Summer 2010

A pilot study in an application of text mining to learning system evaluation

Nitsawan Katerattanakul

Department:

A PILOT STUDY IN AN APPLICATION OF TEXT MINING TO LEARNING SYSTEM EVALUATION

by

NITSAWAN KATERATTANAKUL

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN INFORMATION SCIENCE AND TECHNOLOGY

2010

Approved by

Dr. Wen-Bin Yu, Advisor
Dr. Ronaldo Luna
Dr. Richard H. Hall

ABSTRACT

Text mining concerns discovering and extracting knowledge from unstructured data. It transforms textual data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between documents, and clustering documents into categories.

Given a collection of survey comments evaluating the civil engineering learning system, text mining technique is applied to discover and extract knowledge from the comments. This research focuses on the study of a systematic way to apply a software tool, SAS Enterprise Miner, to the survey data. The purpose is to categorize the comments into different groups in an attempt to identify “major” concerns from the users or students. Each group will be associated with a set of key terms. This is able to assist the evaluators of the learning system to obtain the ideas from those summarized terms without the need of going through a potentially huge amount of data.

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to my thesis advisor, Dr. Wen-Bin Yu, for giving me an opportunity to work on this research. His patience and encouragement carried me through difficult times, and his insights and suggestions helped to shape my research skills. His valuable feedback contributed greatly to this thesis.

Moreover, I am grateful to National Science Foundation (NSF) for the funding of "Introduction of GIS into Civil Engineering Curricula" project (NSF award #: 0717241 Pl: Dr. Ronaldo Luna, "Civil, Architectural and Environmental Engineering Department"), which provided me financial support throughout the research.

In addition, I would like to thank my thesis committee members, Dr. Ronaldo Luna of the Civil, Architectural and Environmental Engineering Department, and Dr. Richard H. Hall of the Business and Information Technology Department for their kindly assistance and comments until the research was completed. This success would have been impossible without their support and encouragement. Also, I would like to thank Dr. Bih-Ru Lea of the Business and Information Technology Department for her constructive comments and suggestions to this research.

Furthermore, thank everyone who completed the surveys assisting me with the evaluation of the study. Besides, I value highly for all lab teaching assistants, Aparna Sukhavasi, Seth P. Lamble, and Protyush Banerjee, who collected the survey data for this research project.

Finally, special thanks to my parents, sister, and friends for their love and support. I especially appreciate their assistance in correcting my thesis writing, and in preparing the thesis defense. I owe all my success to them.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1. CIVIL ENGINEERING LEARNING SYSTEM

The development of the civil engineering learning system (Luna, 2007) was funded by the National Science Foundation (NSF) to introduce Geographic Information System (GIS) to undergraduate students enrolling in a typical civil engineering program. The system consists of five modules in the areas of environmental, geotechnical, surveying, transportation, and water resources, since these topics are standard topics in civil engineering programs nationwide. This civil engineering learning system enables faculty to bring practical applications to the classroom in an effort to enhance traditional instruction.

1.2. SURVEY FOR LEARNING SYSTEM EVALUATION

The civil engineering learning system is improving while being used in the classrooms. The evaluation process is ongoing as part of the iterative system development life cycle for the learning system. Surveys were conducted a week after each class to evaluate effectiveness of instruction and the civil engineering learning system itself. An example of the survey is illustrated in Appendix A. Each survey contained both quantitative and qualitative questions. Quantitative questions (i.e., multiple choices, true/false questions) are relatively easy to be analyzed since statistical and scientific techniques can be applied directly to the results. However, the qualitative part including open-ended questions requires human effort to read all feedback from students in order to come up with conclusions. Text mining is a useful technique which is able to handle unstructured textual data. Therefore, this research aims to apply text mining to the survey comments from students in an effort to assist the evaluation team with the analysis of qualitative data.

1.3. INTRODUCTION TO TEXT MINING

A vast amount of documents are available in the form of books, journals, newspapers, web pages, blogs, databases, and so on. A large amount of knowledge and vital information are embedded into these documents. Humans need to read and analyze substantial documents in order to gain useful insights. In a business environment, faster reactions, more information, and better decision-making are key competitive advantages to becoming successful in business. Thus, it is important to receive right information at the right time and in the right place. However, information is sometimes growing rapidly than we can absorb. In the past, there have been a lot of studies proposing approaches to assist knowledge discovery from data. The well-known concepts include information retrieval (IR), information extraction (IE), natural language processing (NLP), and data mining (DM). IR aims to gather targeted documents which match the specified query from the huge amount of documents and provide them to users. The main focus is on a search engine which searches a collection of documents with keywords. However, users have to read through each of the retrieved documents to locate the information they want. Therefore, without the need of human effort, IE plays an important role to identify and extract a range of specific types of information from texts of interest and present only relevant information. NLP allows users to construct language in a grammatical structure which helps a machine to understand language and interpret documents. Most knowledge discovery systems in NLP are based on such a concept. Data mining (DM) is a knowledge discovery method which uncovers patterns and insightful knowledge from structured data in databases.

Text mining (TM) adopted techniques from well-established scientific fields such as data mining (DM), machine learning, information retrieval (IR), natural language processing (NLP), case-based reasoning, statistics and knowledge management (Sirmakessis, 2004). TM is a variation of DM performing on unstructured data such as textual documents. However, TM have been expanded from traditional IR and IE approaches, and based on NLP. It focuses on identifying patterns and relationships in texts rather than matching and extracting key words. More discussion of these concepts is provided in the next section.

2. LITERATURE REVIEW

2.1. DEFINITION OF TEXT MINING

Text mining is considered a sub-specialty of Knowledge Discovery from Data (KDD) (Liddy, 2000). Feldman et al. (Feldman & Sanger, 2007) broadly defined text mining as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. To be more specific, the process is aimed to understand and interpret semistructured and unstructured data (Sirmakessis, 2004) in order to discover and extract knowledge from them, unlike data mining, which discovers knowledge from structured text (Ananiadou & McNaught, 2006). Text mining retrieves the hidden knowledge and presents it to users in a concise form (Ananiadou et al., 2006). It can be considered as data mining of textual data (Olson & Shi, 2007). Uramoto et al. (2004) added that the technology of text mining enables patterns and trends to be discovered semiautomatically from the huge collections of unstructured data.

Ananiadou and McNaught (2006) mentioned that text mining includes three major activities: information retrieval (IR), information extraction (IE), and data mining (DM). These activities are not explicitly included in the text mining process as steps. Text mining optionally inherits some of their techniques to deal with specific problems in each step throughout the entire process. Moreover, since text mining attempts to discover hidden knowledge in texts, it is essential that the structure and nature of language should be taken into account. Thus, natural language processing (NLP) is usually applied in text mining process. In addition, sentiment analysis is an analyzing approach which has been applied effectively in many text mining applications. These computer science concepts are considered related areas of text mining and described more in detail in the following subsections.

2.1.1. Information Retrieval (IR). This is a process of selecting documents which meet the users' requirements (Nasukawa & Nagano, 2001; Liddy, 2000) and respond to the need of information with the aid of indexes (Ananiadou et al., 2006). IR gathers relevant texts (Ananiadou & McNaught, 2006).

The traditional IR systems usually applied index terms to index and retrieve documents (Baeza-Yates & Ribeiro-Neto, 1999). The systems detect and extract

documents of interest, matching the given query (Feldman & Sanger, 2007), from the huge amount of documents (Nasukawa & Nagano, 2001) and present them to the user (Choudhary et al., 2009), but require the user to read through the documents to locate the relevant information (Feldman & Sanger, 2007). Since it requires users to specify a query to select data that they want, the technology is limited when users do not have clear intention of what they need (Nasukawa & Nagano, 2001).

With word matching approach, the problem is sometimes oversimplified since a lot of semantics in documents or user queries may be lost when the user's intention is replaced with a set of words (Baeza-Yates & Ribeiro-Neto, 1999). Also, matching queries with the index terms, representing documents, may cause the retrieved documents irrelevant to the requests. Therefore, one major concern in IR is to distinguish between relevant and irrelevant documents. Ranking algorithms were applied to assist such decision. The algorithms establish a simple ordering of the retrieved documents, according to the document relevance. IR models have been designed, based on these characteristics, to serve IR purposes. There are three classic models in IR which are Boolean, vector-space, and probabilistic models.

- Boolean Model: It is a simple model based on set theory and Boolean algebra. The requests are represented as Boolean expressions carrying precise semantics. Its advantages are simplicity and the clean formalism behind the model while the major disadvantage is exact matching which may lead to too many or too few retrieved documents.
- Vector-Space Model: This algorithm assigns non-binary weights to index terms. The weights are used to calculate the degree of similarity to partially consider documents which match the queries. Therefore, the resulted ranking is more precise than the Boolean model. The term-weighting scheme improved retrieval performance. Moreover, partial matching allows approximately matched documents are retrieved. Also, the documents are ranked based on their degree of similarity to the requests. However, term dependencies reflecting discrimination of a term to the others in the document might hurt the overall performance.

