-grams), n=3 (tri-grams), n=4 (four-grams). The previous word can be used to predict the probability of the next words in the phrase.

CAINES can accomplish the same goals in its corpus processing tasks. Therefore, the actual CMU-SLM was not used but the `Lingua::En::Tagger` and KWIC files were used to identify frequent n-grams and noun phrases for the extraction rules. CAINES has a part-of-speech tagger and noun phrase extractor that similarly assist with n-gram analysis. These subsystems assist with understanding the semantics of sentences in a corpus. The part-of-speech tagger assigns each word in a sentence a part of speech tag. The noun phrase extractor produces a list of noun phrases found in the corpus. These systems become useful when we want to extract into our subject-predicate-object format. We use `Lingua::En::Tagger`, available at <http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.16/Tagger.pm> to do these two tasks. `Lingua::En::Tagger` is a probability based, corpus-trained tagger that assigns part-of-speech (POS) tags to text. The tagger is based on a lookup dictionary and preset probability values. The tagger will try to assign a POS tag based on the known POS tags for a given word and the

POS tag assigned to its predecessor. In this information extraction study, the terms that appear as noun phrases and verb phrases help us to identify the subjects and the objects of the key verbs that we are trying to find. A list of relevant n-grams used in this study is found in Table 2.

**Table 2. Relevant n-gram Output for Extraction Rules**

Unigrams	Acquisition
	Demand
	Economic
	Indices
	Market
	Merger
	Outlook
	Short-term
Bi-grams	Business segment
	Business activity
	Commodity prices
	Credit environment
	Credit markets
	Credit spread
	Driven by
	Economic activity
	Economic condition
	Enter into
	Entered into
	Emerging markets
	Financial stress
	Long-term borrowings
	Market condition
	Merger between
	Merger includes
	Merger of
Tri-grams	Business segment outlook
	Forward looking statements
	Increase in market
	Increase in net
	Increase in revenues
	Long-term borrowing requirements
	Long-term capital pricing
	Long-term outlook
	Market risk factors
	Market risk framework
	Market risk implications
	Market risk management
	Short- and medium-term
	Weaker economic conditions
n-gram	Current portion of long-term borrowings
	Driven by cash equities
	Driven by challenging conditions
	Driven by lower revenues
	Slowdown in U.S. economic growth
	Unprecedented credit market environment
	U.S. housing market downturn
	U.S. sub-prime residential mortgage

The next step is to implement subsystems to detect patterns in the text of the corpus based on the n-gram vocabulary list. For this process, the KWIC system, SQL, and stemming are used.

#### **4.1.5 KWIC and SQL**

In 1958, Peter Luhn at IBM, developed the Key Word In Context (KWIC) system that uses automatic indexing to recognize word boundaries and frequencies (Luhn, 1960). For CAINES, KWIC is used to analyze word placement in relation to other words in a sentence to determine patterns in the text. The corpus document is further processed to format the text to be compatible with the KWIC system adapted for CAINES. Essentially, the KWIC system parses the text by paragraphs, denoted by a period followed by a new line. The results allow all sentences of the corpus to be included in the KWIC process.

The KWIC system then loads each word of the cleaned file into the first column of each row of a MySQL database table. Each row in the database will consist of additional columns that contain words in the text that follow the word in the first column. What results is a shifting pattern that allows each word of the corpus to appear in each column of the database. Figure 4 shows a sample of text from a Bank of America 10-Q. Table 3 shows the sample text in the KWIC database format.

In our system, SQL queries are used to retrieve, sort, categorize, and analyze word sequence patterns in the database to assist in organizing relationships necessary for our ontology. IBM developed SQL and it is a standard language used for querying database systems.

### First Quarter 2009 Economic Environment

During the first quarter of 2009, credit quality deteriorated further as the economy continued to weaken. Consumers experienced high levels of stress from higher unemployment and underemployment as well as further declines in home prices. These factors combined with further reductions in spending by consumers and businesses and continued turmoil in the financial markets negatively impacted the commercial portfolio. These conditions drove increases in consumer and commercial net charge-offs, and nonperforming assets as well as higher commercial criticized utilized exposure and reserve increases across most portfolios during the three months ended [March 31, 2009](#). For more information on credit quality, see the Credit Risk Management discussion beginning on page 130.

Capital market conditions showed some signs of improvement during the first quarter of 2009 and *Global Markets* took advantage of the favorable trading environment. Market dislocations that occurred throughout 2008 continued to impact our results in the first quarter of 2009 but to a lesser extent as we incurred reduced losses on CDOs and other *Global Markets* exposures (e.g., leveraged finance and CMBS) when compared to the same period in the prior year. We have also reduced certain asset levels in *Global Markets* for balance sheet efficiencies. For more information on *Global Markets*' results and their related exposures, see the discussion beginning on page 103.

Market conditions also continue to impact the ratings of certain monolines. We have direct and indirect exposure to monolines and, in certain situations, recognized losses related to some of these exposures during the first quarter of 2009 which included losses related to a monoline counterparty that restructured its business and subsequently had its credit rating downgraded. For more information related to our monoline exposure, see the Industry Concentrations discussion on page 151.

The above conditions, together with deterioration in the overall economy, will continue to affect many of the markets in which we do business and may adversely impact our results for the remainder of 2009. The degree of the impact is dependent upon the duration and severity of such conditions.

**Figure 4: Example text of Bank of America 3-31-10 10-Q.**



**Table 3. Example Data output from The KWIC Index System.**

First paragraph from the *Economic Conditions* section of the 3-31-09 Bank of America 10-Q. (For this example the database has been reduced to 7 columns rather than the 50 used in the actual database.)

Word1	Word2	Word3	Word4	Word5	Word6	Word7
During	the	first	quarter	of	2009,	credit
the	first	quarter	of	2009,	credit	quality
first	quarter	of	2009,	credit	quality	deteriorated
quarter	of	2009,	credit	quality	deteriorated	further
of	2009,	credit	quality	deteriorated	further	as
2009,	credit	quality	deteriorated	further	as	the
credit	quality	deteriorated	further	as	the	economy
quality	deteriorated	further	as	the	economy	continued
deteriorated	further	as	the	economy	continued	to
further	as	the	economy	continued	to	weaken.
as	the	economy	continued	to	weaken.	Consumers
the	economy	continued	to	weaken.	Consumers	experienced
economy	continued	to	weaken.	Consumers	experienced	high
continued	to	weaken.	Consumers	experienced	high	levels
to	weaken.	Consumers	experienced	high	levels	of

#### 4.1.6 Semantic and lexical analysis

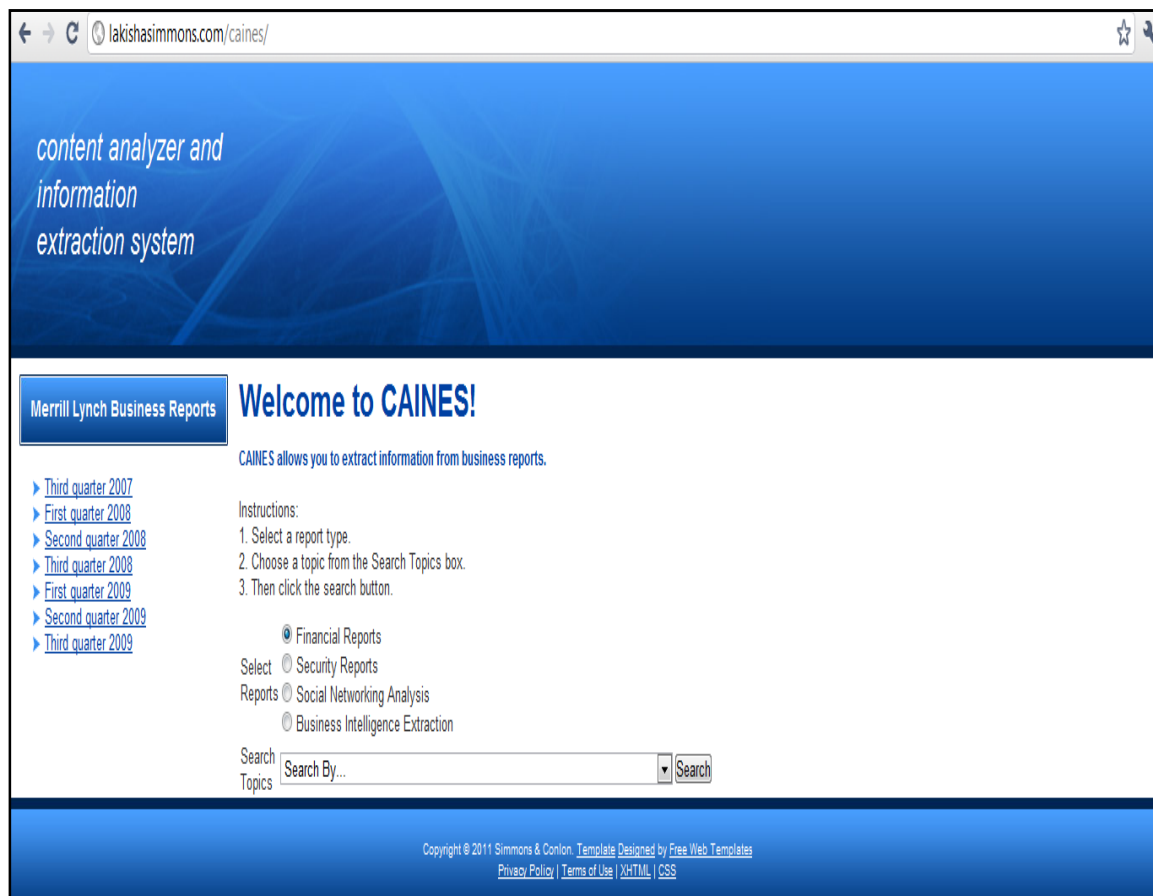
Developed in the early 1990s, WordNet has been used to classify information into hierarchical categories that can be adapted to develop a variety of IE systems (Bagga, J. Chai, A. & Biermann, 1996). The database recognizes and organizes parts of a sentence into machine-readable semantic relations (Miller, 1995). WordNet 3.0 was used in the CAINES corpus as well as other thesauri to determine synonyms for key words determined by the KWIC analysis. CAINES uses it to select synonyms for relationship terms. For example, *decline*, *loss*, *slow*, *weak*, and *lower* can be grouped and coded together in CAINES to extract phrases with predicates denoting some sort of decrease. As another example, *caused*, *driven by*, *resulted in*, *adverse impact*, and *impacted by* were used in rules where CAINES should extract information regarding causation.

#### 4.1.7 Stemming

Stemming is used in IE system development to improve recall. Stemming is the removal of the inflectional ending, such as -ed, -ing, and -s, from words to reduce word forms to its root. Root words are beneficial to CAINES because they can produce more effective results since words have different meanings in different contexts (Xu & Croft, 1998). CAINES incorporates stemming during the corpus development through SQL. Essentially, the roots of the words are combined with wildcard characters for SQL analysis. For example, in the SQL query for the word bank, the wildcard character “%” will be combined with the root word to form bank%. The query could return results such as– “bank”, “banks” and “banking”. Using wildcard characters in queries ensures all relevant references in the corpus are included.

#### 4.1.8 CAINES Extraction Interface

Now that the extraction rules and subsystems are implemented within CAINES, users can extraction information from long reports through the interface. The Web based interface was developed in HTML with PHP: Hypertext Preprocessor (PHP) scripting language. PHP was designed to allow web developers to create dynamic Web pages quickly (php.net). It was primarily chosen for this project because of its uncomplicated interfacing with MySQL and overall ease of programming. CAINES can be accessed from <http://www.lakishasimmons.com/caines/>. See Figure 5 for a screen shot of CAINES.



**Figure 5. CAINES Extraction Interface**

This completes the development methodology of CAINES for semantic based information extraction. The next section discusses the ontology extraction methodology.

#### **4.2 Financial Markets Ontology Extraction Methodology**

CAINES extracts financial ontological information. Ontologies are used for modeling some domain of the world in terms of labeled concepts, attributes, relationships and classifications. Ontologies consist of vocabulary, rules for combining terms, and are reusable structures.

The CAINES financial markets ontology is a formal representation of financial terms, information about finance companies, and definitions of terms in the reports. There are many ontology building methodologies but no clear standard ([http://semanticweb.org/wiki/Ontology\\_Engineering](http://semanticweb.org/wiki/Ontology_Engineering)). In their ontology development guide, Noy and McGuinness (2001) emphasize that there is no one correct way to model a domain and ontology development is an iterative process. However, some of the more popular ontology methodologies include *Enterprise Ontology* (Uschold & King, 1995), the *Unified Process for Ontology building* (De Nicola, Missikoff, & Navigli, 2009), and *METHONTOLOGY* (Fernandez, 1999). Uschold and King's *Enterprise Ontology*, presents a method based on four main activities: purpose identification, building, evaluation, and documentation. De Nicola, Missikoff, & Navigli (2009) created the Unified Process for ONtology (UPON) building, derived in part from the widely used software engineering process, Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999). UPON uses the UML and the use-case driven, iterative approaches of UP to support the preparation of ontology development, which is what distinguishes UPON from other ontology development methodologies.

