

**EDGAR EXTRACTION SYSTEM: AN AUTOMATED APPROACH TO
ANALYZE EMPLOYEE STOCK OPTION DISCLOSURES**

A Dissertation

Presented for the

Doctor of Philosophy in Business Administration

Degree

The University of Mississippi

Gerry H. Grant

April 2005

UMI Number: 3190578

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3190578

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Information Extraction (IE) is the process of finding useful information in unstructured text, extracting specific data, and presenting the data in a summarized, structured format. One potential use of IE is extracting information about stock options that is embedded in the disclosure notes of financial statements. The Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) database is the richest source of financial statement information on the Web. However, the information is stored in text or HTML files making it difficult to search and extract data. This paper examines the development and use of the EDGAR Extraction System (EES), a customized automated system that extracts relevant information about the fair value of employee stock options from financial statement disclosure notes on the EDGAR database.

To the Graduate Council:

I am submitting herewith a dissertation written by Gerry H. Grant entitled "EDGAR Extraction System: An Automated Approach to Analyze Employee Stock Option Disclosures." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration with a major in Management Information Systems.

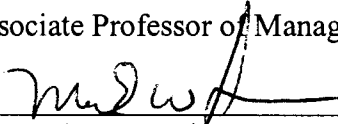


Dr. Sumali J. Conlon, Major Professor
Associate Professor of Management
Information Systems

We have read this dissertation
and recommend its acceptance:



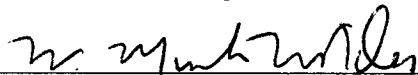
Dr. Milam Aiken
Associate Professor of Management Information Systems



Dr. Mark W. Lewis
Assistant Professor of Management Information Systems



Dr. Brian J. Reithel
Professor of Management Information Systems



Dr. W. Mark Wilder
Associate Professor of Accountancy

Accepted for the Council:



Dean of The Graduate School

Table of Contents

Chapter	Contents	Page
1.	Introduction	1
2.	Information Extraction	
2.0	Origin of Information Extraction	6
2.1	Early Information Extraction Systems	9
2.2	Types of IE Systems	10
2.3	Performance Measures for IE Systems	12
2.4	Information Extraction Systems and the World Wide Web	14
2.5	Information Extraction System Development	15
2.5.1	Approaches to IE Systems	16
2.5.2	Tools and Techniques used in IE Systems	17
2.6	Extensible Business Reporting Language (XBRL)	23
3.	Stock Options	
3.0	Overview	26
3.1	Accounting for Stock Options	27
3.1.1	Consistency Issues of Recent Changes to Stock Option Reporting Requirements	29
3.1.2	Comparability Issues of Prior Requirements for Stock Option Reporting	31
3.2	The EDGAR Database	32
3.3	Other Sources of Freely Disseminated Financial Statement Information on the Web	34
3.4	Current Systems that Extract Information From EDGAR	36
4.	Research Methodology	
4.0	Overview	38
4.1	Corpus Development.	38
4.1.0	Sample Selection	40
4.1.1	Corpus Building	47
4.2	Corpus Analysis	47
4.2.0	CMU-SLM Toolkit	48
4.2.1	KWIC and SQL	49
4.2.2	WordNet	53

4.2.3	Stemming	56
4.2.4	Knowledge Based Analysis	57
4.3	Extractor Development	58
4.3.0	Downloading	58
4.3.1	Parsing	60
4.3.2	Data Separation	61
4.3.3	Wrapper Development	62
4.3.4	Extraction and Results.....	62
5.	Testing and Analysis	
5.0	Overview	70
5.1	Testing	70
5.2	Analysis of Statistical Comparisons.....	78
5.3	Survey Analysis.....	81
6.	Conclusion, Limitations and Future Research	
6.0	Overview	83
6.1	Conclusion	83
6.2	Limitations	84
6.3	Future Research	85
	Bibliography	89
	Vita	97

List of Tables

1.	NASDAQ 100 Companies as of May 4, 2004	41
2.	NASDAQ-100 Index Companies used in Corpus with Corresponding SIC and CIK Codes.....	44
3.	NASDAQ-100 Index Companies used for Testing EES with Corresponding SIC and CIK Codes.....	46
4.	Relevant Bi-grams and Tri-grams used to Construct a Language Model for EES Development.....	50
5.	Data Output from the KWIC Index System	52
6.	SQL Output from the Corpus Database.....	54
7.	List of Synonyms Created by WordNet	55
8.	Recall, Precision and F-Measure for all 76 10-Ks Tested	71
9.	Descriptive Data from Manual and EES Extraction.....	79
10.	Results of Paired T-Tests	80
11.	Survey Results	82

List of Figures

1.	System Architecture of EES	39
2.	An Example of Information Available on Stock Options from a 10-K Annual Report on EDGAR.....	59
3.	An example of the Output File From EES Output.....	64
4.	Information Provided in the Header of a 10-K Annual Report on EDGAR	65
5.	An Example of EES Output when Financial Information is Incorporated by Reference in its Annual Report.....	66
6.	An Example of the Output from EES when the Financial Statements are Incorporated by Reference in the Annual Report to Shareholders	67
7.	Survey Administered to Subjects That Tested EES.....	75
8.	Template Used by Subjects for Manual Data Collection for Testing EES	77

Chapter 1

Introduction

Information Extraction (IE) is the process of finding useful information in unstructured text, extracting specific data, and presenting the data in a summarized, structured format. IE dates back to the Cold War era of the 1960s, but has gained importance due to the explosion of digital information available over the World Wide Web (Web). Considering the wealth of public information available on the Web and the growing number of Web users, it has become increasingly important to develop methods to extract Web information in a format that is easy to use [87].

An extension of natural language processing and artificial intelligence, IE is not an easy task. Accuracy results range from 40% [25] to 70% [16] in most systems. Many methods and techniques have been developed in an effort to make IE systems more accurate and portable. IE systems have numerous applications including underwriting analysis in the insurance industry, extraction of medical systems for diagnoses [50], analysis of news articles [46], classification of legal documents [62], and extraction of financial information from public sources [40].

One potential use of IE is to extract information about employee stock options that is embedded in the disclosure notes of financial statements. Stock options have become extremely controversial over the two decades and have been widely studied by both academics and professional analysts. Finding specific information about company stock options is difficult and time consuming. In addition, recent stock option reporting requirements will hamper consistency in financial reporting and will make future analysis even more complicated.

In December, 2004, The Financial Accounting Standards Board (FASB) issued Statement of Financial Accounting Standard (SFAS) No. 123(R) that requires mandatory expensing of stock options beginning the third quarter of 2005 [36]. This new rule will end a ten year heated debate over the accounting treatment of employee stock options and improve the quality of financial reporting [63]. However, expensing employee stock options beginning in 2005 will present problems with consistency in financial reporting.

Consistency implies that that the same accounting methods have been used over a span of time and is deemed by accounting experts to make financial reporting more useful. Inconsistencies can occur any time an accounting standard changes [31]. The FASB's new accounting standard allows, but does not require, prior financial statements to be restated, therefore, inconsistent accounting methods will hinder trend analysis of a company's financial statements prior to and after adoption [34].

Comparability is another problem associated with accounting for stock options. Comparability implies that there are common characteristics in financial statements that allow users to examine similarities and differences in them [31]. Prior accounting rules allowed companies to choose between expensing and not expensing the fair value of stock options. Companies that chose not to expense stock options prior to the new standard were required disclose in the notes to its financial statements a fair value estimate of its stock options, pro forma net income and earnings per share (EPS), as well as the method used to calculate the fair value and assumptions used in the calculation. Furthermore, companies are allowed, under prior and current standards, to choose among various models to value its options and to make estimates regarding the underlying assumptions used in the model. [32].

Because this disclosure information has been required since 1996, some experts suggest that issues of consistency and comparability can be resolved by analysis of the disclosures in the notes to the financial statements [15]. This is no easy task. Although company financial information can usually be found on the Web, the files are often lengthy and the data may be unstructured and difficult to locate.

The Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval System (EDGAR) is the richest source of freely disseminated U.S. financial information. However, the information is stored in text and HTML files making searching and extracting specific data in a useable format difficult. One solution to the problem of finding financial information from Web sources is to apply IE techniques to files on the EDGAR database.

The purpose of this study is to develop an automated system to extract stock option information from the disclosure notes to financial statements on the EDGAR Database. The research asks three questions:

(1) Can an automated system be designed to accurately extract stock option information?

(2) Can an automated system speed the tedious process of extracting stock option information from large, semi-structured WEB documents?

(3) Is the automated system perceived to be useful?

This paper discusses the development of Edgar Extraction System (EES), a system that extracts information about stock options from the disclosure notes of 10-K annual reports on the EDGAR database. The NASDAQ-100 Index companies were selected as a sample for building a non-annotated, domain specific corpus and for testing

the system. Various machine learning techniques combined with a knowledge based approach were used to analyze the patterns in the corpus. From this knowledge base, algorithms were 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. The system displays the information in a useful, structured format. 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 ($p < 0.001$).

This study makes several contributions to IE research. It builds on the use of corpus machine learning techniques and the knowledge based approach to develop IE systems for specific language domains. It expands current extraction systems and includes the development of financial information extraction from semi-structured Web documents. This study also provides evidence of the usefulness and accuracy of an automated approach to extract specific financial information from files on the SEC's EDGAR Database.

From a practical standpoint, this study provides a valuable tool for financial analysts. EES can help analysts compare financial statements of companies that

expensed stock options with companies that did not expense stock options prior to the adoption of SFAS No. 123(R). EES can be used by analysts to extract information to compare company financial data released before and after SFAS No. 123(R) becomes effective

Chapter 2 of this paper reviews the literature on IE systems, methods, and techniques. Chapter 3 examines the issues related to the current and future accounting treatment of stock options, the impact on users of financial statements, and the availability of financial information on the Web. The research methodology used to develop EES is explained in Chapter 4. Chapter 5 describes the testing of EES and provides statistical analysis. The last chapter, Chapter 6, concludes with a summary, a discussion of the limitations of EES, and suggestions for future research.

Chapter 2

Information Extraction

2.0 Origins of Information Extraction

IE is an outgrowth of Artificial Intelligence (AI) and Statistical Natural Language Processing (NLP). The origins of IE date back to the Cold War era of the 1960s. However, it was not until the late 1980s that research interests began to emerge. IE research has become attractive for several reasons. Extracting data from text presents challenging and interesting problems. Most extraction tasks are well defined since they are normally limited to specific task and domains and are developed for real world situations from real world texts. Also, IE system performance can be measured and compared to human performance on the same tasks [25]. Several organizations, including Message Understanding Conferences (MUC), Text REtrieval Conferences (TREC) and the TIPISTER text program helped define and promote research in IE.

MUC

During the late 1980s, the U. S. Navy sponsored various academic and industrial research projects to develop systems to extract information from naval messages. In order to compare the progress of these research centers and performance of the systems they produced, MUC conferences were held. All participants designed a software program to extract information from text documents. The specific task and topics of study were determined by the organizers of each conference [1].

The first two conferences, MUC-1 (1987) and MUC-2 (1989), focused on extracting information from short naval messages. Many of the first systems that analyze natural language text-based information came from these two conferences [45]. These

first conferences were originally referred to as MUCK-1 and MUCK-2 but changed to the familiar MUC acronym as the conferences gained prominence and attracted the interest of Defense Advanced Research Projects Agency (DARPA) [25].

DARPA is the research and development arm of the U.S. Department of Defense and often supports research and technology in risky projects they feel may provide advances for the military (DARPA, 2004). DARPA began sponsoring MUC in 1991. MUC-3 (1991) and MUC-4 (1992) centered on systems that extracted data about terrorists in Latin America from newspaper and newswire articles. Training text and structured output templates were given to the participants for the research task. A semi-automated scoring system was implemented that independently evaluated each systems score. As the discipline of IE progressed, the participants in MUC came from a more stable environment of IE researchers [25].

The conferences continued in 1993, 1995 and 1997 (MUC-5, MUC-6, and MUC-7) using news articles to extract information about joint ventures, microelectronics, management changes, space vehicles, and missile launches [4]. Many of the systems we have today were developed as a result of these MUC conferences. The success of MUC prompted DARPA's funding of several other programs to encourage IE research [25].

TIPSTER

The TIPSTER text program began in 1991. TIPSTER was jointly sponsored by DARPA and the Central Intelligence Agency (CIA) and was partially managed by the National Institute of Standards and Technology (NIST). The goal of the program was to improve document processing efficiency and focused on three areas, document detection, information extraction, and summarization [66].

Document detection was the main research agenda of the TIPSTER program in the early 1990s and produced major advancements in algorithms used in information retrieval (IR) systems. IR is the process of selecting a subset of relevant documents from a larger domain. Most IR systems rely on key-word searches and often produce poor results due to the ambiguity of specific word use, especially when applied to synonyms and homonyms. Although IR and IE differ in their objective, they are complementary and often use similar processes. IR is a crucial first step in IE systems that extract information from Web documents [1].

The TIPSTER program also helped develop many of the techniques and technology presently used in IE research [67]. TIPSTER encouraged research and design of systems that could be reconfigured and made portable [25]. TIPSTER formally ended in 1998 when funding for the program ceased [67].

TREC

TREC was co-sponsored by DARPA and NIST as part of the TIPSTER program. The program was managed by members of government, academia, and industry to further promote IR and IE research. TREC workshops had various agendas and focused on increased communication between research and industry. Their goal was to speed the use of IR and IE products for commercial use. In 1999, sixteen countries were represented at the TREC-8 conference. Between 1992 and 1999, TREC research succeeded in doubling the efficiency of IR systems [66].

2.1 Early Information Extraction systems

Several IE systems were developed in the 1960s and 1970s. Naomi Sager at New York University developed one of the earliest IE systems in the late 1960s. Her system extracted hospital discharge information from patient records. Sponsored by the American Medical Association (AMA), the output was in a structured form that made it suitable for database management [25]. In the early 1970s Gerald DeJong developed an IE system he named FRUMP. This system was the first to use a data source of unrestricted topics. Using newswire articles the system determined the relevant information using keywords and sentence analysis [25].

Other systems were developed in the early 1980s that extracted data from satellite flight information, guides of plant and animal descriptions, and text that described French historical activities. The earliest IE system to be used for commercial purposes was ATRANS. This system used a simple sentence analysis similar to FRUMP and extracted data from international money transfers [25].

One of the most well known IE systems of this era is the System for Conceptual Information Summarization, Organization, and Retrieval System (SICSOR). SICSOR is a prototype IE system that performs text analysis and question answering in a constrained domain, financial news articles. SCISOR was designed by Paul S. Jacobs and Lisa F. Rau at the General Electric (GE) Artificial Intelligence Lab in the late 1980s. The design of SCISOR was based on the GE NLToolset that used two text-processing domains. The first domain of the NLToolset selects and analyzes stories about corporate mergers and acquisitions in real time as they come across newswires. The second domain presents the extracted output in a template format. The design of NLToolset incorporates AI

methods, NLP techniques, such as lexical analysis and word-based text searches, with knowledge representation and IR [46].

SCISOR is a customized IE system and was unique in its development because it combined a bottom up full parser, language driven interpretation with the top-down skimming parser, expectation-driven process [47]. Another unique feature of SCISOR is its knowledge based design which performed different levels of analysis. The system used knowledge about words and word meanings and applied them to topic analysis text, processing, and response generating. SCISOR processes about six stories per minute with a combined precision/recall of 80-90% [45]. The system performed with 90% accuracy in extracting correct stories and 80% accuracy in extracting correct values [46].

2.2 Types of IE Systems

As described by Cunningham (1999), five types of IE systems have been researched by MUC. These are: name entity recognition systems, co-reference resolutions, template element construction, template relation construction, and scenario template production [26].

Name Entity Systems

A name entity system finds and classifies names, places, organizations, etc. It is the simplest and most reliable IE technology, often performing at 90% accuracy when compared to human extraction [26]. Due to the emphasis placed on this type of system by MUC, name entity systems are the most common systems developed and the most widely studied. While names are widely used for extraction in these systems, other entities, such as dates, times, numbers, and addresses, can also be incorporated [18].

RADA (Radiology Analysis) developed by Johnson, et al. in 1997 is an example of a name entity system. RADA relies heavily on the name entity by matching words or groups of words to a pre-classified glossary. The system was designed to extract structured information from physician dictated radiological reports and performs with recall of 85% and precision of 89% [50].

Co-reference Systems

Co-reference resolution systems identify relations between entities in texts. This method was formally introduced by MUC-6 in 1995 and is used primarily as a building block for other types of IE. Co-references are words in a text that refer to the same thing, such as a pronoun or other noun phrase referring to a proper noun. It is often difficult to determine when two phrases refer to the same entity. Different methods may be necessary to deal with each form [16]. Co-reference systems are used to highlight occurrences with the same object or provide links between them and usually perform at accuracy levels in the 60% to 70% range [26].

RESOLVE [59] and MLR (Machine Learning Based Resolver) [3] are co-reference resolution systems. These systems use a training corpus that is annotated with co-references relations. Machine learning techniques are used to learn specific co-references in the text. These systems perform in the 70% to 80% range. However these results are based on the co-reference task only and do not measure the overall IE tasks. Co-reference systems are domain specific and have limited use beyond their domain content [26].

