2014

A Study Of Data Informatics: Data Analysis And Knowledge Discovery Via A Novel Data Mining Algorithm

Shilpa Balan
University of Mississippi

A STUDY OF DATA INFORMATICS: DATA ANALYSIS AND KNOWLEDGE DISCOVERY VIA A NOVEL DATA MINING ALGORITHM

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

By

SHILPA BALAN

April, 2014

ABSTRACT

Frequent Pattern Mining (FPM) has become extremely popular among data mining researchers because it provides interesting and valuable patterns from large datasets. The decreasing cost of storage devices and the increasing availability of processing power make it possible for researchers to build and analyze gigantic datasets in various scientific and business domains. A filtering process is needed, however, to generate patterns that are relevant. This dissertation contributes to addressing this need. An experimental system named FPMIES (Frequent Pattern Mining Information Extraction System) was built to extract information from electronic documents automatically. Collocation analysis was used to analyze the relationship of words. Template mining was used to build the experimental system which is the foundation of FPMIES. With the rising need for improved environmental performance, a dataset based on green supply chain practices of three companies was used to test FPMIES. The new system was also tested by users resulting in a recall of 83.4%. The new algorithm's combination of semantic relationships with template mining significantly improves the recall of FPMIES. The study's results also show that FPMIES is much more efficient than manually trying to extract information. Finally, the performance of the FPMIES system was compared with the most popular FPM algorithm, Apriori, yielding a significantly improved recall and precision for FPMIES (76.7% and 74.6% respectively) compared to that of Apriori (30% recall and 24.6% precision).

Keywords: Collocation; Information extraction; data mining; FPMIES

DEDICATION

This work is dedicated to my parents:

My mom Sushila and dad Balan

Without their encouragement, support and guidance, I would not have been able to embark on this journey.

This work is a special dedication to my dad. Without his inspiration, the journey of my PhD education would not have been possible.

ACKNOWLEDGMENTS

I am grateful to Dr. Sumali Conlon, who has not only been my advisor but also my mentor. As a graduate student working with Dr. Conlon, the most important virtues I have obtained are self-reliance and patience. I offer my sincere thanks to Dr. Brian Reithel for stressing the need to conduct quality, theory-driven research. Dr. Reithel and Dr. Conlon have given me excellent opportunities to strengthen my research skills and to improve my ability to think and troubleshoot research problems. They have provided me with motivation and guidance throughout the dissertation process.

I would also like to thank Dr. Milam Aiken for giving me opportunities as a graduate assistant and for allowing me to reap the benefits of teaching, which is an invaluable way to learn new subjects and communicate them to students.

I would like to thank Dr. Tony Ammeter for his timely advice and support and as well as his continued encouragement during my coursework.

Finally, I would like to thank my parents who have always guided me and have been there for me during every walk of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Frequent Pattern Mining (FPM) is a core concept in data mining research. As the term suggests, frequent pattern mining mines frequently occurring patterns. FPM was first formalized by Agrawal, Imielinski and Swami (1993) in their paper on mining association rules. Agrawal's original work is based on an analysis of market basket data, where the objective is to determine frequent item sets that customers buy together. In the market basket dataset scenario, each event is a customer transaction which lists the set of items that a customer buys in a supermarket. Since Agrawal's work on market basket datasets, research on FPM became very popular. FPM provides an opportunity to extract interesting and valuable patterns from large datasets. At present, FPM is used in numerous scientific, business and legal application domains where the datasets are large and traditional approaches of data analysis enjoy very limited success.

In the last decade, the field of FPM has expanded immensely. Researchers have proposed various algorithms to mine patterns. FPM generates numerous frequent patterns. With increases in data size, there is an increase in the number of frequent patterns generated by FPM. However, it would take a lot of effort for any individual to analyze a numerous set of frequent patterns. Thus, successful applications of pattern mining require the result set of an FPM algorithm to be a summary. This dissertation is an attempt to overcome the limitation of generating numerous patterns and to generate a smaller set of relevant patterns. Mining frequent patterns provides a

principled path to knowledge discovery in large databases. This chapter summarizes the study, its research contribution, and the hypotheses of the dissertation research model.

The Study

The decreasing cost of storage devices and the increasing availability of processing power make it possible for researchers to build gigantic datasets in various scientific and business domains. Thus, demand for increased use of FPM (Frequent Pattern Mining) is expected to grow even larger in coming years. A major limitation of FPM is that a huge number of irrelevant patterns could be generated. Large data sets are already difficult to analyze, even without the additional burden of irrelevant patterns. This dissertation helps address the need to reduce the number of irrelevant patterns.

Many FPM algorithms have been proposed by previous researchers. The most frequently used is the Apriori algorithm. Apriori discovers all collections of items (called item sets) that occur frequently in a given dataset. It needs to scan all the transactions in a database, which requires a great deal of CPU time. During this process, a vast array of patterns, including irrelevant patterns, are generated. Hence, a filtering process is needed to help identify relevant patterns.

The objective of frequent pattern summarization is to obtain frequent patterns that give meaningful information without losing important content. The discovery of interesting relationships can improve the business decision-making process.

This dissertation presents a solution for generating relevant patterns. The algorithm developed in this dissertation is tested and compared to the Apriori algorithm.

A data sample of three companies is extracted from online sources to test the data mining algorithm presented in this dissertation. The sample data is based on green supply chain information. The importance of green supply chain practices has increased in recent years, generating an expanding number of related electronic documents (Nawrocka, Brorson and Lindhqvist, 2009; Environmental Quality, 1996).

Research Contribution

In this dissertation, an experimental system named "FPMIES" (Frequent Pattern Mining Information Extraction System) has been developed for analyzing content, lexicons, and relationships for extraction of meaningful text. Information extraction (IE) is the process of automatically extracting information from natural language texts. The idea of using a combination of information extraction (IE) techniques and template mining in frequent pattern mining is presented in this dissertation.

Template mining is a particular technique used in IE. In template mining, when text matches a template, the system extracts data according to instructions associated with the template (Chowdhury, 1999). Semantic rules are applied in template mining to help extract information for the FPMIES template.

Collocation analysis identifies sequences of words that co-occur often. Collocation analysis helps identify the important terms in the document that would be of interest to the user. Collocation analysis is also used in the FPMIES algorithm to help extract meaningful information from the corpus data.

Content analysis is a methodology for studying the textual content to determine the presence of certain words, or concepts. The FPMIES algorithm uses content analysis to help

select relevant document statements within a given template. This is similar to the technique used by Balan and Conlon (2012) to extract information from a dataset. Balan et al. used a t-score collocation technique to extract relevant information.

Since the introduction of FPM by Agrawal et al. (1993), many algorithms have been proposed for template mining, such as mining frequent itemsets (Savasere, Omiecinski, and Navathe, 1995; Agrawal and Srikant, 1996; Zaki, Parthasarathy and Ogihara, 1997; Brin, Motwani, Ullman and Tsur, 1997; Zaki and Gouda, 2003). Goethals and Zaki (2003) conducted a survey of the most successful algorithms and techniques used in FPM.

One of the limitations of the standard FPM approach is that when a collection of items appear to be associated, there is a risk of finding many spurious associations. To minimize this risk, this dissertation invents a new FPM algorithm to generate relevant patterns.

The FPMIES algorithm relies upon collocation analysis as a means of pre-processing. The algorithm utilizes semantic rules to generate more relevant patterns. Finally, the FPMIES algorithm uses LCS (Longest Common Subsequence), within the context of template mining, to reduce the number of iterations, thereby yielding more accurate results and increased recall.

The new FPMIES algorithm presented in this dissertation is tested and compared with the widely used FPM algorithm, Apriori. Due to the combination of semantic relationships with template mining, the recall of FPMIES is quite significant.

Hypotheses

The research question studied in this dissertation is: What are the factors that impact the performance of an FPM algorithm? The effectiveness by which a frequent pattern mining

algorithm can generate relevant patterns in a reasonable amount of time determines the performance of the FPM algorithm. The hypotheses presented below are more fully developed in Chapter 3. In order to provide the reader with an appropriate context, however, the test parameters and corresponding hypotheses are introduced below.

Number of database scans. A database stores a collection of patterns. A database pattern is also called an object, event, transaction, or a record. The number of times the database has to be scanned to generate the relevant pattern is an important factor in determining the performance of FPM algorithms to determine its performance. An increase in the number of database scans would reduce the FPM algorithm's performance since the time taken to process the database scans would increase. In this regard, Hypothesis #1 proposes that an increase in the number of database scans will lead to a decrease in algorithm performance.

Algorithm semantic complexity. Semantic relationships connect words, terms, and entities through meaning. Thus, they enable a representation of knowledge with rich semantics. Semantic mining is a critical step in getting useful semantic information for better integration, search, and decision-making. In this regard, Hypothesis #2 proposes that an increase in an algorithm's semantic complexity increases the algorithm's performance.

Scalability. Scalability is the capability of a data structure or algorithm to handle expanding input size without incurring prohibitive complexity or causing inaccuracy or failure. In this regard, hypothesis #3 proposes that an increase in scalability will lead to increase in an algorithm's performance.

Data preparation false negative rate. The data preparation false negative rate is the degree to which the process used to prepare the target data incorrectly excludes a pertinent term

from the data mining target data set. In this regard, Hypothesis #4 proposes that FPM algorithm performance is inversely related to the data preparation false negative rate.

Organization of the Dissertation

Chapter Two provides an exploration of the background literature on data mining and frequent pattern mining. Furthermore, Chapter Two discusses information retrieval and co-occurrence analysis, as well as information extraction and template mining. The research model is presented and discussed in Chapter Three. Chapter Four outlines the research methodology. Chapter Five presents the results of testing the hypotheses. Finally, Chapter Six concludes with a summary of the research results, a discussion of the limitations and assumptions of this study, and presents recommendation for future research.

Definition of Key Terms from Chapter One

Collocation. Collocation has been defined as "the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170).

FPM. This acronym for Frequent Pattern Mining denotes the general task of pattern mining.

FPMIES. This is an acronym for Frequent Pattern Mining Information Extraction System, which is the frequent pattern sampling paradigm that is proposed in this dissertation.

Pattern. The word "pattern" is used within this dissertation as a generic word that represents a set or combination of words occurring together or a set of words and numbers occurring together.

CHAPTER 2

LITERATURE REVIEW

Data Mining and Frequent Pattern Mining Background

Data mining is the science of discovering hidden and useful knowledge from large databases. Data mining draws its roots from machine learning, but exists today at the confluence of machine learning, statistics, and databases. Its focus is the extraction of information. Data mining has emerged as a rapidly growing research field over the last decade (Jagtap and Kodge, 2013). In part, this has been motivated by the growth of the Internet, the increasing availability of computational resources, and the decreasing expense of storage media capable of holding massive datasets.

Frequent Pattern Mining (FPM) is one of the core areas of study in data mining research. FPM was first formalized by Agrawal et al. (1993). Agrawal's original work on market basket data sought to list the set of associated items that customers buy in a supermarket. Frequent pattern mining has been described as "an important data mining paradigm that helps to discover patterns that conceptually represent relations among discrete entities (or items)" (Hasan and Zaki, 2008, p.2). The task of any frequent pattern mining algorithm is to list the frequent patterns without losing important information. The design of any frequent pattern mining algorithm should lead to greater insights into the domain under consideration. At present, FPM is used in numerous application domains where the datasets are generally large and conventional

approaches of data analysis are not very successful (Hasan, 2009). The demand for FPM can be expected to continue to grow as the prices of storage devices decrease.

One of the key challenges of designing data mining algorithms is to find efficient solutions that can scale to meet the demands associated with today's massive datasets. The large set of patterns generated by an FPM algorithm increases the time required for the knowledge discovery task. Finding frequent patterns can play an essential role in mining associations, correlations, and many other interesting relationships among data (Hasan and Zaki, 2008).

The most commonly used FPM algorithm is the Apriori algorithm, proposed by Agrawal and Srikant (1996), for mining frequent item sets. Apriori employs an iterative approach. There are two steps in each of the iterations. The first step generates a set of candidate item sets. Then, in the second step, the frequency of each candidate item set in the database is determined. The limitation of this approach is that as the size of the database increases, a rapidly increasing number of patterns would be generated, making it increasingly computationally expensive to analyze the resulting huge number of patterns.

Ex-Ante: A pre-processing technique

Ex-Ante can be used as a pre-processing technique with any pattern mining algorithm. A pre-processing data reduction technique, Ex-Ante considerably reduces the search space in frequent pattern mining, thereby reducing the data to be analyzed (Bonchi, Giannotti, Mazzanti and Pedreschi, 2003). This pre-processing technique yields a strong reduction of the candidate pattern search space. Ex-Ante is useful for discovering particular patterns.

ExAnte is a preprocessing algorithm that can be used to reduce the time required for the data mining task (Bonchi et al., 2003). Even a small reduction of the number of items to be

analyzed represents a huge pruning of the search space. ExAnte can be combined with any FPM algorithm to start the first iteration of generating candidate item sets by counting the frequency of single items. A limitation of ExAnte is that single items that do not occur frequently are discarded permanently. This runs the risk of eliminating the future combination of those infrequently occurring single items with other more frequently occurring items.

The FPMIES algorithm presented in this dissertation is tested and compared to the Apriori algorithm. FPMIES is primarily built on the concept of an improved pre-processing algorithm and an improved frequent pattern mining algorithm, with the improvements overcoming the aforementioned limitations of Apriori and ExAnte.

The pre-processing technique used in this dissertation does not discard the infrequent one-item sets. Instead, the frequency of the infrequent one-item sets is considered in collocation with other items. Items that collocate frequently with other items are used for further processing during information extraction. The FPMIES algorithm thus eliminates the disadvantage posed by the Ex-Ante technique of permanently discarding the infrequent single items. More details of the new algorithm are discussed in the Research Methodology section of Chapter Three.

Information Retrieval and Co-occurrence Analysis

Information Retrieval (IR) is the process by which a collection of data is searched with the intent of extracting meaningful information as a response to a user request. In an IR environment, a successful search approach is one that is able to provide the most relevant results to the user in a feasible amount of time (Alhenshiri and Blustein, 2012). For this dissertation, collocation analysis is used as a pre-processing technique for information extraction. Collocation has been defined as “the occurrence of two or more words within a short space of each other in a

text" (Sinclair, 1991, p. 170). Other authors have stated that "Collocation analysis is a computational linguistics method for visualizing and analyzing the structure and relationship of words in a large text mass" (Eklund, Toivonen, Vanharanta, and Back, 2011, p.1). Furthermore, according to Vechtomova, Robertson and Jones (2003), Smadja and McKeown (1990), some applications of collocation analysis can be used to improve indexing in information retrieval and natural language generation.

Sinclair (1991), states that words do not occur at random in a text. Instead, words occur together in well-documented collocation pattern. For example, a commonly referenced dictionary of collocations is the Oxford Collocations Dictionary for Students of English (OCDSE) (2002). Collocations consist of common word combinations such as 'green energy'. This combination illustrates an essential building block of English: an adjective-noun pair. The OCDSE contains over 150,000 collocations for nearly 9,000 headwords.

Several approaches have been proposed to retrieve various types of collocations from the analysis of large samples of textual data (Smadja, 1993, Yang and Pedersen, 1997). These techniques strive to produce relevant collocations intended to reflect the associations between words, thus, showing that words cannot occur together at random in a text (Smadja, 1993). According to Bell (2007), collocations are domain specific and can yield interesting information about texts. By using collocations, one can determine how frequently certain words occur together. A particular word under investigation is referred to as the node. Span is the distance, in words, starting from the node word to a related collocation word that occurs either before or after the node word. Previous researchers have used a span of +/- 5 words. For this dissertation, a span of +/- 6 words on either side of the node word is used.

Collocation Techniques

Some of the commonly used collocation techniques are described below.

Mutual information. Mutual information (MI) can be used to perform collocation analysis. Mutual information measures how much one random variable tells about another. The MI score increases with the dependence of the two variables on each other. MI can be thought of as a reduction in uncertainty about one random variable given the knowledge of another. High mutual information indicates a large reduction in uncertainty whereas low mutual information indicates a small reduction. Zero mutual information between two random variables means the variables are independent (Latham and Roudi, 2009).

If *X* and *Y* are independent, then $p(x,y) = p(x)\,p(y)$, and therefore:

$$\log\frac{p(x,y)}{p(x)p(y)} = \log(1)$$

where $p(x,y)$ is the joint probability distribution function of *X* and *Y*, and p(x) and p(y) are the marginal probability distribution functions of *X* and *Y* respectively (Jang, Hyon, and Park, 1999).

Bouma (2005) mentioned that a number of techniques have been applied to mutual information to improve the bias towards rare events. MI is not strictly a measure of significance. It works well in determining the variables' independence, but it is not as accurate in determining the dependence of a variable on another variable. Hence, mutual information is not used to determine collocation for the FPMIES system.

Likelihood ratio. Likelihood ratios are another approach to testing. It was first proposed by Wilks (1938), to test if a given piece of data is a sample from a set of data with a specific distribution described by a hypothesized model. It was later proposed by Dunning (1993) as a

way to determine if a sequence of words came from an independently distributed sample (McInnes et al., 2004, p.3).

In applying the likelihood ratio test to collocation discovery, the following two alternative explanations for the frequency occurrence of a bigram w^1w^2 (Dunning, 1993) are examined.

Hypothesis 1: P ($w^2|w^1$) = p = P ($w^2|$ not w^1)

Hypothesis 2: P ($w^2|w^1$) = p_1 not equal p_2 = P ($w^2|$not w^1)

Hypothesis 1 is a formalization of independence (the occurrence of w^2 is independent of the previous occurrence of w^1). Hypothesis 2 is a formalization of dependence, which is good evidence for an interesting collocation. c1, c2, and c12 are the number of occurrences of w^1, w^2, and w^1 w^2. Hence, the likelihood estimates are estimated using the number of occurrences.

$$p = \frac{c_2}{N}$$

$$p1 = \frac{c_{12}}{c_1}$$

$$p2 = \frac{c_2 - c_{12}}{N - c_1}$$

Where p, p_1, and p_2 are the likelihood estimates.