Considering our goal of using a mature and standardized methodology for creating ontologies, this study adopts METHONTOLOGY. METHONTOLOGY is a structured method to build ontologies and is said to be the most mature methodology when comparing to the IEEE standard for developing software life cycle processes, 1074-1995 (Fernandez, 1999). METHONTOLOGY consists of eight phases: specification, knowledge acquisition, conceptualization, integration, implementation, evaluation and documentation phases (Fernández, Gómez-Pérez, & Juristo 1997). This study does not reuse an existing ontology and therefore omits the integration phase.

#### **4.2.1 Specification**

The first phase, *Specification*, is when the ontology specification document is created. The specification document is created in natural language and outlines the domain and scope of the ontology. It should convey what types of questions the information in the ontology should answer and what the ontology can be used for (Noy & McGuinness, 2001). See Figure 6 for the Financial Markets Ontology Requirement Specification Document.

**Ontology Requirement Specification Document****Domain:** Financial Markets**Date:** January, 6<sup>th</sup>, 2010**Conceptualized-by:** Lakisha L. Simmons**Implemented-by:** Lakisha L. Simmons & Sumali J. Conlon

**Purpose:** The financial market ontology will describe basic company information such as headquarter location and frequent terms commonly used in EDGAR reports. The financial markets ontology is useful in understanding the domain of financial markets.

**Level of Formality:** Semi-formal

**Scope:** Terms with both a high frequency per the noun phrase extraction report and common finance terms per the New York Stock Exchange Glossary are included. Company information from Countrywide Financial, Citicorp, Merrill Lynch, Bank of America, and HSBC Finance Corp is also included. Key concepts as identified by three finance experts as deemed relevant to business managers and finance experts. There are many financial terms that could be included in such an ontology. However, our goal is to identify the primary concepts of the financial market for the purpose of EDGAR financial reports, not to exhaust all the terms contained in the finance domain.

**Figure 6. Financial Markets Ontology: Specification Document**

**Figure 6. Ontology Requirement Specification Document**

#### **4.2.2 Knowledge Acquisition**

The second step in developing an ontology is the *knowledge acquisition* phase. The purpose of this phase is to gain knowledge about the subject matter of the ontology. There is a plethora of ways to gain such knowledge. One of the first would be to search for an existing ontology in the target domain. Libraries of reusable ontologies can be found on the Web and in the literature. Two such sources are the Ontolingua ontology library (<http://www.ksl.stanford.edu/software/ontolingua/>) and the DAML ontology library

(<http://www.daml.org/ontologies/>). Developers can also consult experts, books, handbooks, figures, tables as well as knowledge gathering techniques such as brainstorming, interviews, and analysis of texts among others.

A number of techniques were used in the knowledge acquisition phase of the Financial Markets Ontology. Since no complete current finance ontology could be found, semi-structured interviews were conducted with three finance experts. The experts were asked what concepts they felt were most important in the *Management's Discussion and Analysis of Financial Condition and Results* of a 10-Q. Key concepts were identified and used in building the ontology. Second, the noun phrase extractor results from several reports were used to identify common noun phrases that were discussed in the reports. Then, the main concepts and relationships given in the New York Stock Exchange Glossary were used to build the majority of the ontology. Lastly, the knowledge engineering approach was used to determine high level definitions and properties of terms from 10-Q reports.

#### **4.2.3 Conceptualization**

The fifth stage in the ontology development process is to implement the ontology through proper classification based on all knowledge gained and derived in the previous steps. There are several approaches to the hierarchy layout of the ontology diagram: top-down, bottom-up, and combination or middle-out (Uschold and Gruninger, 1996).

Top-down approach is used when the developer aims to start the ontology with the most general concepts and work down to more specific categorizations. For example, one would start the ontology with the class Wine, and then the subclasses White and Red would follow. White and Red could both then be dissecting into more specific classes at the leaf level. In contrast, the

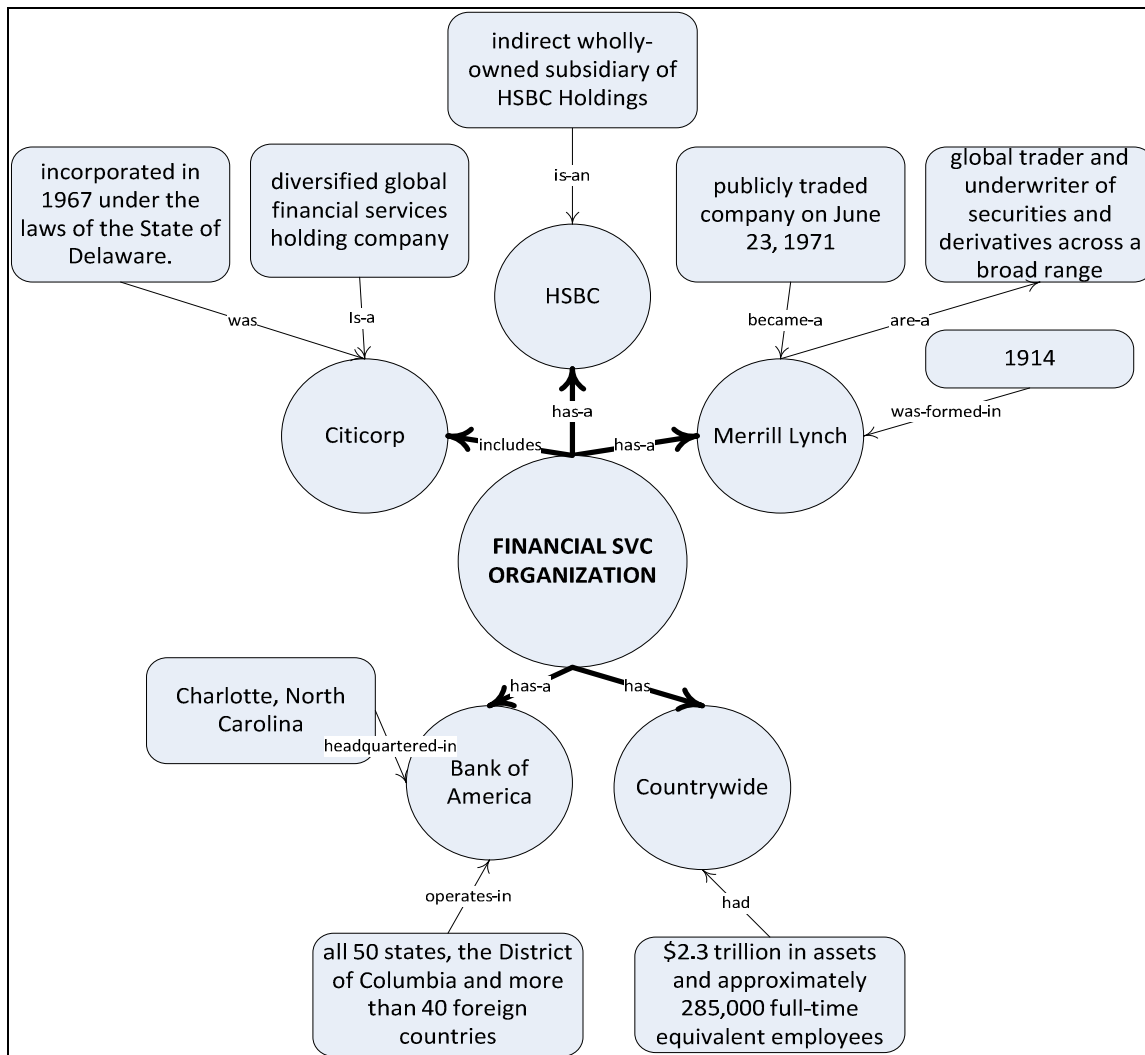
bottom-up approach begins with the leaves, the definitions of the most specific classes, and works to the more general classes. The middle-out approach is a combination of the first two approaches. The developer can start with the most significant or prominent concepts and then derive the subclasses, other significant concepts, or leaves. Uschold and King (1995) recommend the middle-out approach since the most important concepts be identified first, which will then be used to obtain the remainder of the hierarchy by generalization and specialization.

According to Rosch (1978) the classification approach to use in the ontology development depends on the developers view of the domain, thus none of these three methods is inherently better than any of the others. However, the middle-out approach is often chosen by developers since the most important concepts begin the ontology. Uschold & Gruninger (1996) argues that the main advantage of the middle-out approach is that it allows the developer to identify the primary concepts of the ontology early. Which allows the developer to then move on to only generalize those concepts that are necessary and in scope. The middle-out approach is said to require less re-work and effort and more stable concepts.

The Financial Markets Ontology uses the middle-out approach. The concept terms that are independent in nature become the top-level classes and are denoted as circles in the hierarchy. Terms that describe objects are subclasses, denoted as rounded squares, and are connected by a relationship term such as “is-a” or “type-of”. Reports from 5 companies were used to extract basic company information: Countrywide Financial, HSBC, Merrill Lynch, Wachovia, Citigroup, and Bank of America. See Figure 7 for the Financial Markets Ontology diagram of company information.



There are many financial terms that could be included in such an ontology. However, our goal is to demonstrate that ontology information can be extracted using CAINES. Thus, this study identifies the primary concepts of the financial market for the purpose of extracting important information from EDGAR financial reports, not to exhaust the terms contained in the finance domain.



**Figure 7. Financial Markets Ontology Diagram**

#### 4.2.4 Ontology Extraction Implementation

The next step is implementation. Here, CAINES extracts the ontology from the various sources of text. Other studies have codified the ontology in a formal language such as: CLASSIC, Ontolingua, Prolog, or C++ (Noy & McGuinness, 2001). See Table 4 for the company ontology extracted and Table 5 for a partial view of the finance terms ontology. The complete finance ontology can be found in Appendix II.

**Table 4. The Extracted Company Ontology**

Subject	Predicate	Object
Citicorp	is a	diversified global financial services holding company
Citicorp	was	incorporated in 1967 under the laws of the State of Delaware.
Merrill Lynch	was formed	in 1914
Merrill Lynch	became a	publicly traded company on June 23, 1971.
We	are a	global trader and underwriter of securities and derivatives across a broad range
Bank of America	headquartered in	Charlotte, North Carolina,
Bank of America	operates in	all 50 states, the District of Columbia and more than 40 foreign countries.
At 2009, Countrywide	had	\$2.3 trillion in assets and approximately 285,000 full-time equivalent employees.
HSBC	is an	indirect wholly-owned subsidiary of HSBC Holdings

**Table 5. Finance terms ontology A-B (see the Appendix II for the full ontology)**

Accrued interest	is	The interest due on a bond since the last interest payment was made.
American Depositary Receipt (ADR)	is	a security issued by a U.S. bank in place of the foreign shares held in trust by that bank,
American Stock Exchange (AMEX)	is	The second largest stock exchange in the United States, located in the financial district of New York City.
Amortization	is	Accounting for expenses or charges as applicable rather than as paid. Includes
Annual report	is	The formal financial statement issued yearly by a corporation.
Arbitrage	is	A technique employed to take advantage of differences in price.
Assets	is	Everything a corporation owns or that is due to it: cash,
Auction market	is	The system of trading securities through brokers or agents on an exchange
Auditor's report	is	Often called the accountant's opinion, it is the statement of the accounting firm's work
Averages	is	Various ways of measuring the trend of securities prices, one of the most popular of which is the Dow Jones Industrial Average of 30 industrial
Balance sheet	is	A condensed financial statement showing the nature and amount of a company's assets,
Basis point	is	One gradation on a 100-point scale representing 1%; used especially in expressing variations in the
Bear	is	Someone who believes the market will decline. (See: Bull)
Bear market	is	A declining market. (See: Bull market)
Bearer bond	is	A bond that does not have the owner's name registered on the books of the issuer.
Block	is	A large holding or transaction of stock popularly considered to be 10,000 shares or more.
Blue Sky Laws	is	A popular name for laws various states have enacted
Blue chip	is	A company known nationally for the quality and wide acceptance of its products or services,
Book value	is	An accounting term. Book value of a stock is determined from a company's records,
Broker	is	An agent who handles the public's orders to buy and sell

#### **4.2.5 Ontology Evaluation and Documentation**

The evaluation and documentation activities should occur at each of the previous stages. Like testing software, evaluation and documentation are conducted in iterative units throughout the development cycles. When evaluating an ontology, verification and validation are the main activities (Gómez-Pérez, Juristo, & Pazos, 1995). Each of the previous stages has documentation deliverables that are part of the ontology project. These documents have been evaluated by the author of this dissertation and an associate professor knowledgeable about ontologies. The evaluation occurred after each phase for verification and validation before moving on to subsequent steps.