Template Elements, Template Relations and Template Productions

The template element task builds on both the name entity and co-reference systems by associating descriptive information within the entities. These systems are more sophisticated and normally perform with less accuracy. Template relation construction simply finds relations between template entities. 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 positions on the template that are to be filled with the extracted data 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 [26]. The use of templates began in 1991 at MUC-3 and has become a popular IE method used by a majority of IE systems today [4].

2.3 Performance Measurement for IE Systems

The performance of each IE task and the ease it can be developed normally depends on the text type, the domain of the text, and the specific scenario that the user is interested in. Performance measures for IE systems were developed by MUC and refined with each conference task. Measures of the success of IE systems are calculated using Precision, Recall and F-measure and are computed for each slot in the prescribed template.

Precision and Recall, developed for use at MUC-3 and MUC-4, are based on the standard measurements used in IR systems. Precision is calculated by dividing the number of correct answers produced by the number of total answers produced. For

example, if the system produces 15 answers but only 10 answers are correct, the precision rate will be 10/15, or 67%. Precision measures the reliability, or accuracy, of the information extracted [1].

Recall is the number of correct answers produced divided by the total possible correct answers. For example, if 10 correct answers are produced by the system, but there are 20 possible correct answers, recall will be 10/20, or 50%. Recall is a measure of the amount of relevant information that the system extracts [1].

There are normal trade-offs in the two distinct measures. To compensate for the discrepancy in various systems a combined weighted measure, the F-measure, was used at MUC-5 and the final TIPSTER evaluation. 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:

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

[4].

By the mid 1990s TIPSTER and MUC systems showed average recall performance of 40%, with precision performance somewhat better at 50%. Improvements continued with most of the IE systems in the later 1990s improving recall to around 50% and precision to 70% on complex tasks. Some simple systems can reach performance levels in the 90% range [25]. Although these figures may not be impressive at first glance, they are normally compared to human performance of the same extraction tasks. Human performance is usually 79% for recall and 82% for precision. These less than perfect results can be attributed to the length of time it takes humans to extract the data, lack of knowledge on the subject matter, and boredom. Humans often outperform

most IE systems, but they cannot compete with the speed of the computer programs used in automated processes [1].

2.4 Information Extractions Systems and the World Wide Web

Considering the wealth of public information available on the Web, it has become increasingly important to develop methods to extract Web data in a format that is easy to use. Including static web pages, database generated web pages, and e-mail sources, the Web is estimated to be over 530 thousand terabytes. This is roughly 53,000 times larger than the print collection at the U.S. Library of Congress. In 2003, approximately 600 million people worldwide had access to the Web [57].

Extracting useful information from Web documents is not a trivial task. In addition to the vast number of Web pages on the Internet, Web documents are diverse in structure, format, length, and writing style. Web documents often contain spelling errors and are displayed in a number of different languages. Information can be displayed in various forms ranging from text, to visual and audio images and videos. Web pages are dynamic resulting in non-functional links after a period of time. One of the major challenges in extracting information from the Web is finding the right documents with the right information. Another challenge is extracting structured data from unstructured documents [19].

IE programs are helpful because they can identify and extract information from a variety of document types from numerous sources on the Web. The result is a single document containing the condensed data. To accomplish this task, the IE system must select facts from documents that are specifically retrieved for the task. IR techniques are

used to select a subset of documents from the various Web sources; IE systems then extract the relevant information from these retrieved documents. Although the two processes have different objectives, they combine to provide a powerful force to break Web information into smaller pieces containing information that is manageable and meaningful to the user [1]. An example of a Web extraction system is Webfoot.

Webfoot, designed by Stephen Soderland in 1997, uses a preprocessor that parses Web pages into segments based on cues from the page layout. The system uses NLP techniques based on the relationship among the text to be extracted [79]. Other systems use grammar-based approaches, object oriented approaches, and various HTML tools to extract the data [52].

2.5 Information Extraction System Development

An IE system takes input from unstructured, free text, processes the text to extract specific data, and then produces a document in a structured format. The input can come from various types of text and digital sources. The output can be in the form of a text template, a spreadsheet, or database. The development of the extraction process is the critical aspect in the success of the system.

Most IE research has developed around rule-based systems using NLP [21]. NLP is a tedious task involving many complex steps. Sentences are analyzed and tagged according to its parts of speech, nouns, verbs, objects, etc. This syntactic structure is compared to linguistic structure to determine relevant information. In turn, the semantic meaning of the text can be determined by examining patterns of the syntactic structure [1]. Various methods and tools have been developed to aid in the process of extracting

data. NLP techniques include filtering, part-of-speech tagging, lexical semantic tagging, and syntax analysis. NLP techniques are difficult to apply directly to the learning of extraction patterns, co-references, and templates [16].

2.5.1 Approaches to IE Systems

Appelt and Israel (1998) describe two approaches to building IE systems, the Knowledge Engineering Approach and the Automatic Training or Machine Learning Approach. Both approaches rely heavily on the use of a domain specific corpus. A corpus is a set of documents that is annotated and used to train the system. Annotation includes NLP techniques such as parts-of-speech and semantic tagging [4].

In the Knowledge Engineering Approach a person familiar with the IE system and an expert in the domain of the application writes rules for the IE system to extract the data from the text. In this approach, team members use a moderate corpus of text related to the domain. The domain includes the corpus and a set of concepts to be identified in the corpus. Intuition based on the skill and knowledge of the team is used to determine the basic rules of the system. Once the set of rules is written the system must be run over a test corpus, the output examined, and modifications made. This approach is labor intensive and may take several iterations to produce a high performance system [4].

When applying the Automatic Training Approach it is only necessary to have someone with enough knowledge about the domain to annotate the corpus of text used. Once the corpus is annotated, training algorithms are run. When grounded on statistical methods and backed by sound theory, this approach can be effectively measured and hold the promise of domain independence [4].

There are advantages and disadvantages to each approach. The automatic approach does not rely on the skills of a knowledge engineer, but focuses on training data to develop the rules for the IE system. Often referred to as shallow knowledge, the automatic approach has no understanding of the input text. Another limitation is the availability of domain specific training data. At least 1.2 million words are needed to produce a system that performs roughly a linear relationship to the training data. Also, it is often difficult to find errors in a machine-generated process. Although general purpose text understanding is still beyond the reach of current technology, progress is being made to bridge this gap [4].

The Knowledge Engineering Approach 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 [4].

A combination of the two approaches is not uncommon. LEXTER was developed by Didier Bourigault in 1992 to develop terminology for a specific subject. A corpus of text is fed into LEXTER which produces a list of potential terminology units. The lists are then evaluated by an expert to determine their relevance to the subject [11].

2.5.2 Tools and Techniques Used in IE Systems

Various tools and techniques have been tried and tested in developing IE systems. These include wrappers, keyword searches, pattern matching, corpus-based learning techniques, hidden Markov models, AI techniques, and template mining.

Wrappers

One of the more traditional tools used in IE systems are specialized programs called wrappers. Wrappers identify useful information and map them to a suitable format. Wrappers have gained popularity due to the growth of digital information available on the Web and are well suited for HTML documents [38]. Wrappers present many problems for IE researchers. The programs are difficult to perfect and present problems since they must be specifically written and maintained. Due to the dynamics of the Web, the formatting of these wrappers must change frequently [52].

Many documents on the Web rely on structures that may not be well suited to the standard NLP extraction methods used in wrapper development. Also, Web documents, such as e-mails and chat-room transcriptions, often use incorrect grammar and cryptic expressions to convey information that make NLP methods difficult to use [38]. Scalability and portability are other limitations of wrappers. These programs are generally limited to a specific task in a specific domain [52].

Keyword Searches

Many traditional IE systems operate by using keywords to search, index, and extract text. Creators of these systems are free to select any keywords that are valuable to the user. A disadvantage of this method is that it relies on human expertise to index the keywords used making the system expensive, inconsistent, and often inaccurate. In addition, keywords often lose their context when isolated from their source text. Natural language processing has been used to help to overcome some of these problems. NLDB (Natural Language Database) developed by Rau and Jacobs in 1991, made incremental

improvements to the keyword process by using keyword indices to search a text database [72].

Pattern Matching

Beginning in 1995 with MUC-6, pattern matching became a popular technique in the extraction process. Good patterns are patterns that are general enough to be used for the entire domain, but specific enough to eliminate unwanted data. Patterns are developed using machine learning methods on domain corpora [16]. AutoSlog, developed by Ellen Riloff in 1992, was one of the first systems that used a pattern matching technique. AutoSlog creates a dictionary of patterns from a specific domain corpus. It uses the first reference to the targeted information as the most likely site of other description information. AutoSlog was developed from the MUC-4 corpus and produced F-measures in the upper 90% range when applied to the specific domain [73].

Corpus-based Learning

Corpus-based learning techniques are used to develop algorithms to improve the extraction processes. The success of corpus learning depends on the extent and annotation of the corpus used [16]. As the need for faster development cycles in IE grows, machine-learning techniques become increasingly important. The technique requires large text corpus to generate learned algorithms.

One problem associated with corpus-based learning techniques is the relevant content of the corpus used. The Brown Corpus, produced by Nelson Francis and Henry Kucera at Brown University in 1961, was the first general corpus developed to represent the written English language [37]. However, developing algorithms for information extraction from context specific documents require training corpora that target the

relevant domain. For example, in medical text it is important to associate symptoms with specific medical names. In entity recognition systems, performance tends to improve when algorithms are developed from specific corpora [54].

Specific domain extraction systems normally require an understanding of the text. Since machine based corpus learning techniques have fallen short in intellectual understanding of text, system developers often must rely on human expert knowledge to improve performance [62].

Hidden Markov Models

Hidden Markov Models (HMM) are based on the probability of sequences and in IE can be used to predict the probable sequence of words in text. The probabilities are determined from a set of tagged training data and are derived from the words of the text. Probabilities are also determined from the current state of the system. The success of HMM depends on the amount of training data used and the construction of the model. HMM was introduced at MUC-7 in 1997 and has been used in many IE systems [4]. DATAMOLD is an example of a system that uses HMM probabilities.

DATAMOLD is a system that automatically extracts addresses in unstructured form into a structured format using HMM. HMM is used to develop a probabilistic model to determine the sequence of the targeted data. When tested on actual databases, the system produces an accuracy rate of 99% on U.S. addresses and 90% on Asian addresses. The system has been adapted to extract data in a structured form from bibliography text and shows an 87.7% accuracy rate. The use of HMM can often provide a more accurate method than similar systems based on the rule learning methods [8].

AI Techniques

Bottom-up and top-down modeling techniques employ Artificial Intelligence (AI) methods. In a top-down system, the “top” consists of a collection of known situations that try to match incoming documents. If matches can be made, specific information is then extracted. This process works around expectations about what concepts will occur together in a passage of text [44].

More sophisticated systems actively try to build new representations of objects instead of relying on static, pre-existing types developed using a top-down model. To do this, the process had to incorporate distilled knowledge of the texts themselves. This bottom-up approach is more difficult as it begins with the individual words in the text, parses them, and tries to identify the parts of the sentence and their relation [44]. Both methods were used by Jacobs and Rau for their SCISOR system.

SCISOR’s bottom-up system, TRUMP (Transportable Understanding Mechanism Package) uses natural language parsing tools for partial parsing of the texts. This parser combines word phrases and checks syntax to develop a domain specific vocabulary. TRUMPET is SCISOR’s partial-parsing, top-down approach. This approach skims the text and passes over unknown words. The combination of the two methods allows the system to better understand the meaning of the text [46].

Template Mining

Template mining is one of the oldest methods used in IE and can be traced back to the 1980s. It is largely used for the extraction of information from text in a specific domain. A template is a schema of the contents of the source document. The fields of the template are filled by the information directly extracted from the document.

Template mining has been used extensively to extract information from medical records and newswire articles [21]. Numerous examples of template mining systems have been developed for a wide range of interest areas. SCISOR [46] and LOLITA (Large scale Object based Linguistic Interactor Translator and Analyzer) [24] use template mining with predefined slots to extract financial information from news articles.

Other Tools and Techniques

Other methods of IE include the use of tokens, ontology, and modeling based methods. Tokens are a discrete value assigned to items in text, for instance, words, punctuation or numbers. Relating token items to learned rules help identify and generalize features of training examples. Ontological based IE tools locate constants within the Web page and constructs objects with them. This method extracts sections of text containing data items. Modeling based methods use tools that provide targets for objects of interest and try to locate specific web pages that conform to the structure of the object. Modeling methods often use graphic interface tools to develop objects that can then identify other similar objects in a document [52].

The challenge for researchers is to develop better IE techniques and methods to perfect the systems, to experiment with various domains to expand the use of IE, and to promote avenues and funding for future IE research. One potential use of IE is the extraction of financial data from Web sources. A solution to the problem of finding useable, complete, and reliable financial information is to apply Information Extraction (IE) techniques to retrieve data from Web sources that contain financial statement information.

As financial statement information has become readily accessible in digital form and the importance of extracting accurate data has increased, the financial community has begun research to develop methods to aid in financial data extraction. One method being developed uses mark-up tags to identify specific data in the face of the financial statements. Extensible Business Reporting Language (XBRL) is a vehicle that is touted to eliminate many of the problems associated with the transfer and use of financial data [23].

2.6 Extensible Business Reporting Language (XBRL)

The XBRL steering committee was conceived and funded by the American Institute of Certified Public Accountants (AICPA) to develop an Extensible Mark-up Language (XML)-based framework for the exchange of financial information through the Internet. Acknowledging that the process of disseminating financial information does not allow interaction by the user, the SEC adopted a rule allowing registrants to submit voluntary filings using XBRL. The rule is effective as of March 16, 2005. Companies who submit their reports in XBRL must still file in HTML or text format [76].

The XBRL initiative began in Tacoma, Washington when Charles Hoffman, a CPA, began working on the first prototype for XML financial reporting. Hoffman presented his work to the AICPA in October 1998. By August 1999, twelve organizations, including major accounting firms and software companies, joined efforts and formed what became the XBRL Steering Committee [84]. The Steering Committee's Web site, <http://www.xbrl.org>, demonstrates the diversity of its membership and the continuing effort to bring the XBRL project to reality. Like XML, XBRL requires the

use of tags to label each bit of financial data associated with the financial statement and related disclosures. As in all markup languages, these identifying tags are displayed in a document enclosed in a pair of brackets. The start-tag is delimited using a '<' and a '>' character; the end-tag is delimited by '</' and '>'.

Financial reporting taxonomies are used for the preparation of financial statements and to describe specific information associated with each financial fact. A taxonomy is a library, or a vocabulary, of financial facts. On March 7, 2005 the XBRL International Committee published a list of 7 approved U.S. Taxonomies. These taxonomies were approved only after they met with specific criteria and after a period of public review and feedback [85].

Using these taxonomies, instance documents are prepared using a standard set of tags. Instance documents contain specific data elements of a financial statement and their associated value. The document can range in size from one particular data fact to an entire set of financial statements. Using this structured form, the data can be viewed in different formats by applying a style sheet to make it more readable by humans. An endless number of style sheets can be designed using the same instance document. This gives XBRL the potential flexibility for reuse of financial data in a variety of ways [42].

IE can play an important role in the development of XBRL. Data must be extracted from financial statements before the tagging process can begin. IE techniques combined with standard XBRL taxonomies is the framework for software development and allows easy transformation of semi-structured data to a structured set of re-useable financial information. Dextrapi (data extraction API), developed by Leinnemann, et al. in 2000, is a wrapper developed to extract financial data from text documents. Regular

expressions are used to denote keyword identification for extraction. The data is then transformed into machine readable XML syntax [53].

For XBRL to be an effective method to extract financial data, several obstacles must be overcome. The taxonomies developed have met with resistance because they do not reflect current reporting practices and often result in a loss of information. Also, in order to be successful, XBRL must have widespread adoption by the financial community. Issues still are unresolved such as responsibility for maintaining taxonomies and standards for XBRL, software development for the tagging process, and the cost-benefit for company adoption [12]. In the event of widespread adoption of XBRL, it is doubtful that prior financial statements will be recoded to conform to the standard. Thus, XBRL will be of no value in extraction of financial information from prior year financial statements.

Chapter 3

Stock Options

3.0 Overview

The separation of management and ownership in the modern corporation creates the need for incentives to ensure that Chief Executive Officers (CEO) pursue activities that are in the best interest of the shareholders. Compensation policies adopted by the corporation can help mold executive behavior, contributes to the type of executives the company will attract, and can play an important role in the success of the organization [49]. A common scheme to align CEO and shareholder interests is to devise CEO compensation packages that include some form of common stock to induce CEOs to view the corporation from a perspective similar to that of the shareholder. This mix of salary and stock compensation better aligns the preferences of managers with shareholders and reduces the conflict of self-interest [28].

Under stock option compensation plans, management grants the recipient of stock options the non-transferable right, or option, to purchase a fixed number of common shares of the corporation at a specified price (usually the market price at the time of the option grant) for a specified time, commonly ten years. There is typically a waiting period, or vesting period, of three to five years before the options may be exercised. If the recipient leaves the firm before vesting occurs, the options are forfeited [13].