- Probabilistic Model: This model is also known as the Binary Independence Retrieval (BIR) model. It addresses the IR problem within a probabilistic framework. The idea is to iteratively construct a probabilistic description (i.e., index terms with high probability to carry the semantics of the query) of the ideal answer set. Thus, the documents are ranked based on probability of being relevance; however, the algorithm has some disadvantages. First, the method does not take frequency of the index term appearing in the documents into account. Also, it requires initial guess for separation of relevant and irrelevant documents.

2.1.2. Information Extraction (IE). The goal of the process is to identify and extract a range of specific types of information from texts of interest (Ananiadou & McNaught, 2006) without requiring users to read the text (Ananiadou et al., 2006) as opposed to IR which reading is required to locate the information in the document (Feldman & Sanger, 2007). The process focuses on extracting concepts of the entire document rather than extracting the set of tags and keywords alone (Nasukawa & Nagano, 2001). IE application presents only the relevant information of interest (Choudhary et al., 2009) in "machine understandable" form rather than "machine readable" form (Feldman & Sanger, 2007). Basic types of elements which can be extracted from text are entities, attributes, facts, and events (Feldman & Sanger, 2007). Entities are basic building blocks (e.g., people, companies, locations, etc.) founded in text documents. Attributes refer to features of the extracted entities, for instance, title and age. Facts define relations existing between entities. Events are activities which entities participate in.

2.1.2.1 Tasks in information extraction (IE). Feldman and Sanger (2007) described several tasks involved in IE. First of all, name entity recognition (NE or NER) attempts to identify proper names and quantities in the text. In addition, template element tasks (TEs) separate domain-independent from domain-dependent aspects. It identifies only entities, not their relationships. Then, template relationship tasks (TRs) defines a domain-independent relationship between entities. Moreover, scenario templates (STs) explain domain and task-specific entities and relations to test portability to new domains.

Furthermore, coreference tasks (COs) record information which is symmetrical and transitive, especially identity relation (i.e., it marks nouns, noun phrases, and pronouns.)

2.1.2.2 Information extraction (IE) process. According to Feldman and Sanger (2007), IE can be illustrated in Figure 2.1. Most tasks in the process are NLP tasks.

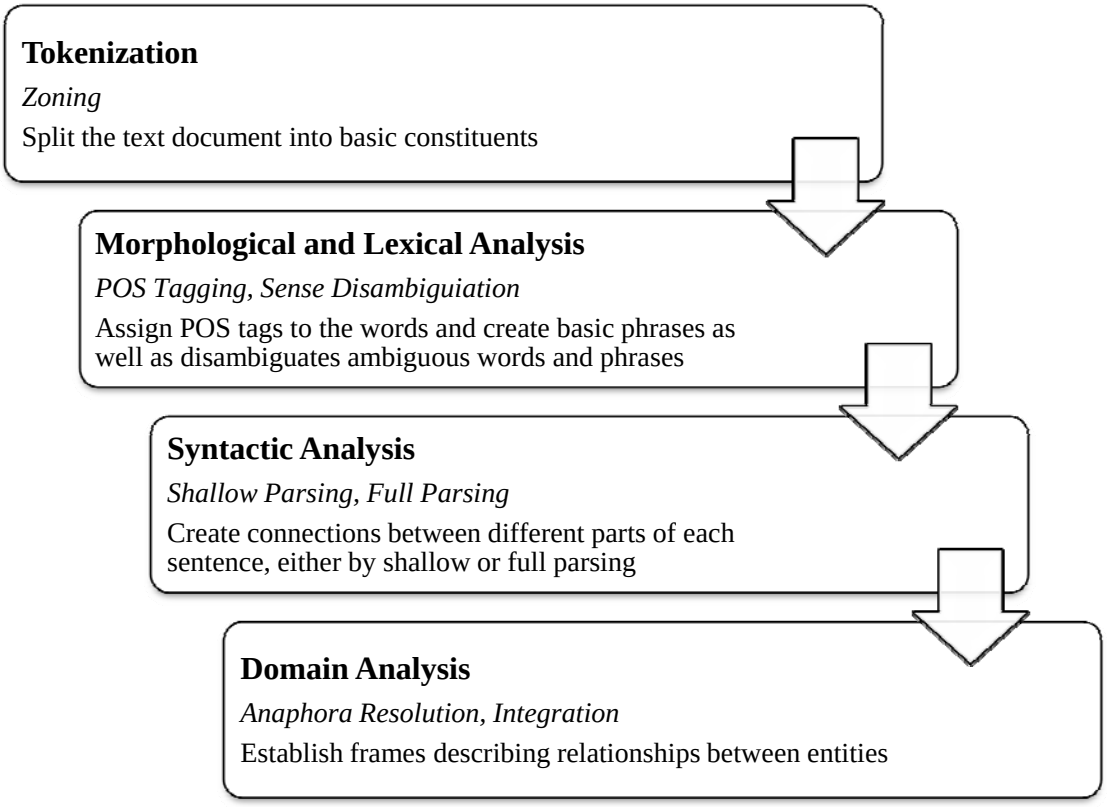

Figure 2.1. Typical Process of Information Extraction (IE)

IE systems mostly employ a bottom-up approach, starting from identifying low-level elements and then defining higher level elements based on the lower ones, for document analysis. For the first two stages, the main concerns are breaking the text into small tokens and tagging each of them by its part of speech. After lexical analysis, syntactic analysis is performed. The syntactic analysis involves some concepts in NLP; it is described more in detail in the NLP subsection. This phase starts from identifying proper names or entity extraction. Entities or keywords extraction are simply performed

based on dictionary lookup (Uramoto et al., 2004). It identifies various entity types (e.g., date, time, location, etc.) with the use of regular expressions using the context around the names, part-of-speech (POS) tags, syntactic features, and orthographic features. After basic entities are defined, the next task is shallow syntactic parsing and identifying nouns and verb groups. Feldman and Sanger (2007) and Sirmakessis (2004) described two common types of parsing: full syntactic parsing and shallow syntactic parsing. Full syntactic parsing performs a full syntactical analysis of sentences based on the certain theory of grammatical rules. As opposed to full parsing, shallow parsing sacrifices depth of analysis to increase speed and robustness since it parses only unambiguous parts, leaving the unclear ones unresolved. Shallow parsing is sufficient and much more preferable for IE purposes due to its speed. Noun and verb groups are constructed based on common patterns developed manually (Feldman & Sanger, 2007). Finally, the last phase, domain analysis, covers relation extraction and inferencing missing values to supplement the meanings. Relations between entities are built using domain-specific patterns which depend on the depth of the language analysis at the sentence level. Coreference or anaphora resolution is one of the main tasks in relation extraction. It matches pairs of NLP expressions referring to the same entity. Two main approaches to anaphora resolution are a knowledge-base, based on linguistic analysis, and machine learning, based on annotated corpus.

2.1.3. Data Mining (DM). Data mining is broadly defined as one of KDD's subtasks (Liddy, 2000). It describes knowledge discovery which applies statistical, mathematical, artificial intelligence, and machine learning techniques to extract insightful information and knowledge from large databases (Turban et al., 2008). According to Ananiadou and McNaought (2006), DM finds associations among the pieces of information extracted from many different texts. Liddy (2000) referred it more narrowly to a step of applying algorithms to detect the hidden patterns which are found among formalized records. These patterns can be rules, trends, correlations, affinities, or prediction models (Turban et al., 2008).

Feldman and Sanger (2007) compared and contrasted data mining with text mining. Data mining assumes that data is already in a structured format; thus, data mining processing focuses on scrubbing and normalizing data and creating extensive numbers of

table joins. By contrast, processing of text mining dominates on the identification and extraction of key terms and transformation of them into more structured format. Regardless of the process in detail, both text mining and data mining are similar in high-level architecture since text mining adopted much inspiration and direction as well as core knowledge discovery operations from research in data mining. Text mining is considered the application of data mining techniques to automate knowledge discovery of unstructured text (Mooney & Nahm, 2005). Therefore, data mining techniques are applied in core-text mining processing and modeling such as classification, clustering, association rules, and decision trees. Clustering is the approach which is used in this research.

Clustering: Clustering is an unsupervised process (Wagstaff et al., 2001) which classifies unlabeled objects into meaningful groups called clusters, without any prior information or pre-defined categories (Feldman & Sanger, 2007). The labels associated with the groups of objects are obtained from the data.

Clustering determines the features which better describe objects in the set, intra-cluster similarity, while distinguish objects in the set from the collection, inter-cluster dissimilarity (Baeza-Yates & Ribeiro-Neto, 1999). Intra-cluster similarity measures a raw frequency of a term k_i inside a document d_j, aka TF factor. Inter-cluster dissimilarity measures the inverse of the frequency of a term k_i among the documents in the collection, aka inverse document frequency or IDF factor. Term weights which were introduced in the previous subsection were derived from this theory. IDF weighting focuses on inter-cluster dissimilarity and tries to reduce the effect when the terms appearing in many documents are not useful for distinguishing documents. The product of TF and IDF (TF-IDF) was proposed as a reasonable measure which tries to balance the two effects, intra-cluster similarity and inter-cluster dissimilarity.