## CHAPTER 5

### HYPOTHESIS DEVELOPMENT & TESTING

CAINES was developed to support users interested in extracting unstructured information from lengthy reports. This occurs through semantic based information extraction. A user could be a financial analyst, investor, small business owner, or anyone else who may be interested in reading the discussion text of financial reports (versus financial data in tables). CAINES can assist a user in extracting more relevant and precise information in less time than manually reading and searching online financial reports. An experiment will compare recall, precision, F-measure, and speed while using CAINES to manual extraction of information. The speed comparison between CAINES and manual extraction is conducted by having students answer questions based on information in the reports. They will answer the questions by manually reviewing the reports and then by reviewing the information extracted by CAINES. The speed is the extraction time *and* the user information processing of the questions. Six hypotheses explain our predication of a user's performance and satisfaction with using CAINES versus manual extraction. The performance, usefulness, ease of use, and user satisfaction were all evaluated after testing. This chapter also discusses the data collection procedures.

## **5.1 IE Performance**

CAINES' performance is evaluated through speed, recall, precision, and F-measure. CAINES was developed to increase a user's efficiency in extracting information from online business reports. The average user processing time is defined as the time the participants spent answering questions with CAINES or by manually reading and extracting information from reports. It is expected that CAINES will allow a user to perform information extraction and answer questions faster than without CAINES.

**H1:** Average processing time is faster when using CAINES than when manually extracting information from EDGAR 10-Q reports.

To test recall, a comparison will be made between the amount of relevant information a user is able to extract using CAINES versus manual extraction. Recall is measured by dividing the correct number of answers given by the total possible number of correct answers. Users of CAINES are expected to receive better recall results than from the manual extraction since it is so difficult to manually extract specific information from such lengthy reports.

**H2:** Average recall is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

Similarly, a comparison is made between the amount of information the users were able to extract using CAINES and manual extraction. Precision is measured by dividing the number of correct user answers by the total answers the user was able to produce for each method. Users of CAINES are expected to receive better precision results than from the manual extraction.

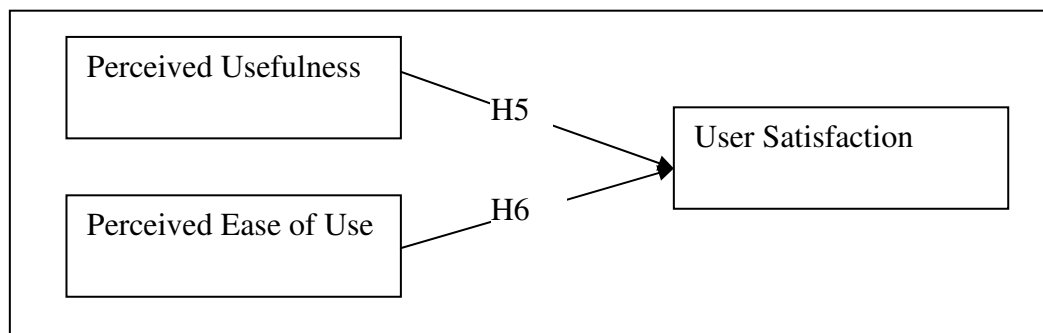
**H3:** Average precision is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

Accordingly, it is expected that the F-measure for CAINES will be greater than the F-measure for manual extraction. The F-measure is a combination of the equal weighted results of precision and recall.

**H4:** Average F-measure is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.

## 5.2 *Technology acceptance measurement*

This section discusses how CAINES is measured for usefulness, ease of use, and user satisfaction. In this study, user satisfaction is the dependent variable hypothesized to be correlated with perceived ease of use and perceived usefulness (Figure 8). It is important that CAINES not only be deemed useful and easy to use, but overall satisfying to the user who is interested in searching long reports for specific information.



**Figure 8. Conceptual model of the impact of CAINES**

There are two hypotheses related to the impact of CAINES on a user's satisfaction. CAINES was developed to be useful in quickly and efficiently finding information in lengthy online financial reports. Hence, perceived usefulness and satisfaction will positively correlate.

**H5:** There is a significant positive correlation between perceived usefulness of CAINES and user satisfaction.



CAINES was designed to be user friendly with a clean and sleek design. The extraction interface is very straightforward. Therefore, the perceived ease of use of CAINES is expected to positively correlate with user satisfaction.

**H6:** There is a significant positive correlation between perceived ease of use of CAINES and user satisfaction.

### 5.3 *Pilot Testing*

For the dissertation proposal, semantic information regarding the ‘economic impact on Bank of America’s operating conditions’ was manually extracted. Recall, precision, and F-measure were the only tests for the pilot study. Twenty-one total relationships and 17 relationships regarding ‘economic impact’ were found. Then, using CAINES the author found 13 correct ‘economic impact’ relationships out of the total 17 semantic relationships that were found.

Metrics were as follows:

$$\begin{aligned}
 \textbf{Recall} &= \frac{\text{Total correctly extracted}}{\text{Total possible correct to extract}} = \frac{13}{17} = \mathbf{76.4\%} \\
 \textbf{Precision} &= \frac{\text{Total correctly extracted}}{\text{Total extracted}} = \frac{13}{14} = \mathbf{92.8\%} \\
 \textbf{F-measure} &= 2 * (\text{recall} * \text{precision}) / (\text{recall} + \text{precision}) = \mathbf{83.8\%}
 \end{aligned}$$

In the pilot, only portions of the full system were used. Perl regular expression, KWIC, SQL, knowledge engineering techniques, and stemming procedures were used to receive the preliminary results. To increase accuracy and make the system more scalable, CAINES included n-gram, noun phrase, and semantic analysis in training and testing the full system.

#### **5.4 Main Study Participants and Task**

Business professionals, undergraduate, and graduate business students participated in this study. Their task was to test the difference between manual and automated information extraction. Qualtrics Survey Research Suite (Qualtrics) was used to administer the survey because it is a quality online tool to conduct research. An online survey tool is appropriate in our study because the respondents were chosen based on some familiarity with using the Internet and or reading online business reports.

First, a pretest was conducted with six business professionals for one week. The professionals participated in the study and provided feedback on question wording, timing, and other flow issues. Once all feedback was incorporated, the main study was conducted.


A total of 54 complete sets of data were downloaded from Qualtrics. Four data sets were deleted because the reverse coded item was not answered appropriately. Six data sets were from the business professionals who pretested the study, and were therefore deleted. Forty-four samples were used in the data analysis.

The undergraduate business students were surveyed while attending a core MIS course that they were enrolled in. Graduate students participated in the study on their own time. The average age of the participants was 24. The minimum and maximum participant ages were 19 and 38 respectively. A little over half of the participants were male. About 35% of the students were majoring in accounting, finance, or economics (Table 6). These three majors may be the most familiar with analyzing financial reports. However, the goal of the testing phase was to compare manual extraction of semantic information to that of extraction with CAINES. Therefore, a broad range of disciplines was acceptable for the sample.

**Table 6. Demographics of Respondents (n=44)**

		<u>Number</u>	<u>Percent</u>
Gender	Male	24	54.5%
	Female	20	45.5%
Major	Marketing	11	25%
	Management	10	22.7%
	Accounting	9	20.5%
	IS/Computers	7	15.8%
	Finance	5	11.4%
	Economics	1	2.3%
	Education	1	2.3%

During the undergraduate classroom administration of the study, the participants received a brief overview of the project. Then they were given access to the experiment which was housed online in Qualtrics. The first page of Qualtrics was a description of the study. The second Web screen was the manual data extraction form noting the information to be extracted manually from the EDGAR database (see Figure 9). This manual extraction page begins with instruction to answer five questions by accessing and reviewing specific financial reports on the EDGAR database. Web links to the specific reports were listed with the questions. As stated in the methodology, the questions were based on what the financial experts said would be important to know. The third Web screen displayed the questions to be answered using CAINES (see Figure 10). The participants used CAINES to query the database that is preloaded with the reports needed to answer the questions. The participants were given 10 minutes for each extraction method. There was a statement displayed on each screen that the page would advance after 10 minutes. This was to ensure that the study would not exceed a 30 minute time frame. Qualtrics was customized to capture the time spent on the manual (screen two) and the CAINES extraction pages (screen three). This data assists in our speed comparison calculation between manual extraction and CAINES extraction.

qualtrics.com

**PLEASE READ...** Answer the 5 questions on this page by searching for the answers in the appropriate Merrill Lynch report. The appropriate Merrill Lynch report will be listed with the question. The report will open in a new window. You will need to come back to this screen to type in your answer.

*After 10 minutes, the survey will automatically go to the next page. Do not rush, just do the best you can.*

The following two questions ask about global and domestic market conditions in 2008.

1. What drove the decline in US economic activity in first quarter 2008?  
[click here for the first quarter 2008 report](#)

2. What country had moderate growth in the first quarter of 2008?  
[click here for the first quarter 2008 report](#)

The following question asks about domestic market conditions in second quarter 2008.

3. What drove the decline in US economic activity in the second quarter 2008?  
[click here for the second quarter 2008 report](#)

The following questions ask about market condition impacts in 2007 and 2008.

4. Market conditions resulted in how much in net losses for the third quarter of 2007?  
[click here for the third quarter 2007 report](#)

5. Market conditions resulted in how much in write-downs for the first quarter of 2008?  
[click here for the first quarter 2008 report](#)

>>

Figure 9. Qualtrics Web Screen - Manual Extraction Instructions and Questions

**Answer the 5 questions below by using a computer system called CAINES.** You will not need to access individual reports. You will still need to come back to this screen to type in your answer.

1. Access the CAINES system: <http://www.lakishasimmons.com/caines>
2. The Financial Reports option is already selected for you.
3. Choose a topic from the *Search Topics* box that corresponds with the questions below.
4. Come back to this page to type in your answer.

The following two questions ask about the business outlook in 2008. Look for business outlook in 2008 in the CAINES search box.

**1. What is the business outlook for Equity Markets?**

Remember, go to CAINES to answer the questions on this page.

**2. What kind of impact will market conditions in the short- and medium- term have on Merrill Lynch?**

The following three questions ask about mergers, acquisitions, and business segment news between 2008 and 2009.

**3. What acquisitions occurred in 2009?**

**4. Were any new business segments created? If so please state the name of the new segment(s) below.**


**5. Is Merrill Lynch ("we") under any examination? choose one:**

- ☐ Yes – by the Securities and Exchange Commission (SEC)
- ☐ No – not under any examination
- ☐ Yes – by the IRS and other tax authorities

**Figure 10. Qualtrics Web Screen - CAINES Extraction Instructions and Questions**

#### ***5.4 Survey Measurement Development***

The user survey was introduced on the last screen in Qualtrics. Acceptance and satisfaction of CAINES were evaluated with a Likert scale survey. To analyze the perceived usefulness of CAINES, participants answered five questions. For example, “Using CAINES increased my productivity”. The next five survey items covered user feelings about the perceived ease of use of CAINES: “Learning to operate CAINES was easy for me”. Lastly, the survey asked questions about the user’s satisfaction with CAINES. For example, “I was satisfied with the time it took to search using CAINES” and “Using CAINES is a good way to search long financial reports”. User satisfaction was measured with three questions and the last question was reverse coded. Reverse coding, phrasing a question opposite to the other questions, was used to help ensure participants were reading the questions and not simply choosing the same response through all the questions. See Figure 11 for the survey Web page.



**Last page! Please indicate the extent to which you agree or disagree with the following statements by selecting a number from 1 to 7:**

	Strongly Agree	Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Disagree	Strongly Disagree
Using CAINES enabled me to search more quickly than manually searching.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using CAINES improved my search performance over manually searching the reports.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using CAINES increased my productivity.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using CAINES enhanced my search effectiveness.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would find CAINES useful if I had to search long financial reports again.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	Strongly Agree	Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Disagree	Strongly Disagree
Learning to operate CAINES was easy for me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would find it easy to get CAINES to do what I wanted it to do.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My interaction with CAINES was clear and understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found CAINES to be flexible to interact with.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It would be easy for me to become skillful at using CAINES.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	Strongly Agree	Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Disagree	Strongly Disagree
Using CAINES is a good way to search long financial reports.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was satisfied with the time it took to search using CAINES.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was better to manually search the reports to find what I was looking for.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Figure 11. Qualtrics Web Screen – User Survey**

What feelings do you have about manually looking for the answers in the financial reports? What are your feelings about CAINES?

What is your age?

Gender:

☐ Male

☐ Female

Major or Field of Expertise

☐ Accounting

☐ Finance

☐ Management

☐ Economics

☐ Information System/Computer Science

☐ Other

Thank you for participating in this study! Click the button below to enter the gift card drawing.

>>

**Figure 11. continued...Qualtrics Web Screen - User Survey**



## CHAPTER 6

### DATA ANALYSIS AND RESULTS

#### **6.1 Overview**

Testing compared extraction and question answering with CAINES with that of manual extraction of question answering from 10-Q reports on the EDGAR database. Precision, recall, and F-measure were provided for the proposal and the main study. In addition, speed, correlation analysis, and descriptive analyses were conducted in the main study. Recall reports, as a percentage, how many correct relationships are extracted. Recall was measured by dividing the number of correct answers produced by the total possible correct answers. Precision measures the accuracy of the extraction of semantic relationships. Precision was calculated by dividing the number of correct answers produced by the number of total answers produced. F-measure combines recall and precision into a single measure and was calculated by using the formula  $F=2RP/R+P$ .