Employee stock options have been the equity incentive method of choice over the last two decades. The use of stock options escalated with the growth of new start-up companies during this period. Stock options became a standard method for cash poor companies, such as e-commerce firms and start up high-tech companies, to lure key

managers [7]. The favorable accounting treatment for stock options during this period helped escalate their widespread use. Most companies could structure their stock option packages to avoid reporting the transaction as compensation expense, thus avoiding a reduction to net income. Another factor that greatly contributed to the widespread popularity of stock options was the unique ability of the granting corporation to simultaneously avoid expense recognition yet still benefit from a corporate tax deduction [17].

3.1 Accounting for Employee Stock Options

Accounting for employee stock options has been extremely controversial for over a decade. Recognizing that stock options represent unrecorded employee compensation expense, in the early 1990s the FASB tried to improve the transparency of financial reporting by issuing an accounting rule that would have required companies to measure and expense the fair value of stock options. This fair value method would have resulted in a reduction of earnings in the company's income statement. Companies that routinely issue large quantities of stock options mounted a successful political campaign that effectively stymied the FASB. Certain members of the United States Congress threatened the continued existence of the FASB if it passed a rule mandating expensing of stock options. As a compromise, in 1995 the FASB issued SFAS No. 123. The new rule only slightly improved financial reporting [55].

Instead of mandating expensing of stock options, SFAS No. 123 allowed companies to choose either expensing the estimated fair value of stock options as described in SFAS No. 123 or to follow the old rule, APB No. 25. APB No. 25 allowed a

company to avoid expense recognition by simply setting the exercise price equal to the market price of the stock on the grant date [32]. Often referred to as the intrinsic method, most companies, especially high technology firms that regularly issue options, continued to follow APB No. 25. However, if the requirements in APB No. 25 were followed, SFAS No. 123 required companies to disclose in the notes to its financial statements the estimated fair value of the stock options and its pro forma impact on net income as if the options had been expensed [32].

In March 2003, the FASB announced that accounting for stock-based compensation would be revisited and added this project to its agenda. One objective of the project was to develop a United States accounting rule that is comparable to international accounting standards which require expensing stock options [34]. In September 2003, Bear Stearns reported that 356 companies had voluntarily adopted the fair value reporting of employee stock options and a significant number of additional companies were expected to follow suit [60].

After extensive deliberation on the issue, the FASB issued SFAS No. 123(R) in December 2004 that requires mandatory expensing of stock options beginning the third quarter of 2005 [36]. This new rule will end a ten year heated debate over the accounting treatment of employee stock options and improve the quality of financial reporting [63]. SFAS No. 123(R) effectively eliminates the ability of companies to use APB No. 25's intrinsic method of accounting for employee stock options and requires companies to adopt the fair value method in SFAS No. 123 [63]. However, expensing employee stock options beginning in 2005 will present problems with consistency and comparability in financial reporting.

3.1.1 Consistency Issues of Recent Changes to Stock Option Reporting

The FASB's Statement of Financial Accounting Concept (SFAC) No. 2 clearly explains the importance of consistency in financial reporting. Consistency enhances the usefulness of financial statements, especially in time series analysis. Consistency does not imply that there is quality in the accounting numbers presented, but implies that there is quality in comparison of the numbers. Beginning in 2005, analyzing company financial statements before and after the adoption of SFAS 123(R) will be more difficult due to inconsistent financial reporting requirements for stock options.

In July 1971, the APB noted in Opinion No. 20, *Accounting Changes and Error Correction*, that consistency in financial reporting greatly enhances the understanding and utility of comparative accounting data to users. APB Opinion No. 20 is an attempt to preserve consistency in reporting while allowing the standard setting process to keep pace with the dynamics of business [30]. APB Opinion No. 20, paragraph 18, requires most accounting changes to include the cumulative effect of the change in the net income of the period when the new rule is adopted, a prospective approach. However, paragraph 27 further states that in certain circumstances there are advantages in retroactive reporting and in these cases all prior periods presented must be restated [30]. The proposed change in the accounting treatment for stock options is subject to the requirements of APB Opinion No. 20 by requiring either a prospective, cumulative effect approach or a retroactive, restating approach.

The FASB considered several approaches for companies to report the transition of adopting the new accounting standard for employee stock options. In SFAS No 123(R) the FASB states that “. . . retrospective application with restatement . . . would be the

best transition method for this Statement because retrospective application would provide the maximum amount of comparability between periods and thus enhance the usefulness of comparative financial statements” [36]. However, the FASB decided that retroactive restatement was impracticable because it would require significant estimates in the current period that reflect conditions that existed in prior periods. The objectivity of these estimates would be impaired by knowledge of existing conditions. For this reason, SFAS No. 123(R) requires a modified prospective approach in reporting the transition effect of this new standard. The standard will allow, but does not require, restatement of prior financial statements [36].

SFAS No. 123 (R) provides a clearer economic impact on the use of employee stock options as compensation. However, the transition to the new requirement may result in a substantial reduction in net income to many companies and result in inconsistent financial reporting. The impact could be material. Operating income in 2001 and 2002 in the Standards & Poor 500 firms are estimated to have been 20% lower if employee stock options had been expensed [80].

Companies that chose not to expense stock options under SFAS No. 123 were required to disclose in the notes to their financial statements information regarding the fair value of stock options. Information required included the fair value of the stock options, pro forma net income, and pro forma earnings per share (EPS), as well as the assumptions and model used to value the stock options [32]. Although deeply embedded in the disclosure notes to the financial statements, this information is available and can aid users to overcome some of the problems of inconsistency associated with the new stock option expensing rule.

3.1.2 Comparability Issues of Prior Requirements for Stock Option Reporting

Comparability between companies is another issue that makes evaluating the impact of stock options on financial statements difficult for users. Financial comparability denotes that entities have some form of similar characteristics in common. The ability to compare financial information between entities is one of the main reasons accounting principles have been developed. The purpose of comparability is to allow users to discover similarities and distinguish differences in the financial position of reporting companies [31].

Often it is difficult to compare two company's financial statements because accounting regulations allowed companies to choose different accounting treatments. This is especially true for the accounting treatment of stock options. SFAS No. 123 allowed companies to choose between the accounting requirements of APB No. 25 and SFAS No. 123. In essence, companies can choose between expensing and not expensing the fair value of stock options it awards. Thus it is difficult to compare the financial statements of companies that chose to expense stock options with those that did not record the expense [32].

Under both SFAS 123 and SFAS 123(R), companies are also allowed to choose the model they use for valuing stock options and are allowed to make estimates for the underlying assumptions associated with the model. The fair value calculation is based on a pricing model, such as the Black-Scholes option pricing model, or a binomial model. The model must consider the exercise price and expected life of the option as well as the current market price of the underlying stock. Other assumptions must be made regarding

the expected volatility of the stock, expected dividends paid on the stock, and the risk-free interest rate for the life of the option [32].

This information about assumptions must be reported in the company's notes to its financial statements [32]. Some experts suggest that comparisons can be easily made using this information from the disclosure notes [15]. However, quickly searching and extracting information from the notes to financial statements to glean this information is no easy task. Although financial information for most companies is available via the Web, the files are often large and difficult to locate.

3.2 The EDGAR Database

The SEC was established by Congress in 1934 to administer the Securities Acts of 1933 and 1934. Although these acts have become intertwined over time, the 1933 Securities Act regulates initial security offerings of a company, while the 1934 Securities Act regulates the secondary trading of these securities [2]. The main objective of the 1933 and 1934 Securities Acts is to protect investors and creditors. 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. Although the SEC requires various types of financial information, the annual report (10-K) includes the most comprehensive collection of financial information.

The SEC brought the first timely, comprehensive financial information to the general public in an accessible and electronic format via the Web. Beginning in 1996, all public domestic companies were required to file their financial statements and other required forms electronically using the EDGAR system. The purpose of EDGAR is to

increase efficiency of the receipt, dissemination, and analysis of corporate information filed with the SEC. EDGAR is one of the largest government filing systems in the world, processing more than 500,000 financial statements annually. On average, 1,500 documents are submitted each day [41]. As a result, detailed financial information from all publicly traded companies is available to the public via the Web. All public domestic companies are required to file financial statements and other required forms electronically using the EDGAR system [41]. As a result, detailed financial information from all publicly traded companies is available to the public via the Web.

Although documents filed on the EDGAR Database are the richest source of financial information available on the Web, they are displayed in formats that have limited use. Before 1999, all files submitted to EDGAR were text documents. Text documents are easily interchanged over the Internet, but it is difficult to extract specific information from a text document [10]. In an attempt to modernize EDGAR in early 1999, the SEC began allowing companies to submit filings that included documents in Hypertext Markup Language (HTML) and Portable Document Format (PDF) [75]. Although these formats provide the user with a version of the financial statement that is easier to read, they also have limited use.

HTML is the standard system for formatting and displaying documents on the Internet. It is a simple language well suited for the display of small documents and provides an excellent method of displaying information. However, HTML tags do not identify information between these tags. Also, it relies on web browser user agents, operated by humans, to search for information on the Internet [9].

PDF also has limited use. PDF is a specific file format developed by Adobe Corporation. It enables users to view and print a file exactly like the original document. In order to view a PDF file, the user must have Adobe Acrobat PDF Reader installed on their computer [33]. The reader is free and can be easily downloaded. However, documents in this format cannot be edited or reused. A PDF document has the same characteristics as a printed document. Specific data is difficult to locate, and once the data is extracted it must be re-keyed into proprietary software for analytical purpose.

These technology breakthroughs in business reporting have enhanced financial information on the EDGAR system. However, there are other sources of financial information available on the Web today. These sources also have limitations.

3.3 Other Sources of Freely Disseminated Financial Information on the Web

Many companies have developed their own Web pages to disseminate more timely financial information directly to interested parties. Third-party providers have emerged in an effort to provide users with financial information that is more concise and easier to use.

Financial Information on Company Commercial Web Pages

The rapid growth of the Web for financial information dissemination began with commercial use of the World Wide Web in 1994. The use of the Web for corporate financial reporting is viewed favorably by the SEC and stock exchange officials, however contents of financial information on corporate web sites is not regulated [29]. Companies use various methods to disseminate financial information on their commercial Web pages. Most large companies now have an “Investor Relations Page” that provides a

wide variety of financial information for public use. The content of financial information can range in context from full electronic copies of the company's Annual Report, to excerpts of financial information, to no financial information at all.

Statutory filed reports and printed annual reports must contain specific information that is regulated by the SEC. There are no requirements regarding financial information provided by companies on their Web pages [33]. Therefore, a primary concern with financial information found on a company's web page is the quality [69] and the completeness [5] of the information provided.

Third party providers of financial information

The difficulty in retrieving information from files on the SEC EDGAR database and the quality of information and inconsistencies of reporting practices on company commercial web sites spurred the growth of commercial third-party providers of financial information. Web sites, such as EdgarScan, Free Edgar, Yahoo! Financial, and MSN Money.com, provide free, but limited, financial data in a common structured format [40].

These third-party providers make financial information available in an easy to read format. The common structure of the information allows users to compare financial information among companies. However, only limited financial data items are available on these sites [40]. Users who require more detailed and expanded information may not find these sites useful. Also, the financial information is normally limited to income statement and balance sheet data and do not allow user access to information in the notes to the financial statements.

The dilemma facing users of financial information from Web sources is two-fold. First, the richest and most reliable source of financial information lies in the volumes of

files on the EDGAR database. The difficulty of searching and reusing this information makes EDGAR an undesirable source. Company commercial Web pages provide financial information in PDF format, which poses the same limitations as text files, or provide summary and incomplete information. Third party sources of financial information provide information in a concise format, but the information is limited to key data items. One solution to the problem of finding useable, complete, and reliable financial information is applying IE techniques to extract financial data from the Edgar database.

3.4 Current Systems That Extract Financial Information from Web Based Financial Statements

Leinneman, et al. (2000) introduced a software agent, Edgar2xml, to detect relevant information in the EDGAR database files. The team applied text-mining tools and used a wrapper, dextrapi (data extraction API), to extract the data items. The format and structure of EDGAR files vary immensely. Some financial statement tables are introduced with Standard General Markup Language (SGML) tags, but the specific data items are in pure ASCII text. Edgar2xml extracts financial information which is then transformed into a machine-readable XML syntax. 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 result shows that it is possible to extract financial information and transform it into XML format [53].

This system was designed only for use on the company balance sheet, and therefore fails to supply the user a method to extract information from other financial

statements tables, or more importantly, the disclosure notes associated with the financial statements. The authors posit that further research is necessary in order to maximize the usefulness of EDGAR's financial information.

EDGAR-Analyzer, developed by John Gerdes (2003), is a tool that automatically analyses EDGAR filings. The system was used to study corporate Year 2000 (Y2K) disclosures in 18,595 10-K filings from 1997 to 1999. However, EDGAR-Analyzer is an all-purpose tool that can search for evidence of user specified subjects. The program uses index files on the EDGAR web site to identify specific files. Once the files are downloaded, the program uses a key-word search to extract the paragraph that contains the specified key word. The information is further processed to extract blocks of data pertinent to the user search. In regard to the Y2K disclosure case study, EDGAR-analyzer extracted an average text block of 11.1 KB from each filing which reduced the amount of text to be manually processed by 96%. Although the amount of text is reduced, the system does not produce a structured output. Instead, it requires the user to sift through the remaining text to manually retrieve data items. The authors suggest that pattern matching techniques using regular expressions can improve their system [40].

These two systems both focus on information extraction from financial statements on the EDGAR database. Both use IE and NLP techniques in their processes. EES builds on these two extraction systems and expands the development of financial information extraction from documents on the EDGAR database.

Chapter 4

Research Methodology

4.0 Overview

The purpose of EES is to extract information related to stock options from annual financial statements on the SEC's EDGAR Database. Targeted information includes pro forma data, fair value of the options awarded, and the method and assumptions used in calculating the fair value. Since 1996, this information has been required by the SEC to be presented in the notes to the financial statements.

The first step in developing EES was to create a training corpus from the notes to the financial statements. Machine learning and knowledge based techniques were applied to the corpus to determine the format of information and to detect patterns in the text. The next step developed algorithms from the corpus analysis. Using these algorithms a wrapper was designed to extract data and display the information. Figure 1 shows the overall structure of EES.

4.1 Corpus Development

Domain-specific vocabularies, such as law, medicine, and accounting, present difficulties when creating heuristics and algorithms for information extraction. The use of corpora and machine learning techniques to develop a domain-specific knowledge base has become increasingly popular since the 1990s [18]. Extraction systems developed from domain-specific corpora report recall and precision of 85% and 89%, Johnson, et al., 1997 [50]; 71% and 92%, Leroy, et al., 2003 [54]; and percentages

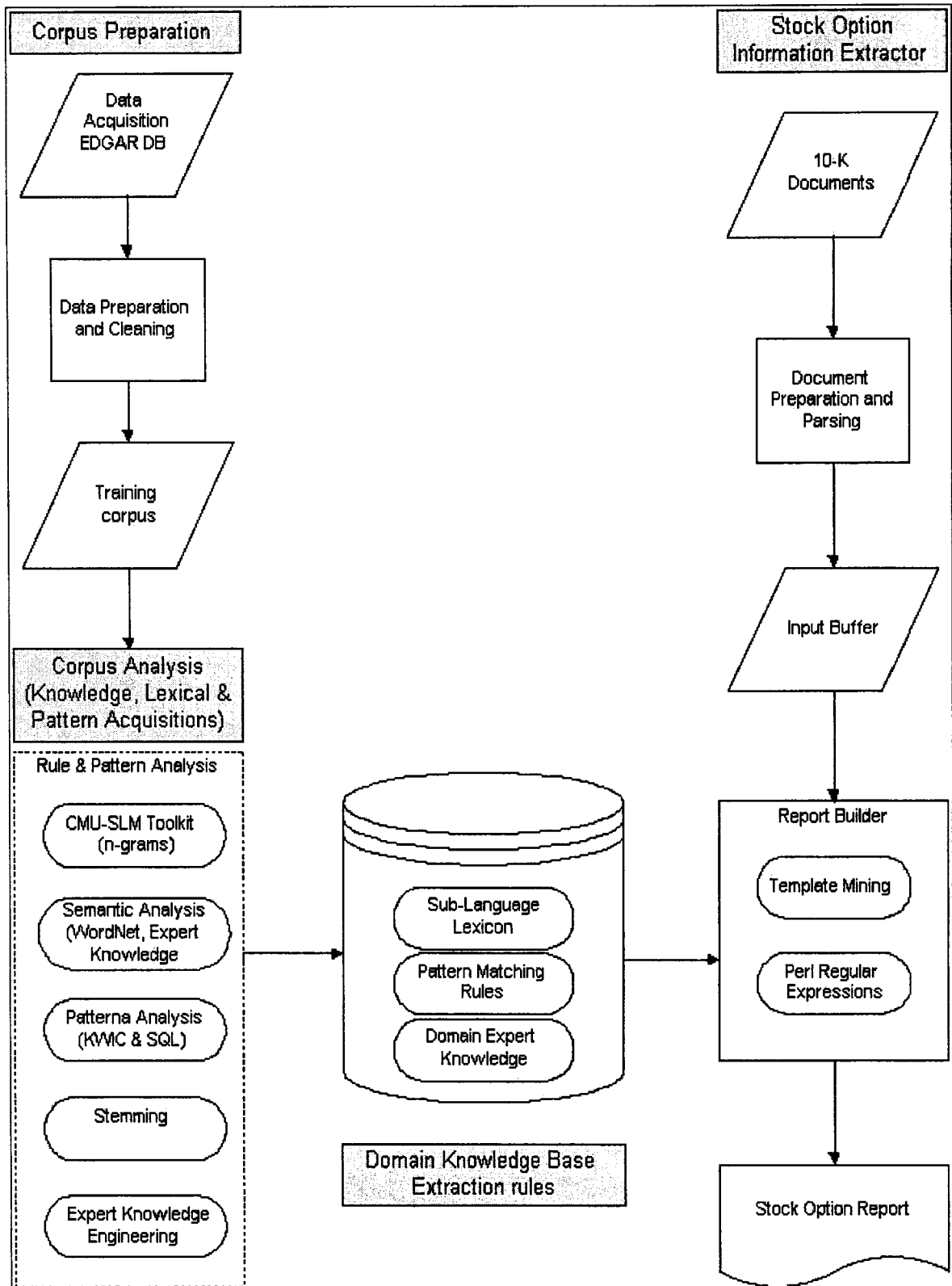


Figure 1. System Architecture of EES.

ranging from the high 70s to the low 50s, Jackson, et al, 1998 [48]. The knowledge based rules and algorithms for EES were developed using a non-annotated, domain-specific corpus.