According to Feldman and Sanger (2007) clustering can have different characteristics. It can be flat if it produces disjoint clusters, or can be hierarchical if the resulted clusters are nested. Clustering will be hard if every object belongs to exactly one cluster, whereas it will be soft when each object may belong to more than one cluster and have a fractional degree of membership in each cluster. There are three common types of clustering algorithms which are agglomerative, divisive, and shuffling. Starting with each

item in a separate cluster, the agglomerative algorithm merges clusters until the criterion for stop is met. By contrast, the divisive algorithm starts with all items stored in one cluster, and then split it into clusters until stopping criterion is satisfied. The shuffling algorithm redistributes objects into clusters.

Turban et al. (2008) mentioned that most clustering techniques are based on a distance between pairs of the items. The distance measures similarity between every pair of the items. It can either be based on true distances or weighted averages of distances. The most commonly used clustering techniques are k-means method and Expectation-Maximization (EM) (Feldman & Sanger, 2007). Both of them are spatial clustering techniques. However, k-means is hard, flat, and shuffling while EM is soft, flat, and probabilistic. Unlike k-means, EM is scalable and allows clusters to be of arbitrary size and shape (Bradley et al., 1998). In addition, EM is suitable when data are incomplete (i.e., missing, truncated, etc.) (McLachlan & Krishnan, 1997) The following subsection gave you an overview of the k-means method. Since EM was the technique being used in this research, it was discussed in more detail in the subsection of core mining processing stage of the text mining process.

K-Means Method: The K-means method is commonly used to partition a set of data into k groups automatically (Wagstaff et al., 2001); where the set of data is represented by a set of vectors. The algorithm starts with selecting k initial cluster seeds (i.e., centers) which can be externally supplied or randomly picked up among the vectors (Feldman & Sanger, 2007). Then, it iteratively refines the centers as the following steps.

- Each vector is assigned to the cluster of the closest center, or seed
- Each cluster center is calculated to be the mean of its current members.
- If no change occurs in cluster assignments, stop the process; otherwise, repeat the process again

The K-means algorithm maximizes quality of the cluster when the center maximizes the sum of similarities (i.e., inverse of a distance function) to all the vectors in the cluster. It does not derive statistical models of the data as well as does not allow clusters to be overlapped (Bradley et al., 1998). Due to its simplicity and efficiency, this method became popular. Nevertheless, if the set of initial seeds is bad, the resulted clusters are often much below the optimal standard. Given unknown number of clusters,

the best number can be computed by running the algorithm with different values of k and selecting the best one (Feldman & Sanger, 2007).

2.1.4. Natural Language Processing (NLP). Ananiadou et al. (2006) defined NLP as "the activity of processing natural language texts by computer to access their meaning."

2.1.4.1 Levels of language processing. There are different levels of language processing: phonological, morphological, lexical, syntactic, semantic, and pragmatic levels (Salton & McGill, 1983).

- Phonological Level: This level does not immediately involve in information retrieval of texts, but deals with understanding and recognizing speech sounds.
- Morphological Level: The main concerns of this level are processing individual word forms and recognizable portions of words. Extensions of this knowledge include stemming as well as prefixes and suffixes recognition and removal.
- Lexical Level: This level involves in operations on full words. The procedures cover common word deletion, dictionary processing of individual words, replacement of words by thesaurus classes, and identifying a set of linguistic features (i.e., noun, adjective, preposition, verb, etc.) for each word.
- Syntactic Level: The level attempts to obtain structural description of a sentence. Thus, it groups words into structural units, such as noun phrases as well as the representation of grammatical structure as subject-verb-object groupings, based on the syntactic features and the structure in which the words are embedded.
- Semantic Level: This level assigns the meaning of the text. It applies contextual knowledge to restructure a text into units which represent its actual meaning.
- Pragmatic Level: The pragmatic level incorporates additional information (i.e., social environment) to define relationships between entities.

For example, given a sentence "John is rowing a boat," the morphological level stems "is rowing" to "row". The lexical level identifies "John" and "boat" are nouns, "row" is a verb, and "a" is an article. The syntactic level recognizes the "John-row-boat"

grouping as a subject-verb-object group which represents grammatical structure of the sentence. The semantic level interprets the sentence by designating “John” as a complementary, differently from the original role which is a subject. Then, the new sentence is constructed as “A boat is rowed by John.” In pragmatic level, additional knowledge is included in the analysis and informs that a thing cannot perform an action on a person; thus, "A boat row John” is impossible.

2.1.4.2 Language processing and information retrieval (IR). There is difference between language processing and information retrieval. Language processing attempts to convey the exact meaning of the text, whereas information retrieval aims to retrieve a particular document (Salton & McGill, 1983). Nevertheless, it is reasonable that retrieving a particular item of a certain subject requires all available related facts from the analysis of meaning of language understanding.

For information retrieval purpose, the syntactic, semantic, and pragmatic levels of language processing are mainly involved. However, syntactic process is the greatest interest in information retrieval; it is covered in more detail in the following subsection.

2.1.4.2.1 Structure of language processing in information retrieval (IR). According to Salton and McGill (Salton & McGill, 1983), the language processing system which is useful for information retrieval consists of three components described below.

- Standardized, Formal Input: The text input must be represented in a standardized, formal representation, based on the meanings and dictionary.
- Stored Knowledge Base: The input is compared with the knowledge base to add descriptions and define additional relationships between entities.
- Task Performance: The required task is performed based on the combination of the input and the knowledge base.

2.1.4.2.2 Syntactic analysis systems. Syntax is an adjunct to language processing and beneficial to retrieval systems for sentence generation and sentence analysis. There are three important kinds of the syntactic analysis systems which are the phrase structure grammars, the transformational grammars, and the transition network grammars (Salton & McGill, 1983).

2.1.4.2.3 Phrase structure grammars. Phrase structure grammars are simple grammars which account for generating and analyzing sentences. They are used to model basic structural properties of the language elements. A rule (1) is an example of the grammar for sentence generation

$$\begin{aligned} S &\rightarrow NP + VP \\ NP &\rightarrow T + N \\ VP &\rightarrow V + NP \end{aligned} \qquad (1)$$

where S, NP, VP, T, N, V are variables, respectively, for "sentence," "noun phrase," "verb phrase," "article," "noun," and "verb. The derivation tree for a sentence "the man rows a boat" exhibits in Figure 2.2. The first line means the variable S can be rewritten as NP followed by VP. The second and third lines are translated in the same way.

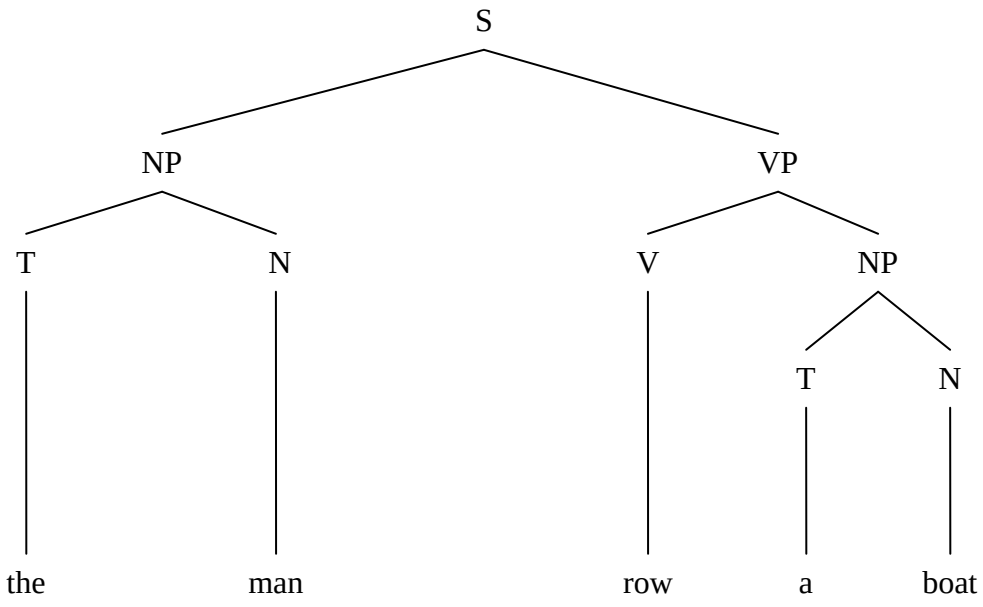

Figure 2.2. Phrase Structure Sentence Generation

Nevertheless, there are problems with this kind of grammars. Discontinuous constituents (e.g., "sign up") and the subject-verb agreement cannot be handled

conveniently by the phrase structure grammars. For example, “the man signed up me” does not fall into the defined rule and it is not recognized the same meaning as “the man signed me up.” More rules have to be added to deal with such specific cases. Therefore, this led to the development of transformational grammars. The analysis part is also simple and straightforward. It rewrites rules until no nonterminal symbol remains.

2.1.4.2.4 Transformational grammars. Transformational grammars syntactically distinct sentences which are semantically equivalent. For instance, they introduce the use of context-sensitive rewrite rules (2)

$$w\ A\ x \rightarrow w\ B\ x \qquad (2)$$

where A is a nonterminal variable and B is a string of terminal or nonterminal characters. The rule means that when the variable A appears in the context w and x, A can be replaced by the string B. From the previous example, one more rule (3) can be added to fulfill the structure.

$$\text{signed} + \text{up} + \text{NP} \rightarrow \text{signed} + \text{NP} + \text{up} \qquad (3)$$

Subject-verb agreement can be handled in the similar way as well. The language analysis consists of two parts: the base component which generates the deep structure representing syntactic and semantic interpretation, and the transformational component which generates the surface structure reflecting the phonetic representation. The sentence analysis process is the reversal of the sentence generation process.