A likelihood ratio is more interpretable. However, it is simply a number that tells how much one hypothesis is more likely than the other. The main disadvantage of this test is that it is often complex to set up, especially if multiple parameters are being examined (Bell, 2007). Hence, likelihood ratio is not used to determine collocation for the FPMIES system.

Z-score. The term collocation was first introduced by J. R. Firth in his paper "Modes of Meaning" published in 1951. Significant collocation in statistical terms is the probability of an item co-occurring with other items as compared to that might be expected by pure chance. Z-score is a good technique to use when there is a huge number of files. Bell (2007, p. 20) states that "the Z-score measure calculates how far and in what direction an element (a word pair), deviates from the distribution's mean, expressed in units of the distribution's standard deviation." The standard deviation is the square root of the word's frequency. The variance is calculated by subtracting each data element from the distribution mean and then squaring the result.

Consider the statistical data represented with the following:

Z total number of words in the text

A given node occurring in the text Fn times

B a collocate of A occurring in the text Fc times

K number of co-occurrences of B and A

S span size, i.e., the number of items on either side of the node considered as its environment.

First, the probability of B co-occurring K times with A is computed. Next, the difference between the expected number of co-occurrences, and the observed number of co-occurrences must be evaluated. The probability (p) of B occurring at any place where A does not occur is expressed as:

$$p = \frac{F_C}{Z - F_n}$$

The expected (E) number of co-occurrences is given by:

$$E = p.F_n.S$$

The problem statement is to decide whether the difference between observed and expected frequencies is statistically significant. This can be done by means of computation of the z-score as a normal approximation to the binomial distribution (Hoel, 1962) using the formula:

$$z = \frac{K - E}{\sqrt{E_q}}$$

where q = 1- p

This formula has proved highly satisfactory. Hence, Z-score is used to determine collocation for the FPMIES system.

Jaccard index. The Jaccard index has been used to find words of co-occurrence in a large corpus such as the Web. To find the most relevant term for a query, the whole document is analyzed. The collocation between two terms means that those two terms co-occur in the same document, paragraph, or sentence. A greater frequency of the co-occurrence of any two terms implies that the two terms occur closer together in the document (Rijsbergen, 1977). The formula for Jaccard is:

$$\frac{df_{xy}}{df_x + df_y - df_z}$$

where df_x is a query term x of document frequency df_x, and the search term y of df_y. The collocation frequency of two terms df_{xy} is computed using the Jaccard index formula stated above (Kim and Choi, 1998). Ogawa, Morita and Kobayashi (1991) developed a fuzzy retrieval

system where a connection matrix has been initialized by the Jaccard co-occurrence measure. Since the Jaccard index works well with a large number of documents, this technique is applied in this dissertation. Hence, Jaccard index is used to determine collocation for the FPMIES system.

Related Work

Xaira. Xaira is an NLP application from Oxford University. It is a general purpose XML search engine that is used with corpus data. It has the ability to show collocations, but a root node and collocated node must be specified in advance. It also has a word span that determines how many collocates to look at around each root word. On Xaira's homepage, a collocation's significance score is displayed. Xaira produces word lists and frequency distributions (Dunning, 1993). Similar to FPMIES, Xaira uses z-scores to compute collocation. However FPMIES uses the collocation results to further extract information on the data set.

BNCweb. BNCweb is a web-based corpus analysis tool based on SARA, a predecessor of Xaira. BNCweb is a licensed product used by the public. BNCweb also uses z-scores, similar to FPMIES. Both word and part of speech collocations are supported. BNCweb is a web-based client program for searching and retrieving lexical, grammatical, and textual data from the British National Corpus (BNC). It relies on the Corpus Query Processor (CQP) to provide a convenient interface. It is considered to be user friendly and fast. It is optimized for use by large groups of users. The original BNCweb was created by Hans Martin Lehmann, Sebastian Hoffmann and Peter Schneider at the University of Zurich. BNCweb was first publicly released in May 2002 (Bell, 2007). Similar to the FPMIES, BNCweb makes use of the open source

software MySQL, for its database. This helps compute collocation results of large corpus data with efficiency.

Collocation Extract. Collocation Extract imports a corpus comprised of either plain text or annotated text in a standard mark-up language. Collocation Extract is designed to provide a list of collocations in the corpus. Users can search for collocates of a particular word in the range of two to five words. When searching for two-word collocations, users can specify the distance between the two words. The search is based on keywords, with two-word collocations. When searching is finished, the collocation windows will display the list of collocates that co-occur with the keyword, sorted by order of significance (Dunning, 1993). In FPMIES, the search terms or candidate items are generated automatically based on their collocation scores. The information for the candidate items is extracted by the FPMIES system based on the selection of the candidate item by the user. FPMIES uses z-score and Jaccard index techniques to compute the significant collocation scores.

Information Extraction and Template Mining

Information Extraction refers to the automatic extraction of structured information. The extraction of structured information from unstructured sources is a challenging task. Information extraction (IE) is the process of automatically extracting information from natural language texts. Template mining is a particular technique used in IE. When text matches a template, the system extracts data according to instructions associated with that template (Chowdhury, 1999). Areas in which template mining has been successfully applied include the extraction of facts from press releases related to company and financial information in systems, abstracting scientific papers, summarizing new product information, and extraction of data from analytical chemistry papers (Hasan, Chaoji and Salem, 2012).

Although different techniques are used for information extraction, template mining is one of the oldest information extraction techniques as described by Cowie and Lehnert (1996), Gaizauskas and Wilks (1998), Vickery (1997). ATRANS (Lytinen and Gershman, 1986), an IE system, was developed and commercially applied. ATRANS used the script approach (Schank and Abelson, 1977; Schank and Riesbeck, 1981) for automatic processing of money transfer messages between banks. JASPER (Andersen, Hayes, Huettner, Schmandt, Nirenburg and Weinstein, 1992; Andersen and Huettner, 1994) was another IE system developed for fact extraction for Reuters. JASPER used a template-driven approach to extract certain key items of information from a limited range of texts such as company press releases. LOLITA (Costantino, Morgan, and Collingham, 1996) was a financial IE system. Chong and Goh (1997) developed a similar template-based financial information extraction system, called FIES, which extracts key facts from online news articles. Template mining has also been used with citation databases. For example, the electronic journal D-Lib Magazine (http://www.dilib.org/dilib) revealed that template mining can be used to develop citation databases automatically from online articles. The simple template mining approach may be used to extract information for each of the fields in the citation database. Such information can later be used for citation analysis and other purposes.

The FPMIES algorithm relies upon the proven template mining technique described above. It uses this technique to help extract information. The information is extracted in the form of the template structure such as {search term}: {summary of information related to search term extracted}.

Semantic Rules

The idea behind a semantic network is that knowledge is often best understood as a set of concepts that are related to one another (Deliyanni and Kowalski, 1979). Semantic relations are meaningful associations between two or more concepts. Semantic relations can refer to relations between words or text segments (Khoo and Na, 2006). Terms in a domain are often related to each other through a range of semantic relations such as hyponymy. X is an instance of Y if X is a specific example of the general concept Y.

According to Harris (1954), similar words tend to have similar contexts. Recall can be improved by substituting with synonyms (Mandala, Tokunaga and Tanaka, 1999). A synonym rule is applied as shown in the example for semantic rules in the Semantic Rules section in Chapter Four.

Wanner (1996) referred to lexical functions as a mapping or relation between two terms. Further, sets of relations can be labeled with the properties of conjunction to indicate relationships among sets of concepts. For example, an airplane has propellers OR jets. "And" is a conjunction of several parts – wings, nose, tail AND door (Khoo et al., 2006).

FPMIES utilizes semantic rules to improve the performance of the algorithm. More information about the semantic rules for FPMIES is presented in Chapter Four.

Performance of IE Systems

The performance of each IE task normally depends on the text type, the domain of the text, and the specific scenario in which the user is interested. Performance measures for IE systems were developed by MUC (Message Understanding Conferences) and refined with each

conference task (Keith, 2006). Measures of the success of IE systems are calculated using Precision, Recall and F-measure (Powers, 2011).

Precision is calculated by dividing the number of correct answers produced by the number of total answers produced. Precision measures the reliability, or accuracy, of the information extracted.

Recall is the number of correct answers produced divided by the total possible correct answers. Recall is a measure of the amount of relevant information that the system extracts.

A combined weighted measure, the F-measure, is often used (Rijsbergen, 1977). 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:

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

IE systems such as TIPSTER and MUC showed average recall performance of 40%, with precision performance somewhat better at 50%. Improvements continued with most of the IE systems in the later 1990's improving recall to around 50% and precision to 70% on complex tasks. Human performance is usually 79% for recall and 82% for precision. Humans often outperform most IE systems, but they cannot compete with the speed of the computer programs used in automated processes.

Summary

To summarize Chapter Two, Frequent Pattern Mining (FPM), a core area in data mining research is the foundation of the new FPMIES system developed in this dissertation. One of the

key challenges of designing an FPM algorithm is the generation of a large number of irrelevant patterns. This inherent limitation of the typical FPM algorithm is addressed and overcome in the FPMIES algorithm.

Template Mining, which is a technique used to extract information according to instructions associated with a template, is used in FPMIES to extract relevant patterns. By using template mining technique, FPMIES is able to extract the relevant patterns in the form of a summary.

The most commonly used FPM Algorithm is the Apriori algorithm. Results of FPMIES are compared to the Apriori algorithm in Chapter Five.

ExAnte is a popular pre-processing technique. However, ExAnte permanently discards infrequently occurring single items. To eliminate this disadvantage of Ex-Ante, collocation is used as a means of pre-processing for FPMIES.

CHAPTER 3

RESEARCH MODEL AND DEVELOPMENT OF HYPOTHESES

This chapter develops the hypotheses in greater detail. The research model is presented in Figure 1. As discussed in Chapter One, the research model is based on four concepts: database scans, semantic complexity, scalability, and data preparation.

A database scan is a parameter discussed in most research (Liu, Schmidt, Voss, Schroeder and Müller-Wittig, 2006). As the number of database scans increases, the performance of the algorithm is adversely affected. The aim of the FPMIES algorithm is to generate relevant patterns. A computational concept, termed LCS (Longest Common Subsequence) is applied within FPMIES to reduce the number of database scans. More detail about LCS is presented later in this Chapter. The details of the LCS implementation in FPMIES are described in Chapter Four. Hypothesis 1 is thus based on the number of database scans used to produce relevant patterns.

Hypothesis 2 focuses on the semantic rules that help generate relevant patterns. WordNet, from the Princeton.edu website, is used to identify synonyms. The synonyms help build the semantic rules used in FPMIES to produce relevant patterns.

Hypothesis 3 deals with scalability. The data mining literature states that an increase in the volume of data input can result in a decline in algorithm efficiency. A content organization

filtering process is applied in FPMIES to address challenges resulting from an increase in data input.

Hypothesis 4 is based on pre-processing. Pre-processing in data mining has proven to decrease time. Ex-Ante is a popular pre-processing algorithm, which eliminates a term based on its frequency. However, the occurrence of a term could be more frequent with its association with another term, which is not taken into consideration by the standard Ex-Ante technique. This limitation is removed by the pre-processing technique used in the FPMIES algorithm.

Research Model

The research question studied in this dissertation is: *What are the factors that impact the performance of the FPM algorithm?* The research model employed to address this question is portrayed in Figure 1.

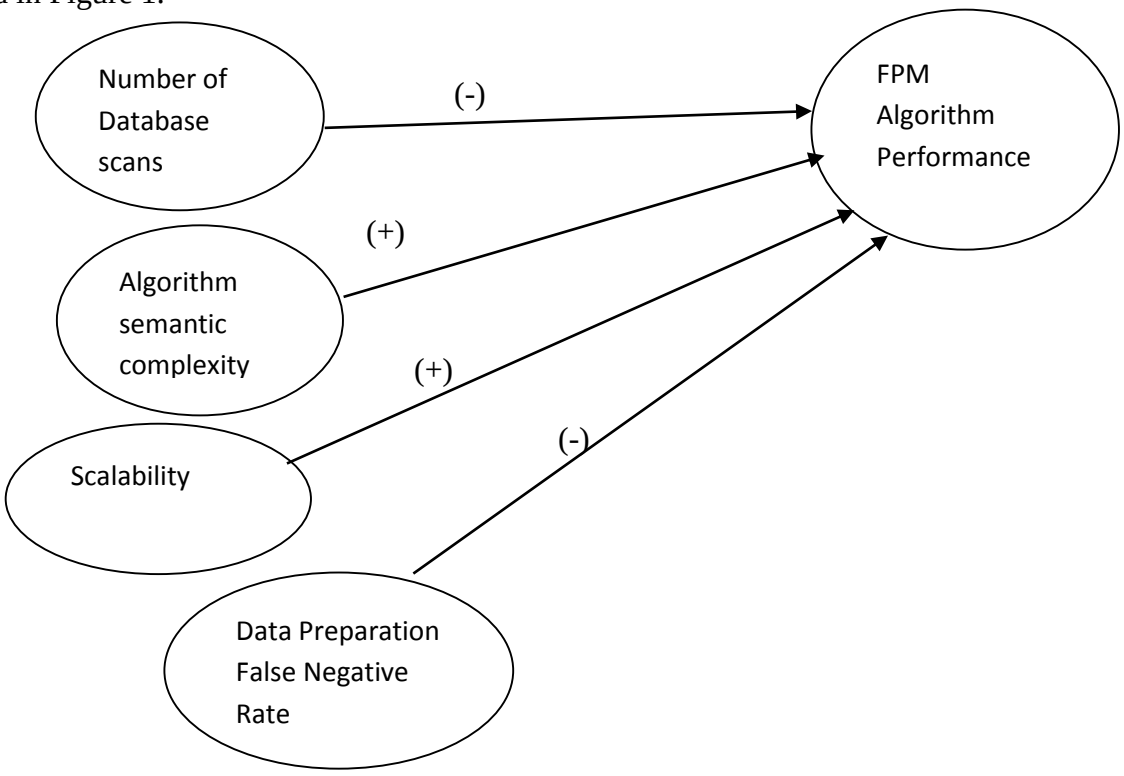


Figure 1: The FPM Algorithm Performance Model

Table 1 presents the definitions of the variables in the research model.

Construct	Definition
Number of Database scans	The number of times the database has to be scanned to generate relevant patterns.
Algorithm semantic complexity	The generation of semantic rules to discover relevant patterns.
Scalability	The capability to handle expanding input size without incurring excessive complexity or causing inaccuracy or failure.
Data Preparation False Negative Rate	The degree to which the process used to prepare the target data incorrectly excludes a pertinent term from the data mining target data set.

Table 1: Research model constructs and definitions

The following subsections discuss the development of the hypothesis.

Number of Database Scans. The number of database scans is defined as the number of times the database has to be scanned to generate relevant patterns (Liu, Schmidt, Voss, Schroeder and Müller-Wittig, 2006). Data are classified based on their values stored in attributes. Typical queries include identifying companies with similar growth patterns, products with

similar selling patterns, stocks with similar price movement, geological features, environmental pollutions, or astrophysical patterns.

A large number of database scans is generally required for a large database to produce information from the database. In the FPMIES system, data is categorized according to the collocation results with the longest common subsequence used to reduce the number of database scans.

The degree of similarity between two sequences can be measured by extracting the maximal number of identical symbols existing in both sequences, known as the longest common subsequence (Taniar and Iwan, 2011). An algorithm's processing time increases with an increase in number of database scans, negatively affecting its performance.

Hypothesis 1: An increase in the number of database scans leads to a decrease in algorithm performance.

Algorithm Semantic Complexity. Semantic relationships connect words through meaning, enabling a semantically-rich representation of knowledge (Chen and Prasanna, 2012). Research in frequent pattern mining has focused on developing efficient algorithms to discover various kinds of frequent patterns. However, little attention has been paid to the important step of interpreting the frequent patterns generated. Generally, the hidden meaning of a pattern can be inferred from words with similar meanings and the words co-occurring with it. In principle, such an annotation can be expected to be well structured and indicative of the pattern's meaning.

Despite its importance, semantic annotation of frequent patterns has not been well addressed in the literature. A novel solution for generating semantic annotations for frequent patterns is presented in this dissertation. A dictionary-like description for a pattern is generated

by FPMIES for finding synonym patterns. The semantic rules for the FPMIES algorithm are described in Chapter Four.

Empirical results on the dataset in Chapter Five show that the FPMIES algorithm is effective for generating semantic pattern annotations. Semantic mining is, therefore, a critical step toward generating useful semantic information.

Hypothesis 2: An increase in an algorithm's semantic complexity will increase the algorithm's performance.

Scalability. Scalability is the capability of an algorithm to handle expanding input size without incurring excessive complexity or causing inaccuracy or failure (Laguna, Gamblin, Supinski, Bagchi, Bronevetsky, Anh, Schulz, Rountree, 2011). As the amount of information available for analysis increases, scalability of data mining applications becomes a critical factor. The success of computerized data management has resulted in the accumulation of vast amounts of data in several organizations (Chen, 1998). Data can consist of text, numeric values, or special characters. Since various types of data can be used in different applications, a knowledge discovery system should be able to perform effective data mining on different kinds of data. A filtering process is required to handle the data input.

In FPMIES, the filtering process relies upon a combination of techniques, including content-based data categorization and applied collocation analysis. The details of the FPMIES algorithm are described in Chapter Four. The processing time of a data mining algorithm must be acceptable.

Hypothesis 3: An increase in scalability will lead to increase in an algorithm's performance.