The corpus was developed in two stages. The first stage involved the selection sample of companies used in the study. The second stage prepared the text documents used in the corpus for analysis.

4.1.0 Sample Selection

The sample used in this study consists of 10-K annual reports from the EDGAR database for companies that comprise the NASDAQ-100 Index as of May 4, 2004. The NASDAQ-100 Index includes 100 of the largest domestic and international non-financial companies listed on the NASDAQ Stock Exchange. This index reflects companies across major industry groups but consists mainly of high-tech, telecommunications, and biotechnology companies [65]. The NASDAQ-100 Index was selected as the sample because of its concentration of high-tech firms. High-tech firms commonly use stock options as a form of compensation [42]. Table 1 provides a list of NASDAQ 100-Index companies as of May 4, 2004.

Four companies in the NASDAQ-100 Index are based in foreign countries and were not used as part of the sample. These foreign companies are ATI Technologies (Canada), Research in Motion LTD (Canada), Ryanair Holdings PLC (Ireland), and Teva Pharmaceuticals Industries (Israel). Although foreign companies must present their financial statements in a format similar to U.S. companies and must reconcile their financial statements to U.S. GAAP, they are allowed to prepare their financial statements

Table 1. NASDAQ 100 Companies as of May 4, 2004.

ADOBE SYSTEMS INC	JUNIPER NETWORKS INC
ALTERA CORP	KLA TENCOR CORP
AMAZON COM INC	LAM RESEARCH CORP
AMERICAN POWER CONVERSION CORP.	LAMAR MEDIA CORP/DE
AMGEN INC	LEVEL 3 COMMUNICATIONS INC
APOLLO GROUP INC	LINCARE HOLDINGS INC
APPLE COMPUTER INC	LINEAR TECHNOLOGY CORP /CA/
APPLIED MATERIALS INC /DE	MARVELL TECHNOLOGY GROUP LTD
*ATI TECHNOLOGIES INC	MAXIM INTEGRATED PRODUCTS INC
BEA SYSTEMS INC	MEDIMMUNE INC /DE
BED BATH & BEYOND INC	MERCURY INTERACTIVE CORPORATION
BIOGEN INC	MICROCHIP TECHNOLOGY INC
BIOMET INC	MICROSOFT CORP
BROADCOM CORP	MILLENNIUM PHARMACEUTICALS INC
C H ROBINSON WORLDWIDE INC	MOLEX INC
CAREER EDUCATION CORP	NETWORK APPLIANCE INC
CDW COMPUTER CENTERS INC	NEXTEL COMMUNICATION INC
CEPHALON INC	NOVELLUS SYSTEMS INC
CHECKPOINT SYSTEMS INC	NVIDIA CORP
CHIRON CORP	ORACLE CORP /DE/
CINTAS CORP	PACCAR INC
CISCO SYSTEMS INC	PANAMSAT CORP /NEW/
CITRIX SYSTEMS INC	PATTERSON DENTAL CO
COMCAST HOLDINGS CORP	PATTERSON UTI ENERGY INC
COMPUWARE CORPORATION	PAYCHEX INC
COMVERSE TECHNOLOGY INC/NY/	PEOPLESOFT INC
COSTCO WHOLESALE CORP /NEW	PETSMART INC
DELL COMPUTER CORP	PIXAR \CA\
DENTSPLY INTERNATIONAL INC /DE/	QLOGIC CORP
DOLLAR TREE STORES INC	QUALCOMM INC/DE
EBAY INC	*RESEARCH IN MOTION LTD
ECHOSTAR COMMUNICATIONS CORP	ROSS STORES INC
ELECTRONIC ARTS INC	*RYANAIR HOLDINGS PLC
EXPEDITORS INTERNATIONAL OF WASHINGTON	SANDISK CORP
EXPRESS SCRIPTS INC	SANMINA-SCI CORP
FASTENAL CO	SCHEIN HENRY INC
FIRST HEALTH GROUP CORP	SIEBEL SYSTEMS INC
FISERV INC	SIGMA ALDRICH CORP
FLEXTRONICS INTERNATIONAL LTD	SMURFIT STONE CONTAINER CORP
GARMIN LTD	STAPLES INC
GENTEX CORP	STARBUCKS CORP
GENZYME CORP	SUN MICROSYSTEMS INC
GILEAD SCIENCES INC	SYMANTEC CORP
INTEL CORP	SYNOPSIS
INTERACTIVE DATA CORP/MA/	TELLABS INC
INTERSIL CORP/DE	*TEVA PHARMACEUTICALS INDUSTRIES
INTUIT INC	VERISIGN INC/CA
INVITROGEN CORP	VERITAS SOFTWARE CORP
JDS UNIPHASE CORP /CA/	WHOLE FOODS MARKET INC
	XILINX INC
	YAHOO INC

* Foreign Companies

following the rules and regulations of their home country. Also, foreign countries do not file the traditional 10-K form used by U.S. companies but instead are required to file a similar form, 20-F [2]. As of December 2000, companies from over 60 countries are registered with the SEC [76]. Due to the numerous and varied accounting rules associated with these foreign countries, EES is limited to extracting information from U.S. companies' 10-Ks. Thus the sample for this study consists of the 96 U.S. companies in the NASDAQ-100 Index.

To limit the scope of the project, the sample contains 10-K annual reports for filing dates between 2001 and 2004. Under SAFS No. 123, companies were required to report specific information about stock options for this four year period.

The 10-K Annual Reports for the sample companies for 2001-2004 were downloaded from the SEC's EDGAR Web site. One company, Biogen, Inc. had not filed a 10-K in 2004 at the date the sample files were downloaded, thus, the entire sample consists of 383 10-K files for the 96 companies. In order to test EES, a portion of the data was held out for testing. Only a small percent of the corpus is normally held out for testing. This is mainly due to the size limitations of most corpora [20]. Five to ten percent is an acceptable level of text to hold out for testing [58]. However, due to the large amount of data in the EDGAR corpus and the quest for reliable results, 20% of the companies were used for testing, with the remaining 80% used for training.

From the sample, 20% of the companies were randomly selected as test companies, while the remaining 80% comprised the training corpus. To insure a more diverse sample of industries, the 96 companies were sorted by Standard Industrial Classification (SIC) codes before the random selection. SIC codes are four digit numeric

codes used by the federal government to group entities into uniform business categories [82]. For example, Microsoft and other companies that develop prepackaged software are identified by the SIC code 7372.

SIC codes are based on a hierarchical structure. The first digit represents a major economic division, such as retail or manufacturing. The second digit designates a group within the division. The third and fourth digits further define the industry group and specific industry [51]. Sorting the sample companies by SIC code helps assure that a variety of industries are represented in both the corpus and test companies.

To separate the sample into corpus and test companies the random number “2” was selected to begin the selection of the test companies. Using the list of 96 companies sorted by SIC codes and beginning with the second company on the list, every fifth company was selected. As a result, 19 companies from fourteen different SIC codes were selected as test companies, while the remaining 77 companies, representing 41 SIC codes, comprised the training corpus. Considering the four year period used for the sample, the final training corpus consisted of 307 10-K annual reports, while the test data consisted of 76 10-Ks.

In order to expedite the download procedure from the EDGAR database for the sample companies, Central Index Keys (CIK) were determined for each company. CIKs are used by the SEC's computer systems to uniquely identify corporations that have filed with the SEC. A list of the 77 NASDAQ-100 Index companies used in the corpus with corresponding SIC and CIK codes is found in Table 2. Table 3 is a list of the 19 companies with corresponding SIC and CIK codes that were used to test EES.

Table 2. NASDAQ-100 Index companies used in Corpus with Corresponding SIC and CIK Codes.

<u>COMPANY NAME</u>	<u>CIK</u>	<u>SIC</u>
ADOBE SYSTEMS INC	796343	7372
ALTERA CORP	768251	3674
AMAZON COM INC	1018724	5961
AMERICAN POWER CONVERSION CORPORATION	835910	3620
APOLLO GROUP INC	929887	8200
APPLE COMPUTER INC	320193	3571
APPLIED MATERIALS INC /DE	6951	3559
BEA SYSTEMS INC	1031798	7372
BED BATH & BEYOND INC	886158	5700
BIOGEN INC	714655	2836
BROADCOM CORP	1054374	3674
C H ROBINSON WORLDWIDE INC	1043277	4731
CAREER EDUCATION CORP	1046568	8200
CEPHALON INC	873364	2834
CHIRON CORP	706539	2834
CISCO SYSTEMS INC	858877	3576
COMCAST HOLDINGS CORP	22301	4841
COMPUWARE CORPORATION	859014	7372
COMVERSE TECHNOLOGY INC/NY/	803014	3661
COSTCO WHOLESALE CORP /NEW	909832	5331
DENTSPLY INTERNATIONAL INC /DE/	818479	3843
DOLLAR TREE STORES INC	935703	5331
ECHOSTAR COMMUNICATIONS CORP	1001082	4841
ELECTRONIC ARTS INC	712515	7372
EXPEDITORS INTERNATIONAL OF WASHINGTON INC	746515	4731
FASTENAL CO	815556	5200
FIRST HEALTH GROUP CORP	812910	6324
FISERV INC	798354	7374
FLEXTRONICS INTERNATIONAL LTD	866374	3672
GARMIN LTD	1121788	3812
GENTEX CORP	355811	3714
GENZYME CORP	732485	2836
GILEAD SCIENCES INC	882095	2836
INTERACTIVE DATA CORP/MA/	888165	6200
INTERSIL CORP/DE	1096325	3674
INTUIT INC	896878	7372
INVITROGEN CORP	1073431	2836
JDS UNIPHASE CORP /CA/	912093	3674
KLA TENCOR CORP	319201	3827
LAM RESEARCH CORP	707549	3559

Table 2 (continued).

<u>COMPANY NAME</u>	<u>CIK</u>	<u>SIC</u>
LAMAR MEDIA CORP/DE	899045	7311
LINCARE HOLDINGS INC	882235	8090
LINEAR TECHNOLOGY CORP /CA/	791907	3674
MARVELL TECHNOLOGY GROUP LTD	1058057	3674
MERCURY INTERACTIVE CORPORATION	867058	7372
MICROCHIP TECHNOLOGY INC	827054	3674
MILLENNIUM PHARMACEUTICALS INC	1002637	2834
NETWORK APPLIANCE INC	1002047	3572
NEXTEL COMMUNICATION INC	824169	4812
NOVELLUS SYSTEMS INC	836106	3559
NVIDIA CORP	1045810	3674
ORACLE CORP /DE/	777676	7372
PACCAR INC	75362	3711
PANAMSAT CORP /NEW/	1037388	4899
PATTERSON DENTAL CO	891024	5047
PATTERSON UTI ENERGY INC	889900	1381
PAYCHEX INC	723531	8700
PEOPLESOFT INC	875570	7372
PETSMART INC	863157	5990
PIXAR \CA\	1002114	7372
QLOGIC CORP	918386	3674
QUALCOMM INC/DE	804328	3663
SANDISK CORP	1000180	3572
SANMINA-SCI CORP	897723	3672
SCHEIN HENRY INC	1000228	5961
SIEBEL SYSTEMS INC	1006835	7372
SMURFIT STONE CONTAINER CORP	919226	2631
STAPLES INC	791519	5940
STARBUCKS CORP	829224	5810
SUN MICROSYSTEMS INC	709519	3571
SYNOPSYS	883241	7372
TELLABS INC	317771	3661
VERISIGN INC/CA	1014473	7371
VERITAS SOFTWARE CORP	1084408	7372
WHOLE FOODS MARKET INC	865436	5411
XILINX INC	743988	3674
YAHOO INC	1011006	7373

Table 3. NASDAQ-100 Index Companies used for Testing with corresponding SIC and CIK Codes.

<u>COMPANY NAME</u>	<u>CIK</u>	<u>SIC</u>
AMGEN INC	318154	2836
BIOMET INC	351346	3842
CDW COMPUTER CENTERS INC	899171	5961
CHECKPOINT SYSTEMS INC	215419	3669
CINTAS CORP	723254	2320
CITRIX SYSTEMS INC	877890	7372
DELL COMPUTER CORP	826083	3571
EBAY INC	1065088	7389
EXPRESS SCRIPTS INC	885721	6411
INTEL CORP	50863	3674
JUNIPER NETWORKS INC	1043604	3576
LEVEL 3 COMMUNICATIONS INC	794323	4813
MAXIM INTEGRATED PRODUCTS INC	743316	3674
MEDIMMUNE INC /DE	873591	2836
MICROSOFT CORP	789019	7372
MOLEX INC	67472	3678
ROSS STORES INC	745732	5651
SIGMA ALDRICH CORP	90185	5160
SYMANTEC CORP	849399	7372

4.1.1 Corpus Building

To build the corpus the downloaded 10-Ks for the 77 companies were converted to text format. Then, using the designated SGML tags <table> and </table> all tables were extracted from the text file. The results produced two separate files. One file contained tables from the 10-K; the other contained the remaining text from the 10-K. The files containing only the text portion of the 10-K were used as the basis of the corpus and were cleaned for processing.

The first step in cleaning the corpus was to remove all HTML tags. HTML tags are formatting elements that enhance the display of the document. For example, the title of the document is commonly found between the bracketed tags <title> and </title> [71]. Since these formatting elements are not part of the text of the document they were eliminated. Care was taken to avoid deleting brackets that contain digits since negative numbers in financial statements are often displayed in brackets similar to the ones used for HTML tags.

The files were then further processed to remove leading and trailing white spaces. Some formatting items, such as solid lines displayed as continuing dashes or equal signs, were also deleted. The 307 corpus files were concatenated into one 57,413 KB file.

4.2 Corpus Analysis

Various tools and techniques were used to analyze patterns in the corpus. These include the CMU Toolkit for Statistical Language Modeling (CMU-SLM), Key Word In Context Index System (KWIC), Structured Query Language (SQL), WordNet, Stemming, and Knowledge Based analysis techniques.

4.2.0 CMU-SLM Toolkit

The CMU-SLM Toolkit is a UNIX based set of software tools developed in 1994 by Philip Clarkson and Ronald Rosenfeld at Carnegie Mellon University. The main purpose of the CMU-SLM Toolkit is to process textual data into n-grams, specifically bi-grams and tri-grams, and provide related statistical data (Clarkson and Rosendfeld, 1997). 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. Using the model, $P(w_n | w_1, \dots, w_{n-1})$, the previous word can be used to predict the probability of the next words in the phrase [58].

The newer version of the CMU-SLM toolkit has been enhanced to process larger corpora and moves beyond tri-grams to support n-gram modeling for any value of n. The toolkit is widely used by universities, governments, and industrial laboratories to model and evaluate large corpora of training text [22]. A system designed by Su, Wu, and Chang (1994) uses bi-gram and tri-gram analysis to translate compound words in technical manuals. The system produced 96.2% recall and 48.2% precision using bi-gram analysis, while tri-grams produced 96.6% recall and 39.6% precision [81].

For EES, the CMU-SLM toolkit was used to construct a language model by creating a list of relevant tri-grams and their frequency. Infrequently used tri-grams, commonly referred to as cutoffs, were removed to reduce the model to a manageable size. To further reduce the number of tri-grams created from the output of the CMU-SLM toolset of the corpus, the same processes were used to generate a tri-gram list of words from the text of SFAS No. 123. Since SFAS No. 123 is the official document that gives guidance on reporting financial information related to stock options, word phrases that

appear both in the corpus and SFAS No. 123 provided a list of relevant word phrases for further analysis. An expert with knowledge of stock option financial reporting reviewed the CMU output files to further reduce the tri-gram list to reflect the most relevant tri-grams for analysis. Three hundred ninety two tri-grams were selected.