2.1.4.2.5 Transition network grammars. Transition network grammars are mostly used in modern automatic language processing systems. People believe that ATN grammars are simpler to handle operations than the other syntactic analysis process. The transition network grammars are so-called augmented transition network (ATN) grammars. Facilities in transformational grammars are inherited to ATN grammars, but more simple and practical. Most systems are based on a finite state machine. A simple network which is able to apply the sentence “the man rows a boat” is illustrated as an example in Figure 2.3.

The ATN grammars actually generate a phrase structure tree along the processing from the first stage to the terminal stage. Each time a word is recognized, a node is created. Given the same sentence, “the man rows a boat,” and the transition network grammar in Figure 2.3, the final phrase structure tree from this process results in the same

tree as the one from the phrase structure grammar process in Figure 2.2. More detailed discussions and examples of the syntactic analysis systems can be researched from Salton and McGill (Salton & McGill, 1983).

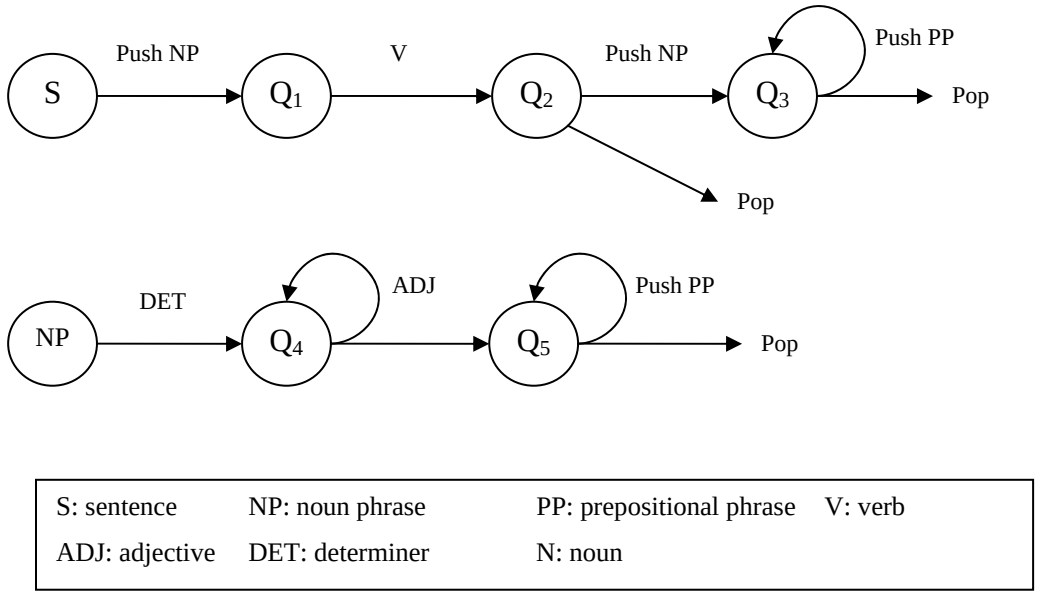

Figure 2.3. A Simple Transition Network Grammar

Given the same sentence, “the man rows a boat,” and the transition network grammar in Figure 2.3, the final phrase structure tree from this process results in the same tree as the one from the phrase structure grammar process in Figure 2.2. More detailed discussions and examples of the syntactic analysis systems can be researched from Salton and McGill (Salton & McGill, 1983).

2.1.4.3 NLP systems. According to Ananiadou et al. (2006), “NLP systems can analyze (parser) natural language using lexical resources (dictionaries), where words have been organized into groups after a grammar (syntactic level) and a semantic layer has assigned meaning to these words or groups of words.” Since the output from this technique is not specific for any particular problem, it is typically employed for domain-independent problem. Most of tasks in pre-processing stage of text mining are based on

NLP such as raw text cleaning, part-of-speech tagging, and stemming. NLP approaches are deductive in nature since they require robust vocabularies and ontologies to support the core mining process (Tremblay et al., 2009).

2.1.5. Sentiment Analysis. Sentiment analysis attempts to identify viewpoints underlying natural language texts (Pang & Lee, 2004). For instance, “thumbs up” and “thumbs down,” attached with movie reviews or videos, reflecting people’s sentiments towards the items. The major concern in sentiment analysis is to determine how sentiments are expressed in texts and what indications from the expressions are (i.e., positive or negative opinions) (Yu et al., 2007). It depends on ability to identify the sentimental terms in the documents (Godbole et al., 2007). Since opinions are usually expressed in complicated ways, they may not be addressed easily by simple text categorization approaches such as keyword identification (Mullen & Collier, 2004). Recognizing semantic of words and phrases is challenging since the textual constituents sometimes do not reflect the actual sentiments. Negative opinions possibly contain many positive words while they convey a strongly negative sentiment, and vice versa.

Sentiment analysis can be implemented by using various techniques such as text mining (Bartlett & Albright, 2008) (Yu et al., 2007), natural language processing (NLP) (Nasukawa & Yi, 2003), or machine learning (Pang & Lee, 2004). Also, areas of its application are varied (Mullen & Collier, 2004), for example, newsgroups and customer trend and feedback tracking for customer relationship management (CRM) applications.

2.2. TEXT MINING PROCESS

Text mining involves techniques from several area, including information retrieval, information extraction, data mining (Choudhary et al., 2009; Ananiadou et al., 2006), natural language processing, and machine learning (Choudhary et al., 2009; Uramoto et al., 2004). Different research papers proposed different processes and used different terminology for text mining process. The same term was sometimes referred to different processes and could be confusing. For instance, Liddy (2000) defined “Text Processing” as a stage where data mining algorithm was applied while Ananiadou et al. (2006) included it in information extraction stage which took place before data mining

was applied. Choudhary et al. (2009) even defined the entire text mining process as Knowledge Discovery in Text (KDT) and referred “text mining” as the last part of KDT when algorithms and tools were applied to the extracted information.

Nevertheless, common text mining process can be broadly described as two main stages: pre-processing and core mining processing. They are the most critical activities for any text mining system (Feldman & Sanger, 2007). After these main stages, some included an analysis stage as well as a post-processing stage. Text analysis is the process to evaluate results whether the knowledge was discovered as well as to estimate its importance (Liddy, 2000). The post-processing is the phase where refinement techniques are used to filter redundant information and cluster closely related data (Feldman & Sanger, 2007). However, these stages are not commonly conducted and are found in rare cases. Figure 2.4 illustrates the entire process of text mining and the following subsections explained activities in each stage more in detail.

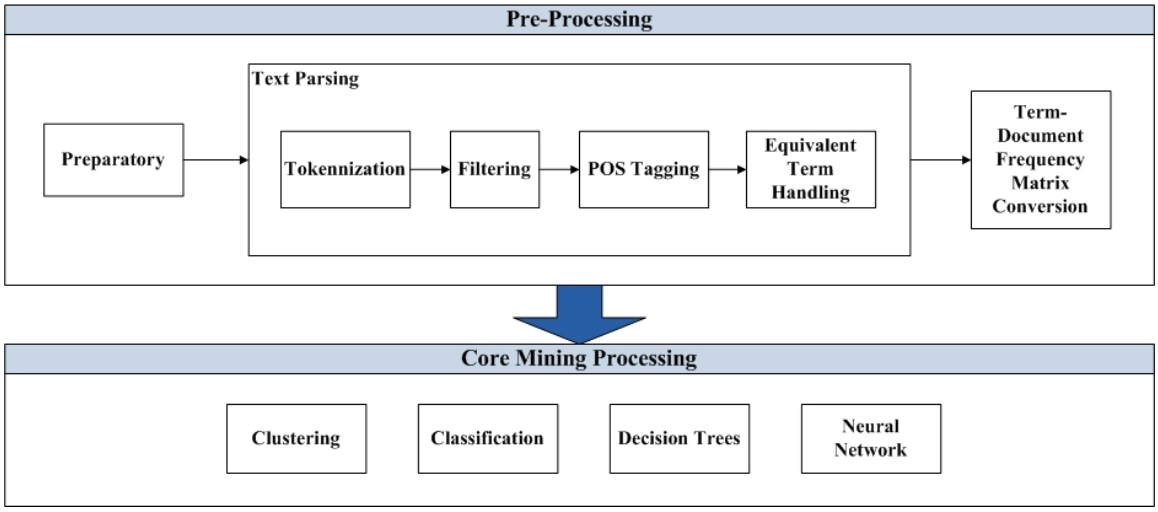

Figure 2.4. Text Mining Process

2.2.1. Pre-Processing. Pre-processing stage (Coussement, 2008; Feldman & Sanger, 2007), text preparation (Liddy, 2000), or data preparation (Becker & Wallace, 2006) retrieve the targeted documents and converts the original texts into suitable format

for the subsequent mining process. It includes all routines required to prepare data for the core mining process (Feldman & Sanger, 2007). Varieties of pre-processing techniques are existed in attempt to structure documents (Feldman & Sanger, 2007). Each starts with a partially structured document, then proceed to enrich structure, and finally end up with the most advanced and meaning-representing features which will be used for the core mining process (Feldman & Sanger, 2007). In order to apply mathematical and scientific methods to analyze the set of parsed terms effectively, the terms have to be converted into a quantitative format. Therefore, the goal of the pre-processing stage is to create a quantitative representation for the documents. Vector-space approach is the conventional model which is applied in this stage of text mining. The final representation of the collection of documents will be in a term-document frequency matrix. This representation is referred to as the "bag-of-words" approach (Sirmakessis, 2004).