#### **6.2 Main Study Analysis and Results**

In this dissertation, a total of 21 10-Q reports, averaging about 100 pages long, were used for information extraction, ontology development, and Semantic Web development. Five companies were used in this dissertation: Countrywide Financial, HSBC, Merrill Lynch, Wachovia, Citigroup, and Bank of America. Ontology extraction was conducted with one report from all five companies.

For the semantic based *information extraction*, six 10-Q Merrill Lynch reports were used for *training*. Training is a technique used in CAINES to learn the patterns in the data and create the specific extraction rules. Merrill Lynch's first and second quarters of 2007, third quarter of 2008, and first, second, and third quarters of 2009 reports were used for training. *Information extraction testing* of CAINES was conducted using the following three reports: third quarter of 2007, first and second quarters of 2008. From the three major extraction rules, specific rules were created to extract information from Merrill Lynch reports (Table 1). Using CAINES, one can extract information about *Global and domestic market conditions in 2008*, *Market conditions in second quarter 2008*, *Market condition impacts in 2007 and 2008*, and information about the *Business outlook in 2008*. Visit <http://www.lakishasimmons.com/caines/index.php> to extract information from reports.

The *Semantic Web training* corpus was comprised of Bank of America reports between 2007 and 2009. *Semantic Web Testing* occurred with Merrill Lynch 2007 to 2009 10-Q reports. The Semantic Web testing was run against all Merrill Lynch reports between 2007 and 2009, a database of 107,533 rows. A Semantic Web was extracted based on *extraction rule 4* (Table 1) and extracts information regarding *mergers, acquisitions, and business segment news between 2007 and 2009*.

### **6.2.1 Performance Results**

Responses to the extraction questions, speed, and survey data were downloaded from Qualtrics. CAINES takes about 3 seconds to extract information from reports. The speed comparison between using CAINES and manually extracting and processing information was conducted by having students answer questions based on information in the reports. They

answered the questions by manually reviewing the reports and then by reviewing the information extracted by CAINES. Recall, precision, and F-measure were calculated based on correct and incorrect answers to the extraction questions. The correct answers to the five questions used during the manual extraction phase are shown in Figure 12. See Figure 13 for the questions and answers that participants answered using CAINES regarding *business outlook in 2008*. Figure 14 displays the actual CAINES screen image of the *business outlook in 2008* output.

Manual Data Collection Questions and Answers	
<b>The following two questions ask about <u>global and domestic market conditions in 2008</u>.</b>	
1. <i>What drove the decline in US economic activity in the first quarter 2008 report?</i>	<u>lower domestic demand and consumer spending</u>
2. <i>What country had moderate growth in first quarter 2008?</i>	<u>Europe</u>
<b>The following question asks about <u>domestic market conditions in second quarter 2008</u>.</b>	
3. <i>What drove the decline in US economic activity in the second quarter of 2008?</i>	<u>driven in part by the difficult conditions in the credit and residential housing markets.</u>
<b>The following questions ask about <u>market condition impacts in 2007 and 2008</u>.</b>	
4. <i>Market conditions resulted in how much in net losses for the third quarter of 2007?</i>	<u>Approximately \$7.9 billion</u>
5. <i>Market conditions resulted in how much in write-downs for the first quarter of 2008?</i>	<u>Approximately \$1.5 billion of net write-downs</u>

**Figure 12. Manual Data Collection Questions and Answers**

## CAINES - Data Collection Questions and Answers

The following questions ask about the business outlook in 2008.

1. *What is the business outlook for Equity Markets?* Look for business outlook in 2008 in the CAINES search box.

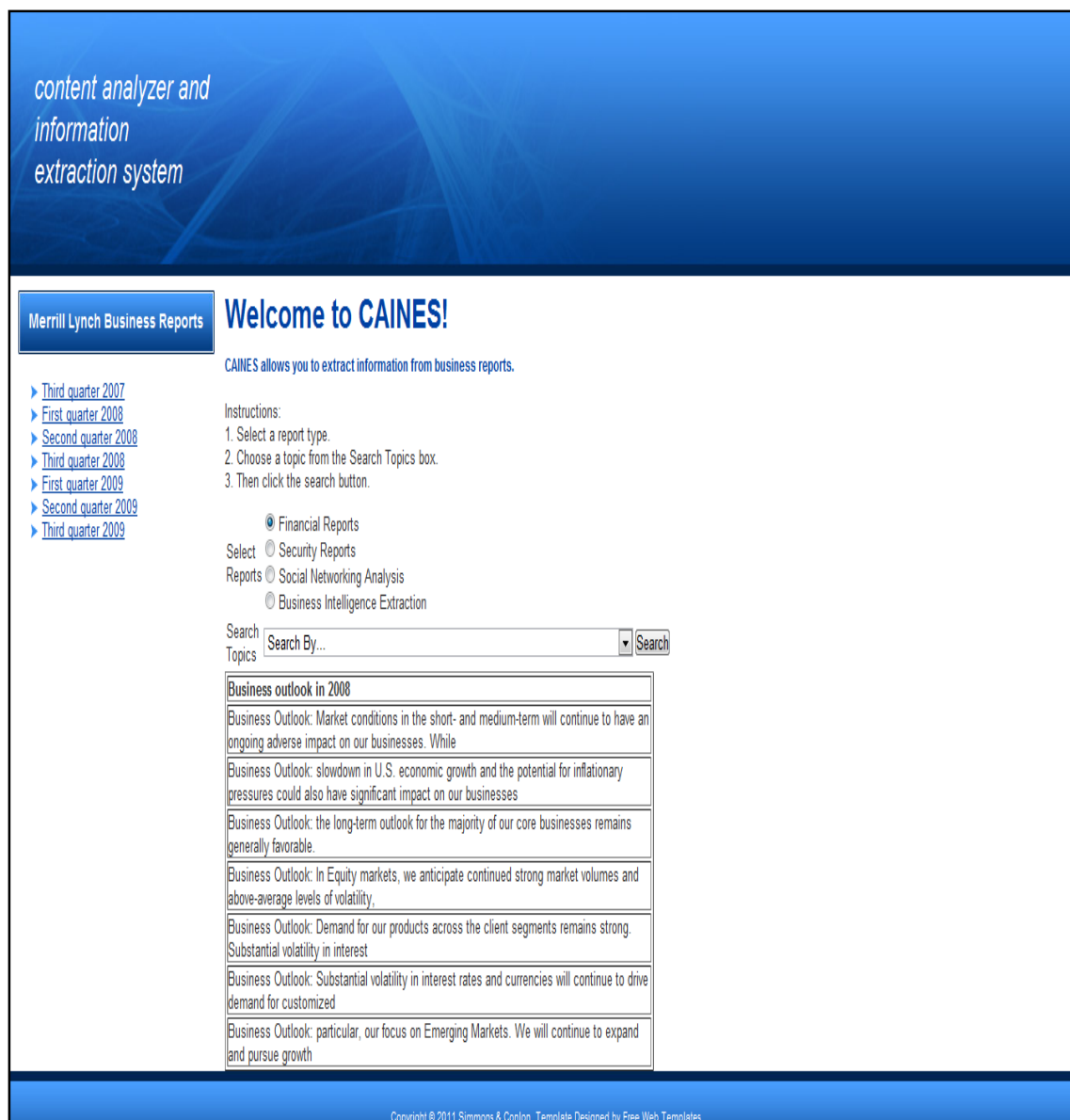
continued strong market volumes and above-average levels of volatility

2. *What kind of impact will market conditions in the short- and medium- term have on Merrill Lynch?*

Adverse impact

Business outlook in 2008
Business Outlook: Market conditions in the short- and medium-term will continue to have an ongoing adverse impact on our businesses. While
Business Outlook: slowdown in U.S. economic growth and the potential for inflationary pressures could also have significant impact on our businesses
Business Outlook: the long-term outlook for the majority of our core businesses remains generally favorable.
Business Outlook: In Equity markets, we anticipate continued strong market volumes and above-average levels of volatility,
Business Outlook: Demand for our products across the client segments remains strong. Substantial volatility in interest
Business Outlook: Substantial volatility in interest rates and currencies will continue to drive demand for customized
Business Outlook: particular, our focus on Emerging Markets. We will continue to expand and pursue growth

Figure 13. Extraction Questions and Answers as output by CAINES



**Figure 14. Information Extracted dealing with the “Business Outlook in 2008”**

Figure 15 displays the Semantic Web questions and answers as output by CAINES.

Figure 16 shows the actual CAINES screen image of the Semantic Web *mergers, acquisitions, and business segment news* output.

The following questions ask about mergers, acquisitions, and business segment news between 2008 and 2009.

3. *What acquisitions occurred in 2009?*

Merrill Lynch was acquired by Bank of America Corporation

4. *Were any new business segments created? If so, please state the name of the new business segment(s) below.*

Global Wealth Management

5. *Is Merrill Lynch ("we") under any examination? Circle one:*

Yes – by the Securities and Exchange Commission (SEC)

No – not under any examination

Yes – by the IRS and other tax authorities

Subject	Predicate	Object
On January 1, 2009, Merrill Lynch	was acquired by	Bank of America Corporation
Effective with the BlackRock merger, MLIM	ceased to exist	as a separate business segment.
a new business segment, Global Wealth Management	was created,	consisting of GPC and Global Investment Management (GIM).
On September 29, 2006, Merrill Lynch	merged	MLIM with BlackRock
MLBUSA	was merged	into Bank of America, N.A., a subsidiary of Bank of America.
MLBT-FSB	also merged	into Bank of America, N.A. At September 30, 2009, the
ML & Co.	became a subsidiary of Bank of America	and established intercompany lending and borrowing arrangements to facilitate centralized liquidity management.
We	had entered into a definitive agreement	with First Republic Bank (First Republic) to acquire all of the outstanding common shares of First Republic
We	have entered into a long-term service	agreement. As consideration for the sale of our interest in Bloomberg L.P., we received notes issued by
We	are under examination	by the Internal Revenue Service (IRS) and other tax authorities in countries including Japan and the United Kingdom, and states in which we have significant business operations,

Figure 15. Semantic Web Questions and Answers as output by CAINES

[First quarter 2008](#)  
[Second quarter 2008](#)  
[Third quarter 2008](#)  
[First quarter 2009](#)  
[Second quarter 2009](#)  
[Third quarter 2009](#)

Instructions:

1. Select a report type.
2. Choose a topic from the Search Topics box.
3. Then click the search button.

☒ Financial Reports  
 Select ☐ Security Reports  
 Reports ☐ Social Networking Analysis  
☐ Business Intelligence Extraction

Search Topics

Subject	Predicate	Object
On January 1, 2009, Merrill Lynch	was acquired by	Bank of America Corporation
Effective with the BlackRock merger, MLIM	ceased to exist	as a separate business segment.
a new business segment, Global Wealth Management	was created	consisting of GPC and Global Investment Management (GIM).
On September 29, 2006, Merrill Lynch	merged	MLIM with BlackRock
MLBUSA	was merged	into Bank of America, N.A., a subsidiary of Bank of America.
MLBT-FSB	also merged	into Bank of America, N.A. At September 30, 2009, the
ML & Co.	became a subsidiary of Bank of America	and established intercompany lending and borrowing arrangements to facilitate centralized liquidity management.
We	had entered into a definitive agreement	with First Republic Bank (First Republic) to acquire all of the outstanding common shares of First Republic
We	have entered into a long-term service	agreement. As consideration for the sale of our interest in Bloomberg L.P., we received notes issued by
We	are under examination	by the Internal Revenue Service (IRS) and other tax authorities in countries including Japan and the United Kingdom, and states in which we have significant business operations.

**Figure 16. CAINES screen of Semantic Web of “mergers, acquisitions, and business segment news between 2008 and 2009”**

The speed comparison between CAINES and manual extraction was conducted by having the participants answer questions based on information in the reports. They answered the questions by manually reviewing the reports and then by reviewing the information extracted by CAINES (Table 7). CAINES, the system itself, takes about 3 seconds to extract information from reports. The average user speed in the manual extraction process was a little over 9

minutes, compared to the average user speed with the help of CAINES about 6 minutes. The 370 seconds is the CAINES extraction time *and* the user completing the questions. Potentially, the manual extraction time would have been longer if there was not a 10 minute time limit.

The amount of recall, or relevant information extracted, was calculated by dividing user correct answers by the total possible correct answers. Recall of the participants for the manual extraction process was 10.45% compared to recall with CAINES of 85.91% (Table 7). Precision was measured by dividing the number of correct answers produced by the number of total answers produced. Precision for the manual extraction was 16.86% compared to 87.16% for CAINES.

Lastly, the F-measure was used to combine the results of precision and recall. The F-measure for the manual process was 12.59% compared to the F-measure for CAINES of 86.46%.