Data Preparation False Negative Rate. Data Preparation False Negative Rate is defined as the degree to which the process used to prepare the target data incorrectly excludes a pertinent term from the data mining target data set. Data pre-processing is an important step in the data mining process. Data gathering methods are often loosely controlled, resulting in unfeasible data combinations, and missing values (Jade, Verma L and Verma K, 2013). Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data comes first before running an analysis.

Data preparation and filtering can take a considerable amount of processing time. Data pre-processing includes transformation, feature extraction and selection, and volume reduction. The product of data pre-processing is the final training set. In the FPMIES algorithm, the data volume is reduced by performing pre-processing using a procedure described in Chapter Four.

Ex-Ante is a popular pre-processing technique for FPM. Based on the frequency of the word occurrence, non-frequent single items are eliminated in the Ex-Ante technique (Bonchi et al., 2003). This is a disadvantage of Ex-Ante, as the occurrence of the term may be more frequent through its association with another term than its occurrence alone. This disadvantage has been eliminated in FPMIES.

Hypothesis 4: FPM algorithm performance is inversely related to the data preparation false negative rate.

Summary

All research models are "attempts by man to model some aspect of the empirical world" (Dubin, 1976). Research models are simplifications of reality that are limited by restrictions such as methodology and prior research. However, research models are an attempt to explain the relationships between various variables of a system.

This dissertation is the first attempt to build a conceptual model that specifically identifies the factors that impact the overall performance of a frequent pattern mining algorithm. While some of the independent variables discussed in the research model have been examined individually in previous research, no other study has combined the factors included in this dissertation's conceptual model of FPM performance. Furthermore, FPMIES is a new FPM algorithm invented for this dissertation. The FPMIES system aims to set a new benchmark for FPM algorithm performance.

The research model presented in this dissertation is applicable in many different content domains. The data set used to test the FPMIES algorithm in Chapter Five includes documents related to green supply chain management. However, FPMIES can be used with other important content domains such as healthcare and finance.

Based on the literature review, four independent variables are considered to influence FPM algorithm performance. A pilot test was performed prior to the final testing. Re-testing confirmed the research results, providing some support for the four hypotheses. The details of the FPMIES algorithm are discussed in Chapter Four. The testing results are presented in Chapter Five.

CHAPTER 4

FPMIES ALGORITHM DEVELOPMENT AND EVALUATION

In this chapter, the detailed internal architecture of the FPMIES algorithm is explored. Its functioning is illustrated through an analysis of a corpus that contains documents related to green supply chain management practices of three companies.

The pseudocode in this chapter introduces the key architectural elements of the new algorithm developed in this dissertation. FPMIES was written using PHP and SQL. The FPMIES system includes three main modules: (1) Pre-Processor, (2) Search Term Presenter, and (3) Information Extraction Engine. The pre-processor loads a collection of documents into the database and generates pre-defined search terms. Relevant information can then be extracted for each search term by FPMIES. The following sections present the modules of FPMIES in detail.

Pre-Processor

An FPM algorithm can generate several patterns from a collection of documents. However, the FPMIES algorithm has the distinctive ability to generate its output in a summary pattern. This makes it easier for the user to obtain essential information from a set of documents.

An FPM algorithm could also generate irrelevant patterns, making the analysis process more difficult. The summary pattern generated by FPMIES leads to enhanced knowledge discovery, allowing the user of the FPMIES system to make better decisions based on the extracted information.

In order to analyze data related to a domain, the system administrator first uploads documents into FPMIES. Once the documents are loaded into the system, pre-processing is performed by means of collocation analysis to generate relevant search terms. The search terms are thus pre-defined by the FPMIES system.

Pre-processing begins with the KWIC index builder. The documents are loaded into the KWIC index file. Stemming is applied while computing collocations from the KWIC index file to reduce the possibility of eliminating potentially-important related terms. For example, for a term ‘green’, another related term such as ‘greening’ is also captured.

KWIC Index Builder. A KWIC index is a file which is created by putting each word of a sentence into separate columns in a row of a database table (Luhn, 1960). Once a complete sentence has been entered into the first row, with one word per column, the second row starts with the second word of the same sentence in the first column and the remaining words of the sentence in subsequent columns (Simmons, Conlon, Mukhopadhyay and Yang, 2011; Grant and Conlon, 2006). This process of generating rows from that same sentence continues until the last word of the sentence is in a row by itself. Therefore, a sentence that contains seven words would result in seven rows in the table. Appendix C contains a KWIC index example.

Stemming Processor. The use of stemming is a distinctive feature of FPMIES that significantly improves the algorithm’s performance compared to other FPM systems. Stemming refers to the process of removing affixes (prefixes and suffixes) from words. In the information retrieval context, stemming can be used to improve recall. For example, assume the corpus included a document entitled “How to Write More Effectively”. If the user issued the query “writing” instead of “write”, no match would be found. However, if the query was stemmed, such that “writing” became “write,” the search would be successful.

The most widely-cited stemming algorithm was introduced by Porter (1980). The Porter stemmer applies a set of rules to iteratively remove suffixes from a word until none of the rules apply. For instance, "general" becomes "gener" and "iteration" becomes "iter." The Porter stemmer ignores prefixes completely, so "reliability" and "unreliability" remain as unrelated tokens (Krovetz, 1993; Xu and Croft, 1998; Arampatzis, van der Weide, Bommel and Koster, 2000).

In FPMIES, the stemming process is applied to the KWIC index file to reduce the possibility of eliminating potentially-important terms. Using key words and phrases, the roots of the words were combined with wildcard characters for SQL analysis. In the SQL query, the wildcard character "%" is combined with the root word. For example, 'hazard%' would return three results – "hazard," "hazardous," and a third compound word "hazard chemical."

Collocation. In FPMIES, the Jaccard index and z-score measures are used to compute collocations. As discussed in the literature review in Chapter Two, span is the distance, in words, starting from the node word to a related collocation word that occurs either before or after the node word. For example, when a span of five is applied to the string "overall material use on reducing **toxics** in their products design products", the word "toxics' is the node. In Figure 2, the total distance between "overall" and "toxics" is called the span and it ranges up to five words.

-5	-4	-3	-2	-1	Node	+1	+2	+3	+4	+5
Overall	material	use	on	reducing	**toxics**	in	their	products	design	Products

Figure 2: Collocation Span Example

In FPMIES, a z-score is used to evaluate the significance of the collocated terms detected by the Jaccard index. The z-score determines if the collocated terms are actually significant.

Generate Search terms. FPMIES identifies search terms based on collocation scores (z-score and the Jaccard index). All terms with negative collocation scores were eliminated, because a negative value indicates the terms were not highly collocated with each other. This process yields a set of more highly collocated search terms. The search terms are also referred to as 'candidate items' elsewhere in this dissertation.

Pre-processor Summary. In FPMIES, pre-processing begins with the KWIC index builder. The documents are loaded into the KWIC index file. Stemming is applied to reduce the possibility of eliminating potentially-important related terms. Collocation score is computed to generate pre-defined search terms. Figure 3 shows the pseudocode for the Pre-processor algorithm.

```
For (count (remaining files) <> 0)
{
        Upload files into database
}
Run KWIC index on the files
Apply stemming on KWIC index
Set collocation span=6
Extract all collocated terms from KWIC index file
For (all collocated terms)
{
        Compute collocation score
        While (collocation score > 0)
        {
                Check if collocation score is significant (based on z-score)
                if(collocation score is significant)
                {
                        Extract collocated term
                        Store collocated term as candidate search term in database
                }
        }
}
```

Figure 3: Pre-processor Pseudocode

Search Term Presenter

In FPMIES, the search terms are generated by the pre-processor and stored in the database. To retrieve information from the FPMIES system, the user simply selects a pre-defined search term from the user-interface. Based on the search term selected by the user, FPMIES displays extracted information corresponding to the search term. The pseudocode for the retrieval of the search terms is presented in Figure 4.

```
while(true)
{
        Extract search terms from database
        Load the search term in the drop-down menu in the front end of the user-interface
}
```

Figure 4: Pseudocode -Search Term Presenter

Information Extraction Engine

After the user selects a search term, all data corresponding to that search term is extracted. The FPMIES system extracts information in the form of a pattern-based summary for the search term. During the retrieval process, WordNet is used as a basis for semantic processing. In the final step of the information extraction process, LCS (Longest Common Subsequence) is applied to generate relevant patterns in the form of a summary. As discussed in Chapter Two, LCS is the degree of similarity between two sequences that can be measured by extracting the maximal number of identical terms existing in both sequences.

Semantic Rules. FPMIES helps improve recall because template mining is combined with WordNet-based semantic relationships. WordNet is considered to be the most important resource available to researchers in computational linguistics and text analysis. It is a lexical

database for the English language. It groups words into sets of synonyms called “synsets” and provides definitions.

The steps of semantic rule application within FPMIES are:

1. Extract nouns and numerals after the candidate item (search term) for each sentence.
2. Extract conjunctions after the candidate items for each sentence. Conjunction List used in FPMIES = {and, but, or, for, nor, yet}
3. Check synonym of each of the item in the search term
4. Repeat Steps 1 and 2 after substitution of synonym if synonym is found

Relevant Information Extraction. In FPMIES, relevant nouns are extracted in the process of mining. This is because nouns are role-identifying. Philips et al. (2007) describe nouns as role identifying expressions that are beneficial for information extraction. Nouns identify the role they play in an event. Charles (2012) states that information extraction systems rely on noun phrases as their primary source of entity identification. For example, in a piece of writing about airplane safety, if the nouns in the text were to include “plane,” and “seatback,” these are informative about the subject matter and context of the text, and, therefore, are quite valuable to the natural language processing task.

Nouns can appear after determiners and adjectives and can be the subject or object of the verb as shown in Table 2. Nouns generally refer to people, places, things, or concepts, e.g. person, Scotland, and book. A noun phrase may be optionally accompanied by a determiner such as numerals (e.g., “one,” “two”). The FPMIES system extracts such determiners as well. Literal

numbers are also extracted. For example: 10, 10%, are numbers that most often represent statistical information.

Word	After a determiner	Subject of the verb
Person	*the* person who I saw yesterday	the person *sat* down
Scotland	*the* Scotland I remember as a child	Scotland *has* five million people
Book	*the* book I bought yesterday	this book *recounts* the colonization of France

Table 2: Noun Extraction Example

Template Mining. Template mining has not been used in conjunction with collocation and semantic rules by data mining researchers in the past. The combination of the powerful technique is one of the contributions of the FPMIES algorithm. In FPMIES, template mining is used in combination with frequent pattern mining, collocation and semantic rules to extract relevant information. This combination helps improve recall and precision.

As discussed in Chapter Two, various algorithms for pattern mining have been developed in the past. Pattern analysis is the task of finding useful patterns from data. Frequent pattern mining is used to find patterns from large sets of data. The aim of this technique is to find patterns that would help the user of the IE system make better decisions on the data. Han, Cheng, Xin and Yan (2007), state that the discovery of frequent patterns plays an essential role in mining interesting relationships among data.

The FPM mining process searches incrementally in the pattern space to enumerate patterns of size 1, 2, 3, and so on, where size represents the number of terms in the pattern. See below for examples of size one, two, and three candidate item sets.

1-candidate

Item set	Frequency
ISO	25
14001	16
Certification	14

2-candidate

Item set	Frequency
ISO 14001	13

3-candidate

Item set	Frequency
ISO 14001 certification	12

The FPMIES algorithm is an improved version of FPM that generates relevant patterns. The algorithm scans the database and counts the frequency of the number of items extracted for each search term. The pattern or information extracted for each scan is stored in the database. Longest common subsequence (LCS) is then applied to the stored patterns in the database. Since the patterns are stored in the database and the longest common subsequence is applied to the stored patterns, the number of scans is significantly reduced, thereby substantially improving processing speed.

Information Extraction Engine Summary. In FPMIES, relevant nouns are extracted in the process of mining. Template mining is used in combination with frequent pattern mining, collocation analysis and semantic rules to extract relevant information. Longest common subsequence (LCS) is then applied to extract relevant patterns. Figure 5 shows the pseudocode for the Information Extraction Engine.

```
While (length of file>0)
{
        Locate position of terms in search term in the file
        For each term in search term
        {
                While (! End of Sentence)
                {
                        Extract nouns and numerals following each term of the search term
                        Apply Semantic Rules
                        Apply conjunction rule

                        Store pattern in database
        }
        }
}
Retrieve stored patterns from database
Apply LCS
Generate results in a template form (with the search term as the label on the left side and
the information extracted on the right side)
```

Figure 5: Pseudocode –Information Extraction Engine

FPMIES Algorithm. By combining the functionality from the three main modules of the FPMIES system that have already been presented, it can be seen that the system relies upon the *Pre-processor* to populate a database that can later be used for information retrieval and extraction. During retrieval and extraction, the user will select a previously-generated search term via the *Search Term Presenter* that the *Information Extraction Engine* will then use as a basis for operation. See Figure 6 for an illustration of the FPMIES architecture.

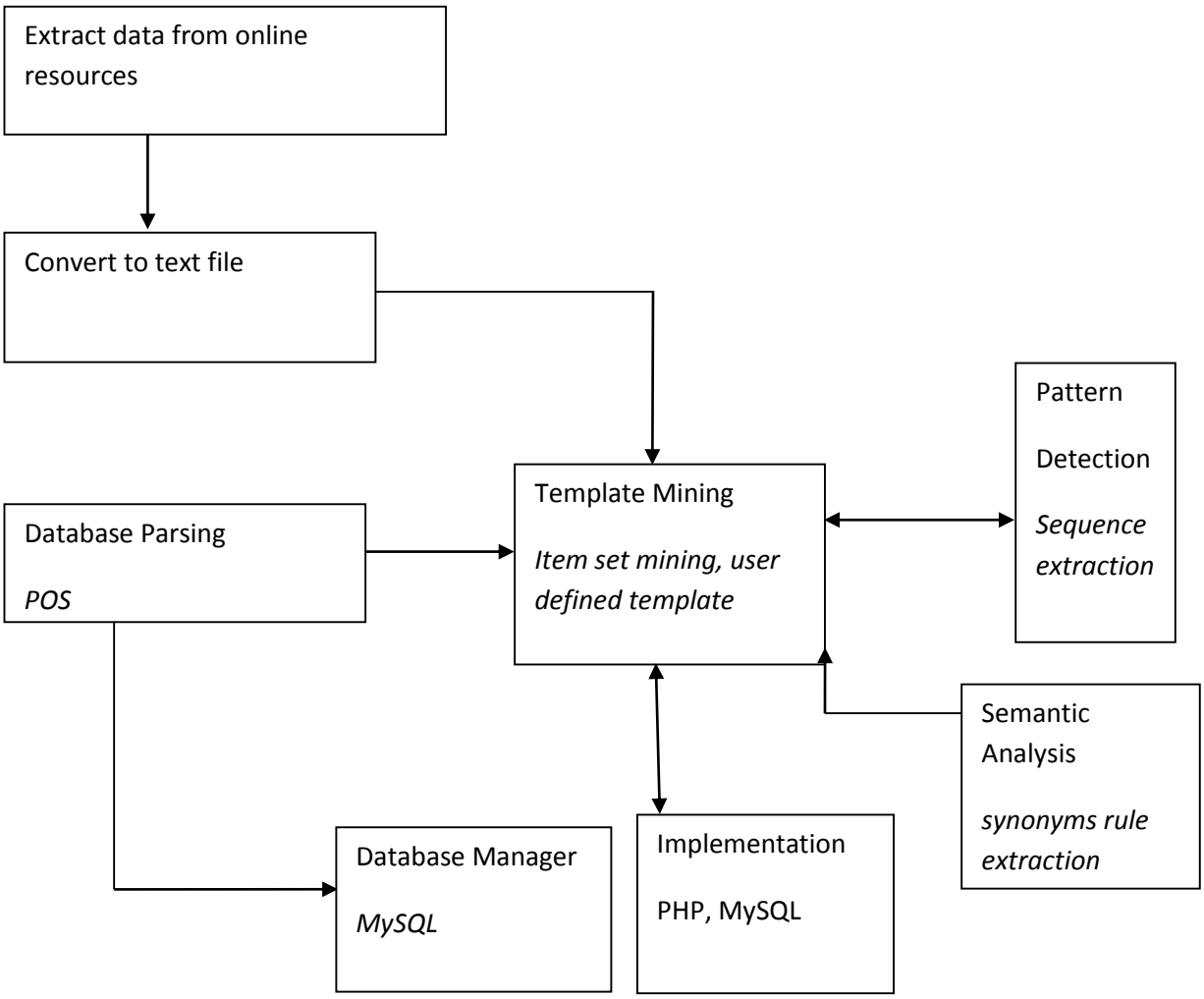


Figure 6: FPMIES Architecture

An Example of FPMIES Implementation

This section describes how FPMIES was implemented for a particular domain. Given the increasing importance of environmental awareness, the area of "green supply chain" was selected to test FPMIES.

Documents were collected regarding three major companies – IBM, HP and Dell – and their green supply chain practices. These three companies are respected for their efforts related to green supply chain management in the electronics industry. FPMIES will be used to analyze these documents and provide detailed information on the steps taken by the three companies to implement green supply chain management practices.

Data

The sources of data are company web sites (IBM, HP, Dell), news reports and related articles from the web. The news reports primarily included CNET News and Computer Weekly. The web sources included greensupplychain.com/news and greenbiz.com. The source files that were collected contained data regarding: the list of chemicals restricted by the companies, various green programs implemented by the companies, efforts by the companies to reduce greenhouse gas/carbon emissions, programs to improve waste management, initiatives related to green packaging, and new or improved recycling strategies. Forty-five documents were collected for each company.

Pre-Processing

A collection of data files on green supply chain management was uploaded into the database. Pre-processing was applied to extract the important search terms from the documents.