Further analysis reduced the 392 tri-grams to about 40. For this reduction step, tri-grams that were unique in the original list solely because of punctuation, capital letters or other trivial matters were combined. Some tri-grams that were unique only because of differences in the third word of the phrase were reduced to bi-grams. A list of relevant bi-grams and tri-grams used in this study is found in Table 4.

The next step used to develop EES applied machine learning and knowledge based techniques to detect patterns in the text of the corpus based on the bi-gram, tri-gram vocabulary list created. For this process, the KWIC system, SQL, and knowledge based techniques were used.

4.2.1 KWIC and SQL

Developed by Has Peter Luhn at IBM in 1958, KWIC uses automatic indexing to recognize word boundaries and frequencies [56]. For EES development, KWIC was used to analyze word placement in relation to other words in a sentence to determine patterns that exist in the text. For this procedure, the corpus document was further processed to format the text to be compatible with the KWIC system adapted for EES. The KWIC system parses by paragraphs, denoted by a period followed by a new line. Thus, carriage returns were added to the corpus document to designate the end of each sentence as one

Table 4. Relevant Bi-grams and Tri-grams used to Construct a Language Model for EES Development.

1998,	risk-free	interest
Average	risk-free	interest
Black	Scholes	valuation
Black-	Scholes	model
Black-	Scholes	pricing
Black-Scholes	formula	
Black-Scholes	method	
Black-Scholes	model	
Black-Scholes	multiple	option
Black-Scholes	option	pricing
Black-Scholes	option-pricing	model
Black-Scholes	pricing	model
Black-Scholes	single	
Black-Scholes	valuation	
Compensation	cost	
Compensation	expense	
Dividend	yield	
Dividend	yields	
Estimated	Life	
Expected	dividend	yield
Expected	life	
Expected	lives	
Expected	volatility	
Fair	amount	value
Fair	cost	losses
Fair	value	of
Forma	Net	income
Forma	Results	-
Option	Plan	
Option	Plans	
Pro	forma	loss
Pro	forma	
Pro-forma	earnings	per
Pro-forma	earnings	
Pro-forma	income	
Pro-forma	results	of
Risk-free	interest	rate
risk-free	interest	
risk-free	rate	
Risk-free	weighted	average

paragraph. The results allowed all sentences of the corpus to be included in the KWIC process.

The KWIC system loaded each word of the cleaned file into the first column of each row of an Oracle database table. Each row in the database consisted of 65 additional columns that contain words in the text that follow the word in the first column. The result was a shifting pattern that allowed each word of the corpus to appear in each column of the database. The database table contained 8,838,914 rows of data representing roughly one row for each word in the corpus file. An example of the database format of the KWIC output is in Table 5.

SQL queries were used to sort, categorize, and analyze word sequence patterns in the database. SQL was developed by IBM research and is now the standard language used for querying database systems. Using the bi-grams and tri-grams developed from the CMU-SLM toolkit as target word phrases, simple SQL statements were designed to place key words in specific columns of the database table. From the SQL outputs, words that surrounded the key phrases were analyzed for similar patterns to develop algorithms for the extraction process. The analysis was done by importing the SQL output data into an Excel spreadsheet.

Various sorting tools were used in Excel to analyze patterns in the text. For example, the data was sorted and arranged on various columns using the basic Sort feature in Excel. In addition, using the Auto Filter feature in Excel, a list of unique words for each column was easily determined and used for additional classifications. Auto Filter was also used to determine the format of numbers and words in the text. For

Table 5. Data output from The KWIC Index System.

An example from Item 8, Note F Notes to the Financial Statements from Ross Stores, Inc.'s 10-K for fiscal 2001. (For this example the database has been reduced to 10 columns rather than the 65 used in the actual database.)

Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
The	fair	values	for	each	option	granted	were	estimated	on
fair	values	for	each	option	granted	were	estimated	on	the
values	for	each	option	granted	were	estimated	on	the	date
for	each	option	granted	were	estimated	on	the	date	of
each	option	granted	were	estimated	on	the	date	of	grant
option	granted	were	estimated	on	the	date	of	grant	using
granted	were	estimated	on	the	date	of	grant	using	the
were	estimated	on	the	date	of	grant	using	the	Black-Scholes
estimated	on	the	date	of	grant	using	the	Black-Scholes	option
on	the	date	of	grant	using	the	Black-Scholes	option	pricing
the	date	of	grant	using	the	Black-Scholes	option	pricing	model
date	of	grant	using	the	Black-Scholes	option	pricing	model	with
of	grant	using	the	Black-Scholes	option	pricing	model	with	the
grant	using	the	Black-Scholes	option	pricing	model	with	the	following
using	the	Black-Scholes	option	pricing	model	with	the	following	assumptions
the	Black-Scholes	option	pricing	model	with	the	following	assumptions	for
Black-Scholes	option	pricing	model	with	the	following	assumptions	for	2000,
option	pricing	model	with	the	following	assumptions	for	2000,	1999,
pricing	model	with	the	following	assumptions	for	2000,	1999,	and
model	with	the	following	assumptions	for	2000,	1999,	and	1998,

example, the value for risk-free interest was most often shown using a percent sign, as 6.12%, but also occurred frequently without the sign. The most frequent patterns found were used to develop the algorithms for EES. An example of a SQL output from the corpus database is found in Table 6.

4.2.2 WordNet

WordNet is a lexical database comprised of sets of synonyms (synets) that represent a concept. The database recognizes and organizes nouns, verbs, adverbs, and adjectives into machine-readable semantic relations. The semantic relations are represented by pointers between words and synets [61]. Developed by George A. Miller in the early 1990s, WordNet has been used extensively in NLP research. Bagga, et al. used WordNet in 1996 to classify information into hierarchical categories that can be adapted to develop a variety of IE systems [6].

For EES, WordNet was used in the corpus analysis to determine synonyms for key words determined by the CMU-SLM and KWIC analysis. The synonym list created by WordNet was evaluated by a domain expert who determined that the list created by WordNet did not provide relevant synonyms for the domain-specific vocabulary used in financial reporting. Furthermore, WordNet provided no synonyms for commonly used compound words, such as fair-value, pro-forma, and risk-free. This is consistent with Vorhees' (1994) findings where synonyms were used to test query expansion [83]. The study found that using synonyms derived from WordNet did not effectively improve the results of queries [83]. An example of the output from WordNet is found in Table 7.

Table 6. SQL Output from the Corpus Database.

In this example the SQL query selects the word in column 2 as ‘risk-free’ and the word in column 3 as “interest”. The output was imported into Excel and was sorted by word 5, then word 6 and word 1. (For this example the spreadsheet has been reduced to 11 columns rather than the 65 used in the actual database.)

Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10	Word11
years,	risk-free	interest	rate	of	0.96%	to	2.49%,	expected	volatility	of
life,	risk-free	interest	rate	of	3%	and	a	market	value	of
life,	risk-free	interest	rate	of	3%	and	a	market	value	of
2000):	risk-free	interest	rate	of	5.11%	and	5.75%,	respectively	no	dividend
average	risk-free	interest	rate	of	6.20%	and	an	expected	volatility	of
average	risk-free	interest	rate	of	6.20%	in	2000,	6.00%	in	1999
average	risk-free	interest	rate	of	6.20%	in	2000,	6.00%	in	1999
average	risk-free	interest	rate	of	6.20%	and	an	expected	volatility	of
2000:00:00	risk-free	interest	rate	of	1.67%,	5.11%,	and	5.75%	respectively	no
0%,	risk-free	interest	rate	of	2.0%,	expected	life	of	4	years,
a	risk-free	interest	rate	of	2.49%.					
a	risk-free	interest	rate	of	2.49%.					
a	risk-free	interest	rate	of	2.72%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	2.72%;	and	(iv)	no	dividends.	
assumptions:	risk-free	interest	rate	of	3.25%,	contractual	life	of	5	years,
a	risk-free	interest	rate	of	3.4%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	3.4%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	3.71%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	3.71%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	3.71%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	4.0%;	and	(iv)	no	dividends.	
a	risk-free	interest	rate	of	4.0%;	and	(iv)	no	dividends.	
2001:00:00	risk-free	interest	rate	of	4.25%,	1.67%,	and	5.11%,	respectively	dividend
follows:	risk-free	interest	rate	of	4.3%;	dividend	yield	0%;	expected	life
a	risk-free	interest	rate	of	4.35%,	volatility	of	86.88%	and	no
a	risk-free	interest	rate	of	4.35%,	volatility	of	86.88%	and	no

Table 7. List of Synonyms Created by WordNet.

Stock

placental	handle	stem
eutherian	support	working_capital
eutherian_mammal	render	accumulation
placental_mammal	lumber	security
cravat	timber	certificate
merchandise	soup	repute
product	animal_group	reputation
provide	genealogy	have
wares	family_tree	have_got
handle	flower	supply
grip	flower	render
handgrip	plant_part	furnish
hold	stalk	provide

Option

action	deciding	derivative
decision_making	derivative_instrument	

Compensation

correction	defense_mechanism	defense
rectification	defence_reaction	defence_mechanism
recompense	defence	defense_reaction

Volatility

emotionality	irresoluteness	chemical_property
emotionalism	irresolution	

Dividend

Bonus	net	profits
incentive	profit	lucre
net_income	earnings	net_profit
number		

Expense

detriment	cost	expenditure
hurt	outgo	outlay

Cost

value		
outgo	outlay	involve
expenditure	necessitate	take
	demand	call_for

4.2.3 Stemming

Stemming is another method used in IE system development to improve recall. Stemming involves the removal of the inflectional ending, such as –ed, -ing, and -s, from words to reduce word forms to its common root. The process can be as simple as handling plurals, but can be expanded to handle complex word structures. Stemming can the extraction process by expanding queries with related word forms [86].

M. F. Porter developed one of the most well known stemming programs in 1980. The Porter Stemmer automatically removes endings from words until conditions are met for termination of the process. The process proceeds through several steps bit by bit until the desired result is obtained. A problem with stemming is that often the removal of the suffix changes the meaning of the word [70]. Although the Porter Stemmer produces results quickly, the result is often aggressive and produces stems that are not actually words [86].

Corpus-based stemming is a process designed to stem words to suit a given text based corpus. Stemming in relation to a given corpus can improve the effectiveness of query expansion for the specific domain. Root words that can be determined from the text of the corpus will produce more effective results since words have different meanings in different contexts [86]. Thus stemming is often corpus specific. The result of the Porter Stemmer does not apply universally.

The stemming process was incorporated in EES during the corpus development through SQL, and in the wrapper development, through Perl. Using key words and phrases determined by CMU-SLM analysis, the root of these words were combined with wildcard characters for SQL analysis. For example, in the SQL query for options, the

wildcard character “%” was combined with the root word “option” (option%). The query returned three results – “option”, “options” and a third compound word “option-stock”. Using wildcard characters in queries expanded the selection to include all relevant references in the corpus to the key words and phrases.

Wrapper development wildcards were also used to improve recall and target all references to the key word. In Perl various characters are used as wildcards in regular expressions. A dot (.) is used to match any single character, while the asterisk (*) will match any number of characters [74]. The use of wildcard characters with root words in the wrapper helped expand the pattern matching process to improve recall of the system.

4.2.4 Knowledge-based Analysis

An expert with knowledge of the domain periodically assessed rules generated from the analysis and made modifications necessary to increase performance of the system. For example, n-grams and frequencies from the CMU-SLM toolkit output were analyzed by the expert to determine the most relevant words and phrases for further processing. Expert knowledge is essential to wrapper development to anticipate patterns that may not be evident in the corpus. Although this trial and error method is time consuming and labor intense, the knowledge engineering approach often produces systems that outperform automatically trained systems [4]. EES combined automatic approaches with the enhancement of expert domain knowledge. This dual approach to EES helped reduce the time of a hand crafted system, while increasing the performance level associated with the machine learning process.

4.3 EES Extractor Development

EES extracts information about the fair value of stock options and the method and assumption used for the valuation from the notes to the financial statements for companies that file with the SEC. An example of the information available in a typical 10-K document is in Figure 2. The example is taken from Item 8, Note F of the Notes to the Financial Statements from Ross Stores, Inc's 10-K for fiscal year ended February 3, 2001.

The steps involved in the process included (1) downloading the 10-K files into a buffer directory on the local hard drive, (2) parsing the files into text format and extracting the sections of the files that contain the financial statement and accompanying notes, (3) separating the tagged tables from the text part of the file creating two new files, and (4) developing the wrapper to extract the specific information using the algorithms developed from the training corpus.

4.3.0 Downloading 10-K Files

To speed the extraction process, EES first locates and downloads the specific 10-K files from the SEC's EDGAR database to a directory on the local system. For this study, 19 companies for the years 2001-2004 were tested, thus 76 10-K files were downloaded. For this study, all 76 10-Ks were successfully downloaded from the SEC's EDGAR database in approximately 8 minutes using an 11.0 Mbps cable modem. The average time to download each 10-K is about 6.5 seconds.

Stock-Based Compensation Plans. At February 3, 2001, the company had five stock-based compensation plans, which are described below. Statement of Financial Accounting Standards No. 123 (SFAS 123), "Accounting for Stock-Based Compensation," establishes a fair value method of accounting for stock options and other equity instruments. Had compensation cost for these stock option and stock purchase plans been determined based on the fair value at the grant dates for awards under those plans consistent with the methods of SFAS 123, the company's net income and earnings per share would have been reduced to the pro forma amounts indicated below:

		2000	1999	1998
(\$000, except per share data)				
Net income	As reported	\$ 151,754	\$ 150,106	\$ 133,843
	Pro forma	\$ 143,399	\$ 142,800	\$ 128,820
Basic earnings per share	As reported	\$ 1.84	\$ 1.66	\$ 1.42
	Pro forma	\$ 1.74	\$ 1.58	\$ 1.37
Diluted earnings per share	As reported	\$ 1.82	\$ 1.64	\$ 1.40
	Pro forma	\$ 1.74	\$ 1.57	\$ 1.36

At year-end 2000, 1999 and 1998, there were 6.6 million, 4.4 million and 5.7 million shares, respectively, available for future issuance under these plans.

The weighted average fair values per share of options granted during 2000, 1999 and 1998 were \$8.19, \$7.85 and \$6.21, respectively. For determining pro forma earnings per share, the fair values for each option granted were estimated on the date of grant using the Black-Scholes option pricing model with the following assumptions for 2000, 1999 and 1998, respectively: (i) dividend yield of 0.8%, 0.7% and 0.6%; (ii) expected volatility of 56.0%, 46.1% and 45.8%; (iii) risk-free interest rate of 6.3%, 5.9% and 5.2%; and (iv) expected life of 3.4 years, 3.2 years and 3.3 years. The company's calculations are based on a multiple option approach, and forfeitures are recognized as they occur.

Figure 2. An Example of Information Available on Stock Options from a 10-K Annual Report on EDGAR.

A sub-directory for each company is created and the 10-K files are stored in the company sub-directory by year. The EES wrapper takes advantage of the Library for World Wide Web in Perl (LWP) module. LWP is a set of Perl modules that automates finding and downloading files on the Web [14]. The company's CIK number is used to assure that the proper file is downloaded. Two versions of the download script were developed. One script allows user input from the keyboard, that is, the user determines the company and year to download and process. This version of the script was used by the subjects testing the system. The second version develops an array of company names and CIK codes from a separate text file. This version allows a number of companies and years to be downloaded and processed at one time and was used in this study for the preliminary analysis.

4.3.1 Parsing

The next step of the extraction process parses the 10-Ks forms that are in the download directory and converts the files to text. The resulting text, or parsed, files are stored in a separate data directory for further processing. From these parsed files, sections, or items, that contain the financial statements and accompanying notes to the financial statements are extracted to a second file in the same directory.

For most 10-K files, the financial statements and accompanying notes are found in Item 8 of the 10-K. However, in some cases, a statement found in Item 8 makes reference to the financial statements and accompanying notes being in another part of the file, usually a later item that also contains other exhibits and schedules. These other exhibits and schedules are often found in either Item 14, Item 15, or in some other section

not designated by an item number. In order to capture of all the information needed in the EES process, Items 14 through the end of the 10-K file are extracted from the parsed file to a separate item file. The resulting file, containing the financial statements and accompanying notes from the items in the 10-K, is further processed to separate the formatted tables from the text portion of the file.

4.3.2 Data Separation

A visual inspection of the corpus 10-Ks indicated that the stock option information to be extracted could be presented in either a table or text format. To speed the extraction process, data in the item files, containing the financial statements and accompanying notes, are separated and copied into two files- tables and text. The table data is identified by the SGML tags<Table> <Table>. Data found between these two tags become the table file. The remaining text is copied to the text file.

All 76 10-K files were successfully parsed into the four separate files needed for processing and placed into the input buffer. Of the four files, one file contains the data remaining from the 10-K after the section, or item, containing the financial statements information is parsed. This file is used to extract the company name and fiscal year end from the heading of the 10-K. The second file consists of the items containing the financial statements. The third file contains the tables from the financial statements and accompanying notes, while the fourth file consists of the text portion of the items of the financial statements and accompanying notes. Creating four files increases the speed of the extraction process by allowing more efficient search patterns to be created.