**Table 7. Speed, Recall, Precision, and F-measure results (n=44)**

	User Processing Speed (seconds)		Recall		Precision		F-measure	
	Manual	w/CAINES	Manual	w/CAINES	Manual	w/CAINES	Manual	w/CAINES
Mean	556.39	370.61	.105	.859	.169	.872	.1259	.865
Min	163	182	.00	.40	.00	.40	.00	.40
Max	600	600	.60	1.0	1.0	1.0	.75	1.0
Std Dev	96.26	110.989	.180	.181	.309	.167	.220	.173

### 6.2.2 Performance Hypotheses Results

A paired sample t-test was conducted to test the speed, recall, precision, and F-measure hypotheses (Table 8). All four hypotheses were supported as detailed in Table 9. H1 which suggested that CAINES would be faster, is strongly supported,  $t(43) = 8.861, p = <.01, \alpha = .05$ . Participants spent statistically significantly more time extracting semantic information manually ( $M = 556.39, SD = 96.26$ ) than they did using CAINES ( $M = 370.61, SD = 110.989$ ).



In regards to higher recall and precision with CAINES, both were supported, H2,  $t(43) = -19.940, p = <.01, \alpha = .05$  and H3,  $t(43) = -13.286, p = <.01, \alpha = .05$ . Thus, there was evidence that a difference in recall existed between the manual extraction ( $M = .1045, SD = .1804$ ) and CAINES ( $M = .8591, SD = .1809$ ). There was also significant evidence that precision of the manual processing ( $M = .1686, SD = .3085$ ) was less than for CAINES ( $M = .8716, SD = .1668$ ). Similar results supported the F-measure hypothesis, H4,  $t(43) = -17.645, p = <.01, \alpha = .05$ . The F-measure revealed significant differences between the manual extraction ( $M = .1259, SD = .2202$ ) and CAINES extraction ( $M = .8646, SD = .1735$ ).

**Table 8. Results of Paired T-Test**

Paired Samples T-Test					
Paired Comparisons	Mean Difference	Standard Deviation	T-Value	Degrees of Freedom	P-Value
Time w/CAINES versus Manual	185.777	139.072	8.861	43	<.001
Recall w/CAINES versus Manual	-.755	.251	-19.940	43	<.001
Precision w/CAINES versus Manual	-.703	.351	-13.286	43	<.001
F-Measure w/CAINES versus Manual	-.739	.278	-17.645	43	<.001

**Table 9. Performance Hypotheses Results Summary**

<b>Hypothesis</b>	<b>T-Value</b>	<b>Supported</b>
<b>H1:</b> Average processing time is faster when using CAINES than when manually extracting information from EDGAR 10-Q reports.	8.861*	Yes
<b>H2:</b> Average recall is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-19.940*	Yes
<b>H3:</b> Average precision is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-13.286*	Yes
<b>H4:</b> Average F-measure is higher when using CAINES than when manually extracting information from EDGAR 10-Q reports.	-17.645*	Yes

\*Significant at the .05 level

### **6.2.3 Survey Measurement Analysis**

The survey administration was the last step in the study. Participants completed the survey after participating in the manual and CAINES extraction processes. Survey responses were downloaded from Qualtrics and analyzed in SPSS 17. Scale reliability, descriptive statistics, and correlation statistics of the survey are discussed next.

Internal consistency was assessed by calculating Cronbach's alpha. The values were 0.82, 0.95, and 0.98 for satisfaction, perceived ease of use, and perceived usefulness respectively (Table 10). The scales are deemed reliable since they were all greater than the accepted threshold of 0.70 as recommended in the literature (Nunnally, 1967).

As shown in Table 10, participants believed CAINES was useful for fast and effective searching of long reports and that it improved performance and productivity. Participants agreed that it was easy to learn to use and maneuver CAINES. They agreed that CAINES was clear and understandable and that they could become skillful with CAINES. They somewhat agreed that CAINES was flexible to interact with. Further, they were satisfied with using CAINES for long

reports and the CAINES search time. They agreed that manually extracting was not the better search method (wording reverse coded).

**Table 10: Survey Response Summary (strongly disagree (1) to strongly agree (7))**

Construct	Items	Mean	StdDev
PU (.98)	PU1- faster search	6.43	1.065
	PU2-improved search performance	6.39	1.017
	PU3-increased productivity	6.36	1.080
	PU4-enhanced search effectiveness	6.41	1.085
	PU5- would be useful again	6.45	1.066
	Grand Mean	6.41	
PEOU (.95)	PEOU1-learning to use was easy	6.14	1.112
	PEOU2-easy to maneuver	5.98	1.191
	PEOU3-clear and understandable	6.07	1.129
	PEOU4-flexible to interact with	5.71	1.374
	PEOU5-can be become skillful with	6.16	1.219
	Grand Mean	6.01	
SAT (.82)	SAT1-good for long reports	6.23	1.118
	SAT2-satisfied with search time	6.36	1.036
	SAT3-manual search was not better	6.18	1.040
	Grand Mean	6.30	

Cronbach's alpha in parenthesis

PU=Perceived Usefulness

PEOU=Perceived Ease of Use

SAT=User Satisfaction

To test the relationships between PU and satisfaction and PEOU and satisfaction, the correlations between the constructs were analyzed. The purpose of correlation analysis is to measure the linear or nonlinear strength of a relationship. PU and PEOU of CAINES were both positively correlated with satisfaction (Table 11). The relationship between PU and satisfaction was significantly correlated,  $r = .795$ ,  $p < .01$ . PEOU and satisfaction were also significantly correlated,  $r = .831$ ,  $p < .01$ . Thus, H5 and H6 are supported (Table 12).

**Table 11. Correlations between Constructs**

	PU	PEOU	SAT
PU	1	.795**	.776**
PEOU		1	.831**
SAT			1

\*\*Correlation is significant at the 0.01 level (2-tailed).

**Table 12. Satisfaction Hypotheses Results**

Hypothesis	Path	Correlation (r-value)	Supported
<b>H5:</b> Perceived usefulness (PU) will positively influence user satisfaction (US).	PU → US	.776**	Yes
<b>H6:</b> Perceived ease of use (PEOU) will positively influence user satisfaction (US).	PEOU → US	.831**	Yes

\*\*Correlation is significant at the 0.01 level (2-tailed).

### 6.3 Discussion

Overall, greater levels of recall, precision, and F-measure, were achieved with CAINES than with manual extraction of information from 10Q reports. CAINES was also advantageous in terms of speed. Extraction with CAINES was faster than the manual extraction process. The time difference between CAINES and the manual was not very large probably because users were limited to 10 minutes for each extraction method. Thirty two out of the forty four participants spent the entire 10 minute period on the manual extraction questions. Only five participants required the entire 10 minute period using CAINES. In addition, one participant commented that they used the search feature in their Internet browser (Control + F) to find the answers to the manual questions. This could be another explanation for the small gap in processing time between the two methods. Lastly, the sample consisted of business students and

therefore they may have answered some of the questions based on their common business knowledge. Thus, they may have answered the questions based on financial knowledge and not looked at the reports in detail.

Users extracted more relevant information with CAINES in its 6 minutes than they did during the manual process in 9 minutes. This time savings seemed to result in a larger amount of relevant information being processed, thus recall was 85.91% with CAINES versus 10.45% with the manual extraction process.

Regarding precision or accuracy of the manual process, three participants were able to achieve a 100% level. One participant answered three questions and those three were correct; two participants answered two questions and both were correct. The minimum precision for CAINES was 40%. Twenty-four participants were able to obtain 100% recall and precision levels using CAINES. CAINES clearly outperformed manual processing and users were satisfied with their experience with CAINES.

Thirty-five out of the forty-four participants left comments about their experience. Many participants alluded that the manual extraction process was “arduous”, “tedious”, “cumbersome”, and “frustrating” and that CAINES “saves a lot of time”, “more efficient and a productive use of time”, and that it made the search “more condensed”. Two participants commented:

*“Manually looking for the answers was very time consuming and difficult. CAINES made breaking down the financial reports very simple. I wouldn't evaluate a financial report without CAINES again.”*

*“With technology being as fast as it is today, manually looking through will keep you one step behind your competition. While using CAINES I felt a step ahead.”*

Some participants thought that CAINES could be more flexible to interact with. This was the main critique of CAINES. Two participants commented:

*“I feel that caines is a bit limited in what you can search for in the financial reports”*

*“Manually searching was difficult because there is a large amount of data and the information that I need is contextually embedded in sentences and paragraphs that are difficult to dissect. CAINES was easier because it had the information categorized in search categories. However, if CAINES did not have a category that I felt captured the information that I was looking for, I believe CAINES would become quite difficult to use. On the other hand, Manually searching the information made me feel as though I would never find it, whereas with CAINES I believed that I would find it eventually if I clicked on the right link or selected the correct search topic.”*

Even with a few critiques, users were very satisfied with their experience with CAINES. CAINES brings a new level of performance and satisfaction because it is more advanced than similar systems. Like FIRST, CAINES extracts information from online business reports. FIRST extracted short articles that were about half a page long that consisted of structured financial data and converted it into XML for use in business applications. CAINES is more advanced than FIRST in that it can extract unstructured information from discussion text in lengthy reports. With high accuracy, CAINES can return relevant semantic phrases via a Web based user interface.

EES is a system that extracts text from the Edgar database like CAINES. However, EES extracts structured information mainly from tables such as pro-forma information, earnings per share, the fair value of the options, and the model used to calculate the fair value. CAINES extracts unstructured information. EES recall, precision, and F-measure were 82.71%, 72.62%, and 77.34%, respectively. CAINES outperforms EES with recall, precision, and F-measure at 85.91%, 87.16%, and 86.46% respectfully.

Contrary to semantic and ontology based studies, CAINES follows the W3C suggested is-a, has-a type ontology triples. DySE uses a query structure of a list of terms (subject\_keyword) and a domain of interest (domain\_keyword). RelExt extracts domain specific verbal relations. DySE recall was 15% and precision was 90% and when recall was high at 90%, precision was merely 40%. RelExt recall and precision were only 36% and 23.9% respectfully. CAINES outperforms all of these systems with recall, precision, and F-measure at 85.91%, 87.16%, and 86.46% respectfully.

In today's fast moving business environment, professionals need to scan many types of online reports quickly and efficiently. Our study suggests that users would find value in a system, such as CAINES, that quickly extracts relevant semantic information from online reports.

## CHAPTER 7

### CONCLUSION

#### 7.1 *Implications*

Extracting semantic information from online business reports is a challenging but important application for IE and the future of Semantic Web Agents. It will only become more important as the size of Web documents continue to grow with overlapping and conceptually equivalent facts. This work demonstrates that semantic relationships and ontology information can be extracted from online financial reports. Semantic and ontological information can then be discovered and search by computer agents such as CAINES. CAINES can be useful to business managers, analysts, lenders, shareholders, and potential investors who want to quickly process online financial data. CAINES can help decision makers because CAINES can find relevant information from the most recent reports quickly. Table 13 summarizes the results of this study. Table 14 displays a comparison of CAINES' performance to that of similar extraction systems.

**Table 13. Summary of Results**

	Manual	CAINES	Difference
User Information Processing Speed	556.39	370.61	Significant
Recall	10.45%	85.9%	Significant
Precision	16.86%	87.2%	Significant
F-measure	12.59%	86.5%	Significant



**Table 14. Comparison of CAINES to Similar Extraction Systems**

	EES	DySE	RelExt	CAINES
System Speed	36 seconds	unknown	unknown	3 seconds
Recall	82.71%	15%	36%	85.9%
Precision	72.62%	90%	23.9%	87.2%
Usefulness	69%	unknown	unknown	91.5%
Ease of Use	unknown	unknown	unknown	85.9%
Satisfaction	unknown	unknown	unknown	90%

In response to the research questions, CAINES was able to accurately extract semantic based information from online financial reports and allowed users to perform better than people extracting Semantic Web information manually. With CAINES, a Semantic Web was created. To make the Web of data a reality, two things are important according to W3C. First, the data that is currently on the Web must be made into a standard format that Semantic Web tools can reach and manage. Second, the Semantic Web needs access to the relationships among the data, to create a Web of Data, not a collection of datasets. CAINES has demonstrated the usefulness of an IE system in following the standard format, extracting relationships, and searching the semantic based information. In addition, users overwhelmingly agreed that CAINES was useful and easy to use, and they were satisfied with the system.

## **7.2 *Limitations and Future Work***

This dissertation found that CAINES was more accurate in retrieving relevant information and in less time than manually searching and extracting or keyword searching. The target population of CAINES in the business world is those who need to access financial reports. Although statistical power tests showed that the sample size was adequate, larger samples with a population of financial professionals could be used to replicate the results. However, this study shows that CAINES is great even for novice users. One could expect greater results with financial professionals who are well versed in the financial domain.

CAINES could be useful in extraction of other important information in domains such as health care. CAINES can assist the health care community with deep content analysis of treatment databases or semantic based extraction of specific health information. In sum, CAINES is a system that can be customized and scaled to accommodate extraction and analysis of many sources of online text.