After generating the KWIC index file, the collocation scores were calculated. Table 2 shows the Jaccard index for the most highly collocated terms in the example data set. A collocation span of six was used. The Jaccard index values ranged between 0 and 1. Previous research has shown that the terms tend to collocate more as the value of the Jaccard index approaches 1. For instance, in Table 3, ISO and 14001 are the most frequently collocated terms in the IBM document. Any negative collocation score values were eliminated because a negative value represents terms that are not highly collocated with each other.

Word	Collocated Term	IBM	HP	Dell
Carbon	Footprint	1	0.42	0.45
ISO	14001	0.91	0.72	1
EICC	Practice	0.84	0.73	0.68
Greenhouse	Gas	0.52	0.125	-
Green	IT	0.361	-	-
Green	Innovations	0.142857	-	-
Green	Technology	0.689	-	-
Green	Supply chain	0.145	-	-
Restricted	Chemicals	0.62	0.71	-
Electronic	Waste	-	0.13	-
Packaging	Waste	-	0.51	-
Green	Energy	-	-	0.44
Recyclable	Packaging	-	-	0.25
Green	Packaging	-	-	0.55
Hazardous	Substance	-	-	0.77

Table 3: Jaccard index for highly collocated terms in FPMIES

The significant collocations generated by z-score (at 5% significance level) are listed in Table 4. For example, Carbon-Footprint and EICC-Practice are the most significantly occurring collocated terms in the table.

Word	Collocated Term	z-score
Carbon	Footprint	10.902
ISO	14001	10.776
EICC	Practice	11.956
Greenhouse	Gas	7.356
Green	IT	7.783
Green	Innovations	3.573
Green	Technology	9.422
Green	Supply chain	5.134
Restricted	Chemicals	8.035

Table 4: Z-score for IBM (p=0.05, z=1.64)

Figure 7 presents a chart of the significant z-scores from the IBM document collection. It is clear that Carbon-Footprint and EICC-Practice are far most significant than Green Innovations and Green Supply Chain in the IBM document collection.

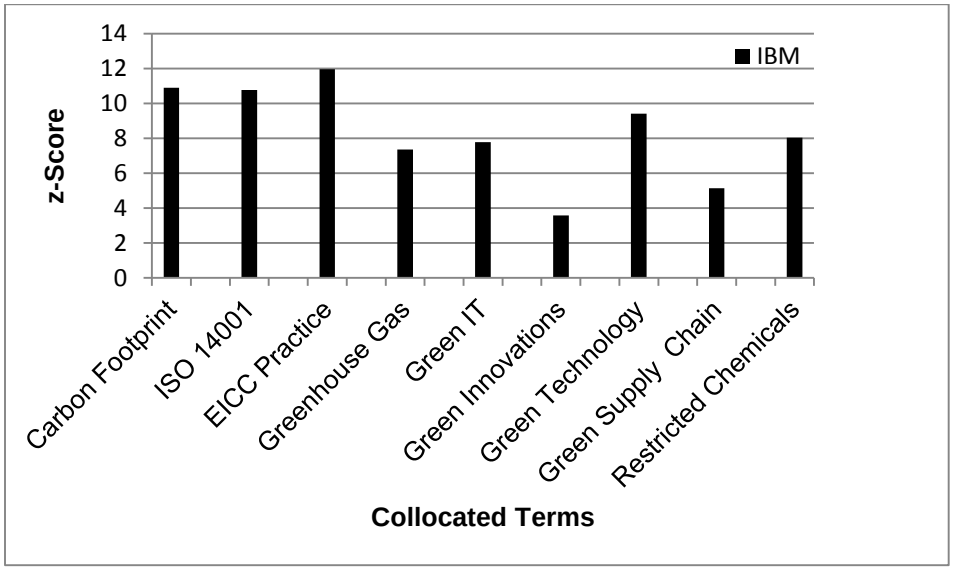


Figure 7: Z-score chart for IBM

Table 5 lists z-scores for the HP document collection. In Table 5, ISO-14001 and EICC-Practice are the most collocated terms. All scores listed in Table 5 are significant at the 5% significance level.

Word	Collocated Term	z-score
Electronic	Waste	3.147
ISO	14001	11.437
Greenhouse	Gas	3.246
EICC	Practice	9.754
Packaging	Waste	7.817
Restricted	Chemicals	8.923
Carbon	Footprint	5.368

Table 5: Z-scores for HP (p=0.05, z=1.64)

Figure 8 presents a chart of the significant z-scores from the HP document collection. The chart clearly indicates that the term pairs Electronic-Waste and Greenhouse-Gas occur less frequently than the other significant term pairs.

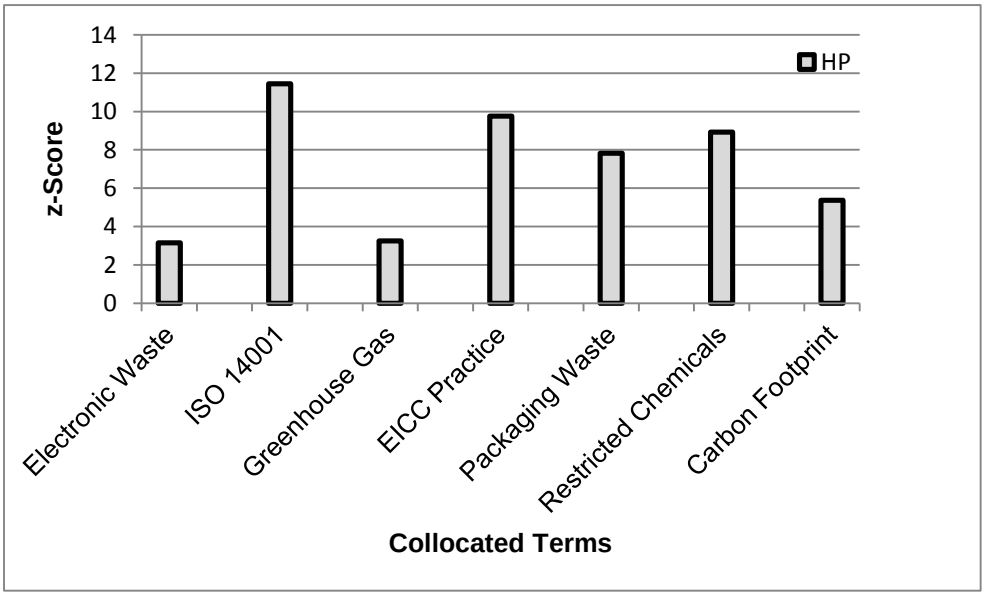


Figure 8: Z-score chart for HP

Table 6 lists the significant z-scores from the Dell document collection. The pattern of results is noteworthy in its similarity to the results for HP.

Word	Collocated Term	z-score
Green	Energy	7.225
Recyclable	Packaging	5.984
Green	Packaging	8.333
EICC	Practice	9.611
ISO	14001	11.854
Hazardous	Substance	7.824
Carbon	Footprint	5.668

Table 6: Z-scores for Dell (p=0.05, z=1.64)

Figure 9 presents the chart of the z-scores from the Dell document collection. The chart shows that the term pair ISO-14001 is the most collocated term from this document collection.

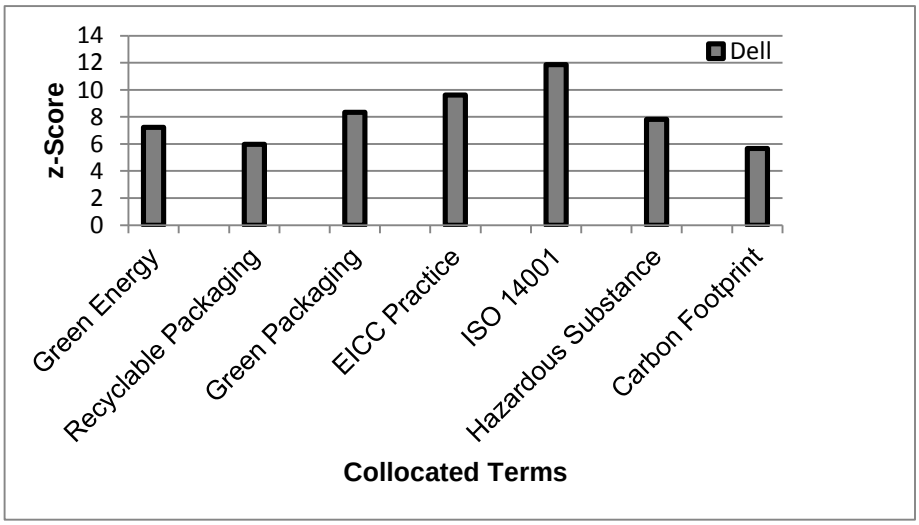


Figure 9: Z-score chart for Dell

Search Term Presenter

In FPMIES, the significant collocated phrases generated during pre-processing can be used as search terms. FPMIES presents the user with a list of these pre-defined search terms and allows the user to select a search term for information extraction. Table 7 lists the search terms generated for all three companies. 'Y' in Table 7 stands for 'yes' indicating that the collocated score is significant for that organization.

Collocated Term	IBM	HP	Dell
Carbon Footprint	Y	Y	Y
ISO 14001	Y	Y	Y
EICC Practice	Y	Y	Y
Greenhouse Gas	Y	Y	-
Green IT	Y	-	-
Green Innovations	Y	-	-
Green Technology	Y	-	-
Green Supply chain	Y	-	-
Restricted Chemicals	Y	Y	-
Electronic Waste	-	Y	-
Packaging Waste	-	Y	-
Green Energy	-	-	Y
Recyclable Packaging	-	-	Y
Green Packaging	-	-	Y
Substance	-	-	Y

Table 7: Search Terms generated by FPMIES

Information Extraction Example

The information extraction process begins with the user selecting one of the generated search phrases. Semantic rules are applied and then template mining & LCS is subsequently applied to extract relevant information.

Semantic Rules. The synonyms of the relevant terms related to green supply chain were extracted from WordNet. Appendix A lists the synonyms provided by WordNet for the example data set.

Figure 10 presents an example of the application of a semantic rule to the search term (candidate item) 'Hazardous Substances'.

> Global concerns over the human health and environmental risks associated with the use of certain environmentally-sensitive materials in electronic products has led the Directive on the restriction of the use of certain **Hazardous Substances**, designed to restrict the use of **cadmium, chromium, lead,** and **mercury** in electronic products. Dell understands the environmental risks associated with the **substances** covered by the RoHS Directive. Dell has also established public goals to phase-out the use of lead and other non-regulated brominated flame retardants in our products in advance of legal requirements. Delivering RoHS compliant products is a significant challenge for the electronics industry and involves a complex set of technical attributes.

Figure 10: Data Sample from Dell file

The steps followed to apply semantic rules to the search term in Figure 10 are:

1. "Hazardous Substances" is set as the search term.
2. Synonyms of the word *hazardous* are located.

   Synonyms retrieved from WordNet = {*peril, risk, endangerment, stake}.*
3. Synonyms of the word *substances* are then located.

   Synonyms retrieved from WordNet = *{core, essence, consistency}*
4. Since no other item is found after synonym substitution, the synonym rule in this example extracts "cadmium, chromium, lead, mercury". These extraction results indicate

the major chemicals that were mentioned by Dell in the company's documents related to green supply chain management that were included in this example.

Template Mining and LCS. A data sample from the IBM document collection is shown in Figure 11-A. In FPMIES, the candidate item "restricted chemicals" will result in "polyvinylchloride" and "retardants products" being extracted. Figures 12-A and 12-B show the FPMIES algorithm implementation for the candidate item "restricted chemicals" in the IBM document collection.

A

IBM **restricted polyvinylchloride**. This specification applies to materials, parts, assemblies and products used in the fabrication of IBM logo or non logo products when this specification is referenced. This specification does not apply to product packaging or shipping materials, which are covered by other IBM specifications. **Restricted chemical contents** were provided in their **list**. For IBM part numbers controlled by IBM Microelectronics Division, the controlling environmental specifications IBM engineering specification applicable requirement stated in this document. The industry standards cited are IBM anti-smoke, flammability, qualification requirements. **Restricted standards** for halogenated flame **retardants** set by EPA were further implemented by IBM. IBM thus **restricted chemicals polyvinylchloride and retardants.**

Figure 11 A: Data Sample 1 from IBM file

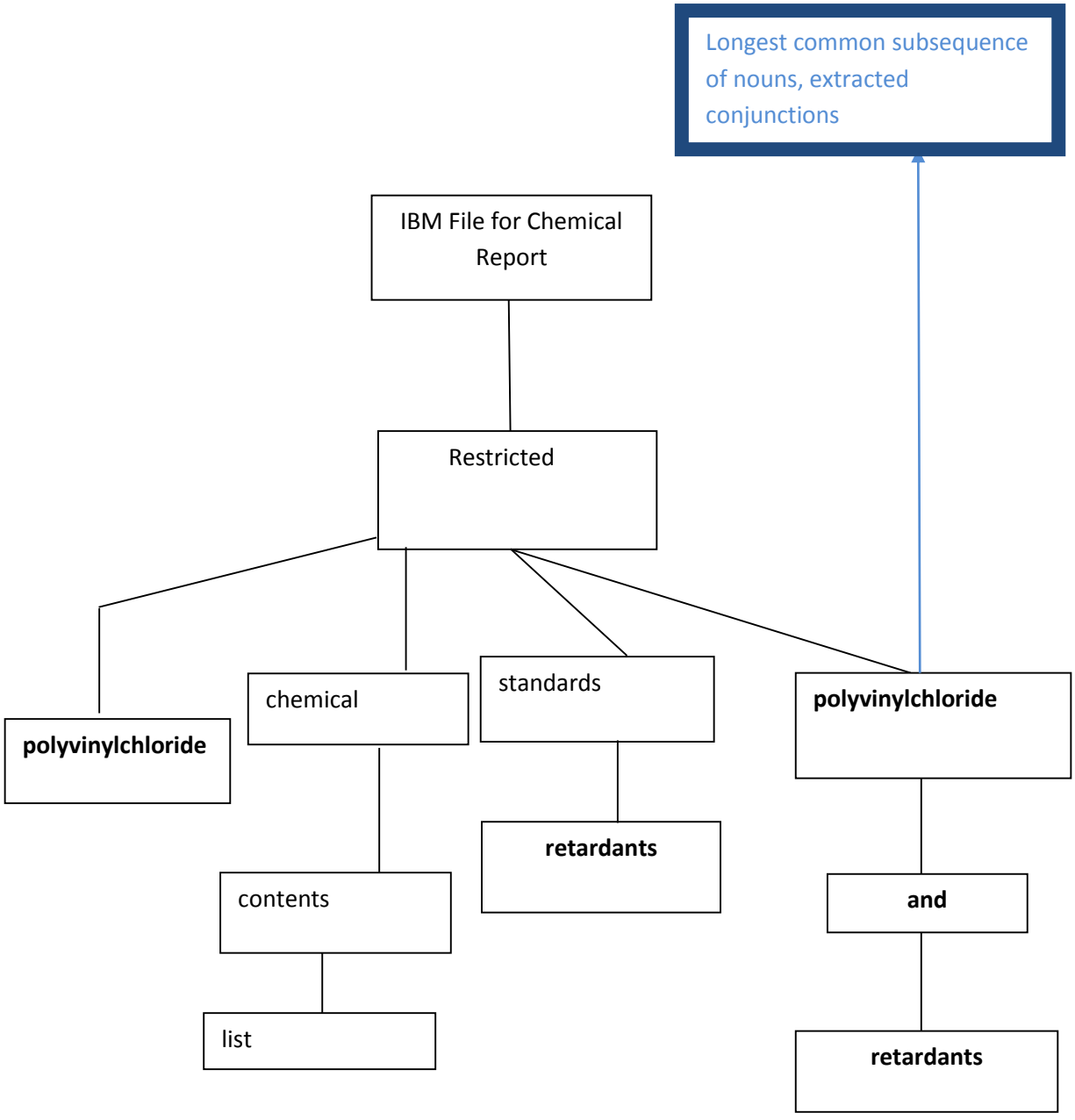


Figure 12-A: Example of FPMIES algorithm implementation

The search term or candidate item serves as the label on the left side. The right-hand side presents the output of FPMIES for the given search term.

Restricted Chemicals:	Polyvinylchloride and retardants products

A snapshot of the FPMIES database for the IBM document collection, as mentioned in Figure 12-A, is shown in Table 8.

Object	Pattern
Restriction of chemicals	Polyvinylchloride (Pattern 1)
	chemical contents list (Pattern 2)
	standards retardants (Pattern 3)
	polyvinylchloride and retardants (Pattern 4)

Table 8: Snapshot of the FPMIES Database

Another example for the item "chemicals" of the candidate item "restricted chemicals" is shown in Figure 11-B. The FPMIES information extraction results is shown in Figure 12-B.

IBM has a long history of continually taking steps to evaluate the **chemicals** used in its **products**, identifying potential **substitutes** that may have less impact on the **environment**. IBM is eliminating, restricting and prohibiting the use of **chemicals** for which a more preferable alternative is available that is capable of meeting quality and safety requirements of its **products**. Alternative materials used in place of PVC or halogenated flame retardants shall meet IBM's assessment requirements. An alternative can be considered viable if it eliminates the halogenated flame retardant or PVC to less than 1000ppm.