4.3.3 Wrapper Development

The wrapper is written in Perl. Perl is a popular program for system designers and for web development. It was originally designed for text processing but has grown into a sophisticated system. Much of its growth is due to the free availability of the Perl program and the numerous Perl modules and libraries. Although developed on a Unix system, it has been adapted to operate on Windows and Macintosh platforms [68].

Perl soared in popularity because it is adept at creating, managing, and extracting information from the Web using its LWP. Perl also supports regular expressions. Regular expressions work like a mini program to describe and parse text. Regular expressions can be used to isolate specific passages, find and replace text, and perform various types of text and data manipulation with just a few lines of code. When combined with the grep command, Perl can extract and print text for the regular expression patterns that are matched [39]. Even though Perl is open source code, it is quite stable and normally performs as expected [14].

4.3.4 Extraction and Results

The actual extraction process occurs in four steps. The first step extracts the company name and fiscal year end from the header of the parsed text file to an HTML output file. The file name, containing the CIK and filing year, is also included in output file. The second step determines if the financial statements and accompanying notes are included in the 10-K file. The third step extracts information regarding two sets of information. Pro-forma and fair value information is the first set of information extracted. EES first searches for the information in the table file then searches for the

information in text file. EES then searches for the second set of information, the assumptions used to value the stock options. The search for assumption information uses the same procedure. Finally, the option pricing model is extracted. All searches in EES are case insensitive. An example of the output from Express Scripts, Inc., 10-K for fiscal year end December 31, 2003 is shown in Figure 3.

Precision, Recall, and the F-measure were calculated for the output of the actual data extraction process. Recall was measured by dividing the number of correct answers produced by EES by the total possible correct answers. Precision was calculated by dividing the number of correct answers produced by EES by the number of total answers produced by EES. The F-measure for EES analysis assumes an equal weight of recall and precision using the formula $F=2RP/R+P$. EES's overall recall was 82.71%, precision was 72.62% and the F-measure was 77.34%.

Step one – Extracting Header Information

Name entity recognition, key word searches, and pattern matching techniques were combined with a knowledge-based approach to extract information from the heading of the 10-K. EES takes advantage of the HTML formatting in the original tables of the 10-K. An example of the header information found in a typical 10-K file is Figure 4. The example is taken from Ross Stores, Inc.'s 10_k for fiscal year ended February 3,

CIK: 0000885721_2004

Name: EXPRESS SCRIPTS, INC.

FYE: FOR THE FISCAL YEAR ENDED DECEMBER 31, 2003

Stock Option Pro Forma Information

(in thousands, except per share data)	2003	2002	2001
Stock-based compensation, net of tax			
As reported	\$ 4,437	\$ 5,102	\$ 5,553
Pro forma	16,294	16,479	15,424
Net income			
As reported	\$ 249,600	\$ 202,836	\$124,700
Pro forma	237,743	191,458	114,937
Basic earnings per share			
As reported	\$ 3.21	\$ 2.60	\$ 1.60
Pro forma	3.05	2.46	1.48
Diluted earnings per share			
As reported	\$ 3.16	\$ 2.55	\$ 1.56
Pro forma	3.00	2.39	1.44

Assumptions

	2003	2002	2001
Expected life of option	3-10 years	3-5 years	2-5 years
Risk-free interest rate	1.6%-3.7 %	1.4%-5.0 %	1.7%-4.9 %
Expected volatility of stock	52%-53 %	54 %	55 %
Expected dividend yield	None	None	None

Option Pricing Model

Black Scholes

Figure 3. An example of the Output File from EES.

**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549**

FORM 10-K

(Mark One)

- ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE
SECURITIES EXCHANGE ACT OF 1934
For the fiscal year ended February 3, 2001**
OR
- ANNUAL REPORT PURSUANT TO SECTION 13 or 15 (d) OF THE
SECURITIES EXCHANGE ACT OF 1934 [NO FEE REQUIRED]
For the transition period from _____ to _____**

Commission file number 0-14678

ROSS STORES, INC.

(Exact name of registrant as specified in its charter)

Delaware
(State or other jurisdiction of
incorporation or organization)

94-1390387
(I.R.S. Employer Identification No.)

8333 Central Avenue, Newark, California
(Address of principal executive offices)

94560-3443
(Zip Code)

Registrant's telephone number, including area code:

(510) 505-4400

Securities registered pursuant to Section 12(b) of the Act:

None

Figure 4. Information Provided in the Header of a 10-K Annual Report on EDGAR.

2001. For this portion of the extraction process the average recall, precision, and F-measure were 92%, 97%, and 95% respectively.

Step two – Determining the Existence of the Financial Statements

The second step determines if the file actually contains the financial statements and accompanying disclosure notes in the 10-K document. The SEC allows financial statement information and/or the Auditor's Report to be referenced in the Annual Report to Shareholders, which may be in another document associated with the 10-K [77]. If a company chooses to reference its financial statements, the company must state that fact in Item 8 of its 10-K. In these cases the information regarding stock option compensation expense is not included in the 10-K document. Figure 5 is an example of Intel Corporation's 10-K that incorporates its financial statements by reference to its Annual Report to Shareholders.

EES is designed to search for stock option information only in a 10-K. If the information is not available because it is incorporated by reference in the company's Annual Report to Shareholders, EES returns a string in the output file stating that the financial statements are not available. If the information is available, EES refers back to the knowledge based domain and proceeds with the extraction process. An example of

<p>ITEM 8. FINANCIAL STATEMENTS AND SUPPLEMENTARY DATA</p> <p>Consolidated financial statements of Intel at December 29, 2001 and December 30, 2000, and for each of the three years in the period ended December 29, 2001 and the Report of Independent Auditors thereon, and our unaudited quarterly financial data for the two-year period ended December 29, 2001 are incorporated by reference from our 2001 Annual Report to Stockholders, on pages 20 through 37.</p>

Figure 5. Reference to Annual Report to Stockholders.

CIK: 0000050863_2002
Name: INTEL CORPORATION
FYE: FOR THE FISCAL YEAR ENDED DECEMBER 29, 2001

The financial statement and notes are referenced in the Annual Report to Shareholders. They are not included in this 10-K, therefore cannot be found using this system.

Figure 6. An example of the Output from EES when the Financial Statements are Incorporated by Reference in the Annual Report to Shareholders.

an EES output when the financial statements are reference to the Annual Report to Shareholders is presented in Figure 6. The example is from the output of Intel Corporation, fiscal year ended December 29, 2001.

Several search patterns were developed from the training corpus to determine if the financial statements and accompanying notes were part of the 10-K. The most common phrase found in the corpus, “incorporated.by.reference”, was used in the first pattern search. The dots serve as a wildcard for any one character in the string. However, a total of 13 files were searched incorrectly using this pattern. Several iterations using phrases found in the training corpus were tried, but the phrase “annual report.*holder” produced the best results with recall of 80%, precision of 100%, and a F-measure of 88.89%. The “.*” quantifier serves as a wildcard to match any number of characters between report and holder and allows the system to identify stockholder, as well as shareholder.

Step Three – Extracting Stock Option Information

The third step searches the table and text buffer input files to extract pro-forma and fair value information as well as the assumptions used by the company to value the

stock options. The extraction process begins by extracting tables that contain the targeted information from the table files of the input buffer.

The SGML tags, <table> <\table>, are used to identify tables in the file. Key word phrases and patterns derived from the corpus analysis are used to extract the entire table where the phrase is found. Since the information is already in a structured format, no additional formatting was needed. The two patterns used to produce the best results are “as reported|pro.forma.*net” to extract the pro forma and fair value information and “risk-free|dividend yield|volatility|expected life” to extract the assumptions. The first pattern “as reported|pro.forma.*net” uses wild cards and the “or” operator, “|”, to search for tables containing the fair value of stock options and the pro-forma information.

The second search pattern, “risk.free|dividend yield|volatility|expected life” is used to search for tables that contain the assumptions. Although each of the assumptions appears in different forms throughout the 10-Ks, the “or” operator allowed EES to extract the table that contains any of the four specified assumptions in the pattern. Searching for each assumption individually frequently produces duplicates of the same table.

Next EES searched the text files for the same two sets of information, pro-forma and fair value information and assumptions used. The data in the text file contained semi-structured as well as unstructured text and was the most difficult of the extraction tasks. Various key word searches were used to extract from the text portion of the files. A key word search for “fair*value” was combined with the term “as reported”. In this search, the system searched for the key word ‘fair value’, but only printed the extracted text if “as reported” appeared in the same block. A similar search for the assumptions used the key word “risk*free” and printed the extracted text only if a percent sign (%)

appeared in the same block. For the pro-forma and fair value search, overall recall was 78.59% with precision of 64.02% and an F-measure of 70.56%. The result of the assumptions extraction was greater with recall of 90.61%, precision of 75.69%, and an F-measure of 82.48%.

Step 4 – Extracting the Option Pricing Model

The last step in the EES process extracted the option-pricing model used by the company to value the stock options. Various forms of key word searches were used. Recall, precision, and the F-measure were all 100% when extracting the option pricing model.

Chapter 5

Testing and Analysis

5.0 Overview

Testing compared EES extraction results to manual extraction of stock option information from 10-K documents on the EDGAR database. For the test, the randomly selected 19 companies from the NASDAQ-100 Index were used. The companies were tested to extract information for a four-year period, 2001-2004. One data extraction form was incomplete and was eliminated from the test sample. Thus, a total of 75 10-K files were tested. Detailed precision, recall, and F-measure results from EES are in Table 8.

5.1 Testing

Forty-two accounting students enrolled in “Accounting Information Systems” participated in testing EES. Subjects were given an evaluation questionnaire to provide information about demographics, accounting courses previously taken or enrolled in, and comments on their perception of the usefulness of EES. The survey questionnaire is shown in Figure 7. Twelve subjects were classified as juniors, 22 were seniors, and 6 were classified as graduate students. Two subjects did not report their college classification level. Thirty-eight of the 41 subjects had taken or were currently enrolled in Intermediate Accounting I, while 24 subjects had taken or were currently enrolled in Intermediate Accounting II. The average number of accounting courses taken by the test subjects was 4.36 courses. All courses included in the survey were junior level or above.

Thirty-two students tested two company 10-Ks; eleven subjects tested only one company. The test was conducted in a classroom setting using a DSL Web interface, Pentium 4 processors, with 128 MB of RAM.

Table 8. Recall, Precision, and F-measure for all 76 10-Ks Tested.

	CIK	EES Recall	EES Precision	EES F-Measure	EES Time in Seconds	Manual Recall	Manual Precision	Manual F-Test Measure	Manual Time in Seconds
AMGEN	318154								
2001		81.25%	74.29%	77.61%	34	93.55%	90.63%	92.06%	1920
2002		90.63%	82.86%	86.57%	33	100.00%	100.00%	100.00%	600
2003		81.82%	85.71%	83.72%	36	95.35%	95.35%	95.35%	1980
2004		100.00%	87.76%	93.48%	37	54.76%	85.19%	66.67%	1980
BIOMET	351346								
2001		100.00%	96.77%	98.36%	32	86.21%	83.33%	84.75%	1080
2002		63.64%	61.76%	62.69%	32	9.38%	16.67%	12.00%	960
2003		100.00%	97.78%	98.88%	52	58.14%	89.29%	70.42%	1980
2004		100.00%	97.78%	98.88%	29	90.70%	100.00%	95.12%	840
CDW	899171								
2001		85.37%	100.00%	92.11%	31	37.50%	100.00%	54.55%	1080
2002		97.50%	79.59%	87.64%	32	74.36%	76.32%	75.32%	900
2003		100.00%	56.79%	72.44%	36	48.89%	53.66%	51.16%	1080
2004		100.00%	56.79%	72.44%	30	93.33%	91.30%	92.31%	1320
CHECKPOINT	215419								
2001		80.56%	78.38%	79.45%	30	91.43%	88.89%	90.14%	780
2002		80.56%	78.38%	79.45%	31	71.43%	59.52%	64.94%	780
2003		37.78%	94.44%	53.97%	25	90.91%	100.00%	95.24%	480
2004		37.78%	94.44%	53.97%	30	59.09%	61.90%	60.47%	900
CINTAS	723254								
2001		100.00%	100.00%	100.00%	30	0.00%	0.00%	0.00%	900
2002		100.00%	100.00%	100.00%	30	100.00%	100.00%	100.00%	300
2003		100.00%	100.00%	100.00%	81	100.00%	100.00%	100.00%	120
2004		100.00%	100.00%	100.00%	39	100.00%	100.00%	100.00%	540
CITRIX	877890								
2001		100.00%	100.00%	100.00%	35	10.26%	10.00%	10.13%	900
2002		100.00%	100.00%	100.00%	32	100.00%	100.00%	100.00%	480
2003		83.78%	56.36%	67.39%	36	100.00%	100.00%	100.00%	480
2004		100.00%	61.54%	76.19%	30	94.87%	97.37%	96.10%	420

Table 8 (continued).

	CIK	EES Recall	EES Precision	EES F-Measure	EES Time in Seconds	Manual Recall	Manual Precision	Manual F-Test Measure	Manual Time in Seconds
DELL	826083								
2001		62.50%	86.96%	72.73%	37	87.10%	100.00%	93.10%	600
2002		62.50%	86.96%	72.73%	36	70.97%	64.71%	67.69%	540
2003		100.00%	58.90%	74.14%	36	64.29%	75.00%	69.23%	1200
2004		100.00%	87.76%	93.48%	36	100.00%	100.00%	100.00%	780
EBAY	1065088								
2001		100.00%	30.77%	47.06%	38	100.00%	100.00%	100.00%	1020
2002		100.00%	100.00%	100.00%	38	100.00%	100.00%	100.00%	1800
2003		100.00%	34.96%	51.81%	41	23.81%	58.82%	33.90%	1740
2004		100.00%	100.00%	100.00%	41	100.00%	100.00%	100.00%	2220
EXPRESS SCRIPTS	885721								
2001		100.00%	100.00%	100.00%	31	100.00%	97.62%	98.80%	1920
2002		100.00%	100.00%	100.00%	32	97.56%	100.00%	98.77%	2520
2003		100.00%	100.00%	100.00%	31	93.33%	93.33%	93.33%	1320
2004		100.00%	100.00%	100.00%	31	60.00%	90.00%	72.00%	1200
INTEL	50863								
2001		75.00%	75.00%	75.00%	34	100.00%	9.09%	16.67%	900
2002		100.00%	100.00%	100.00%	33	100.00%	100.00%	100.00%	1020
2003		100.00%	74.14%	85.15%	37	71.43%	76.92%	74.07%	2220
2004		100.00%	74.14%	85.15%	37	66.67%	65.12%	65.88%	780
JUNIPER	1043604								
2001		33.33%	100.00%	50.00%	32	100.00%	50.00%	66.67%	900
2002		33.33%	100.00%	50.00%	71	81.82%	21.43%	33.96%	1020
2003		93.48%	46.24%	61.87%	43	88.89%	93.02%	90.91%	1200
2004		93.48%	84.31%	88.66%	40	95.56%	100.00%	97.73%	1200
LEVEL 3	794323								
2001		75.00%	75.00%	75.00%	31	100.00%	100.00%	100.00%	1200
2002		50.00%	40.00%	44.44%	31	100.00%	100.00%	100.00%	1200
2003		66.67%	40.00%	50.00%	33	69.57%	53.33%	60.38%	2400
2004		100.00%	40.00%	57.14%	37	73.33%	34.38%	46.81%	2340

Table 8 (continued).

	CIK	EES Recall	EES Precision	EES F-Measure	EES Time in Seconds	Manual Recall	Manual Precision	Manual F-Test Measure	Manual Time in Seconds
MAXIM	743316								
2001		100.00%	100.00%	100.00%	31	33.33%	3.03%	5.56%	2040
2002		90.32%	57.14%	70.00%	32	93.33%	100.00%	96.55%	720
2003		88.10%	63.79%	74.00%	33	73.17%	69.77%	71.43%	1140
2004		85.71%	62.07%	72.00%	32	100.00%	97.62%	98.80%	1020
MEDI-MMUNE	873591								
2001		42.86%	100.00%	60.00%	32	100.00%	100.00%	100.00%	480
2002		35.71%	100.00%	52.63%	31	100.00%	100.00%	100.00%	480
2003		11.90%	83.33%	20.83%	34	31.71%	34.21%	32.91%	600
2004		100.00%	100.00%	100.00%	33	90.48%	100.00%	95.00%	480
MICROSOFT	789019								
2001		47.06%	84.21%	60.38%	74	100.00%	100.00%	100.00%	720
2002		11.76%	66.67%	20.00%	39	87.88%	87.88%	87.88%	1200
2003		88.89%	83.33%	86.02%	37	93.18%	93.18%	93.18%	660
2004		20.00%	66.67%	30.77%	36	78.95%	60.00%	68.18%	300
MOLEX INC	67472								
2001		100.00%	100.00%	100.00%	32	100.00%	100.00%	100.00%	18600
2002		100.00%	100.00%	100.00%	34	100.00%	7.14%	13.33%	300
2003		100.00%	100.00%	100.00%	37	100.00%	7.14%	13.33%	600
2004		37.50%	94.74%	53.73%	31	93.62%	100.00%	96.70%	540
ROSS STORES	745732								
2001		82.93%	73.91%	78.16%	34	22.50%	21.95%	22.22%	600
2002		100.00%	72.73%	84.21%	31	100.00%	100.00%	100.00%	480
2003		97.83%	73.77%	84.11%	37	91.11%	100.00%	95.35%	900
2004		97.83%	73.77%	84.11%	37	88.89%	100.00%	94.12%	900

Table 8 (continued).