## BIBLIOGRAPHY

## BIBLIOGRAPHY

- Adams, C. (2001). The web as a database: New extraction technologies and content management, *Online*, 25, 27-32.
- Adams, D. A., Nelson, R. R., & Todd, P. A. (1992). Perceived usefulness, ease of use and use information technology, *MIS Quarterly* 16(2) 227-250.
- Adam, N., Gangopadhyay, A., & Clifford, J. (1994). A form-based approach to natural language query processing. *Journal of Management Information Systems*, 11(2), 109-135.
- Aggarwal, P., Vaidyanathan, R., & Venkatesh, A. (2009). Using lexical semantic analysis to derive online brand Positions: An application to retail marketing research. *Journal of Retailing*, 8(2), 145-158.
- Andersen, P., Hayes, P., Huettner, A., Schmandt, L., Nirenburg, I., & WeinStein, S. (1992). Automatic extraction of facts from press releases to generate news stories. *Processing of the 3<sup>rd</sup> Conference on Applied Natural Language Processing*, 170-177.
- Appelt, D. & Israel, D. (1999). *Introduction to information extraction technology*. Retrieved February 19, 2006, from <http://www.ai.sri.com/~appelt/ie-tutorial/>.
- Bagga, A., Chai, J. & Biermann, A. (1996). The role of WordNet in the creation of a trainable message understanding system. *Proceedings of the 13th National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, 941-948.
- Bardi, A., Calogero, R., & Mullen, B. (2008). A new archival approach to the study of values and value--behavior relations: Validation of the value lexicon. *Journal of Applied Psychology*, 93(3), 483-497.
- Baziz, M., Boughanem, M., Aussenac-Gilles, N., & Chrisment, C. (2005). Semantic cores for representing documents in IR. *Proceedings of the ACM Symposium on Applied Computing (SAC'05)*, 1011-1017.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 34-43.
- BBC NEWS. (2008). *Timeline: Sub-prime losses*. Retrieved from <http://news.bbc.co.uk/2/hi/business/7096845.stm>
- Burke, S. (2002). Perl and LWP, Ed. Nathan Torkington. O'Reilly & Associates, Sebastopol, CA.

- Cady, S. & Hardalupas, L. (1999). A lexicon for organizational change: Examining the use of language in popular, practitioner, and scholar periodicals, *Journal of Applied Business Research*, 15(4), 81.
- Cahn, S.M. (2002). *Classics of Western Philosophy* (Sixth ed.). Hackett Publishing Company.
- Ceci, M., Malerba, D., & Tanca, L. (2007). *DB2OWL: A tool for automatic database-to-ontology mapping*. Proceedings of the Fifteenth Italian Symposium on Advanced Database Systems, SEBD 2007, 17-20 June 2007, Torre Canne, Fasano, BR, Italy.
- Chen, H. (2003a). Introduction to the JASIST Special topic section on web retrieval and mining: A machine learning perspective, *Journal of the American Society for Information Science and Technology*, 54(7), 621-624.
- Chen, H. (2003b). Web retrieval and mining, *Decision Support Systems*, 35(1), 1-5.
- Clarkson, P. & Rosendfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2707-2710.
- Conlon, S., Hale, J., Lukose, S., & Strong, J. (2008). Information extraction agents for service-oriented architecture using web service systems: A framework. *Journal of Computer Information Systems*, 48(3), 74-83.
- Conlon, S., Lukose, S., Hale, J., & Vinjamur, A. (2007). Automatically extracting and tagging business information for e-business systems using syntactic and semantic analysis. *Semantic Web Technologies and eBusiness: Virtual Organization and Business Process Automation*. A. F. Salam and Jason Steven edited. Idea Group Inc, 101-126.
- Cooksey, R., Gates, G., & Pollock, H. (1998). 'Unsafe' business acts and outcomes: A management lexicon. *Business Horizons*, 41(3), 41.
- Corby, O., Dieng-Kuntz, R., Gandon, F., & Faron-Zucker, C. (2006). Searching the semantic Web: approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21(1), 20-27.
- Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1), 80-91.
- Cunningham, H. (1999). *Information extraction - A user guide* (second edition). Retrieved from: <http://www.dcs.shef.ac.uk/~hamish/IE/userguide/main.html>.
- Davis, F. D. (1986). A technology acceptance model for empirically testing new end-user information systems: Theory and results. (*Doctoral dissertation, Sloan School of Management, Massachusetts Institute of Technology*).

- Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, 13(3), 319-340.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.
- de Kunder, M. (2010). *The size of the world wide web*. Retrieved from <http://www.worldwidewebsite.com/>.
- De Nicola, A., Missikoff, M. & Navigli, R. (2009). A software engineering approach to ontology building, *Information Systems*, 34(2), 258-275.
- Etzioni, O., Banko, M., Soderland, S., & Weld, D. (2008). Open Information Extraction from the Web. *Communications of the ACM*, 51(12), 68-74.
- Etzioni, O., Cafarella, M., Downey, D., Popescue, A.M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A., (2004). Methods for domain-independent information extraction from the web: An experimental comparison, *Aaai Conference On Artificial Intelligence*, 391-398.
- Fernández, M., Gómez-Pérez, A., & Juristo, N. (1997). METHONTOLOGY: From ontological art towards ontological engineering. *Symposium on Ontological Engineering of AAAI*. Stanford (California).
- Fishbein, M. & Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley Publishing Company, Reading, MA.
- Frakes, W. & Terry, C. (1996). Software reuse: metrics and models. *ACM Computing Surveys*, 28, 415-435.
- Gaizauskas, R. & Wilks, Y. (1998). Information extraction: Beyond document retrieval. *Journal of Documentation*, 54(1) 70-105.
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Inexperience and experience with online stores: The importance of TAM and trust. *IEEE Transactions on Engineering Management*, 50(3), 307-321.
- Gerdes, J. (2003). Edgar-Analyzer: Automating the analysis of corporate data contained in the SEC's Edgar Database, *Decision Support Systems*, 35, 7-9.
- Goodwin, N.C. (1987). Functionality and usability. *Communications of the ACM*, 30, 229-233.
- Grant, G. & Conlon, S. (2006). Edgar extraction system: An automated approach to analyze employee stock option disclosures. *Journal of Information Systems*, 20 (2), 119-142.
- Gregg, D.G. & Walczak, S. (2006). Adaptive web information extraction. *CACM*, 49(5).

- Gresser, J., Haemmerli, B., Morrow, S. Sullivan, K., & Arend, H. (2010). Parsifal D2.1 draft ontology of financial risks & dependencies within and without the financial sector V2.0 (ontologies), *Telecommunications Software and Systems Group*, retrieved February 21, 2011 from <http://www.tssg.org/archives/2009/07/parsifal.html>
- Grishman, R. & Sundheim, B. (1996). Message understanding conference - 6: A brief history. In: *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 466–471.
- Hendler, J. & Berners-Lee, T. (2010). From the semantic web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence*, 174(2), 156-161.
- Hendler, J. (2009). Web 3.0 emerging. *IEEE Computer* 42(1).
- Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM*. 51(12), 58-67.
- Jacobs, P. & Rau, L. (1990a). The GE NLtoolset: A software foundation of intelligent text processing, *Proceedings of the 13th International Conference on Computational Linguistics*, 3 373-375.
- Jacobs, P. & Rau, L. (1990b). Extracting information from on-line news. *Communications of the ACM*, 3(11) 88-97.
- Jacobs, P. & Rau, L. (1998). Natural language techniques for intelligent information retrieval. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval*, ACM Press, 85-99.
- Jacobson, I., Booch, G. & Rumbaugh, J. (1999). *The Unified Software Development Process*, Addison Wesley, USA.
- Jain, A. & Ipeirotis, P. (2009). A quality-aware optimizer for information extraction. *ACM Transactions on Database Systems*, 31(1), 5-5:48.
- Karahanna, E & Straub, D. (1999). The psychological origins of perceived usefulness and perceived ease-of-use, *Information & Management*, 35, 237-250.
- Lee, Younghwa, Kozar, K. A. & Larsen, K. R.T. (2003). The Technology Acceptance Model: Past, Present, and Future, *Communication of the AIS*, (12), 725 -780
- Lehnert, G. (1991). Symbolic/subsymbolic sentence analysis: Exploiting the best of two worlds. *Advances in Connectionist and Neural Computation Theory*, 1, 135-164.
- Leinmann, C., Schlottmann, F., & Seese, D., Stuempert, T. (2001). Automatic extraction and analysis of financial data from the EDGAR database, *South African Journal of Information Management*, 3(2).

- Luhn, H. (1960). Keyword-in-context index for technical literature (KWIC Index), *American Documentation*, 11, 228-295.
- Lytinen, S. & Gershman, A. (1993). ATRANS: Automatic processing of money transfer messages. *Proceedings of the 5<sup>th</sup> National Conference of the American Association of Artificial Intelligence*, IEEE Computer Society Press, pp. 93-99.
- Manning, D. & Schutse, H. (2002). *Foundations of Statistical Natural Language Processing* (5th ed.). Cambridge, MA: The MIT Press.
- Manning, C.D., Raghavan, P. & Schütze, H., (2008). *Introduction to information retrieval*. Cambridge University Press.
- McCarthy, J. & Lehnert, W. (1995). Using decision trees for coreference resolution. *Proceedings of the 14th International Conference on Artificial Intelligence*. Ed. C. Mellish, 1050-1055.
- Miller, G. (1995). Wordnet: A lexical database for English. *Communication of the ACM*, 38 (11) 39-41.
- National Institute of Standards and Technology Association, *NIST Overview*. (2000). Retrieved February 22, 2010 from <http://trec.nist.gov/overview.html>.
- National Institute of Standards and Technology Association, *Tipster Text Program*. (2001). Retrieved February 3, 2010 from [http://www-nlpir.nist.gov/related\\_projects/tipster/](http://www-nlpir.nist.gov/related_projects/tipster/)
- Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., & Swartout, W.R. (1991). Enabling technology for knowledge sharing. *AI Mag.* 12(3), 36–56.
- Noy, N. F. & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology, *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*.
- Pennebaker, J. W. & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77, 1296–1312.
- Pentland, B. (1995). Grammatical models of organizational processes. *Organization Science*, 6(5), 541-556.
- Rinaldi, A. M. (2009). An ontology-driven approach for semantic information retrieval on the Web. *ACM Trans. Internet Technology*, 9(3).
- Rosch, E. (1978). *Principles of Categorization. Cognition and Categorization*. R. E. and B. B. Lloyd, editors. Hillside, NJ, Lawrence Erlbaum Publishers: 27-48.



- Sager, N., Friedman C., & Lyman, M. (1987). *Medical Language Processing: Computer Management of Limited Data*. Addison Wisely, Reading, MA.
- Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, 1(3).
- Schwartz, S. H. (1992). *Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries*. In M. Zanna (Ed.), *Advances in experimental social psychology*, New York: Academic Press, 25, pp. 1–65.
- Sheikh, M. (2009). A Methodology to Improve the Performance of Extracting Information from Financial Documents, doctoral dissertation, The University of Mississippi.
- Schutz, A. & Buitelaar, P. (2005). *RelExt: A Tool for Relation Extraction from Text in Ontology Extension*. The Semantic Web – ISWC 2005. Springer Berlin / Heidelberg. 3279/2005.
- Simmons, L.L., Yang, J., Mukhopadhyay, S., & Conlon, S. J. (2009). Driving forces of consumers' online reviews: An empirical study of the movie industry. *Decision Sciences Institute*, New Orleans, LA, November 14-17, 2009.
- Simperl, E. (2009). Reusing ontologies on the Semantic Web: A feasibility study. *Data & Knowledge Engineering*. (68), 905–925.
- Staab, S. & Studer, R. (2004). *Handbook on Ontologies*. International Handbooks on Information Systems. Springer-Verlag.
- Straub, D., Keil, M., & Brennan, W. (1997) Testing the technology acceptance model across cultures: a three country study. *Information & Management* 33(1), 1–11.
- Strickland, J. (2008). *How Web 3.0 Will Work*. Retrieved March 1, 2010 from HowStuffWorks.com. <http://computer.howstuffworks.com/web-30.htm>
- Uschold, M. & Gruninger, M. (1996). Ontologies: Principles, methods and applications, *Knowledge Engineering Review* 11(2).
- Uschold, M. & King, M. (1995). Towards a methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing in IJCAI 1995*, Montreal, Canada.
- U.S. Securities and Exchange Commission. (2010). *The investor's advocate: How the SEC protects investors, maintains market integrity, and facilitates capital formation*. Retrieved February 1, 2010 from: <http://www.sec.gov/about/whatwedo.shtml#create>
- U.S. Securities and Exchange Commission. (2008). *2008 Performance and accountability report*. Retrieved from <http://www.sec.gov/about/secpar/secpar2008.pdf#sec1>
- U.S. Securities and Exchange Commission. (2009). *2009 Performance and accountability report*. Retrieved from <http://www.sec.gov/about/secpar/secpar2008.pdf#sec1>

Venkatesh, V. & Davis, F.D. (1996). A model of the antecedents of perceived ease of use: Development and test, *Decision Sciences*, 27(3), 1996, 451-481.

What is the Semantic Web? Retrieved from [http://www.altova.com/semantic\\_web.html](http://www.altova.com/semantic_web.html).

W3C Mission. (2009). Retrieved from <http://www.w3.org/Consortium/mission>

Wilks, Y. (1997). *Information Extraction as a Core Language Technology*. M-T. Pazienza (ed.): Springer, Berlin.

Xu, J. & Croft, W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, 16(1) 61-81.