Figure 11-B: Data Sample 2 from IBM file

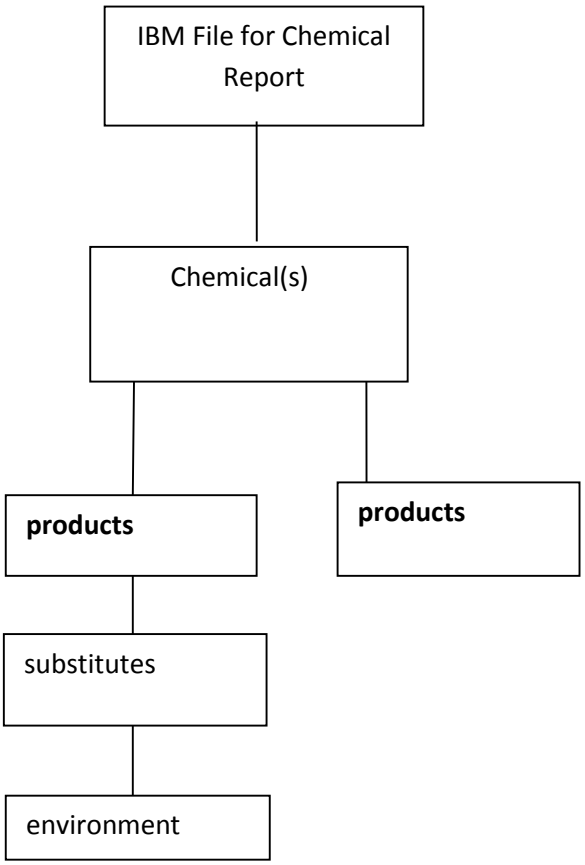


Figure 12-B: Example of FPMIES Algorithm implementation

Comparison of FPMIES with Apriori

Because the FPMIES algorithm is a new approach, its performance will be compared with the most frequently referenced algorithm from the IE literature, Apriori. This comparison will be used to highlight the advantages of FPMIES over its predecessor. Figure 13 presents a pseudocode for the Apriori algorithm.

```
while (length of file <> 0 )
{
        Do for each word in file
        {
                Set item= word
                Set count = 0

                For (i=1; i<numberofwords in file; i++)
                {
                        Generate i-frequent item set
                        Set count = count + 1 if item found
                }

        }
        Display frequent items
}
```

Figure 13: Pseudocode –Apriori Algorithm

For the search term 'restricted chemicals' in the IBM document collection, as presented in Figure 12-A, Apriori starts with the generation of one item and increases incrementally to 2-items and then to 3-items. The Apriori results are shown in Table 9.

1-Item Result	2-Item Results	3-Item Results
Chemical	Chemical, product	Chemical, product, standard
Polyvinylchloride	Chemical, specification	Chemical, product, specification
Retardants	Chemical, substance	Restricted, Chemical, substance
Phthalates	Restricted, Chemical	Restricted, Chemical, standard
Restricted	Restricted, Polyvinylchloride	Restricted, Polyvinylchloride, product
Specification	Restricted, phthalates	Restricted, phthalates, product
Standard	Restricted, standard	
Flame	Restricted, substance	
Retardant	Product, specification	
Product	Product, Polyvinylchloride	
Substance	Product, Phthalates	
	Product, standard	
	Flame, retardants	

Table 9: Apriori Generation

The steps involved in executing the Apriori algorithm can require extensive CPU time when compared to the processing time for FPMIES. More detailed analysis of performance related results will be presented in Chapter Five. Due to space and time limitation, it was beyond the scope of this dissertation to present additional n-item sets (4-item, 5-item, etc.) for Apriori. In reality, Apriori would generate n-item sets where 'n' stands for the maximum number of frequent item sets that Apriori can generate. Apriori thus generates a huge number of patterns, making it difficult for a user to analyze the data.

The FPMIES algorithm makes it easier to extract key insights from a document collection, as shown in Appendix G, where the complete result from the IBM, HP, and Dell document collection is presented. FPMIES is useful because the output generated from FPMIES is in the form of a summary giving more detailed information on the search term.

Chapter Five explores the performance of the FPMIES algorithm in comparison to manual information retrieval. FPMIES is also compared to automated retrieval using Apriori. The manual comparison was performed using a sample of 44 university students. The comparisons are presented in terms of precision and recall as well as time required for the information extraction tasks.

CHAPTER 5

TESTING AND ANALYSIS

Testing is performed to compare FPMIES with Apriori. An experiment was also performed to compare recall, precision, F-measure, and speed of FPMIES with manual extraction. The subjects who participated in the experiment were provided a set of survey questions based on the data set. In manual extraction, the questions were answered by manually reviewing the reports. The users answered the same questions by extracting information using FPMIES. The survey questionnaire is shown in Appendix D.

Comparison of FPMIES with Manual Testing

The FPMIES system was tested with a sample of 44 university students. The testing sample consisted of undergraduate and graduate students. Some of the graduate students, primarily the Ph.D. students were pursuing degrees in Chemical Engineering, so they can be reasonably expected to have some familiarity with toxic chemicals. The undergraduate students were surveyed from a business school at a university in the southern United States. Undergraduate students tested the FPMIES system in a classroom. Graduate students participated in the study at their convenience. Subjects were asked to provide information on demographics, supply chain courses previously taken, and their feedback on the usefulness of the FPMIES system. Results from Table 10 indicate that 50% of the respondents had taken supply chain courses before.

		Number	Percent
Gender	Male	24	54.6%
	Female	18	40.9%
	Non-respondents	2	4.5%
Major	MIS	8	18.2%
	Management	9	20.5%
	Accounting	13	29.6%
	Finance	4	9.1%
	Mktg	5	11.4%
	Journalism	1	2.3%
	Chemical Eng	3	6.8%
	Non-respondents	1	2.3%
Taken Supply Chain Course before	Y	22	50%
	N	21	47.7%
	Non-respondents	1	2.3%

Table 10: Demographics of Respondents (n=44)

The amount of relevant information extracted by the subjects is measured using the standard recall formula of IE systems. Recall is measured by dividing the number of correct answers produced by the total possible correct answers. On ten tasks, the recall for the FPMIES system was found to be 83.4% compared to the recall of manual extraction of 27.3%.

Precision is measured by dividing the number of correct answers produced by the number of total answers produced. The precision for manual extraction is 38.7% compared to 84.6% precision for the FPMIES system.

The standard F-measure combines 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 FPMIES is 83.9% compared to the F-measure for manual process of 30.7%.

Subjects were timed on the manual and automated processes. The average amount of time needed to complete the ten queries using a manual extraction process was 354.54 seconds compared to the average speed of 128.54 seconds for the human subjects to complete the same ten queries using FPMIES.

Statistical Analysis

Statistical analysis from Table 11 reveals that the subjects spent significantly more time extracting information manually (mean = 354.54) compared to the time taken to extract information using FPMIES (mean = 128.54). This is a substantial performance improvement. The t-test results for manual and FPMIES extraction is shown in Table 13. It is clear from these results that individuals can process a large amount of information in much less time using FPMIES, when compared to manual extraction.

Time in Seconds	N	Minimum	Maximum	Mean	Std. Deviation
Manual time	44	180.00	900.00	354.545	148.566
FPMIES time	44	22.00	266.00	128.545	69.700

Table 11: Speed comparison between manual system and FPMIES (n=44)

Table 12 indicates that average recall and precision for FPMIES are higher compared to manual extraction. The table further indicates that the mean of F-measure for FPMIES is 0.8393, which is higher than the F-measure mean of 0.3068 for manual extraction.

	Recall-Mean	**Recall Std Dev**	**Precision-Mean**	**Precision-Std Dev**	**F-Measure-Mean**	**F-Measure-Std Dev**
Manual	0.273	0.104	0.387	0.188	0.307	0.116
FPMIES	0.834	0.135	0.846	0.131	0.839	0.131

Table 12: Recall, Precision and F-measures for Manual and FPMIES system

Table 13 shows the t-test results for recall and precisions for manual and FPMIES extraction. The t-values in the table are found to be significant.

Paired Samples T-Test					
Paired Comparisons	Mean	Standard Deviation	T-Value	Degrees of Freedom	P-Value
Manual Time-FPMIES Time	226.00	142.601	10.513*	43	<.001
Manual Recall-FPMIES Recall	-.561	.171	-21.716*	43	<.001
Manual Precision-FPMIES Precision	-.458	.250	-12.135*	43	<.001
Manual F-measure-FPMIES F-measure	-.532	.186	-18.943*	43	<.001

*Significant at the .05 level.

Table 13: Results of paired t-test

Statistical Power and Effect size. Although there are no formal standards for power (sometimes referred as π), most researchers assess the power of their tests using π=0.80 as a standard for adequacy. The study must be of adequate size, relative to the goals of the study. The size of the study must be big enough for the effect to be of scientific significance. Obtaining a size of scientific importance requires obtaining meaningful input.

Sample size (N) is the number of observations (cases) in a sample. Beta (ß) is defined as the probability of committing a Type II error (failure to reject a false null hypothesis). The power of the test (1-ß) is the probability of correctly rejecting a false null hypothesis. The computation of statistical power depends on a model (or test). A power value is between 0 and 1. If the power is less than 0.8, the sample size would have to be increased.

To know if an observed difference is not only statistically significant but also important or meaningful, its effect size would have to be calculated as well. Effect size is the difference in means between the two groups divided by the standard deviation.

To interpret the resulting number, most social scientists use this general guide developed by Cohen:

- < 0.1 = trivial effect
- 0.1 - 0.3 = small effect
- 0.3 - 0.5 = moderate effect
- > 0.5 = large difference effect

The means for processing time for automatic and manual extraction are 354.55 and 128.545 seconds respectively. For sample size of 44 students, alpha =0.05 and two-tail test, power is ~0.9 and effect size is 0.6. By looking at the T distribution table (df = 43 at the .05 level, 2-tail test), it is determined that the probability of being greater than or equal to 2.021 is .025. The statistical power is computed as ~0.9 which is verified from the power table. This high statistical power

indicates that the T-test is highly likely to detect an effect with a magnitude of 0.6. The null hypothesis will be rejected if it is false at .05 level 2-tail test. Hence it can be concluded that the sample size is adequate to draw conclusions of scientific significance.

Comparison of FPMIES with Apriori

Table 14 presents a comparison of the characteristics of FPMIES, Apriori, and Ex-Ante.

Algorithm	Number of Database Scans	Algorithm Semantic Complexity	Scalability	Data Preparation False Negative rate
Apriori	High	Low	Low	-
FPMIES	Low	High	High	Low
Ex-Ante (Pre-processing algorithm)	-	-	-	High

Table 14: Comparison of FPM Algorithms

The scales of the independent variables listed in Table 14 are:

Number of Database Scans (per File):

Low: 0 to 20
Medium: 20 to 40
High: > 40

Scale for Algorithm semantic complexity:

Low: No algorithm semantic complexity is applied or a custom dictionary built for synonyms of certain words based on the programmer's selection of important words.
High: Scanning the Princeton word net for synonyms

Data Preparation False Negative rate:
(The elimination of a pertinent term)

High: Elimination of a monotone term due to its infrequency
Low: Non-elimination of monotone terms. Checking for the frequency of 2 or more terms occurring together.

Scalability (only text, no graphics):

Low: 500-1500 KB
Medium: 1500-3500 KB
High: >3500 KB

The scalability for FPMIES and Apriori is shown in Figure 14. As size of data input increases, the processing time of the algorithm increases.

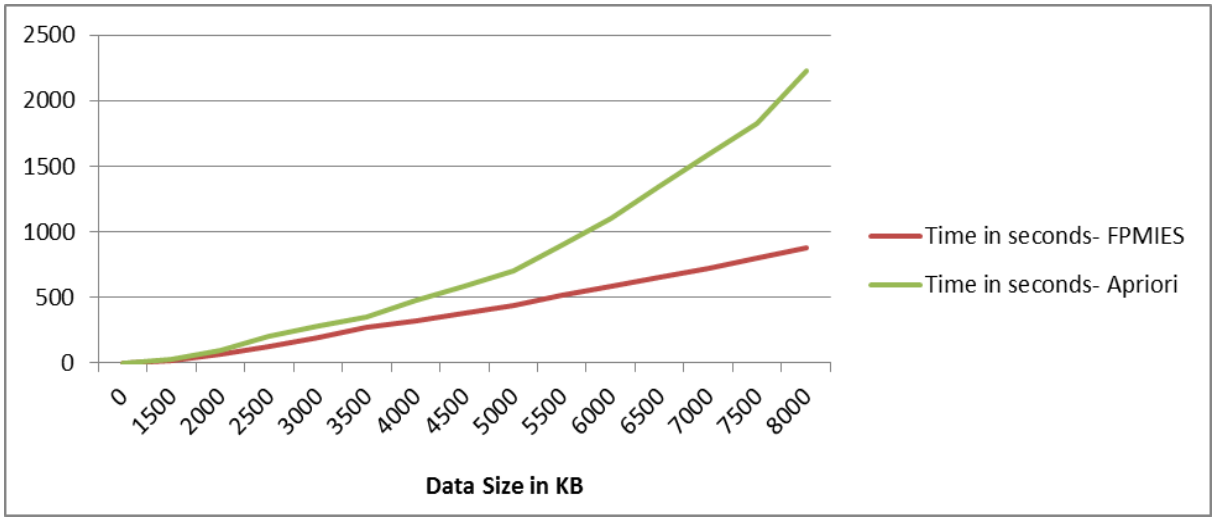


Figure 14: Scalability chart of FPMIES vs Apriori

The amount of data stored in a database has an impact on the performance of a system. The system becomes slower with additional data in the database since greater amount of data has to be scanned by the system to generate the output. Scalability illustrates the dependency of performance on factors such as data volume. Scalability refers to the maximum data input the system can handle yielding a suitable performance efficiency. For a given data input, a system that provides the output in a reasonable period of time is considered to perform reasonably well. It is thus important to determine the maximum amount of data FPMIES can handle.

In FPMIES, data is filtered by content-organization of the data. Collocation analysis is further applied as a means of pre-processing to generate more relevant data. In the process of

filtering, semantic rules are applied to yield meaningful patterns. The FPMIES system is tested for performance efficiency with different sizes of data input. As seen in the figure, scalability for FPMIES is high, compared to the scalability of Apriori.

Recall for Apriori is measured by dividing the number of relevant items retrieved by the total number of relevant items. Precision for Apriori is measured by dividing the number of relevant items retrieved by the total number of retrieved items. Recall and precision for the Apriori example in Table 9 in Chapter 4 are computed using the respective formula.

Recall for Apriori (for example in Table 9) = 11/30= 36.67%

Precision for Apriori (for example in Table 9) = 11/96 = 11.45%

F-measure for Apriori (for example in Table 9) = (2RP/R+P) = 0.1749 ~ 17.5%

These results provide evidence that an increase in the number of database scans decreases the FPM algorithm performance. With increased input data to Apriori, the algorithm performance will keep diminishing. This is because the number of database scans increases with increased input data. Hence, the level of scalability in Apriori is low. Recall and precision of other candidate items that are computed in a similar manner, is shown in Table 15.

Candidate Item	**Recall-Apriori**	**Precision-Apriori**	**Recall-FPMIES**	**Precision-FPMIES**
Green Energy	32.1	30.2	78.2	71.4
Recyclable Packaging	33.4	22.6	67.4	68.9
Green Packaging	35.6	12.5	72.3	65.6
EICC Practice	23.1	21.6	75.4	72.6
ISO 14001	19.1	33.1	79.9	78.3
Hazardous Substance	25.4	31.1	83.4	84.3

Carbon Footprint	37.7	18.9	71.8	69.9
Green IT	12.4	32.6	67.4	75.2
Greenhouse Gas	34.5	29.4	84.5	78.3
Green innovation	24.6	22.9	84.3	72.1
Green Technology	33.7	27.6	77.4	76.3
Green Supply Chain	34.8	35.5	75.6	87.3
Restricted Chemicals	36.6	11.4	78.2	66.7
Electronic waste	33.9	23.4	74.3	79.8
Packaging waste	34.1	15.7	81.2	72.5