	CIK	EES Recall	EES Precision	EES F-Measure	EES Time in Seconds	Manual Recall	Manual Precision	Manual F-Test Measure	Manual Time in Seconds
SIGMA ALDRICH	90185								
2001		50.00%	40.00%	44.44%	32	100.00%	10.00%	18.18%	900
2002		100.00%	100.00%	100.00%	35	100.00%	16.67%	28.57%	1200
2003		75.00%	50.00%	60.00%	38	100.00%	7.89%	14.63%	900
2004		75.00%	50.00%	60.00%	35	100.00%	7.32%	13.64%	1200
SYMANTEC	849399								
2001		52.63%	48.78%	50.63%	42	75.68%	84.85%	80.00%	1740
2002		57.14%	57.14%	57.14%					
2003		100.00%	78.18%	87.76%	34	100.00%	100.00%	100.00%	840
2004		100.00%	75.41%	85.98%	32	100.00%	100.00%	100.00%	300
Totals		82.71%	72.62%	77.34%	2695	80.66%	73.78%	77.06%	95880

Survey Evaluation

Please answer the questions below.

Age _____

Level in college (Sophomore, Junior, Senior, Graduate, Special)

Gender (M, F) _____

Please indicate accounting courses you have taken or are currently enrolled in (check all that apply).

_____ Intermediate 1

_____ AIS

_____ Advanced 2

_____ Intermediate 2

_____ Tax 1

_____ Cost

_____ Tax 2

_____ Auditing

_____ Advanced 1

_____ Other accounting courses taken or currently enrolled in (specify)

Do you think EES is useful for extracting stock option disclosure information from the EDGAR Database?

Briefly, please explain why.

Figure 7. Survey Administered to Subject that Tested EES.

After a brief overview of the SEC's EDGAR system, each subject was given a data extraction form noting the information to be extracted manually from the EDGAR database. Each student's beginning and ending time was recorded for the extraction process for each 10-K. The data collection form used by the students is provided in Figure 8.

The subjects then used EES to extract the information from the same 10-K. Comparisons between the manual and automated extraction process were made to evaluate the speed, precision, recall, and the F-measure.

After the data collection processes were complete, four hypotheses were tested as described below.

Recall

The amount of relevant information extracted by the subjects for the manual extraction process was measured using the standard recall formula used in IE. Recall was measured by dividing the number of correct answers produced by the total possible correct answers. Overall, recall for the manual extraction process was 80.66% compared to recall from EES of 82.71%.

H1: There is no difference in the population means of recall between the automated EES process and the manual process of extracting information about stock options from 10-K Annual Reports on the SEC's EDGAR Database.

Precision

Precision was measured by dividing the number of correct answers produced by the number of total answers produced. Precision for the manual extraction was 73.78% compared to 72.62% for EES.

Data Collection Form

For the following company and year use the company's 10-K annual reports from the EDGAR database to find the information needed to fill the following blank lines.

Company _____ CIK Code _____

Filing Year _____ Beginning Time _____

Company Name (Exact name of registrant as specified in its charter):

Fiscal Year End: _____

Stock Option Pro-forma Information:

Years Reported	_____	_____	_____
Income (loss) as Reported	_____	_____	_____
Income (loss) Pro Forma	_____	_____	_____
Basic EPS as Reported	_____	_____	_____
Basic EPS Pro Forma	_____	_____	_____
Diluted EPS as Reported	_____	_____	_____
Diluted EPS Pro Forma	_____	_____	_____
Fair Value of Option	_____	_____	_____

Assumptions Used to Value Stock Options:

Years Reported	_____	_____	_____
Dividend Yield	_____	_____	_____
Volatility	_____	_____	_____
Risk Free Interest Rate	_____	_____	_____
Expected Life	_____	_____	_____
Model Used	_____		

If the information is unavailable, please indicate why.

Ending Time _____

Figure 8. Template Used by Subjects for Manual Data Collection for Testing EES.

H2: There is no difference in the population means of precision between the automated EES process and the manual process of extracting information about stock options from 10-K Annual Reports on the SEC's EDGAR Database.

F-measure

In order to compare the two systems, the standard F-measure was used to combine the results of precision and recall. The F-measure used for this analysis assumes an equal weight of recall and precision. The F-measure for the manual process was 77.06% compared to the F-measure for EES of 77.34%.

H3: There is no difference in the population means of the F-measure between the automated EES process and the manual process of extracting information about stock options from 10-K Annual Reports on the SEC's EDGAR Database.

Speed

Subjects were timed on both the manual and automated process. The average speed in the manual extraction process was approximately 21.5 minutes, compared to the average speed EES extraction of approximately 36 seconds.

H4: There is no difference in the in the population means of the time between the automated EES process and the manual process of extracting information about stock options from 10-K Annual Reports on the SEC's EDGAR Database.

5.2 Analysis of Statistical Comparisons

Table 9 provides descriptive data collected from the manual and EES data extractions. Paired sample t-tests were conducted to test the four hypotheses. Table 10 shows the results of the t-test. There was no evidence (p-value = .9970) that a difference in recall exists between the manual extraction (mean = .8211) and EES (mean = .8209). Thus, the statistical results failed to reject H1 indicating that EES extracted relevant information comparable to the manual extraction of stock option information. The results

Table 9. Descriptive Data from Manual and EES Data Extraction.

Descriptive Data						
Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Sample Size
EES Recall	.8209	.9783	.2509	.1176	1.00	75
EES Precision	.7967	.8333	.2040	.3077	1.00	75
EES F-Measure	.7722	.7945	.2100	.2000	1.00	75
EES Time in Seconds	35.93	34.00	9.0530	25.00	81.00	75
Manual Recall	.8211	.9333	.2534	.0000	1.00	75
Manual Precision	.7454	.9302	.3373	.0000	1.00	75
Manual F-Measure	.7368	.9206	.3158	.0000	1.00	75
Manual Time in Seconds	1292.80	900.00	2101.00	300.00	18600.00	75
Accounting Courses Taken	4.36	4.00	2.2520	1.00	11	75

Table 10. Results of Paired T-Tests.

Paired Samples T-Test					
Paired Comparisons	Mean Difference	Standard Deviation	T-Value	Degrees of Freedom	P-Value
EES Recall/ Test Recall	-.0002	.3553	-.0040	74	.9970
EES Precision/ Test Precision	.0512	.3736	1.1870	74	.2390
EES F-Measure/ Test F-Measure	.0354	.3590	.8530	74	.3960
EES Time/ Test Manual Time	-1256.870	2101.1750	242.6230	74	0.0000

also failed to reject H2 since the precision of the manual system (mean = .7454) and EES (mean = .7967) showed no significant differences (p-value = .2390). Similar results failed to reject H3 as the F-measure revealed no significant differences between the manual extraction (mean = .7368) and EES (mean = .7722).

Statistical analysis rejected H4 since the subjects spent significantly more time extracting stock option information manually (mean = 1292.8) than they did using EES (mean = 35.93). The mean time savings for EES was 1,257 seconds, resulting in a p-value < 0.001. EES performed approximately 36 times faster than the subjects in the manual extraction. This is believed to be a substantial improvement and can result in a larger amount of information being processed.

Further analysis revealed that there was no significant correlation between the number of accounting courses the subjects had taken and their F-measure (p = .4600) or the time they took for the manual extraction (p = .1430). Even though there was substantial improvement in the time spent by the subject on their second 10-K extraction (p < 0.001), EES (mean = .3753) still significantly out performed (p < 0.001) the manual extraction (mean = 817.50). In addition, there was no significant improvement in the F-measure by the subjects on their second extraction attempt (p=.3760).

5.3 Survey Analysis

All 42 subjects completed a brief questionnaire regarding the usefulness of EES. Twenty nine of the subjects rated EES “useful,” 5 rated EES “both useful and not useful,” and 4 rated EES “not useful.” Four subjects did not respond to the question. The reason most cited for EES being useful was the speed of the automated system. Lack of

information extracted or wrong information extracted was the comment seen most when subjects responded that the system was not useful. Many comments that thought the system was useful also noted that some of the information was missing or incorrect. A summary of the survey results with open ended comments categorized is in Table 11.

One subject commented on the usefulness of EES in research:

I have been involved in a data look-up project. (I had to look up information in 10-Q filings regarding 404 disclosures and internal control weaknesses.) It was extremely time consuming and I had to limit the number of companies I looked at. This type of software would be extremely useful and would have made the work faster. I also would have been able to increase my population.

Table 11. Survey Results

Question 1	
Do you think EES is useful for extracting stock option disclosure information from the EDGAR Database?	Number of Responses
EES was useful	29
EES was useful and not useful	5
EES was not useful	4
No comment	4
Question 2	
Briefly, please explain why.	
EES is faster than manual extraction	15
EES provided inaccurate information or no information	13
Most of the information EES extracted was accurate	8
EES was convenient and easy to use	9
EES provided complete and accurate information	2
Errors can occur in a manual extraction	2
EES separates the information into categories	2
EES can be helpful in research	1
Not knowledgeable enough about stock options to determine if EES is useful	1
Computers do not get bored with tedious tasks	1
Downloading the 10-Ks to the hard drive is useful	1
EES may be useful for extracting other types of information	1
EES provided feedback about the availability of the information	1

Chapter 6

Conclusion, Limitations and Future Research

6.0 Overview

EES can be useful to extract stock option information from the 10-K files on the EDGAR database. Sixty-nine percent of the subjects surveyed perceived EES to be useful; twelve percent perceived to EES be both useful and not useful; and only 9.5% reported that EES would not be useful to extract stock option information from the disclosure notes of financial statements. EES extracted stock option information from the disclosure notes of 10-K files with recall, precision, and F-measures comparable to manual extraction. EES's speed greatly exceeded the manual process. EES is easy to use and provides a convenient method for a difficult extraction task.

6.1 Conclusion

Extracting information from the text of financial statements is a challenging but important application for IE. Methods developed in these systems, such as EES, can foster IE research by building on past systems and developing new techniques for future research.

EES has several advantages over other methods to extract information from financial statements on EDGAR. Unlike free third party services and EDGAR2xml, EES searches beyond the main financial statements and focuses on the semi-structured, information-rich notes to the financial statements. Also, EES does not rely on XML

tagging or taxonomies necessary for XBRL implementation, but uses a training corpus and NLP techniques to build a knowledge database of key phrases and patterns in the free text. Although EDGAR-Analyzer extracts information from the notes to the financial statements, it extracts paragraphs of text and does not display the information in a structured format.

The EES wrapper is written in PERL. PERL is an opened-source software, is available in windows or UNIX format, and is freely available to download from the Web. The key phrase and pattern matching approach allows users to easily adapt the wrapper for various extraction tasks. EES provides a valuable tool for financial analysts and other users to compare financial statements when companies use different accounting methods and assumptions.

6.2 Limitations

As in any system design and testing there are limitations inherent in this study. The corpus was developed from NASDAQ-100 Index companies. Corpus analysis of other indices, or sets of companies, may produce different algorithms, thus different results. Also, EES was tested on 19 NASDAQ-100 index companies. Had the test been conducted on other NASDAQ-100 companies, or companies from other indices, the results may be different from the results of this study. The test was conducted on 10-K filings over a four-year period, 2001-2004. Testing of earlier years, 1995-2000, may have produced different results.

Tests for EES were conducted using high-speed internet connections on computer systems running Microsoft Windows XP Operating system, Intel 4 Processors, with

either 128 or 512 MB of RAM. EES used on a different system may result in slower or faster processing.

EES is designed for a single user to extract stock option information from 10-K filings on the EDGAR database. The current algorithms used in EES do not search other documents on the EDGAR database. EES extracts stock option information only if it appears in the 10-K. EES algorithms are not designed to point to other files on the EDGAR database if the information is not found in the 10-K.

Also, EES does not determine if the company currently uses the fair value of accounting for stock options. If pro-forma information is extracted, an assumption is made that the company does not expense stock options. If the information is unavailable or not extracted, further manual analysis is needed to make the determination. Another limitation of EES is that it extracts some text that is not pertinent to the search. Also, EES does not extract information from amended 10-Ks. Trend analysis must make use of data in the latest date the 10-K is found.

6.3 Future Research

The efficiency of our capital markets depends on accurate and timely financial information. A system, such as EES, that can sift through the vast amount of text associated with financial statements and extract specific, relevant information can benefit securities analysts, lenders, shareholders, and potential investors. Further research is needed to improve the accuracy and reliability of these types of extraction systems.

Better algorithms are needed to reduce the extraction of unwanted information while improving the overall accuracy and speed of the system. Since a natural trade-off

exists between precision and recall, care must be taken to increase precision while maintaining or increasing recall of the system.

EES can be expanded to search and extract a variety of information from the 10-K. Further, different algorithms can be developed from the existing corpus to extract other financial information from 10-K documents. Pension benefits and assumptions, capital and operating lease information, and details of segment performance are examples of interesting research topics with pertinent data available in the 10-K.

Current user input to EES relies on DOS based commands. Improvements in the system might include a Windows or HTML user interface for more intuitive input commands. For maximum potential use, EES must break beyond a single user approach to encompass multi-users on a server platform.

Corpus development from groups of companies other than the NASDAQ-100 Index can help develop algorithms for specific types of financial information. For example, a corpus of companies from the oil and gas industry can be used to develop algorithms to extract financial information specific to that industry, such as oil and gas reserves and methods used to calculate their value. The procedures used for corpus development in this study can be applied for any corpus development of 10-K documents. Using industry designated SIC codes and SEC CIK codes, any number of companies can be downloaded by industry or index and processed into a database corpus for analysis.

Corpus development can also be expanded to include other forms and documents on EDGAR. Proxy statements, for example, contain a wealth of information about company directors and executive compensation and are often used in research. Form 8-K is required by the SEC for companies that have a change of control, have filed for

bankruptcy, or that have changed auditors [2]. Automated information extracted from these various forms can help encourage research well beyond the normal financial statement analysis.

Corpora can be developed and shared to foster rapid expansion of future extraction systems that focus on financial statement information. A repository of corpora available through the Web or other distribution methods can further stimulate the development of new extraction systems and encourage improvements to existing systems.

In an effort to improve the usability of the financial information on EDGAR, the SEC continuously updates the required format of its electronic filings as technology permits. Beginning in 2005, the SEC allows companies to submit required test documents in an XBRL format [78]. EES must continue to evolve by adding new algorithms that incorporate methods that allow extraction of XBRL data. In addition, the SEC continues to expand the amount of information available by requiring more forms to be filed electronically. EES algorithms must also be updated to address the dynamics of the EDGAR database.

Information Extraction research is still in its infancy. The major challenges facing IE is the accuracy of the systems and the cost of producing more accurate systems. As the amount of digital information increases on the Web, the demands for IE systems will also increase. Demand for IE systems is prevalent in industry, government, and education. As a result, the field has potential in many directions, especially in the area of financial information extraction.

Bibliography

Bibliography

- [1] K. Adams, The Web as a database: new extraction technologies and content management, *Online* 25 (2) (2001) 27-32.
- [2] A. Afterman, *SEC Regulation of Public Companies*, Prentice Hall, New Jersey, 1995.
- [3] C. Aone, W. Bennett, Evaluating automated and manual acquisition of anaphora resolution strategies, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* 1995, 122-129.
- [4] D. Appelt, D. Israel, Introduction to information extraction technology, <http://www.ai.sri.com/~appelt/ie-tutorial/> March 22, 2005.
- [5] H. Ashbaugh, K. Johnstone., Corporate reporting on the internet, *Accounting Horizons* 13 (3) (1999) 241-258.
- [6] A. Bagga, J. Chai, A. Biermann, 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*, 1996, 941-948.
- [7] A. Bianco, To expense or not to expense, *Business Week* (July 29 2002) 44-47.
- [8] V. Borkar, S. Sarawag, Automatic segment of text into structured records, *ACM SIGMOD*, (May 21-24 2001), <http://www.cs.washington.edu/homes/kd/papers/sigmod01.pdf> March 22, 2005.
- [9] J. Bosak, XML, Java, and the future of the Web, (1997), <http://www.ibiblio.org/pub/sun-info/standards/xml/why/xmlapps.htm> March 22, 2005.
- [10] J. Bosak, T. Bray, XML and the Second-Generation Web, (1999), <http://www.sciam.com/article.cfm?articleID=0008C786-91DB-1CD6-B4A8809EC588EEDF> March 22, 2005.
- [11] D. Bourigault, Surface grammatical analysis for the extraction of terminological noun phrases, *Proceedings from the 15th International Conference on Computational Linguistics* 3 (1992) 977-981.
- [12] M. Bovee, M. Ettredge, R. Srivastava, M. Varsarhelyi, Does the year 2000 XBRL taxonomy accommodate current business financial-reporting practice?, *Journal of Information Systems* 16 (2) (2002) 165-182.