Consequently, pre-processing activities can be broadly divided into three stages which are preparatory processing, text parsing, and term-document frequency matrix conversion. Natural Language Processing (NLP) is the main focus in this phase: thus, its activities involve techniques from NLP, along with statistical and machine learning (Sirmakessis, 2004). The detail of each activity is described in the following subsections.

2.2.1.1 Preparatory processing. Preparatory processing converts the raw representation of a document such as PDF file, scanned pages, and speech into text stream for further processing (Feldman & Sanger, 2007). Various techniques are existed such as optical character recognition (OCR), speech recognition, conversion of electronic files, and perceptual grouping (Feldman & Sanger, 2007). Tseng et al. (Tseng et al., 2007) proposed methods which can be considered preparatory processing to prepare text for patent analysis. The tasks were included in the first step which was called text summarization. They separated subsections by using a regular expression matcher. Then, since most patents do follow the standard set of title for subsections, it is able to use Perl expressions to match the name of segments to identify each segment. Preparatory processing is rarely discussed in research areas of text mining since it is more likely beyond the scope.

In some cases, this stage may be used to identify the documents of interest from the document universe. IR techniques may be adopted to help identifying those target

documents. Tools for IR include general-purpose search engines such as Google (Ananiadou et al., 2006). Besides, there are many tools designed specifically for a particular application. For example, Textpresso, Query Chem, iHOP, EBIMed, and PubMed are used in biological, biomedical, and chemical areas (Ananiadou et al., 2006).

2.2.1.2 Text parsing. The objectives of parsing are to break documents into smaller, syntactic chunks as well as assign them syntactic structure (Sirmakessis, 2004). Concepts and activities of text parsing are adopted from information extraction (IE) process. Recalling parsing in IE, shallow parsing performs well enough; it is widely used in many applications, including text mining. Moreover, NLP can be applied here to assist concept extraction (Nasukawa & Nagano, 2001). Text parsing activities can be grouped into four major categories being described below.

2.2.1.2.1 <u>Tokenization.</u> This step breaks the continuous stream of characters in a document into meaningful constituents (Feldman & Sanger, 2007). In text mining, the task mostly involves breaking text into words and sentences although a document can be broken up into chapters, sections, paragraphs, sentences, words, and syllables (Feldman & Sanger, 2007). White space characters can be used to separate the words or sentences (Coussement, 2008). A period is used to signal the end of the sentence and identify the sentence boundary (Feldman & Sanger, 2007).

2.2.1.2.2 <u>Filtering.</u> This step separates special characters (i.e., alpha numeric text and numerals) and punctuations (i.e., commas, apostrophes, exclamation marks) from the original texts (Coussement, 2008) since they are “unwanted texts” which do not help text differentiation (Choudhary et al., 2009).

Not only special characters, removing irrelevant terms can increase the efficiency of the text mining process. Rare words which are useless to further analysis should be left out (Coussement, 2008). Coussement (2008) eliminated words appearing in the document less than three times.

Also, removing common words frequently occurring from the text is able to reduce redundancy and improve the accuracy of the results from text mining process since it reduce size of the document and avoid information to be overloaded (Choudhary et al., 2009). The words are referred to words which are non-informative parts of speech as well as high frequency words which do not contain essential information within the

text (Choudhary et al., 2009). These non-informative words are included in a “stop list”. However, it is important to be careful not to remove the relevant words.

In some rare cases, it is possible to use a “start list” to control words to be included in the analysis. This allows text miners to examine just only a certain set of words of interest and ignore the undesirable ones.

2.2.1.2.3 Part-of-speech (POS) tagging. Feldman et al. (2007) defined POS tagging as “the annotation of words with the appropriate POS tags based on the context in which they appear”. Each POS tag provides the semantic content of each word, for instance, prepositions represent relationships among things (Feldman & Sanger, 2007). Words can be either an informative or non-informative part of speech (Coussement, 2008). Non-informative parts include determiners, conjunctions, auxiliaries, preposition, pronouns, negative articles or possessive markers, interjections, proper nouns, abbreviations, and numbers (Coussement, 2008). Informative parts are nouns, verbs, adjectives, and adverbs (Coussement, 2008). The most common set of POS tags contains article, noun, verb, adjective, preposition, number, and proper noun (Feldman & Sanger, 2007). They belong to the informative parts of speech which are meaningful and useful for knowledge discovery.

2.2.1.2.4 Equivalent term handling. Terms which mean the same thing should be treated in the same way. There are various types of situation when these terms appear in the documents. Several techniques can be applied to reduce the ambiguity based on the use of NLP and ontology. The subsections below are tasks which are commonly performed in text mining.

2.2.1.2.4.1 Stemming. Stemming performs morphological analysis of words (Choudhary et al., 2009) (Feldman & Sanger, 2007) to conflate word variations into a single, simplified representative form which is called a stem (Coussement, 2008). For instance, “write”, “writes”, “wrote”, “written”, “writing” can be stemmed to “write”. It helps reduce document size and represent the document more concisely (Choudhary et al., 2009). Moreover, stemming significantly increases retrieval performance as well as reduces the corpus dictionary (Coussement, 2008). Dictionary-based stemmer may be used to perform the task (Coussement, 2008). The stemmer compares all morphological variations with a reference dictionary and applies standard decision rules to suggest the

correct stem when a term is not recognized (Coussement, 2008). In addition, an algorithm proposed by Porter (1980) can be use in stemming. The algorithm is aggressive in removing words' suffix and made the stemmed words hard to read; thus Tseng et al. (2007) modified the algorithm to remove only simple plurals and general suffixes.

2.2.1.2.4.2 Synonym and entity recognizing. Providing relationships among words that have similar meaning during the pre-process is able to aid the rationalization of their relationships during the subsequent mining process (Becker & Wallace, 2006). A "synonym list" should be constructed to store the relationships of these equivalent words. Not only synonyms, specific names, jargons and compound nouns are also included in the list to relate to their simplified term. For example, "gas" and "Freon" mean the same thing (Coussement, 2008). There are also spelling variations such as "color" and "colour". Moreover, multiword terms and compound noun can be written in different ways, for instance, "thyroid hormone receptor" is sometimes shortened to "thyroid receptor". Same as synonyms, these words can be defined in the synonym list to handle ambiguity. Tseng et al. (2007) designed an algorithm for a document concentrating on a topic which was likely to mention a set of strings a number of times. They found that a correct combination of words was a longest repeated string. Besides, language aspects have been changing overtime, especially for jargons and slangs. In the past ten years, no one would recognize the words like "facebook", "hi5", or "iPhone". A synonym list provided an opportunity for analysts to update these special words from time to time.

IE techniques can be adopted in text mining to handle domain-specific knowledge (Feldman & Sanger, 2007). Some domain-specific documents such as biological, chemical, biomedical, and medical documents usually contain special entities (e.g., names of protein, enzyme, etc.) Ananiadou et al. (2006) explained the difficulty of entity recognition as the naming of entities is often consistent and imprecise. Variety of names may be used to denote the same concept. Orthography may be varied such as the use of hyphens and slashes as well as upper and lower cases. Some suggested that characters are all converted to lower case (Coussement, 2008). Also, abbreviations and acronyms have to be handled (e.g., "AIDS" has its canonical form as "Acquired Immune Deficiency Syndrome.") Other than the synonym list, heuristics and/or scoring rules, machine learning, and statistical methods are existing methods for acronym recognition.

2.2.1.2.4.3 Misspelling handlings. Moreover, misspelling errors are handled. One traditional way to deal with misspellings is comparing all words in the document with words in the dictionary (Coussement, 2008; Becker & Wallace, 2006). Unmatched words are misspelled ones. According to Becker and Wallace (2006), each misspelled term was compared to all correctly spelled terms. Fuzzy matching function can be used to suggest which correctly spelled word was the best match to the misspelled one. Then, analysts should review the match list and override the recommendation where it is needed. This can also be handled by the synonym list.

2.2.1.3 Term-document frequency matrix conversion. A vector-space approach is commonly employed to convert qualitative representation of documents into quantitative one since it is simple as well as has been proved that it is superior or as good as the known alternatives (Baeza-Yates & Ribeiro-Neto, 1999). Coussement (2008) described the approach as "the mean that original documents are converted into a vector in a feature space based on the weighted term frequencies. Each vector component reflects the importance of the corresponding term by giving it a weight if the term is present or zero otherwise." The final vector is represented as a term-document frequency matrix.