Table 15: Comparison of FPMIES with Apriori

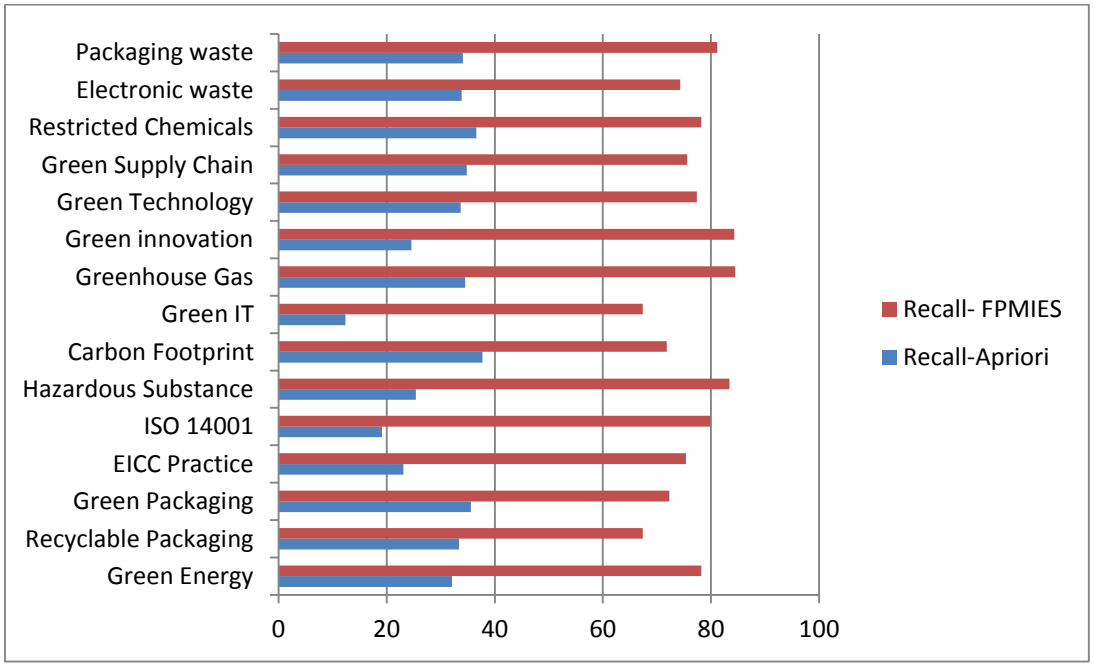


Figure 15-A: Recall of Apriori vs FPMIES

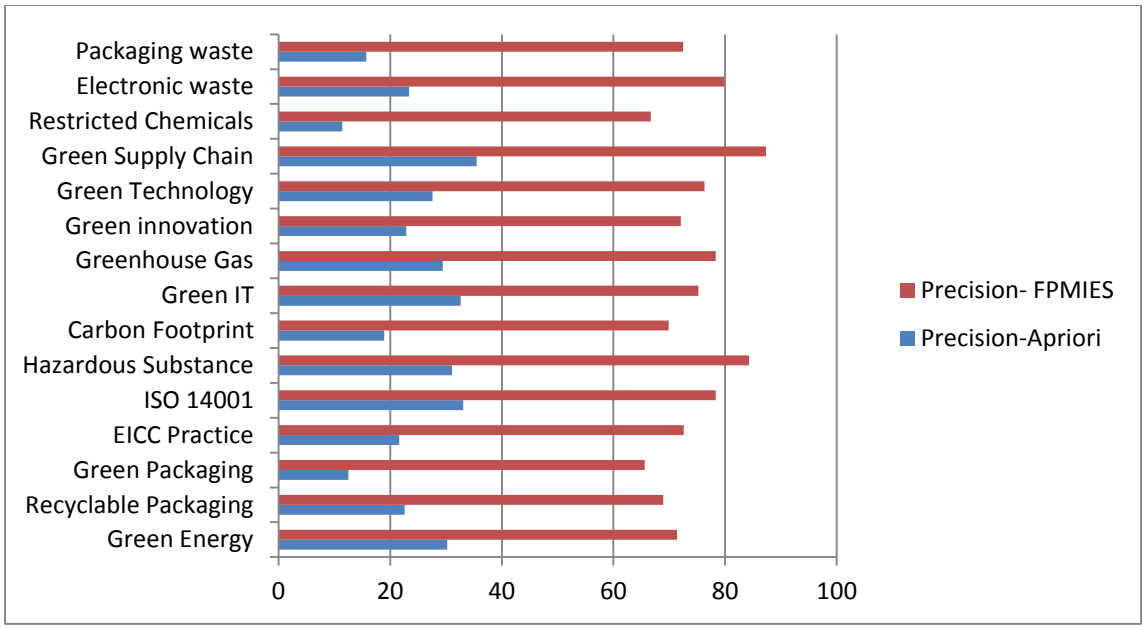


Figure 15-B: Precision of Apriori vs FPMIES

Figure 15 shows the recall and precision of Apriori and FPMIES. The recall and precision of Apriori are low as compared to those of FPMIES. Average recall of Apriori is approximately 30% as compared to the high recall value of FPMIES of 76.7%. Average precision of Apriori is 24.6% as compared to the high precision value of FPMIES of 74.6%. Since recall and precision are high for the FPMIES system, the F-measure for FPMIES is substantially higher than that of Apriori. F-measure of FPMIES is 75.2% and F-measure of Apriori is 54.6 %.

Summary

The FPMIES system was tested with a sample of 44 university students. The amount of relevant information extracted by the subjects was measured using the standard recall formula commonly used to evaluate information extraction and retrieval systems. The F-measure for FPMIES is 83.9% compared to the F-measure for manual process of 30.7%.

The performance of the FPMIES system was also compared to Apriori. The recall and precision statistics for Apriori were low when compared to those of FPMIES. The F-measure of FPMIES was 75.2% while the F-measure of Apriori was 54.6 %. This represents a substantial performance improvement for FPMIES when compared to the benchmark algorithm of Apriori.

CHAPTER 6

DISCUSSION AND CONCLUSION

Discussion

The focus of this dissertation was an examination of the factors that impact FPM algorithm performance. An improved FPM algorithm was developed that uses a combination of advanced techniques such as collocation analysis, semantic rules and template mining.

The research model presented in Chapter Three was tested and analyzed using both human subjects and computer-based benchmarking against the commonly used Apriori algorithm. The FPMIES system was used by the human subjects to complete an information retrieval task and their computer-supported results were compared to their performance in a manual information retrieval task. FPMIES was also compared to Apriori. All four hypotheses suggested by the research model were supported, as shown in Figure 16.

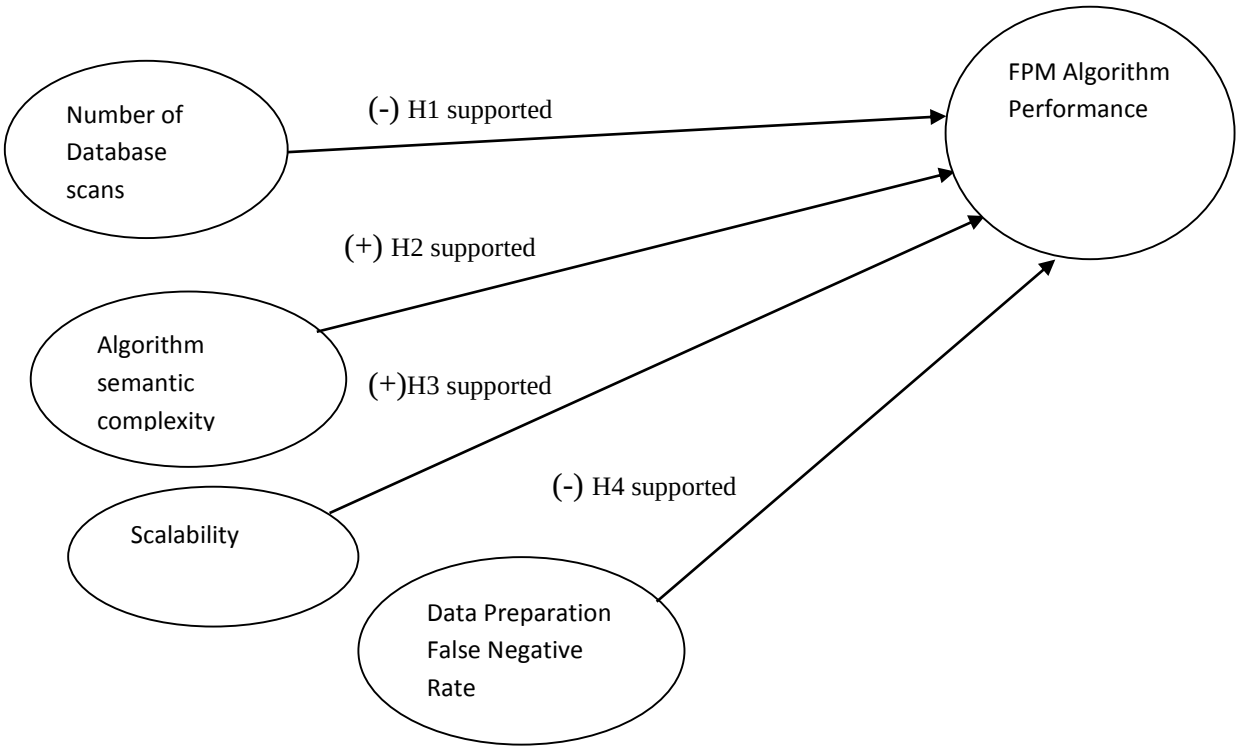


Figure 16: FPM Algorithm Performance Model

Data analysis demonstrated that the number of database scans affected the FPM algorithm performance. As seen earlier in Tables 14 and 15, a lower number of database scans resulted in improved FPM algorithm performance. Thus hypothesis 1 was supported, confirming that an increase in the number of database scans leads to a decrease in algorithm performance. Semantic rules applied within an FPM algorithm helped generate more relevant patterns. This resulted in improved FPM algorithm performance. Hypothesis 2 is thus supported, as presented earlier in Tables 14 and 15. Therefore, an increase in an algorithm's semantic complexity can increase the algorithm's performance.

Furthermore, Tables 14 and 15 also provide support for Hypothesis 3. FPMIES has a higher scalability compared to Apriori, when considering the amount of data that can be processed by the system without incurring a substantial time penalty, thus improving FPM algorithm performance. In other words, an increase in scalability will lead to an increase in the algorithm's performance.

Finally, Hypothesis 4 is supported from Tables 14 and 15, which indicate that FPM Algorithm Performance is inversely related to the Data Preparation False Negative Rate. The risk of eliminating a pertinent term is eliminated in the FPMIES algorithm. This is because even if single-occurring terms are not frequent by itself, their frequency of co-occurrence with another term is taken into consideration in FPMIES unlike the way in which the terms are deleted under Ex-Ante processes.

Table 16 provides a summary of the hypotheses results.

Hypotheses	**Result**
H1- Increase in number of database scans will lead to decrease in algorithm performance.	Supported
H2- Increase in algorithm semantic complexity increases the algorithm performance.	Supported
H3- Increase in scalability will lead to increase in algorithm performance.	Supported
H4- FPM Algorithm Performance is inversely related to the Data Preparation False Negative Rate.	Supported

Table 16: Hypotheses Results

Based on the statistical analysis presented in Chapter Five, there was some evidence to support all four hypotheses. Furthermore, when considering statistical power of 0.9 and the anticipated effect size of 0.6, an adequate number of observations for the experiment involving

human subjects were determined to be at least 27 observations. Because a total of 44 subjects were involved in the study, appropriate standard of scientific significance were met.

Conclusion

Methods developed using FPMIES can foster for future IE research. FPMIES has several advantages, including great speed and producing more relevant searches. FPMIES was rated by users in this study to be a satisfactory system. Due to the combination of semantic relationships with template mining, the recall of FPMIES is quite significant. The overall recall of FPMIES was 83.4% when tested with users.

Appendix G presents the output generated by the FPMIES algorithm. FPMIES enables a user to extract relevant information in relatively less time compared to manually reading and searching from online reports.

An improved method of frequent pattern mining has been presented in this dissertation with some unique solution methods. A combination of collocation analysis, semantic rules and template mining yielded high recall and precision for FPMIES. With the number of database scans reduced, and increase in scalability and semantic complexity, FPMIES produced better results compared to the Apriori algorithm. Results indicated 76.7% recall for FPMIES and 30% for Apriori. The precision for FPMIES resulted as 74.6% and 24.6% for Apriori.

Limitations and Future Research

This section discusses the limitations and the scope for further research based on this study. The data set used to test the FPMIES algorithm contained documents related to the green supply chain practices of three companies. However, the algorithm can also be tested using data sets for other application domains such as healthcare, finance, etc.

The span selected for identifying collocated terms was six (6) in this study. Search terms were generated based on the frequency of collocated terms within that span. A different span size could yield different search results, so a method of determining optimal span size for a given data set might be useful in future research.

The use of students as human subjects is often problematic; however, for the relatively simple information retrieval tasks required in this study, student subjects were adequate. Future studies could explore the efficacy of the FPMIES system with other user groups.

A longitudinal study could be performed involving usage of the FPMIES system within a particular problem domain over time. Such a study would help eliminate user needs as they became more familiar with the system through repeated usage.

Further, the FPMIES algorithm was compared to the Apriori algorithm, which is the most popular FPM algorithm. The FPMIES algorithm could also be compared with other FPM algorithms such as Lofia and DFPMT.

BIBILIOGRAPHY

Agrawal, R., Imielinski, T., Swami, A. (1993). Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, Special Issue on Learning and Discovery in Knowledge-Based Databases. 5(6). pp. 914-925

Agrawal, R., Srikant, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. *Advances in Database Technology (EDBT)*. pp. 1-17

Alhenshiri, A., Blustein, J. (2012). Exploring Visualisation in Web Information Retrieval. *International Journal of Internet Technology and Secured Transactions*. pp. 320-330

Andersen, P., Hayes, P., Huettner, A., Schmandt, L., Nirenburg, I., Weinstein, S. (1992). Template Mining for Information Extraction from Digital Documents. *Third Conference on Applied Natural Language Processing*. pp. 170-177

Andersen, P., Huettner, A. (1994). Knowledge Engineering for the JASPER Fact Extraction System. *Integrated Computer-Aided Engineering.* 1(6). pp. 473-493

Arampatzis, A., Van der Weide, P., Van Bommel, P., Koster, C. (2000). *Linguistically-Motivated Information Retrieval*. Encyclopedia of Library and Information Science. Marcel Dekker, Inc. New York

Balan, S., Conlon, S. (2012, August). Evaluating Best Practices in Green Supply Chain. *AMCIS 2012 Proceedings*. Paper 60

Bell, E. (2007). *Collocation Statistical Analysis Tool: An Evaluation of the Effectiveness of Extracting Domain Phrases via Collocation*. Lancaster University. B.Sc. Dissertation. pp. 1-56

Bonchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D. (2003). ExAnte: Anticipated Data Reduction in Constrained Pattern Mining. *Knowledge Discovery in Databases*. pp. 59-70

Bouma, G. (2005). Normalized Pointwise Mutual Information in Collocation Extraction. *Proceedings of Biennial GSCL Conference*. pp. 31-40

Brin, S., Motwani, R., Ullman, J., Tsur, S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 26(2). pp. 255–264

Cheng, H., Yan, X. (2007). Frequent Pattern Mining: Current Status and Future Directions. *Data Mining Knowledge Discovery*. pp. 56-85

Chen, N., Prasanna, V. (2012). Learning to Rank Complex Semantic Relationships. *International Journal on Semantic Web and Information Systems.* pp. 1-9

Chen, Y. (1998). An Efficient Parallel Algorithm for Mining Association Rules in Large Database. *CiteSeer*. pp. 1-10

Chong, U., Goh, A. (1997). FIES: Financial Information Extraction System. *Information Services and Use*. 17(4). pp. 215-223

Chowdhury, G. (1999). Template Mining for Information Extraction from Digital Documents. *Library Trends*. 48(1). pp. 183-208

Costantino, M., Morgan, R., Collingham, R. (1996). Financial Information Extraction Using Pre-defined User-definable Templates in the LOLITA System. *Journal of Computing and Information Technology*. 4(4). pp. 241-255

Cowie, J., Lehnert, W. (1996). Information Extraction. *Communications of the ACM*. 39(1). pp. 80-91

Croft, U. (1998). What do People Want from Information Retrieval: The Top 10 Research Issues for Companies that Use and Sell IR Systems. D-Lib Magazine. Retrieved December 7, 1998 from http://www.dlib.org/dlib/november95/11croft.html

Deliyanni, A., Kowalski, R. (1979). Logic and Semantic Networks. *ACM*. pp. 184-192

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*. pp. 61-74

Eklund, T., Toivonen, J., Vanharanta, H., Back, B. (2011). Customer Feedback Analysis Using Collocations. *AMCIS 2011 Proceedings*. pp. 1-8

Environmental Quality (1996). The 25th Anniversary Report of the Council on Environmental Quality. Retrieved Feb, 2012 from http://ceq.hss.doe.gov/ceq_reports/annual_environmental_quality_reports.html

Firth, J. (1951). Modes of Meaning. *Papers in Linguistics*. Oxford University Press. pp. 190-215

Gaizauskas, R., Wilks, Y. (1998). Information Extraction: Beyond Document Retrieval. *Journal of Documentation*. pp. 70-105

Goethals, B., Zaki, M. (2003). Advances in Frequent Itemset Mining Implementations: Introduction to FIMI03. *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementation.* pp. 1-10

Grant, G., Conlon, S. (2006). Edgar Extraction System: An Automated Approach to Analyze Employee Stock Option Disclosures. *Journal of Information Systems*. 20(2). pp. 119-142

Harris, Z. 1954. Distributional Structure. *Word* 10(23). pp. 146–162
Hasan, M. (2009). *Mining Interesting Subgraphs by Output Space Sampling*. Thesis Submitted to the Graduate School at Rensselaer Polytechnic Institute. pp. 1-112

Hasan, M., Chaoji, V., Salem, S. (2012). DMTL: A Generic Data Mining Template Library. Retrieved June, 2012 from http://www.cs.rpi.edu/~zaki/PaperDir/PS/LCSD05.pdf

Hasan, M., Zaki, M. (2008). A Novel and Effective Approach for Mining Representative Graph Patterns. *Statistical Analysis and Data Mining*. pp. 67-84

Hoel, P. (1962). Introduction to Mathematical Statistics. Wiley. New York

Jade, R., Verma, L., Verma, K. (2013). Intelligent Data Mining Techniques for Coal Mining Data. *IJET*. pp. 1-4

Jagtap, S., Kodge, B. (2013). Census Data Mining and Data Analysis using WEKA. *International Conference in Emerging Trends in Science, Technology and Management (ICETSTM)*. Singapore. pp. 35-40

Jang, M., Hyon, S., Park, Y. (1999). Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. pp. 224-229

Khoo, C, Na, C. (2006). Semantic Relations in Information Science. *Annual Review of Information Science and Technology*. pp. 157-228

Keith, B. (2006). *Encyclopedia of Language and Linguistics.* Elsevier. Oxford

Kim, M., Choi, K. (1998). A Comparison of Collocation Based Similarity Measures in Query Expansion. *Information Processing and Management*. pp. 19-30

Krovetz, R. (1993). Viewing Morphology as an Inference Process. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 191-202

Laguna, I., Gamblin, T., Supinski, B., Bagchi, S., Bronevetsky, G., Anh, D., Schulz, M., Rountree, B. (2011). Large Scale Debugging of Parallel Tasks with AutomatDeD. *ACM*. pp. 1- 10

Latham, P., Roudi, Y. (2009). Mutual Information. Retrieved Jan, 2013 from http://www.scholarpedia.org/article/Mutual_information

Liu, W., Schmidt, B., Voss, G., Schroder, A., Müller-Wittig, W. (2006). Bio-Sequence Database Scanning on a GPU. *IEEE*. pp. 1-8