## APPENDIX

**Appendix I.:** Biographies of financial experts interviewed for this dissertation

Expert 1	Willy Johnson is in the Retail Banking Division with SunTrust Bank. His position is Area Manager which means he is responsible for the growth and profitability of branch offices in the Memphis, TN Region. This includes evaluating and implementing strategies for profitable growth and client development. As a part of these duties he reviews balance sheet and income statement performance on a regular basis to determine performance to plan. He is also responsible for understanding what economic and market conditions are impacting the growth of SunTrust's balance sheet and earnings.
Expert 2	Marvin Green has been in banking (mostly commercial) for just over 13 years. He started on the management training program of a Mississippi based bank with about \$8 billion in assets. During his 5 1/2 year stint there, he was placed in the Commercial Credit Analysis Department and left the bank as a Senior Commercial Credit Analyst. After that, he spent a little over a year at a southeastern regional bank (based out of Tennessee) with about \$40 billion in assets as Commercial Loan Review Officer. His current institution is a Tennessee based bank with about \$38 billion in assets and also does business in the contiguous states. He spent about 5 1/2 years as a Commercial Credit Review Officer with his current institution until last fall. At the present, he is employed as a Senior Audit Officer in the Internal Audit Department and work on the Capital Markets/Credit "team". He holds the Commercial Credit Risk Certification "CRC".
Expert 3	Colleen Mpofu has over 9 years of experience in Accounting and Finance. Throughout her career she has gained solid experience in accounting for mid to large manufacturing and service companies as well as small businesses. She is currently a Senior Cost Accountant at Pepsico Beverages and Foods focusing on inventory management, cost analysis, budgeting, forecasting, new employee training and development and SAP implementation. In 2005, she founded Elite Pro Solutions to provide tax preparation, bookkeeping, and start-up services to individuals and small businesses in the beauty, retail, and other service industries. Previously, she worked as a Senior Accountant for Enterprise Rent-A-Car preparing financial statements and expense analysis for management. Colleen has an MBA from Indiana Wesleyan University, a BS in Accounting and Finance from Indiana University, and is currently studying to become a Certified Public Accountant.



## Appendix II.: Finance terms ontology

Accrued interest	is	The interest due on a bond since the last interest payment was made.
American Depositary Receipt (ADR)	is	a security issued by a U.S. bank in place of the foreign shares held in trust by that bank,
American Stock Exchange (AMEX)	is	The second largest stock exchange in the United States, located in the financial district of New York City.
Amortization	is	Accounting for expenses or charges as applicable rather than as paid. Includes
Annual report	is	The formal financial statement issued yearly by a corporation.
Arbitrage	is	A technique employed to take advantage of differences in price.
Assets	is	Everything a corporation owns or that is due to it: cash,
Auction market	is	The system of trading securities through brokers or agents on an exchange
Auditor's report	is	Often called the accountant's opinion, it is the statement of the accounting firm's work
Averages	is	Various ways of measuring the trend of securities prices, one of the most popular of which is the Dow Jones Industrial Average of 30 industrial
Balance sheet	is	A condensed financial statement showing the nature and amount of a company's assets,
Basis point	is	One gradation on a 100-point scale representing 1%; used especially in expressing variations in the
Bear	is	Someone who believes the market will decline. (See: Bull)
Bear market	is	A declining market. (See: Bull market)
Bearer bond	is	A bond that does not have the owner's name registered on the books of the issuer.
Block	is	A large holding or transaction of stock popularly considered to be 10,000 shares or more.
Blue Sky Laws	is	A popular name for laws various states have enacted
Blue chip	is	A company known nationally for the quality and wide acceptance of its products or services,
Book value	is	An accounting term. Book value of a stock is determined from a company's records,
Broker	is	An agent who handles the public's orders to buy and sell
Brokers' loans	is	Money borrowed by brokers from banks or other brokers for a variety of uses.
Bull market	is	An advancing market. (See: Bear market)

Buy side	is	The portion of the securities business in which institutional orders originate.
Callable	is	A bond issue, all or part of which may be redeemed
Capital gain or capital loss	is	Profit or loss from the sale of a capital asset. The capital
Capital stock	is	All shares representing ownership of a business, including preferred and common. (See: Common stock, Preferred
Capitalization	is	Total amount of the various securities issued by a corporation. Capitalization
Cash flow	is	Reported net income of a corporation plus amounts charged off for depreciation, depletion, amortization, and
Cash sale	is	A transaction on the floor of the stock exchange that calls for delivery of the
Certificate	is	The actual piece of paper that is evidence of ownership of
Equipment trust certificate	is	A type of security, generally issued by a railroad,
Equity	is	The ownership interest of common and preferred stockholders in a company.
Ex-dividend	is	A synonym for "without dividend." The buyer of a stock selling
Ex-rights	is	Without the rights. Corporations raising additional money may do so by
Extra	is	The short form of "extra dividend." A dividend in the form
Face value	is	The value of a bond that appears on the face of the bond, unless
Face value	is	ordinarily the amount the issuing company promises to pay at maturity. Face value is
Face value	is	not an indication of market value. Sometimes referred to as par value. (See: Par)
Fiscal year	is	A corporation's accounting year. Due to the nature of their particular business, some companies do
Fixed charges	is	A company's fixed expenses, such as bond interest, which it has agreed to pay whether
Flat income bond	is	This term means that the price at which a
Floor	is	The huge trading area - about the size of a football
Floor broker	is	A member of the stock exchange who executes orders on the floor of the Exchange
Formula investing	is	An investment technique. One formula calls for the shifting of funds from common shares to
Free and open market	is	A market in which supply and demand are freely expressed in terms of

Fundamental research	is	Analysis of industries and companies based on such factors as sales, assets, earnings, products or
Funded debt	is	Usually interest-bearing bonds or debentures of a company. Could include long-term bank loans. Does not
General mortgage bond	is	A bond that is secured by a blanket mortgage
Gilt-edged	is	High-grade bond issued by a company that has demonstrated its ability
Give-up	is	A term with many different meanings. For one, a member of
Gold fix	is	The setting of the price of gold by dealers (especially in a twice-daily London meeting
Good delivery	is	Certain basic qualifications must be met before a security sold on the Exchange may be
Government bonds	is	Obligations of the U.S. Government, regarded as the highest grade securities issues.
Growth stock	is	Stock of a company with a record of growth in earnings at a relatively rapid
Holding company	is	A corporation that owns the securities of another, in most cases with voting control.
Hypothecation	is	The pledging of securities as collateral - for example, to secure
IRA	is	Individual retirement account. A pension plan with tax advantages. IRAs permit
Income bond	is	Generally income bonds promise to repay principal but to pay interest only when earned. In
Indenture	is	A written agreement under which bonds and debentures are issued, setting
Independent broker	is	Member on the floor of the NYSE who executes orders for other brokers having more
Index	is	A statistical yardstick expressed in terms of percentages of a base
Index	is	The composite index covering price movements of all common stocks listed
Initial public offering	is	(See: Primary distribution)
Institutional investor	is	An organization whose primary purpose is to invest its own assets or those held in
Interest	is	Payments borrowers pay lenders for the use of their money. A
Intermarket Trading System (ITS)	is	An electronic communications network now linking the trading floor of seven registered exchanges
Interrogation device	is	A computer terminal that provides market information - last sale price, quotes, volume, etc. -
Investment	is	The use of money for the purpose of making more money,



Investment banker	is	Also known as an underwriter. The middleman between the corporation issuing new securities and the
Investment company	is	A company or trust that uses its capital to invest in other companies. There are
Investment counsel	is	One whose principal business consists of acting as investment advisor and rendering investment supervisory services.
Issue	is	Any of a company's securities, or the act of distributing such
Keogh plan	is	Tax-advantaged personal retirement program that can be established by a selfemployed individual. (See: IRA)
Legal list	is	A list of investments selected by various states in which certain institutions and fiduciaries, such
Leverage	is	The effect on a company when the company has bonds, preferred
Liabilities	is	All the claims against a corporation. Liabilities include accounts, wages and
Liquidation	is	The process of converting securities or other property into cash. The
Liquidity	is	The ability of the market in a particular security to absorb a reasonable amount of buying or selling at reasonable price changes.
Liquidity	is	one of the most important characteristics of a good market.
Listed stock	is	The stock of a company that is traded on a securities exchange.
Load	is	The portion of the offering price of shares of open-end investment
Locked in	is	Investors are said to be locked in when they have profit on a security they
Long	is	Signifies ownership of securities. "I am long 100 U.S. steel" means
Manipulation	is	An illegal operation. Buying or selling a security for the purpose
Margin	is	The amount paid by the customer when using a broker's credit
Margin call	is	A demand upon a customer to put up money or securities with the broker. The
Market order	is	An order to buy or sell a stated amount of a security at the most
Market price	is	The last reported price at which the stock or bond sold, or the current quote.(See:
Maturity	is	The date on which a loan or bond comes due and
Member corporation	is	A securities brokerage firm, organized as a corporation, with at least one
Member firm	is	A securities brokerage firm organized as a partnership and having at least
Member organization	is	The term includes New York Stock Exchange member firms and

		member corporations.
Money market fund	is	A mutual fund whose investments are in high-yield money
Mortgage bond	is	A bond secured by a mortgage on a property. The value of the property may
Municipal bond	is	A bond issued by a state or a political subdivision, such as county, city, town
Mutual fund	is	(See: Investment company)
NASD	is	please refer to the details listed above for FINRA.
NYSE Composite Index	is	The composite index covering price movements of all common
NYSE common stocks	is	based on 1965 as 50. An index is not
Nasdaq	is	An automated information network that provides brokers and dealers with price quotations on securities traded over-the-counter.
Negotiable	is	Refers to a security, the title to which is transferable by delivery.
Net asset value	is	Usually used in connection with investment companies to mean
Net change	is	The change in the price of a security from the closing price on one day
New York Futures Exchange (NYFE)	is	A subsidiary of the New York Stock Exchange devoted to the trading
New York Stock Exchange (NYSE)	is	The largest organized securities market in the United States, founded in 1792.
New issue	is	A stock or bond sold by a corporation for the first time. Proceeds may be
Noncumulative	is	A type of preferred stock on which unpaid dividends do not
Odd Lot	is	An amount of stock less than the established 100-share unit. (See: Round lot)
Off-board	is	This term may refer to transactions over-the-counter in unlisted securities or
Offer	is	The price at which a person is ready to sell. Opposed
Open order	is	(See: Good 'til canceled order)
Open-end investment company	is	(See: Investment company)
Over-the-counter	is	A market for securities made up of securities dealers who may
Overbought	is	An opinion as to price levels. May refer to a security
Oversold	is	The reverse of overbought. A single security or a market which,
Paper profit (loss)	is	An unrealized profit or loss on a security still
Par	is	In the case of a common share, par means a dollar

Participating preferred	is	A preferred stock that is entitled to its stated dividend and to additional dividends on
Passed dividend	is	Omission of a regular or scheduled dividend.
Penny stocks	is	Low-priced issues, often highly speculative, selling at less than \$1 a share. Frequently used as
Point	is	In the case of shares of stock, a point means \$1.
Portfolio	is	Holdings of securities by an individual or institution. A portfolio may
Preferred stock	is	A class of stock with a claim on the company's earnings before payment may be
Premium	is	The amount by which a bond or preferred stock may sell
Price-to-earnings ratio	is	A popular way to compare stocks selling at various price levels. The P/E ratio is
Primary distribution	is	Also called primary or initial public offering. The original sale of a company's securities. (See:
Prime rate	is	The lowest interest rate charged by commercial banks to their most creditworthy customers; other interest
Principal	is	The person for whom a broker executes an order, or dealers
Prospectus	is	The official selling circular that must be given to purchasers of
Proxy statement	is	Information given to stockholders in conjunction with the solicitation of proxies.
Prudent Man Rule	is	An investment standard. In some states, the law requires
Quote	is	The highest bid to buy and the lowest offer to sell
Rally	is	A brisk rise following a decline in the general price level
Real Estate Investment Trust (REIT)	is	An organization similar to an investment company in some respects but concentrating
Record date	is	The date on which you must be registered as a shareholder of a company in
Red herring	is	A registration statement filed with but not yet approved by the Securities and Exchange Commission
Redemption price	is	The price at which a bond may be redeemed before maturity, at the option of
Refinancing	is	Same as refunding. New securities are sold by a company and
Registered bond	is	A bond that is registered on the books of the issuing company in the name
Registered competitive market maker	is	Members of the New York Stock Exchange who trade on the floor for
Registered representative	is	The man or woman who serves the investor customers of a broker/dealer. In a New