In the first two steps, the most informative term were selected. Thus, the current set of terms is ready to be converted. Base on the term assignment array of Salton and McGill (1983), the vector representation of documents can be represented as a term-document frequency matrix as shown in Table 2.1. Terms are rows and documents are columns. Each cell contains a frequency value of the term in the document. In the matrix, $f_{i,j}$ is the number of times that term i appears in document j. Albright (2004) described this model in detail. The model ignores the context of the documents while provides their quantitative representation.

The resulted matrix is generally sparse and will become much sparse when the size of document collection increases since few terms contain in any single document. Also, only hundreds of documents can yield thousands of terms. Huge computing time and space are required for the analysis. Therefore, reducing dimensions of the matrix can improve performance significantly. Singular Value Decomposition (SVD) is a popular

technique to deal with dimensional reduction. It projects the sparse, high-dimensional matrix into smaller dimensional space.

Table 2.1. M x N Term-Frequency Matrix Representing a Collection of Documents

Term	ID	Document											
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	…	Dn
T1	1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	$f_{1,6}$	$f_{1,7}$	$f_{1,8}$	$f_{1,9}$	$f_{1,10}$	…	$f_{1,n}$
T2	2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{2,5}$	$f_{2,6}$	$f_{2,7}$	$f_{2,8}$	$f_{2,9}$	$f_{2,10}$	…	$f_{2,n}$
T3	3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{3,5}$	$f_{3,6}$	$f_{3,7}$	$f_{3,8}$	$f_{3,9}$	$f_{3,10}$	…	$f_{3,n}$
…	…	…	…	…	…	…	…	…	…	…	…	…	…
Tm	M	$f_{m,1}$	$f_{m,2}$	$f_{m,3}$	$f_{m,4}$	$f_{m,5}$	$f_{m,6}$	$f_{m,7}$	$f_{m,8}$	$f_{m,9}$	$f_{m,10}$	…	$f_{m,n}$

In addition, another way to improve retrieval performance of the analysis is to apply weighting methods (Berry & Browne, 1999). According to Berry and Browne (1999), the performance refers to the ability to retrieve relevant information while dismiss irrelevant information. Each element of the matrix ($a_{i,j}$) can be applied weighting and represented as

$$a_{i,j} = L_{i,j}G_iD_j, \quad (4)$$

where $L_{i,j}$ is the frequency weight for term i occurring in document j, G_i is the term weight for term i in the collection, and D_j is a document normalization factor indicating whether document j is normalized. This equation was originally applied from information retrieval for search engine where longer documents have a better chance to contain terms matching the query than the shorter ones. Therefore, the document normalization factor was included to equalize the length of the document vectors from documents which vary in length (Salton & Buckley, 1988). Since this paper focused on text mining and the

lengths of the documents in the collection were not varied, the third factor was unnecessary and ignored by replacing the variable with 1. Then, the final equation is

$$a_{i,j} = L_{i,j}G_i \quad (5)$$

The element $a_{i,j}$ will replace $f_{i,j}$ in Table 1. Defining the appropriate weighting depends on characteristics of the document collection. The frequency weights and term weights are popular weighting schemes which are described in more detail in the following subsections.

2.2.1.3.1 <u>Frequency weights.</u> Frequency weight is used to adjust the frequencies in the term-by-document matrix to prevent high-frequency, commonly-occurring terms from dominating the analysis. Because unique, often rare terms can play a significant role in distinguishing between different types of documents, it is normal to try to adjust rare term frequencies with a weight factor to give them an opportunity to contribute more to the analysis.

They are functions of the term frequency ($L_{i,j}$). This factor measures the frequency of occurrence of the terms in the document by using a term frequency (TF). Common methods include binary and logarithm. Three common weighting schemes are shown below where $f_{i,j}$ represents the original frequency of term i appears in document j.

Binary: $$L_{i,j} = \begin{cases} 1 \text{ if term i is in document j} \\ 0 \text{ otherwise} \end{cases} \quad \text{or } X(f_{i,j}) \quad (6)$$

Logarithm: $$L_{i,j} = \log_2(f_{i,j} + 1) \quad (7)$$

None or Term Frequency: $$L_{i,j} = f_{i,j} \quad (8)$$

Sometimes, a term is repeated in a document for a lot of time; thus, it reflects high frequency in the document collection as a whole even though it appears in only one document. To reduce the effect from the repetitive terms, Binary and Logarithm can be

applied to the term frequency. The Binary method takes no repetitive effect into account while Logarithm reduces the effect, but still maintains it in some degree. Therefore, the Logarithm is a method in between Binary and None. Moreover, taking log of the raw term frequency reduces effects of large differences in frequencies (Dumais, 1991).

According to Berry and Browne (1999), the selection of appropriate weighting methods depends on the vocabulary or word usage patterns for the collection. The simple term frequency or none weighting term frequency is sufficient for collection containing general vocabularies (e.g., popular magazines, encyclopedias). Binary term frequency works well when the term list is short (e.g., the vocabularies are controlled).

2.2.1.3.2 <u>Term weights.</u> Term weights are statistical measures used to evaluate how important a word is to a document in a collection or corpus. They take word count in the document into account. Common methods are

Entropy: $G_i = 1 + \sum_j \frac{p_{i,j} log_2(p_{i,j})}{log_2(n)}$ (9)

GF-IDF: $G_i = (\sum_j f_{i,j}) / \sum_j X(f_{i,j})$ (10)

IDF: $G_i = \log (n / \sum_j X(f_{i,j}))$ (11)

Normal: $G_i = 1 / \sqrt{\sum_j f_{i,j}^2}$ (12)

None: $G_i = 1$ (13)

where $f_{i,j}$ represents the original frequency of term i appears in document j, n is number of documents in the collection, as well as

$p_i = f_{i,j} / \sum_j f_{i,j}$ (14)

$$X(f_{i,j}) = \begin{cases} 1 \text{ if term i is in document j} \\ 0 \text{ otherwise} \end{cases} \quad (15)$$

According to Berry and Browne (1999), the choice for an appropriate term weight depends on the state of the document collection, or how often the collection is likely to change. This weighting scheme responds to new vocabulary and affects all rows of the matrix. Thus, it is useful when updating of new vocabulary is acceptable or rare such as static collections whereas it is disregarded when updating needs to be avoided by using none weighting. All of the formulas emphasize words that occur in few documents whereas give less weight to terms appearing frequently or in many documents in the document collection. Entropy takes the distribution of terms over documents into account. Normal is the proportion of times the words occurring in the collection. These weighting methods are developed from clustering theory which will be explained more in the next subsection.

2.2.2. Core Mining Processing. The stage inherits analysis methods from data mining such as classification, decision trees, and clustering. Since the goal was to cluster comments into several clusters without pre-defined categories, this research only focuses on clustering. Clustering method being used in this research was expectation-maximization (EM).

Expectation-Maximization (EM): The Expectation-Maximization (EM) algorithm is generally a framework for estimating the parameters of distribution of variables in data (Feldman & Sanger, 2007). It is adapted to the clustering problem as a probabilistic clustering technique which is not based on distance unlike the k-means method. According to Bradley et al. (1998), EM performed superior to other alternatives for statistical modeling purposes. It attempts to group items similar to each other together. In general, data is not distributed in the same pattern; thus, some combinations of attributes are more preferable than the others. The concept of density estimation is applied to EM, in order to identify the dense regions of the probability density of the data source. The goal of EM is to identify the parameters of each of k distributions that meet the probability of the given items belonging to the cluster. Initially, parameters of k

distributions are randomly or externally selected. Then, the algorithm proceeds iteratively as described in the following steps (Feldman & Sanger, 2007).

- Expectation: Compute probability of the item belonging to the cluster by using the current parameters of the distributions, and then relabel all items accordingly to the probability.
- Maximization: Using current labels of the items, reestimate the parameters of the distributions to maximize the likelihood of the items
- If the change in log-likelihood after each iteration becomes small, stop the process; otherwise, repeat the process again

Finally, clustering results are labels of the items, generated clusters, attached with estimated distributions.

After text mining process is done, a set of clusters is generated, along with assignments of each document to clusters.

2.3. ISSUES OF TEXT MINING

Nowadays, many researches show some common issues in text mining which can be problematic and text miners have to be concerned during text mining process.

2.3.1. A Large Document Collection. To handle the vast amount of textual domain-specific data, the text-mining approach must be highly scalable and robust (Uramoto et al., 2004).

2.3.2. Noninactive Information Extraction (IE) Systems. It is difficult to apply text mining process as well as use both mining functions and trial-and-error approach on noninteractive IE systems iteratively to discover hidden knowledge (Uramoto et al., 2004).