Luhn, P. (1960). Keyword-in-Context Index for Technical Literature (KWIC index). *American Documentation*. pp. 288–295

Lytinen, L., Gershman, A. (1986). ATRANS: Automatic Processing of Money Transfer Messages. *The Fifth National Conference on Artificial Intelligence*. Philadelphia, Pennsylvania. pp. 1089-1093

Mandala, R., Tokunaga, T., Tanaka, H. (1999). Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. *ACM*. pp. 191-197

McInnes, B. (2004). Extending the Log Likelihood Measure to Improve Collocation Identification. *ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 191–197

Nawrocka, D., Brorson, T., Lindhqvist, T. (2009). ISO 14001 in Environmental Supply Chain Practices. *Journal of Cleaner Production*. pp. 1435-1443

Ogawa, Y., Morita, T., Kobayashi, K. (1991). A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. Fuzzy Sets and Systems. *Special Issue on Applications of Fuzzy Systems Theory*. 39(2). pp. 163-179

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program.* 14(3). pp. 130-137

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*. pp. 37-63

Rijsbergen, C. (1977). A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*. 33(2). pp. 106-119

Savasere, A., Omiecinski, E., Navathe, S. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*. pp. 432–444

Schank, R., Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum. Hillsdale

Schank, R., Riesbeck, C. (1981). *Inside Computer Understanding: Five Programs Plus Miniatures*. Lawrence Erlbaum. Hillsdale

Sinclair, J. (1991). *Corpus, Concordance, Collocation: Describing English language.* Oxford University Press. Oxford

Smadja, F. (1993). Retrieving Collocations from Text. *Computational Linguistics*. 19(1). pp. 143-177

Smadja, A., McKeown, R. (1990). Automatically Extracting and Representing Collocations for Language Generation. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252-259

Simmons, L., Conlon, S., Mukhopadhyay, S., Yang, J. (2011). A Computer Aided Content Analysis of Online Reviews. *Journal of Computer Information Systems*. 52(1). pp. 43-55

Taniar, D., Iwan, L. (2011). *Exploring Advances in Interdisciplinary Data Mining and Analytics*. IGI. Indonesia

Vechtomova, O., Robertson, S., Jones, S. (2003) Query Expansion with Long-Span Collocates, *Information Retrieval*. pp. 251-273

Vickery, B. (1997). Knowledge Discovery from Databases: An Introductory Review. *Journal of Documentation*. 53(2). pp. 107-122

Wanner, L. (1996). *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins Publishing. Amsterdam

WordNet (2012). A Lexical Database for English. Retrieved June, 2012 from http://wordnet.princeton.edu/

Xu, J., Croft, B. (1998). Corpus-Based Stemming Using Co-occurrence of Word Variants. *ACM Transactions on Information Systems*. 16(1). pp. 61-81

Yang, Y., Pedersen, J. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*. pp. 412-420

Zaki, M., Parthasarathy, S., Li, W., Ogihara, M. (1997). Evaluation of Sampling for Data Mining of Association Rules. *7th IEEE International Workshop on Research Issues in Data Engineering.* pp. 1-9

Zaki, M., Gouda, K. (2003). Fast Vertical Mining Using Diffsets. *ACM*. pp. 1-10

APPENDIX

APPENDIX A: SYNONYMS CREATED BY WORDNET

Word	Synonym
Chemical	chemical substance chemic (relating to or used in chemistry) chemical (of or made from or using substances produced by or used in reactions involving atomic or molecular changes) "chemical fertilizer"
Hazard/hazardous	peril risk endangerment stake
Supply	provision render furnish
Waste	barren emaciate rot
Environment	surroundings surrounding conditions

Carbon	carbon paper (a thin paper coated on one side with a dark waxy substance (often containing carbon) carbon copy(a copy made with carbon paper)
Electronic	electronic devices (relating to electronics; concerned with or using devices that operate on principles governing the behavior of electrons) electronic energy
Energy	vigor vitality
Green	Greenness Greenish
Substance(s)	core essence consistency
Restricted	control limit bound confine

Standard	criteria measure touchstone monetary standard
EPA	Environmental Protection Agency
Recycle/Recyclable	reprocess reuse
Packaging	material used to make packages business of packaging

APPENDIX B: GREEN SUPPLY CHAIN KEY TERMS

Glossary of Key Terms	
Business Process	A set of logically related tasks or activities performed to achieve a desirable business outcome.
Disposition	Process of discarding waste materials or obsolete products.
Environmental Burden	A release or modification to the environment, due to an industrial process, that may have adverse effects.
Eco-Efficiency	The ability to meet cost, quality, and performance goals; reduce environmental impacts; and conserve valuable resources.
Hazardous Waste	A waste such as chemicals or nuclear material that is hazardous to humans or animals and requires special handling. Hazardous waste costs are typically substantially higher than other waste costs due the special handling, training, and recording as well as higher disposal fees that are required.
Life Cycle	A sequence of stages spanning the lifetime of a product, process, service, facility, or enterprise from inception to final use and disposition; in the case of materials, includes extraction, acquisition, manufacturing, and ultimate reuse or disposal.
Logistics	Activities to move incoming materials and distribute finished products to the proper place, at the desired time, and in the optimal quantities.
Materials Handling	Process of developing and implementing manual, mechanized, and automated systems to provide movement of materials throughout a facility.
Materials Management	The grouping of management functions supporting the complete cycle of material flow, from the purchase and internal control of production materials to the planning and control of work-in process to the warehousing, shipping, and distribution of the finished product.
Materials Recovery	Activities to prevent the release of materials into air, water, or solid waste streams and incorporate these materials back into the manufacturing process.
Materials Tracking	Assessment of what, where, why and how much material is acquired, incorporated into products and co-products, and channeled into waste streams throughout the materials life cycle.
Net Present Value	The discounted value of future earnings for a given number of time periods. The discount rate reflects the company's time value of money and commonly ranges from 10% to 15% per year.
Purchasing	Process of determining specifically which materials, supplies and services must be procured, and then obtaining those resources from suppliers.
Recovery	Process of obtaining a valuable resource from a potential waste material.
Sourcing	Process of determining the types of products and services required and establishing purchasing relationships with capable suppliers.
Supply Chain	The functions inside and outside a company that enable the value chain to make products and provide services to the customer.

Value Analysis	A systematic approach that identifies a required function of a product or service, establishes a value for that function.

APPENDIX C: EXAMPLE OUTPUT OF KWIC INDEX FILE

Bamboo	Promotes	healthy	Soil	and	using	Bamboo	also	Makes
Promotes	Healthy	soil	And	using	bamboo	also	makes	
healthy	Soil	and	Using	bamboo	also	makes		
Soil	And	using	Bamboo	also	makes			
And	Using	bamboo	Also	makes				
Using	Bamboo	also	Makes					
bamboo	Also	makes						
Also	Makes							
Makes								

APPENDIX D: SURVEY QUESTIONS

Start Time:

End Time:

State the company that you selected:

The following question is on ISO –

1. Does the company have ISO 14001 Certificate?

Yes No

The following question is on harmful chemicals –

2. Which of the following hazardous substances or chemicals are restricted by the company?
   a. PVC and Brominated retardants
   b. Polyvinylchloride and retardants
   c. PVC only
   d. None of the above

The following question is on greenhouse gas emission –

3. How many tones does the greenhouse gas emissions amount to?
   a. 1 billion
   b. 3 billion
   c. 6 billion
   d. Perfect score of 100
   e. Not applicable

The following questions are on carbon emission –

4. If the company does not follow the required standards, what did they report?
   a. Total carbon footprint of 850 kg
   b. Total carbon footprint of 1000kg
   c. Total carbon footprint of 350 kg
   d. Not applicable

5. By which device/means, is the carbon footprint produced:
   a. calculator
   b. E6400 laptop
   c. Television
   d. None of the above

The following question is on EICC –

6. Which year did the company start practicing Electronic Industry Code of Conduct (EICC)
   a. 2001
   b. 2000
   c. 2004
   d. 1998

The following questions are on Green innovation and Green Technology:

7. To expand Green innovation, the following was done:
   a. A innovation company was found
   b. A $86 million data center was built
   c. A green strategy was initiated
   d. None of the above

8. By means of green technology, the following was achieved:
   a. Food purification
   b. Green plants
   c. Water purification by removing arsenic
   d. None of the above

The following questions are on Green Packaging and Packaging waste:

9. Which of the following material is used for green packaging:
   a. Wood
   b. Stick
   c. Bamboo
   d. None of the above
10. Which of the following is a source of producing solid waste?
   a. Metal
   b. Plastic
   c. Iron
   d. None of the above

Demographics

Year of Graduation:

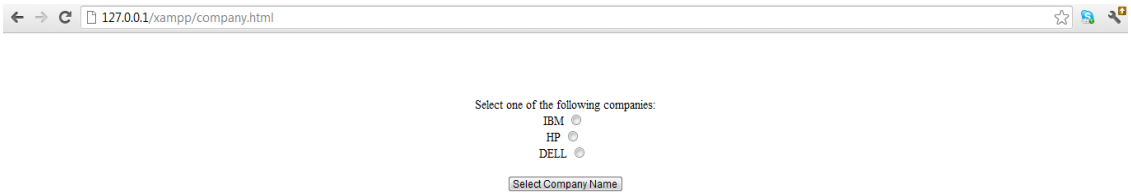
Major/Degree sought after/Area of Specialization (if any):

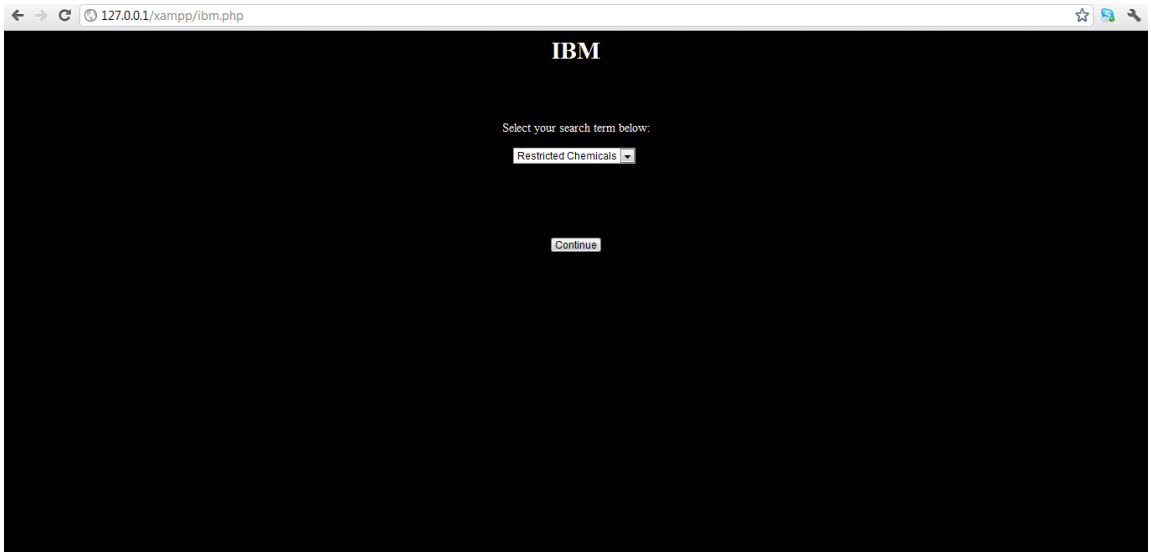
Have you taken supply chain course before?

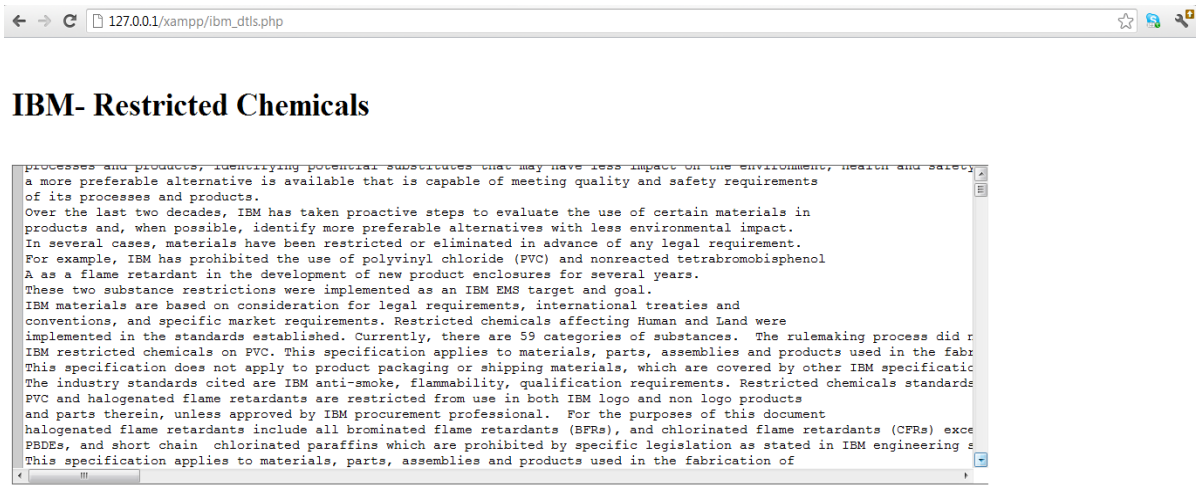
Gender:

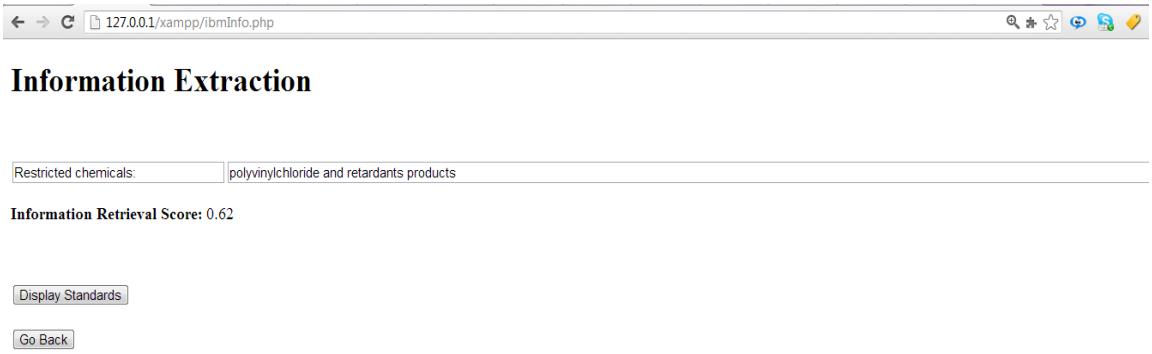
Age: Under 20 21-25 26-30 31-35 36-40 above 40

APPENDIX E: FPMIES SCREENSHOTS

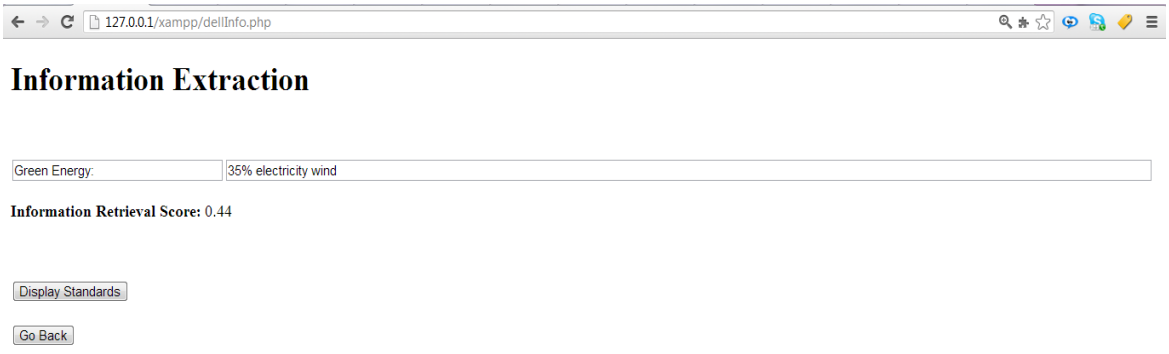

127.0.0.1/xampp/company.html
Select one of the following companies:
IBM
HP
DELL
Select Company Name


127.0.0.1/xampp/ibm.php
IBM
Select your search term below:
Restricted Chemicals
Continue


127.0.0.1/xampp/ibm_dtls.php
IBM- Restricted Chemicals
processes and products, identifying potential substitutes that may have less impact on the environment, health and safety
a more preferable alternative is available that is capable of meeting quality and safety requirements
of its processes and products.
Over the last two decades, IBM has taken proactive steps to evaluate the use of certain materials in
products and, when possible, identify more preferable alternatives with less environmental impact.
In several cases, materials have been restricted or eliminated in advance of any legal requirement.
For example, IBM has prohibited the use of polyvinyl chloride (PVC) and nonreacted tetrabromobisphenol
A as a flame retardant in the development of new product enclosures for several years.
These two substance restrictions were implemented as an IBM EMS target and goal.
IBM materials are based on consideration for legal requirements, international treaties and
conventions, and specific market requirements. Restricted chemicals affecting Human and Land were
implemented in the standards established. Currently, there are 59 categories of substances. The rulemaking process did r
IBM restricted chemicals on PVC. This specification applies to materials, parts, assemblies and products used in the fabr
This specification does not apply to product packaging or shipping materials, which are covered by other IBM specificatic
The industry standards cited are IBM anti-smoke, flammability, qualification requirements. Restricted chemicals standards
PVC and halogenated flame retardants are restricted from use in both IBM logo and non logo products
and parts therein, unless approved by IBM procurement professional. For the purposes of this document
halogenated flame retardants include all brominated flame retardants (BFRs), and chlorinated flame retardants (CFRs) exce
PBDEs, and short chain chlorinated paraffins which are prohibited by specific legislation as stated in IBM engineering s
This specification applies to materials, parts, assemblies and products used in the fabrication of
Information Extract