Registrar	is	Usually a trust company or bank charged with the responsibility of
Registration	is	Before an initial public offering may be made of new securities
Regular way delivery	is	Unless otherwise specified, securities sold on the New York
Regulation T	is	The federal regulation governing the amount of credit that may be advanced by brokers and
Regulation U	is	The federal regulation governing the amount of credit that may be advanced by banks to
Rights	is	When a company wants to raise more funds by issuing additional
Round lot	is	A unit of trading or a multiple thereof. On the NYSE, the unit of trading
SEC	is	The Securities and Exchange Commission, established by Congress to help protect
Scale order	is	An order to buy (or sell) a security, that specifies the total amount to be
Scripophily	is	A term coined in the mid-1970s to describe the hobby of
Seat	is	A traditional figure of speech for a membership on an exchange.
Secondary distribution	is	Also known as secondary offering. The redistribution of a block of stock some time after
Sell side	is	The portion of the securities business in which orders are transacted. The sell side includes retail
Seller's option	is	A special transaction on the NYSE that gives the seller the right to deliver the
Serial bond	is	An issue that matures in part at periodic stated intervals.
Serial bond	is	An issue that matures in part at periodic stated
Settlement	is	Conclusion of a securities transaction when a customer pays a broker/dealer
Short covering	is	Buying stock to return stock previously borrowed to make delivery on a short sale.
Short sale	is	A transaction by a person who believes a security will decline and sells it,
Sinking fund	is	Money regularly set aside by a company to redeem its bonds, debentures or preferred stock
Specialist	is	A member of the New York Stock Exchange who has two
Speculation	is	The employment of funds by a speculator. Safety of principal is
Speculator	is	One who is willing to assume a relatively large risk in
Spin off	is	The separation of a subsidiary or division of a corporation from its parent company by

Split	is	The division of the outstanding shares of a corporation into a
Stock Exchange (AMEX)	is	The second largest stock exchange in the United States,
Stock Exchange (NYSE)	is	The largest organized securities market in the United States,
Stock dividend	is	A dividend paid in securities rather than in cash. The dividend may be additional shares
Stock exchange	is	An organized marketplace for securities featured by the centralization of supply and demand for the
Stop limit order	is	A stop order that becomes a limit order after
Stop order	is	An order to buy at a price above or sell at a price below the
Street name	is	Securities held in the name of a broker instead of a customer's name are said
Swapping	is	Selling one security and buying a similar one almost at the
Syndicate	is	A group of investment bankers who together underwrite and distribute a
Technical research	is	Analysis of the market and stocks based on supply and demand. The technician studies price
Tender offer	is	A public offer to buy shares from existing stockholders of one public corporation by another
Third market	is	Trading of stock exchange-listed securities in the over-the-counter market by non-exchange member brokers.
Ticker	is	A telegraphic system that continuously provides the last sale prices and
Trader	is	Individuals who buy and sell for their own accounts for short-term
Trading floor	is	(See: Floor)
Trading post	is	The structure
Transfer	is	This term may refer to two different operations. For one, the
Transfer agent	is	A transfer agent keeps a record of the name of each registered shareowner, his or
Treasury stock	is	Stock issued by a company but later reacquired. It may be held in the company's
Turnover rate	is	The volume of shares traded in a year as a percentage of total shares listed
Underwriter	is	(See: Investment banker)
Unlisted stock	is	A security not listed on a stock exchange. (See: Over-the-counter)
Up tick	is	A term used to designate a transaction made at a price
Up tick	is	A term used to designate a transaction made at a price higher than

		the preceding
Variable annuity	is	A life insurance policy where the annuity premium (a set amount of dollars) is immediately
Volume	is	The number of shares or contracts traded in a security or
Volume	is	usually considered on a daily basis and a daily average is
Voting right	is	Common stockholders' right to vote their stock in affairs of a company. Preferred stock usually
Warrants	is	Certificates giving the holder the right to purchase securities at a
When issued	is	A short form of "when, as and if issued." The term indicates a conditional transaction
Working control	is	Theoretically, ownership of 51% of a company's voting stock is necessary to exercise control. In
Yield	is	Also known as return. The dividends or interest paid by a
Yield to maturity	is	The yield of a bond to maturity takes into
Zero coupon bond	is	A bond that pays no interest but is priced,

## VITA

### EDUCATION

PhD Management Information Systems, University of Mississippi, 2011

MBA Technology Management, University of Phoenix, 2004

BBA Information Systems, Tennessee State University, 2002

### JOURNAL PUBLICATIONS

1. **Simmons, L. L.**, Conlon, S.J., Mukhopadhyay, S., & Yang, J. (forthcoming). A computer aided analysis of eWOM reviews: Their content and impact. *Journal of Computer Information Systems*.
2. **Simmons, L.**, Thomas, C., & Tsuma, C. Mbarika, V. (2011, in press). TeleEducation initiatives for sub-saharan Africa: The case of the African virtual university in Kenya. *Journal of STEM Education: Innovations and Research*.
3. Aiken, M., **Simmons, L. L.**, & Balan, S. (2010). Automatic interpretation of English speech. *Issues in Information Systems*, 11, 1, 129-133.
4. **Simmons, L. L.** Simmons, C. B., Ammeter, A.P., Ghosh, K. (2010). Understanding the benefits of social exchange in B2B communities, *Issues in Information Systems*, 11, 1, 134-141.
5. **Simmons, L.**, & Aiken, M. (2010). Group support system meeting termination: Allowing participants to vote, *Business Research Yearbook Global Business Perspectives*, 17, 1, 94-99.
6. Aiken, M. & **Simmons, L.** (2010). Automatic transcription of spoken English to German and Spanish text, *Business Research Yearbook Global Business Perspectives*, 17, 1, 106-111.
7. **Simmons, L. L.** & Clayton, R. (2010). The impact of small business B2B virtual community commitment on brand loyalty. *International Journal of Business and Systems Research*, 4, 4. (Acceptance level: 11-20% per Cabell's).
8. Simmons, C.B. & **Simmons, L. L.** (2010). The changing requirements for computer science graduates. *The Journal of Computing Sciences in Colleges*, 25, 5.
9. Aiken, M., Park, M., **Simmons, L.**, & Lindblom, T., (2009, July). Automatic translation in multilingual electronic meetings. *Translation Journal*, 13, 4.

## CONFERENCE PRESENTATIONS

1. **Simmons, L.L.** & Simmons, C.B. (2010). Task-technology fit and self-efficacy: Impacts on trusting beliefs in eLearning. Proceedings of the 41<sup>st</sup> Annual Meeting (San Diego, CA) of the Decision Sciences Institute, Atlanta: Decision Sciences Institute, 2010, pp. 1121-1126.
2. Mukhopadhyay, S., Conlon, S.J., & **Simmons, L.L.** (2010). Online consumer feedback reviews – Mood categorization on content analysis. Abstract accepted for presentation at the 2010 annual meeting of the Decision Sciences Institute, San Diego, CA, November 20-23, 2010.
3. **Simmons, L.L.** Simmons, C. B., Ammeter, A.P., Ghosh, K. (2010). Understanding the benefits of social exchange in B2B communities. Referred paper presented at the 2010 International Association for Computer Information Systems, Las Vegas, NV, October 6-9, 2010.
4. Aiken, M., **Simmons, L.L.**, & Balan, S. (2010). Automatic interpretation of English speech. Refereed paper presented at the 2010 International Association for Computer Information Systems, Las Vegas, NV, October 6-9, 2010.
5. **Simmons, L.L.**, & Aiken, M. (2010). Group support system meeting termination: Allowing participants to vote. Refereed paper presented at the International Academy of Business Disciplines, Las Vegas, Nevada, April 8-10, 2010.
6. Aiken, M. & **Simmons, L.L.** (2010). Automatic transcription of spoken English to German and Spanish text. Refereed paper presented at the International Academy of Business Disciplines, Las Vegas, Nevada, April 8-10, 2010.
7. Simmons, C.B. & **Simmons, L.L.** (2010). The changing requirements for computer science graduates. Refereed paper presented at the annual meeting of the Consortium for Computing Sciences in Colleges Mid-South Conference, Searcy, AR, March 26-27, 2010.
8. **Simmons, L.L.**, Yang, J., Mukhopadhyay, S., Conlon, S. J., (2009). Driving forces of consumers' online reviews: An empirical study of the movie industry. Abstract presented at the annual meeting of Decision Sciences Institute, New Orleans, LA, November 14-17, 2009.
9. **Simmons, L.L.**, & Ammeter, A.P. (2008). The impact of social interaction on economic benefits of participation in B2B virtual communities. Refereed paper presented at the annual meeting of the Academy of Management - OCIS Division, Anaheim, CA, August 8-13, 2008.
10. Wilkerson, D., **Simmons, L.**, & Mbarika, V. (2008). TeleEducation initiatives for sub-Saharan Africa: The case of the African virtual university in Kenya. Refereed paper presented at the annual meeting of Multimedia Educational Resource for Learning and Online (MERLOT) International Conference, Minneapolis, MN, August 7-10, 2008.
11. **Simmons, L.L.** & Clayton, R. (2008). The impact of small business B2B virtual community commitment on brand loyalty. Refereed paper presented at the annual meeting of The PhD



Project - Information Systems Doctoral Student Association, Toronto, ON, August 11-13, 2008.

## **RESEARCH IN PROGRESS**

1. Bradley, R. V., Byrd, T. A., **Simmons, L. L.**, Pratt, R. Enterprise architecture: The panacea for healthcare IT value woes? Targeted for: MISQ Executive. Status: Data Analysis.
2. **Simmons, L. L.**, & Simmons, C. B. Culture and IT Artifact Trust in eLearning. Targeted for: CAIS. Status: Data Analysis.
3. **Simmons, L. L.** & Conlon, S. C. Ontology Extraction of Semantic Relationships using CAINES. Targeted for: Decision Support Systems. Status: Data Analysis.

## **WORK EXPERIENCE**

### **Academic Positions**

2009 – 2010 *Research Assistant*, International Center for IT and Development.

2008 – 2009 (spring), *Graduate Instructor*, University of Mississippi.  
Undergraduate Course: Management Information Systems II, two sections

2008 – 2009 (fall), *Graduate Teaching Assistant*, University of Mississippi.  
Undergraduate Course: Management Information Systems II, two sections

2007 – 2008 (fall), *Graduate Teaching Assistant*, University of Mississippi. Undergraduate Courses: Systems Analysis and Design and Web Application Programming

2006 - 2007, *Adult Education Teacher*, Nashville Davidson County Schools.  
Taught math, grammar, and literacy to students preparing to take the GED. Evaluated student learning, test readiness, and recommended additional learning.

### **Industry Positions**

2006 – 2007, *6 Sigma Black Belt*, Caterpillar Financial, Nashville, TN  
Responsible for facilitating and applying the 6 Sigma methodology to diverse improvement and implementation projects. Managed Sponsor relationships and communicated with all levels of management. Worked with Financial Representatives from accounting for project benefit approval. Trained project team members on 6 Sigma Green Belt methodology.

2004 – 2006, *IT Business Analyst*, Caterpillar Financial, Nashville, TN  
Responsible for gathering business requirements for IT development projects. Responsible for facilitating the communication between customers and IT developers.

2002 – 2004, *Programmer Analyst*, Caterpillar Financial, Nashville, TN  
Developed application integration interfaces using CrossWorlds. Created functional and technical design documents. Performed string, load, and component testing. Created service contracts, support plans, and test plans for assigned projects.

## **FELLOWSHIPS AND AWARDS**

2009- current ICITD Research and E-Learning Fellowship

2007- current University of Mississippi Graduate School Fellowship

## **HONOR SOCIETIES**

2008 Beta Gamma Phi

2002 Beta Gamma Sigma

2001 Phi Kappa Phi

## **ACADEMIC & PROFESSIONAL MEMBERSHIPS**

Association for Information Systems

Decision Sciences Institute

PhD Project Information Systems Doctoral Student Association (current Past-President)

Tennessee State University Memphis Alumni Council

## **ACADEMIC SERVICE**

1. *2010 Reviewer*, Decision Sciences Institute (DSI), Nov. 20, 2010  
Track: Cross-Functional Interfaces (Mrk/OM/Fin/IS/Acct)
2. *2010 Session Chair*, International Association of Business Disciplines (IABD)  
Track: Communication and Technology
3. *2010 Reviewer*, Consortium for Computing Sciences in Colleges Mid-South (CCSC- MS)
4. *2010 Reviewer*, International Conference on ICT for Africa, (ICIA-2010)
5. *2009-2010 President*, PhD Project Information Systems Doctoral Student Association
6. *2009 Reviewer*, European Conference on Information Systems (ECIS - 2010),  
Track: ICT in Africa and other economically developing regions
7. *2009 Session Chair*, Decision Sciences Institute (DSI), Nov. 15, 2009
8. *2009 Reviewer*, American Conference on Information Systems (AMCIS - 2009),
9. *2009 Reviewer*, Multimedia Educational Resource for Learning and Online Teaching (MERLOT) Tracks: Creative Collaborations and Innovations in Online Teaching and Learning
10. *2008-2009 Secretary*, PhD Project Information Systems Doctoral Student Association

## **INVITED TALKS**

*2009 Panelist*, PhD Project Conference, Nov. 19, 2009

Panel: Life as a Student

*2009 Panelist*, PhD Project Conference, Nov. 20, 2009

Panel: MIS Information Session

*2008 Panelist*, PhD Project Conference, Nov. 21, 2008

Panel: MIS Information Session

*2007 Speaker*, Tennessee State University Memphis Alumni Council

Talk: Secrets of the Successful College Student