2.3.3. Ambiguities in Natural Language. Term variant and ambiguity causes difficulty in entity identification (Ananiadou et al., 2006). The same word can express different meanings in different contexts (Nasukawa & Nagano, 2001). For example, “Lincoln” is a polysemous word which possibly means a city, person, street, school, etc. Conversely, different terms can mean similarly or even the same thing (Nasukawa & Nagano, 2001). Synonyms (Becker & Wallace, 2006), homonyms (Ananiadou et al.,

2006), acronyms (Ananiadou et al., 2006), abbreviations, as well as jargons are needed to be handled properly. For instance, “NE” represents northeast or northeastern while “search” can refer to seek or look for. Compound nouns, which are adjacent words meaning one word, also need to be considered as synonyms (Becker & Wallace, 2006). In text mining, it is necessary to treat those similar terms identically to avoid sparseness of data (Nasukawa & Nagano, 2001). Furthermore, tokenization can be challenging. To identify sentence boundaries, it is important to distinguish between a period which closes a sentence and the one which part of a previous token such as Mrs. and Sr. (Feldman & Sanger, 2007)

2.3.4. Relationships and Dependencies among Terms. Co-occurrence words may be misinterpreted in their relationships (Nasukawa & Nagano, 2001). For example, “He left when she arrived” may lead to misinterpretation of the relationships between “he” and “arrived” as well as “she” and “left”. Moreover, “Joe punched Jim” means differently from “Jim punched Joe”. Thus, the sequence of terms has to be taken into consideration (Nasukawa & Nagano, 2001).

2.4. APPLICATIONS IN TEXT MINING

Text mining is an emerging technology which can be applied to many applications. Application areas of text mining include analyzing biomedical documents, patents, financial reports, news articles, customer relations management, and medical records (Choudhary et al., 2009). Some of these applications are discussed more in detail in the following subsections since they relate to the focus of this research.

2.4.1. Patent Analysis. Patent documents contain a lot of technical and legal terminology as well as are lengthy in size. Thus, the analysis requires heavy effort and expertise. Since professional analysts are costly to find and train, the automated systems to assist the analysis become in great demand (Tseng et al., 2007). Patent analysis has wide applicability to varieties of business regardless of its narrow focus on patent-related documents; thus, its solution might be considered a “horizontal” application (Feldman & Sanger, 2007).

Patent documents are semi-structured since some parts are uniform and formatted while some are unstructured containing free texts which are various in length and contents (Tseng et al., 2007). Traditionally, patent analyses were based on structured information such as filing dates, assignees, or citations (Archibugi & Pianta, 1996). Major approaches were bibliometric methods, data mining, and database management tools such as OLAP (On-Line Analytical Processing) applications.

By focusing on simplicity, Tseng et al. (2007) proposed a text-mining methodology specialized for full-text patent analysis. Since the patent documents were semi-structured, only unstructured data, including title, abstract, claims, and description of the invention, were focused. These sections were classified and processed separately. The methodology followed the typical process of text mining. They defined a similarity function to extract term pairs relevant to the same topics as well as defined a weighting function based on term frequencies to select key terms for further analysis. Terms were then clustered based on co-occurrence and document were classified based on KNN (K-Nearest Neighbor) algorithm since it allows effectively clustering on large volume of documents, from concepts to topics, and topics to categories. Since single-step clustering often resulted in skewed document distributions among clusters, they applied multi-stage clustering method which results from the previous stage were considered as super-documents at the current stage.

2.4.2. Customer Relations Management (CRM). Text Mining can be applied to assist companies and organizations to better understand their customers and develop better relationship with the right customers. The key challenge for many companies is developing and increasing more profitable customers, for example, retaining the right customers, making them buy more or better products, and making sure that they are satisfied with the complaint handling system (Coussement, 2008). Customers' e-mails, online reviews, and call center logs are useful sources to track their opinions and perceptive towards a company's products. However, the valuable information is included in a large amount of unstructured textual contents which the company might not have much time to continuously monitor. Therefore, text mining plays an important role to assist analysts to extract knowledge from unstructured documents containing valuable information from customers. The knowledge can be classified by category or sentiment.

As examples for applications in CRM, Bartlett and Albright (2008) presented the way to implement sentiment classification using text mining to classify movie reviews on the website. In addition, Becker and Wallace (2006) combined text mining, statistical process control and a balanced scorecard to eliminate inaccurate manual warranty claim coding and reduce the time required to identify the root causes. Furthermore, Segall et al. (2009) applied text mining in hotel customer survey data and its data management.

2.4.3. News Articles Analysis. News articles are essential sources containing huge amounts of social information. Unlike typical text documents, the news reports implies special characteristics such as social interests and behavior which are continuously changing and have impact of each other. Analysis of the news collection provides understanding of the current situation as well as opportunity to forecast the future event from the rich information in the news.

Montes-y-Gómez et al. (2001) presented a text mining method to analyze news collections, including newspapers, newswires, and mass media, in order to discover interesting facts: trends, associations, and deviations. Trend analysis reflects general trends of the societal interests. For instance, inflation is a disappearing topic while interest rate is an emerging topic. Ephemeral association discovery defines influence of the peak topic on the other topics. Deviation detection discovers irregularities which differ from the typical cases. These deviations may convey interesting societal implications. Montes-y-Gómez et al. (2001) employed simple statistical representations (i.e., frequencies and probability distribution of topics) and statistical measures (i.e., the average of the median, standard deviation, and correlation coefficient) to perform the analysis. Their research could be applied broadly and did not focus on a particular subject. However, there were some analysis systems which were designed for a specific area. For example, Pham et al. (2008) demonstrated the analysis framework for mining financial news. Since financial analysis involves input from the financial domain experts, the systems were designed to allow the experts express in which aspects of the data they are interested as well as build a categorical description of news corresponding to their interests.

Not only discovering valuable knowledge which could not be captured by traditional analysis, but text mining also provides ability to forecast the future. Yu et al.

(2007) performed sentiment analysis on news to present impact of a special event on energy demand. The analysis extracted sentiments hidden in a news article and used them to compose a time-series pattern or event pattern to indicate short-term demand drivers. Sentiments were estimated from positive and negative words used in the news article. The event pattern was constructed from plotting the cumulative magnitude, the difference of positive and negative sentiments, associated with the news report according to the time interval. It presented duration and degree of the impact. Then, the pattern was transformed to demand time series adjustment to forecast energy demand fluctuation due to the special event such as earthquake and hurricane.

Besides, competitive intelligence (CI) is highly beneficial in business environment. It is a process which the companies inform themselves their rivals' activities and performance (West, 2001) through organized, structured information gathering, analysis and processing (Cook & Cook, 2000). Organizing data is a part of CI process. One way to organize a large amount of news articles is clustering the articles based on their topics. Mogotsi (2007) explored the main topics addressed in the news stories published in a daily newspaper in a particular time period by using clustering algorithms. He finally found out that agglomerative clustering performed poorly. Moreover, Shah and ElBahesh (2004) proposed a clustering system specific to media organization and public relations department. The system automatically grouped related news articles based on the content of the entire article, rather than on specific keywords or document popularity. They employed three clustering techniques which were k-nearest neighbor (KNN), single-link, and hybrid algorithms and found that the hybrid algorithm outperformed the others. Similarly, Arora and Bangalore (2005) combined classification and clustering to group related sport articles into a tight and accurate cluster. In their research, classification took out the articles which did not belong to the domain of interest and clustering formed subgroups among the classified articles using the modified version of k-nearest neighbor (KNN) and single-link clustering algorithm. In addition, real-time information and multi-language understanding are also important to the competitive environment as such. To satisfy the needs, Atkinson and Van der Goot (2009) presented a new real-time multilingual news monitoring and analysis system which extracted information of 'what' is happening to 'whom' and 'where' in the world.

2.5. EVALUATION OF THE CIVIL ENGINEERING LEARNING SYSTEM

Some past studies have been conducted to evaluate the civil engineering learning system introducing GIS to civil engineering students enrolling in a typical civil engineering program. Hall et al. (2005) evaluated the effectiveness of the learning system as well as identified factors contributing the overall effectiveness. They separated students into two groups. One group worked on the interactive learning system while the other used a board game designed for the lab. Both groups carried out the same lab activity under the same experimental condition. They completed a quiz and a post experimental questionnaire comprising open-ended questions. The analyses showed that the overall effectiveness of the learning system appeared to be good. The students using the learning system scored higher on the quiz than those in the other group. Also, the qualitative results indicated that the students found the lab was strongly related to "real world" engineering as well as motivational and engaging. Moreover, the students suggested that additional guidance and context should be provided as well as a more elaborate introduction could be added. They also indicated that additional features (i.e., options and additional components) could be included in the learning system.

Tandon et al. (2008) conducted a similar study with the same objectives as the study introduced previously. This research collected data from one set of students participating a regular lab session by using the learning system. The results corresponded to the other research. The students rated the lab session as more motivational, and more effective than the class texts. They were even more applicable to the real world learning than the class lecture and text books. Besides, quantitative analysis indicated that the prime factors contributing the effectiveness of the learning system were holistic learning, real world applicability, engagement, and, motivation. The students wanted to understand the big picture of what and why they were doing. They preferred the tasks to be challenging and practical as they could relate them to the real world. With the real data, the students felt more engaged and motivated. The students required a sufficient explanation for completing each task.

2.6. GROUNDED THEORY (GT)

In qualitative analysis, the evaluators of the system read, summarized the whole document collection, and then used the discovered insights to improve the learning system. One emerging approach for qualitative analysis is the Grounded Theory (GT).

The Grounded Theory (GT) is a qualitative research methodology which enables analysts to develop theory explaining the main concern of the population of substantive area and how the concern is resolved or processed (Scott, 2009). As opposed to a scientific method which generally begins with hypothesis or theory, the GT process starts from gathering data, analyzing data, and finally ending with generating theory from the collected data. According to Scott (2009), the stages of developing the GT are listed below.