127.0.0.1/xampp/ibmInfo.php
Information Extraction
Restricted chemicals:
polyvinylchloride and retardants products
Information Retrieval Score: 0.62
Display Standards
Go Back


127.0.0.1/xampp/dell_dtls.php
Dell- Green Energy
Yesterday marked the one-year anniversary of Dell's commitment to be the greenest technology company on the planet. We ha
First, they reported their largest single-year product recycling volume, recovering a massive 102 million pounds of IT ec
A second accomplishment to celebrate is their becoming the first major computer manufacturer to offer Silver 80 PLUS-cert
of the Climate Savers Computing Initiative. This is nothing new for Dell - since 2005, improvements to their desktops alc
And a third item for celebration is progress in Dells global zero-carbon initiative. Partnering with The Climate Group's
Back in September, they decided the company operations should be carbon neutral by the end of '08, and they're on track w
Already powering corporate headquarters with 100% green energy using electricity sources such as wind, showing themselves
While we know there is more than simple do-good motivation behind the progress, it is nonetheless encouraging to see such
Dell is running facilities on green energy that are renewable. With green energy conservation sources as wind, Dell is ma
Dell has a broad focus with its sustainability initiatives -- from energy efficiency to purchasing green power to shippin
The company is touting its energy-efficiency and e-waste achievements as top-line results from the report:
Dell has improved the performance-per-watt of its servers by 3100 percent within the last five years. Its desktops and la
Its e-waste collection efforts grew by 16 percent over last year, with nearly 68 million kilograms (just about 150 millic
The company has also surpassed its goal to reduce the amount of packaging used by 10 percent, and is three-quarters of th
The companys reported energy use holds a few interesting items, as well. It should be noted that Dells net revenue jumpec
And along with that big boost in sales, Dells emissions also grew: Direct Scope emissions were up nearly 10 percent, to 3
Information Extract


127.0.0.1/xampp/dellInfo.php
Information Extraction
Green Energy:
35% electricity wind
Information Retrieval Score: 0.44
Display Standards
Go Back

APPENDIX F: STANDARDS (SOURCE: EPA)

Search Term	Standards (Source: EPA)
Green Energy	Using green power helps reduce the environmental impacts of electricity use and supports the development of new renewable generation capacity. Organizations can meet EPA Partnership requirements using any combination of three different product options: (1) Renewable Energy Certificates, (2) On-site generation, and (3)Utility green power products. Among the green power resources are wind, solar and small-hydro.
Recyclable/Green Packaging	The U.S. federal government primarily has taken an advisory role in supporting the recycling and reuse of nonhazardous wastes. Recycling and composting prevented 85.1 million tons of material away from being disposed of 2010. This prevented the release of approximately 186 million metric tons of carbon dioxide equivalent into the air in 2010.
EICC Practice	In 2004, a group of multinational electronics manufacturers launched the EICC to help incorporate common sustainability tenets across their industry. In 2009, members of the EICC developed a common platform for suppliers to report their GHG emissions data. Through the EICC Carbon Reporting System, suppliers common to multiple customers enter their data only once and specify which of their customers are permitted to access the information.
ISO 14001	The major requirements of an EMS under ISO 14001 include a policy which includes commitments to prevention of pollution, continual improvement of the EMS leading to improvements in overall environmental performance, and compliance with all applicable statutory and regulatory

	requirements.
Carbon Footprint	The Environmental Protection Agency (EPA) proposed a Carbon Pollution Standard. This step under the Clean Air Act would set national limits on the amount of carbon pollution power plants can emit. EPA's proposal would ensure that this progress toward a cleaner, safer and more modern power sector continues. Power plants are the largest individual sources of carbon pollution in the United States.
Green IT/ Green innovation/ Green Technology	Green information technology, or Green IT, brings environmental thinking to the world of information technology. EPA is undertaking a number of initiatives to enable end users to green their IT use, from desktop to data center. These initiatives will provide valuable tools to data center operators as they improve the energy efficiency of their facilities.
Greenhouse Gas	The EPA in 2009 found that by causing or contributing to climate change, green house gases endanger both the public health and the public welfare of current and future generations. The proposed require new fossil fuel-fired EGUs (electric utility generating units) greater than 25 megawatt electric (MWe) to meet an output-based standard of 1,000 pounds, based on the performance of widely used natural gas combined cycle technology.
Green Supply Chain	The Green Suppliers Network is a component of the EPA Initiative. Green Suppliers Network works with large manufacturers to engage their small and medium-sized suppliers in low-cost technical reviews that focus on process improvement and waste minimization. EPA provides program support. The United States Environmental Protection

	Agency (EPA) has written a guide called the "The Lean and Green Supply Chain: A Practical Guide for Materials Managers and Supply Chain Managers to Reduce Costs and Improve Environmental Performance." This is an outstanding guide that provides a systematic approach to implementing a Green Supply Chain. It's a four step decision making process. The first step is to identify environmental costs within your process or facility. The next step is to determine opportunities which would yield significant cost savings and reduce environmental impact. The third step is to calculate the benefits of proposed alternatives. The last step is to decide, implement and monitor the improvement solutions.
Restricted Chemicals/ Hazardous Substance	EPA is concerned about phthalates because of their toxicity and the evidence of human and environmental exposure to these chemicals. Phthalates are used in many industrial and consumer products, many of which pose potentially high exposure. Phthalates have been detected in food and also measured in humans. Phthalates are high production volume chemicals used primarily as plasticizers in polyvinylchloride (PVC) products. A number of brominated and chlorinated organic chemical compounds are used as flame retardants (BFRs and CFRs).The bound chemicals are not released from products, but residual, unreacted flame retardant present in the product can be released and lead to human exposure. These have been now banned and can be substituted by alternatives. As an industrialized nation, the United States produces, transports, stores, uses, and disposes of millions of tons of hazardous substances per day. Hazardous substances

	are found in many consumer products and services that we use every day. Hazardous substances take many forms, and are defined under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA). EPA's Emergency Response program provides quick response to the release, or threatened release, of hazardous substances wherever and whenever they occur.
Electronic waste	According to the Consumer Electronics Association (CEA), Americans now own approximately 24 electronic products per household. Donating used electronics for reuse extends the lives of valuable products. Recycling electronics prevents valuable materials from going into the waste stream. Consumers now have many options to recycle or donate for reuse their used electronics. Many computer, TV, and cell phone manufacturers, as well as electronics retailers offer some kind of take back program or sponsor recycling events. About half of the states currently have laws on disposal and recycling of electronics. EPA encourages all electronics recyclers to become certified and all customers to choose certified recyclers.
Packaging waste	Greening the Government through Waste Prevention, Recycling, and Federal Acquisition, requires all federal agencies to promote cost-effective waste prevention in all of its facilities. EPA National Programs/Projects: • Pay-As-You-Throw is a system where households pay for garbage collection by the amount of trash collected rather than a flat fee. Households save money by throwing away less garbage and

	recycling. • Waste Wise is a voluntary EPA program that helps businesses, government agencies, and nonprofit organizations save money by reducing their garbage and recycling. • Climate Change and Waste is an EPA program that encourages people to reduce waste in order to help slow global warming.

APPENDIX G: OUTPUT OF FPMIES

(FOR THE GREEN SUPPLY CHAIN DATA SET)

	Outcome of FPMIES Algorithm		
Search Term	**IBM**	**HP**	**Dell**
Green Energy	-	-	35% electricity wind
Recyclable Packaging	-	-	products 0 waste
Green Packaging	-	-	Material bamboo notebook
EICC Practice	Code standards member 2004	Code conduct 2004	Standards 2004 industry code
ISO 14001	Products environment standard	Certificate environment standard	products environment certification
Hazardous Substance	-	-	PVC and Brominated retardants
Carbon Footprint	transportation carbon 12.6%	tools calculator	E6400 laptop 350kg CO2
Green IT	Environment and Cost savings	-	-
Greenhouse Gas	carbon dioxide and perfluorocompounds 6 billion tones	Air emission score 100	-
Green innovation	$86 million data center	-	-
Green Technology	lab water arsenic	-	-
Green Supply Chain	nationwide logistics network	-	-
Restricted Chemicals	polyvinylchloride and retardants products	Material PVC	-
Electronic waste	-	programs and operations	-
Packaging waste	-	Solid waste plastic content	-

VITA

Shilpa Balan

Assistant Professor of Management Information Systems Flagler College St. Augustine, Florida 32085	Aug 2014 Onwards
Assistant Professor of Computer Information Systems University of Dubuque Dubuque, Iowa 52001	Aug 2013- July 2014

Revised Date: April, 2014

Education

Ph.D. Business Administration. University of Mississippi, 2014
Major: Management Information Systems
Dissertation Title: A Study on Data Informatics: Data Analysis and Knowledge Discovery via a Novel Data Mining Algorithm

Masters. Management Information Systems. Boise State University, 2006

Masters. Computer Applications. University of Mumbai, India, 2005

B.Sc. Computer Science (Bachelor of Science in Computer Science). University of Mumbai, India, 2002

Journal Publications

Balan,S., Aiken,M., Hazarika,B.(2011, October). Exploring the Feasibility of Mobile Multilingual Electronic Meetings. *Issues in Information Systems*. pp. 16-22.

Aiken,M., Balan,S. (2011). An Analysis of Google Translate Accuracy. *Translation Journal*. http://translationjournal.net/journal/56google.htm

Aiken,M., Park,M., Balan,S.(2010). A Prototype System for Machine Interpretation. *Translation Journal*. http://translationjournal.net/journal/53mi.htm

Aiken,M., Balan,S., Vanjani,M., Garner,B.(2010). The Effect of Comment Errors in Multilingual Meetings. *Communications of the IIMA*. pp. 49-60

Aiken,M., Simmons,L., Balan, S. (2010). Automatic Interpretation of English Speech. *Issues in Information Systems*. pp. 129-133

Conference Presentations

Balan, S. (2013, November). Sentiment Analysis of Academic Plagiarism Detection Tools. ISECON Proceedings

Balan,S., Ammeter,T.(2013, April). Impact of Social Networks on Healthcare Innovation. GSC Research. University of Mississippi.

Balan,S., Conlon,S.(2012, November). Comparing Green Supply Chain Practices in Healthcare. DSI Proceedings.

Balan,S., Conlon,S.(2012, November). Content Analysis of E-book Reviews. DSI Proceedings.

Balan,S., Conlon,S.(2012, August). Evaluating Best Practices in Green Supply Chain. AMCIS Proceedings.

Conlon,S., Strong,J., Balan,S., Sheikh,M.(2012). Automatic Extraction of Information on Terrorist Incidents from the Internet. DSI Proceedings.
Conlon,S., Balan,S., Hazarika, B.(2011, October). Extracting Online Information about the Impact of Natural Disasters on Supply Chains. IACIS.

Aiken,M., Ghosh,K., Balan,S.(2010,October). Automatic Translation of Multilingual Comments in Electronic Meetings. Advances in Business Research.

Academic Experience

Flagler College
St. Augustine, FL (Aug 2014 Onwards)
Position: Assistant Professor of Management Information Systems

Responsible for teaching IS courses and conducting research.

Courses Assigned to Teach:
Data Management for Business
Management Information Systems
Introduction to Computers and Management Applications

University of Dubuque
Dubuque, IA (Aug 2013- July 2014)
Position: Assistant Professor of Computer Information Systems

Responsible for teaching CIS courses, coaching students for programming contests and conducting research.

Courses Assigned to Teach:
<u>Fall</u>
CIS 103 - Computer Applications in Business
CIS 215- Programming 1 (Java)
CIS 332- Database Systems (MySQL, Microsoft SQL Server 2008)

Spring
CIS 103 - Computer Applications in Business
CIS 209- Introduction to Programming
CIS 315- Programming- 2 (Advanced Java)
CIS 405- Project Management

University of Dubuque
Dubuque, IA (June 2013- Aug 2013)
Position: Adjunct Faculty of Computer Information Systems

Courses Assigned to Teach:

CIS 215- Programming 1 (Java)
CIS 103 - Computer Applications in Business (Online Class)

CMS
Online (May 2013- Aug 2013)
Position: MBA Exam Writer/Reviewer

Joint Project between CMS and EDMC.
Responsible to write/review test questions for MBA- Technology/ IT Management

University of Mississippi
Oxford, MS (Aug 2009-May 2013)
Position: Graduate Teaching and Research Assistant
Graduate Assistant in the School of Business, University of Mississippi from Fall 2009.

- Involved in conducting various research in Data mining, Business Intelligence, Health IS, GDSS
- Involved with student grading.
- Provided special lectures in Information Systems, including video classes

University of Mississippi
Oxford, MS (Summer 2012)
Position: Graduate Instructor
Taught students course BUS 400- "Introduction to ERP (Enterprise Resource Planning) and Business Intelligence using SAP".
Covered various Business modules-Purchasing, Pricing, Sales, HR, Accounting, etc. in ERP. Also covered Business Intelligence using BI tool (Query designer and BEX Analyzer) and Business Objects.

University of Mississippi
Oxford, MS (Aug 2010 -May 2011)
Position: Graduate Instructor
Taught course MIS 309- "Introduction to Information Systems" to undergraduate students. This course introduces students to the introductory concepts in MIS including databases, web programming, systems analysis and design, E-R diagrams, etc.

Boise State University
Boise, ID (Jan 2006- Dec 2006)
Position: Graduate Assistant
Assisted in a Website Usability Study for Boise State University's Broncoweb. Analyzed existing design, conducted user survey and prepared report with recommendations for redesign. Involved in programming using Perl and MySQL.

Boise State University
Boise, ID (Fall 2005)
Position: Math Grader
Responsible for grading Probability and Statistics homework of undergrad students.

Industry Experience

FNC
Oxford, MS (June 2012- July 2012)
Position: Database Developer (Internship)

FNC provides software that streamlines and automates the parts of the mortgage process that deal with collateral assessment.
Developed triggers, stored procedures and functions using SQL Server 2008. Involved in text pattern recognition. Involved with Business Intelligence tools such as SSIS. Worked on SSIS package with Change Data Capture and Change Tracking.

IT Symbiotics
Oxford, MS (May 2011- Aug 2011)
Position: Systems Analyst (Internship)

Web designing coding using PHP, HTML and MySQL

Micron Technology
Boise, ID (2007- 2008)
Position: Software Engineer
Micron is one of the world's leading providers of advanced semiconductor solutions. Provided IS support in the assembly and test operations of wafer/chip manufacturing in Micron. Required understanding the processes of wafer/chip manufacturing. Required communication with customers to gather requirements. Followed the SDLC methodology. Developed stored procedures using SQL Server, Performance Tuning. Web programming using Cgi, HTML, Javascript. Also used technologies Perl, XML,VB. Exposed to C#. Used tools such as Documentum, Rational Clearquest, Subversion

Cougar Mountain Software, Inc.
Boise, ID (Summer 2006)
Position: Software Programmer-Intern

Cougar Mountain software is designed to manage financial information and comply with accepted accounting practices and government regulations.

Worked on bank reconciliation for accounting software. Developed custom relational tables, SQL Queries and stored procedures, reports and forms for bank reconciliation. Required understanding the business processes of the functional users, developing and applying written specifications, proficiency in writing queries, extensive use of Crystal Reports 11.5, and analytical and problem solving skills. Required working with end users to understand their needs. Required use of SQL Server 2000/2005.

Awards and Achievements

Awarded Faculty Professional Development fund from University of Dubuque, Fall 2013

Awarded Travel Grant from University of Dubuque, Fall 2013

Awarded Certificate for valued contribution as MIS Graduate Student Council Senator by University of Mississippi for the academic year 2011-2012.

Awarded Travel Grant from Department of MIS and Graduate School Office, University of Mississippi for Fall 2010, Fall 2011, and Fall 2012.

Graduate Assistantship at University of Mississippi from Fall 2009 onwards.

Graduate/Research Assistantship at Boise State University for Spring 2006 and Fall 2006.

Gem Scholarship (Non-Resident Tuition Waiver) by Boise State University for the academic year 2005-2006

Ranked 3rd at Bombay University for MCA-SEM II in 2003

Awarded a merit certificate from the Homi Bhabha Young Scientist competition in 1992, Mumbai, India

Won a prize at the District Level for the Science Exhibition in 1991, Mumbai, India

Academic Service

- Coached a group of students at University of Dubuque for the Pearson Student Coding Contest, United States, 2013. Students were accepted for the final round of the contest.
- Coached students at University of Dubuque for the MICS Coding Contest (Midwest Instructions and Computing Symposium). Verona, WI.
- Editorial Review Board: IACIS, 2013
- Graduate Student Council Senator for MIS Dept., University of Mississippi, 2011-2012

Professional Memberships

- Education Special Interest Group (EDSIG), Association of Information Technology Professionals
- Americas Conference on Information Systems (AMCIS)
- America's SAP Users Group

Technical Skills

Programming Languages	C, C++, Java, C#, MS SQL/T-SQL, PL/SQL Visual Basic 6.0, Visual Basic.Net
Database	SQL Server 2000/2005/2008, MySQL, MS Access, Oracle 10g
Web Technologies	HTML, CSS, XML, Javascript, Perl, PHP, Dreamweaver
Business Intelligence Tools	SAP-BI, SAP-BO, SQL Server 2008-SSIS
SAP	SAP ERP- Fitter Snacker, SAP Global Bike, SAP BI (Query Designer, BEX Analyzer), SAP Business Objects
Microsoft Office	Word, Excel, PowerPoint, Access, Outlook
Statistical Tools	SPSS, SAS, Small Stata, R
Tools	Microsoft Project , Microsoft Visio, Rational Rose, Crystal Reports
Online Instructional Tools	Blackboard, Moodle
Operating Systems	Windows, Linux