- [13] S. Bryan, S. Lilien, CEO stock-based compensation: an empirical analysis of incentive-intensity, relative mix, and economic determinants, *Journal of Business* 73 (4) (2000) 661-693.
- [14] S. Burke, Perl and LWP, Ed. Nathan Torkington. O'Reilly & Associates, Sebastopol, CA, 2002.
- [15] Business Week, How Expensive Will Expensing Be?, (April 26 2004) 116.
- [16] C. Cardie, Empirical methods in information extraction, *AI Magazine*, winter (1997) 65-79.
- [17] CCH, 2002 U. S. Master Tax Guide, CCH, Inc., Chicago, 2001.
- [18] H. Chen, 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) (2003) 621-624.
- [19] H. Chen, Web retrieval and mining, *Decision Support Systems* 35 (1) (2003) 1-5.
- [20] S. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, *Proceedings of the 34th Conference of the Association for Computational Linguistics*. 1996, 310-318.
- [21] G. Chowdhury, Template mining for information extraction from digital documents, *Library Trends* 48 (1) (1999) 182-219.
- [22] P. Clarkson, R. Rosendfeld, Statistical language modeling using the CMU-Cambridge Toolkit, *Proceedings from the ESCA Eurospeech Conference*, 1997.
- [23] Z. Coffin, The top ten effects of XBRL: the future of internet reporting, *Strategic Finance* June (2001) 64-67.
- [24] M. Costantino, R. Collingham, Financial information extraction using pre-defined user-definable templates in the LOLITA system, *Journal of Computing and Information Technology* 4 (4) (1996) 241-255.
- [25] J. Cowie, W. Lehnert, Information extraction, *Communications of the ACM* 39 (1) (1996) 80-91.
- [26] H. Cunningham, *Information extraction - a user guide (second edition)*, (1999), <http://www.dcs.shef.ac.uk/~hamish/IE/userguide/main.html> March 22, 2005.
- [27] DARPA, Defense Advanced Research Projects Agency 2004, <http://www.darpa.mil/> March 22, 2005.

- [28] K. Eisenhardt, Agency theory: an assessment and review, *The Academy of Management Review* 14 (1) (1989) 57-75.
- [29] M. Ettredge, M. Richardson, S. Scholz, Timely financial reporting at corporate web sites?, *Communications of the ACM* 45 (6) (2002) 67-71.
- [30] Financial Accounting Standards Board, Accounting Principles Board Opinion 20, FASB, Stamford, CT, 1971.
- [31] Financial Accounting Standards Board, Qualitative Characteristics of Accounting Information, Statement of Financial Accounting Concept No. 2, FASB, Stamford, CT, 1980.
- [32] Financial Accounting Standards Board, Statement of Financial Accounting Standards No. 123: Accounting for Stock-Based Compensation, FASB, Norwalk, CT, 1995.
- [33] Financial Accounting Standards Board, Business Reporting Research Project: Electronic Distribution of Business Reporting Information 2000, White Paper of the Steering Committee of the Financial Accounting Standards Board, <http://www.fasb.org/brrp/brrp1.shtml> March 22, 2005.
- [34] Financial Accounting Standards Board, Project Updates: Equity-Based compensation (2003), http://www.fasb.org/project/equity-based_comp.shtml October 29 2003.
- [35] Financial Accounting Standards Board, FASB Publishes Proposal on Equity-Based Compensation to Improve Accounting and Provide Greater Transparency for Investors (May 3 2004), <http://www.fasb.org/news/nr033104.shtml> March 22, 2005.
- [36] Financial Accounting Standards Board, Statement of Financial Accounting Standards No. 123 (Revised 2004), Share Based Payment, <http://www.fasb.org/pdf/fas123r.pdf> March 1 2005.
- [37] W. Francis, H. Jucera, Manual of Information to Accompany a Standard corpus of Present-day edited American English, for use with Digital Computers, (Revised Edition 1979), <http://helmer.aksis.uib.no/icame/brown/bcm.html> March 22, 2005.
- [38] D. Freitag, D. Kushmerick, Toward General-Purpose Learning for Information, (2000), <http://acl.ldc.upenn.edu/P/P98/P98-1067.pdf>. March 22, 2005.
- [39] J. Friedl, Mastering Regular Expressions, second ed., O'Reilly & Associates, Cambridge, MA, 2002.

- [40] J. Gerdes, Jr., Edgar-Analyzer: automating the analysis of corporate data contained in the SEC's Edgar Database, *Decision Support Systems* 35 (2003) 7-9.
- [41] C. Goldfinger, D. Valschaerts, Edgar System: Critical Link for US Equity Markets, (2003), http://www.fininter.net/internet_securities/edgar_system.htm March 22, 2005.
- [42] C. Hoffman, C. Strand, XBRL Essentials, American Institute of Certified Public Accountants, New York, 2001.
- [43] House Panel Set to Rein in FASB on Option Rule, *Wall Street Journal*, June 16 2004, C3.
- [44] IR Systems That Are Not Based on the Keyword Vector Model, (2003), http://pi0959.kub.nl/Paai/Onderw/V-I/Content/non_KVM.html March 22, 2005.
- [45] P. Jacobs, L. Rau, The GE NLtoolset: a software foundation of intelligent text processing, *Proceedings of the 13th International Conference on Computational Linguistics*, 1990: 3 373-375.
- [46] P. Jacobs, L. Rau, SCISOR: extracting information from on-line news, *Communications of the ACM*, 3 (11) (1990) 88-97.
- [47] P. Jacobs, L. Rau, Natural language techniques for intelligent information retrieval, *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval*, 1998, ACM Press, pp. 85-99.
- [48] P. Jackson, K. Al-Kofahi, C. Kreilick, B. Grom, Information extraction from case law and retrieval of prior cases by partial parsing and query generation, *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 1998, 60-67.
- [49] M. Jensen, K. Murphy, CEO incentives - it's not how much you pay, but how. *Harvard Business Review*, 86 (3) (1990) 138-148.
- [50] D. Johnson, R. Taira, A. Cardenas, D. Aberle, Extracting information from free text radiology reports, *International Journal of Digital Libraries*, 1 (1997) 297-308.
- [51] G. Klein, What Is an SIC Code and What Do We Do with It?, <http://library.willamette.edu/agsm/sic.htm> March 22, 2005.
- [52] A. Laender, B. Ribeiro-Neto, A brief survey of web data extraction tools, *SIGMOD Record* 31 (2) (2001) 84-93.

- [53] C. Leinmann, F. Schlottmann, D. Seese, T. Stuempert, Automatic extraction and analysis of financial data from the EDGAR database, Proceedings of the 2nd Annual Conference On World-Wide Web Application, 2000.
- [54] G. Leroy, H. Chen, J. Martinez, A shallow parser based on closed-class words to capture relations in biomedical text, Journal of Biomedical Informatics 36 (2003) 145-158.
- [55] A. Levitt, Take on the Street, Pantheon Books, New York, 2002.
- [56] H. Luhn, Keyword-in-context index for technical literature (KWIC Index), American Documentation 11 (1960) 228-295.
- [57] P. Lyman, H. Varian, How Much Information?, (2003), <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/> March 22, 2005.
- [58] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, 5th ed., The MIT Press, Cambridge, MA, 2002.
- [59] J. McCarthy, W. Lehnert, Using decision trees for coreference resolution, Proceedings of the 14th International Conference on Artificial Intelligence. Ed. C. Mellish, 1995, 1050-1055.
- [60] P. McConnell, J. Pegg, C. Senyek, D. Mott, More Companies Voluntarily Adopt Fair Value Expensing of Employee Stock Options, Bear Sterns Equity Research, (September 4, 2003), <http://www.capclaw.com/resource-files/StockOptionExpensing.pdf>. March 22, 2005.
- [61] G. Miller, Wordnet: A lexical database for English, Communication of the ACM, 38 (11) (1995) 39-41.
- [62] M. Moens, C. Uyttendaele, J. Dumortier, Intelligent information extraction from legal texts, Information and Communications Technology Law 9 (1) (2000)17-26.
- [63] T. Mulligan, M. Kristof, T. Y. Jones, Tougher Accounting Rule Expected for Stock Options, Los Angeles Times (March 30 2004) A1+.
- [64] T. Mulligan, Options Proposal Draws Heat, Los Angeles Times (April 1 2004) C1+.
- [65] NASDAQ-100 Index, (2003), http://dynamic.NASDAQ.com/dynamic/NASDAQ100_activity.stm March 22, 2005.

- [66] National Institute of Standards and Technology Association, NIST Overview August 1, 2000, <http://trec.nist.gov/overview.html> March 22, 2005.
- [67] National Institute of Standards and Technology Association, Tipster Text Program, (2001), http://www.itl.nist.gov/iaui/894.02/related_projects/tipster March 22, 2005.
- [68] N. Patwardban, E. Siever, S. Spainbour, Perl in a Nutshell, Ed. Linda Mui, Second ed., O'Reilley & Associates, Sebastopol, CA, 2002.
- [69] S. Petravick, J. Gillett, Distributing earnings reports on the internet, Management Accounting, 80 (4) (1998) 54-56.
- [70] M. Porter, An algorithm for suffix stripping, Program, 14 (3) (1980) 130-137.
- [71] D. Raggett, Getting Started with HTML, (February 13, 2002), <http://www.w3.org/MarkUp/Guide> March 22, 2005.
- [72] L. Rau, P. Jacobs, Creating segmented databases from free text for text retrieval, Proceedings of the 14th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), 1991, 337-346.
- [73] E. Riloff, Automatically constructing a dictionary for information extraction tasks, Proceeding of the 11th National Conference on Artificial Intelligence, AAAI Press/The MIT Press, 1993, 811-816.
- [74] R. Schwartz, T. Phoenix, Learning Perl, 3rd ed., O'Reilly & Associates, Sebasatopol, CA, 2001.
- [75] Securities and Exchange Commission, Securities Act Release, No. 33-7684, (May 17, 1999), <http://www.sec.gov/rules/final/33-7684.txt> March 22, 2005.
- [76] Securities and Exchange Commission, Companies Registered and Reporting with the U.S. Securities and Exchange Commission, (2000), <http://sec.gov/divisions/corpfin/internat/geo2000.htm> March 22, 2005.
- [77] Securities and Exchange Commission, Private communication with Jim Dailey, SEC Internet Support Staff, October 21, 2004..
- [78] Securities and Exchange Commission, Release Nos. 33-8529, 34-51129, 35-27944, 39-2432, IC-26747, XBRL Voluntary Financial Reporting Program on the EDGAR System, <http://sec.gov/rules/final/33-8529.htm> March 22, 2005.
- [79] S. Soderland, Learning to extract text-based information from the World Wide Web, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997.

- [80] D. Stafford, Statistics on CCO Compensation Show Opposite of 'Pay for Performance', (Aug. 26 2002), <http://www.kansascity.com/mld/kansascity/business/3936881.htm> March 22, 2005.
- [81] M. Su, J. Wu, A. Chang, A corpus-based approach to automatic compound extraction, Proceedings of the 32nd Conference on Association for Computing Linguistics, Annual Meeting of the ACL, 1994, 242-247.
- [82] U. S. Department of Labor, Bureau of Labor Statistics, BLS Glossary 2004, <http://www.bls.gov/bls/glossary.htm> March 22, 2005.
- [83] E. Voorhees, Query expansion using lexical-semantic relations, Proceedings of ACMSIGIR, 1994, 61-69.
- [84] What is the History of XBRL?, <http://web.bryant.edu/~xbrl/xbrl/history.htm> March 22, 2005.
- [85] Financial Reporting Taxonomies – Approved, <http://www.xbrl.org/FRTApproved/> March 22, 2005.
- [86] J. Xu, W. Croft, Corpus-based stemming using co-occurrence of word variants, ACM Transactions on Information Systems 16(1) (1998) 61-81.

VITA

Place of Birth:

Gulfport, Mississippi

Degrees:

Master of Professional Accountancy, University of Southern Mississippi 1996
Bachelor of Science, University of Southern Mississippi 1969

Certification:

CPA, Active License

Experience:

2004-2005 Lecturer - California State University Fullerton

2002-Present Doctoral Student – University of Mississippi
Major – Management Information Systems
Minor – Accounting, Management
Expected date of completion – May 2005

2000-2002 Accounting Instructor - Mississippi College, School of Business
Adjunct Instructor – Mississippi College, School of Law

1997-2000 Accounting Instructor – University of Southern Mississippi Gulf Coast

1996-1997 Alexander, Van Loon, Sloan, Levens, & Favre, LLC - Staff Accountant

1994-1996 Graduate Student. University of Southern Mississippi Gulf Coast

Publications:

“Analyzing the Cash Flow Impacts of Employee Stock Options,” co-authored with Conrad S. Ciccotello and C. Terry Grant, *Financial Analysts’ Journal*, 2004. Vol. 60, No. 2, pp. 39-46.

“The Evolution of Corporate Governance and Its Impact on Modern Corporate America,” *Management Decision*, 2003 Vol. 41, No. 9, pp. 923-934.

“An Analysis of GAAP-Based and Operational Earnings Management Techniques,” co-authored with William R. Ortega. *Strategic Finance*, July 2003, Vol. LXXXV, No. 1, pp. 50-56.

“Earnings Management and Auditor Materiality,” co-authored with C. Terry Grant and Chauncey M. DePree, Jr., *Journal of Accountancy*, September, 2000, Vol. 190, No. 3, pp. 41-44.

“Impact of the 150 Hour Requirement on Mississippi Universities,” co-authored with Jerry L. Levens, Mississippi Society of Certified Public Accountants, *CPA Newsletter*, Vol. 41, October 1999, pp. 5-6.

Conference Presentations:

“The Role of Information Extraction Systems in Improving E-Business,” Co-Authoring with Sumali J. Conlon, Susan Lukose, Frank Mathew, and Mahmudul I. Sheikh, Decision Science Institute Annual Conference, Boston, MA, November 20-23, 2004.

“Analyzing the Cash Flow Impacts of Employee Stock Options,” Co-authored with Conrad S. Ciccotello and C. Terry. Grant, Corporate Reporting & Governance Conference, Fullerton, CA, January 2004.

“The Evolution of Corporate Governance and Its Impact on Modern Corporate America” Academy of Management Meeting, August 4-6, 2003, Seattle, WA.

“An Analysis of GAAP-Based and Operational Earnings Management Techniques”, Co-authored with William R. Ortega. Southeastern American Accounting Association Regional Meeting, March 27-29, 2003, Charleston, SC.

Awards:

“The Evolution of Corporate Governance and Its Impact on Modern Corporate America” selected to be included by Emerald Group Publishing in its 2004 issue of *Strategic Direction*, a condensed collection of outstanding articles published by Emerald in 2004.

Institute of Management Accountants National Student Competition Case for the 2003-2004 Academic year. “An Analysis of GAAP-Based and Operational Earnings Management Techniques”, Co-authored with William R. Ortega.

The University of Mississippi Graduate Council Research Grant, Spring 2003.

Classes Taught:

Accounting Information Systems
Cost Accounting
Business Software Applications
Principles of Accounting
Intermediate Accounting
Field Experience in Accounting
Accounting and Finance for Lawyers

Professional Organizations:

American Accounting Association
Information Systems Section - AAA
American Institute of Certified Public Accountants
Institute of Management Accountants
Institute of Management Accounts, Orange County Chapter
Association for Information Systems
Association for Computing Machinery

Conferences Attended:

USC Accounting Forum, Sponsored by John Wiley & Sons, Inc., March 4, 2005
Corporate Reporting & Governance Conference - CSUF – Costa Mesa, CA -
September
2004
USD Accounting Forum, Sponsored by John Wiley & Sons, Inc., Spring 2004
AAA Western Regional Meeting – Newport Beach, CA – May 2004
Academy of Management National Meeting – Seattle, WA – August 2003
AAA Southeastern Regional Meeting – Charleston, SC – March 2003
AAA National Conference – San Antonio, TX – August 2002
AAA Western Regional Meeting – San Diego, CA – May 2002
AAA Western Regional Meeting – San Jose, CA – May 2001
AAA National Conference – Philadelphia, PA – August 2000
AAA National Conference – San Diego, CA – August 1999
AICPA/AAA Accounting Educators' Conference, Washington, D.C. – November
1998
Mississippi Society of CPAs Annual Meeting – Destin, FL – June 1998,
2000 and 2001
Eastern Finance Association Annual Meeting – Williamsburg, VA – April 1998
Eastern Finance Association Annual Meeting – Panama City, FL – April 1997

Service Activities:

Tutor – Intermediate Accounting Cal State Fullerton, El Toro Campus 2004-05
Participant – Professor for a Day, Cal State Fullerton 2004
CBE Assessment Committee, Cal State Fullerton 2004-05

Ad Hoc Reviewer, *Financial Analyst Journal*, 2004
Mississippi Society of CPAs CPE Committee - 2000-2002
Faculty Advisor for Beta Alpha Psi – USM Gulf Coast - 1998-2000
Established University of Southern Mississippi Gulf Coast Student Internship
Program - 1999
Instructor Becker-Conviser CPA Review Course, 2000-2002
Area Coordinator and Instructor, Conviser-Duffy CPA Exam Review Course,
1998-99