- Identify substantive area
- Collect data and open code them as collecting where open coding is “the analytic process through which concepts are identified and their properties and dimensions are discovered in data” (Strauss & Corbin, 1998)
- Write memos throughout the entire process to define codes and their relationships
- Conduct selective coding and theoretical sampling to recognize the core category and main concern
- Sort memos and find the theoretical code
- Read the literature
- Write up theory

“Microanalysis” (Strauss & Corbin, 1998) is analysis applied at the early stages of developing the GT. According to Strauss and Corbin, microanalysis is “the detailed line-by-line analysis necessary at the beginning of a study to generate initial categories (with their properties and dimensions) and to suggest relationships among categories; a combination of open and axial coding,” where axial coding is “the process of relating categories to their subcategories” at “the level of properties and dimensions,” properties are “characteristics of a category, the delineation of which defines and gives it meaning,” and dimensions are “the range along which general properties of a category vary, giving specification to a category and variation to the theory.”

In practical, the evaluators first read comments one by one and wrote down an initial note for each one of them; where the initial note is the summarized concept of the comments. During this step, it required a lot of time and efforts to go through all comments in the collection. Then, they performed open coding which defined subcategories along with their possible properties and dimensions of each comment based on the initial note. Next, axis coding was done to form the high-level categories from the subcategories at the level of the defined properties and dimensions. Finally, a category was assigned to each comment and all categories defined up to this phase would be used to build a model for further analysis and theory creation. The GT will be used for analyzing the mental model of the students how they process their learning, which learning paths they would prefer, and what they expect from each path. Thus, based on their regular paths and expectations, the analysts can redesign the module to make it easy for them to understand and use. Figure 2.5 (Macri et al., 2002) illustrates an example of a Grounded Theory for resistance to change.

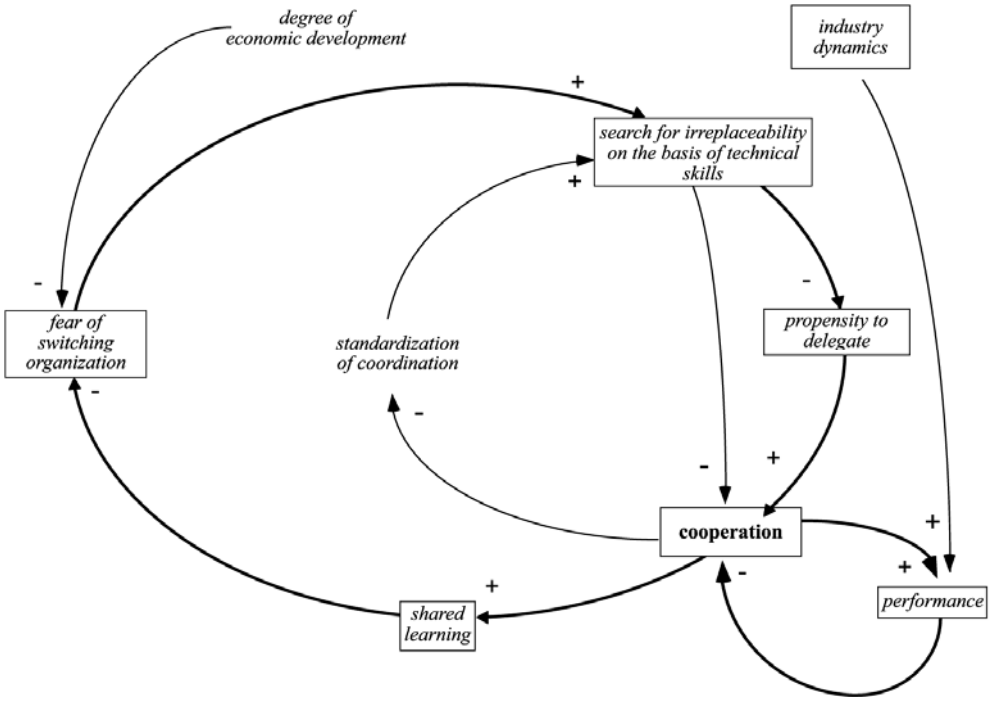

Figure 2.5. Example of the Grounded Theory (Source: Macri et al., 2002)

3. METHODOLOGY AND RESULTS

3.1. TEXT MINING PROCESS IN SAS ENTERPRISE MINER

The goal of this research is to extract knowledge from surveys conducted after students had accomplished each lab, performing on the civil engineering learning system. Feedback from open-ended questions in the surveys was taken as documents. Thus, each document of the collection to be analyzed in this research contained a few sentences. The nature of the data was not too general since the questions were based on the GIS labs. This led the answers possibly contain jargons as well as specific terms used in the learning system. On the other hand, students could freely express their opinions in the answers which could be anything and then made vocabulary uncontrolled. SAS Enterprise Miner was a tool being used throughout the research.

3.1.1. Pre-Processing. The original data was gathered electronically as texts and no data type conversion was needed. However, it was prepared into a suitable format for feeding into the text miner program during preparatory processing. A comma-separated values (CSV) file was constructed. This file type is one of the formats supported by SAS. It is used to store data in a table where rows contained a list of documents and columns represented the author showing the author ID of each text and the text itself. In the table, blank cells represented missing comments from students. The sample of the input data was shown in Table 3.1. Only the ‘Text’ column was used in the analysis.

After the input data file was prepared, the model for text mining was created as a diagram in SAS Enterprise Miner, displayed in Figure 3.1. The left node was an Input Data node where the data file was imported into and the right node was a Text Miner node where text mining process would be performed to explore information in the document collection. Both nodes were connected via a line. The direction of the arrow represented the flow of data. The input data was fed into the text mining process. SAS Enterprise Miner took care of tasks automatically based on parameter settings. In text parsing, some key parameter settings for the Text Miner node were shown in Table 3.2.

Table 3.1. Data fed into SAS Enterprise Miner

Author	Text (comments)
A	Better directions, Clear.
A1	Re-Do
A2	Go more in depth, spend more than one lab on the program.
A3	Make it shorter, you don't want much when you are bored out of your mind.
A4	Make it easier to understand
A5	Make it shorter, may be have it over a couple of days than just once. Make the learning curve more graduate.
A6	Make it shorter, make it slightly easier, make it more interesting.
A7	Introduction to the program, not just here and the link good luck.
A8	The movie's that showed each step were good and so were the test directions. I was unclear about what I was looking at on each step.
A9	Cover air pollution sources and transport better, make the GIS part much simpler so it can actually be finishe on time.
A10	Too in depth for a starting tutorial.
A11	Liked step by step guidelines. Wanted to know why I was doing the steps, what was it showing me? What is the relevance?
A12	The program needs a better explanation as to why you are inputing certain information.
A13	Explain what we were doing to us.
A14	NA
A15	Don't do this lab in this class or make it shorter. Start out slow, don't try to make us learn the whole program at once.
A16	Less technical steps, more focus on air pollution.
A17	No answer
A18	Simplify- this is our first time working with this program and it was really confusing.
A19	No answer
A20	Use a program that is easier to understand.
A21	Better organized and set up. More qualitative info. Not as long.
A22	More descriptive instructions. More description of what is being done.
A23	Weakness: No explanation of the meaning of what each step means.
A24	Don't make so complicated
A25	The video simulation could be correct. It would be helpful to see whats going on and the results we are looking for.

Figure 3.1. Model for Text Mining

Table 3.2. Parameter Settings of the Text Miner Node for Text Parsing Stage

Property	Value
Language	ENGLISH
Stop List	SASHELP.STOPLIST
Start List	
Stem Terms	Yes
Punctuation	No
Numbers	No
Different Parts of Speech	Yes

The documents were in English. No start list was being used since the comments were open-ended; thus, it was hard to define all possible key terms. Moreover, the document collection was small; it was reasonable to analyze all terms. The default stop list is used. Stemming technique was applied. Punctuation and numbers were not included in the analysis. Each term was tagged its part-of-speech (POS). Same terms having different POSs were recognized as different terms. A synonym list has been modified from the default provided by SAS Enterprise Miner. A part of the list is shown in Table 3.3.

Table 3.3. A Part of the Modified Synonym List

	TERM	PARENT	CATEGORY
1	application	software	noun
2	computer	software	noun
3	program	application	noun
4	gi	gis	noun
5	hands on	hands-on	adj
6	hands-on	interactive	adj
7	interactive	practical	adj
8	borrowsites	borrow site	noun
9	borrow sites	site	noun
10	borrow site	site	noun

Computer, program, application, and software were defined as synonyms. Moreover, the list handled misspellings by defining the correct spelling as the parent of each misspelled word. For example, the fourth and fifth rows of the list in Table 3.3 were created since students sometimes spelled ‘gis’ to ‘gi’ or forgot a hyphen in ‘hands-on’. Also, a term-document frequency matrix was derived based on the parameters set for the Text Miner node. Table 3.4 shows some of key parameter settings.

To improve performance, dimensional reduction technique was applied; thus, the “Compute SVD” was set to “Yes”. Singular Value Decomposition (SVD) is a popular approach which was also used in this research. It was computed with high resolution. The higher resolution yields more SVD dimensions, which summarizes the data set better while require more computing resources. The number of SVD dimensions should not be

too small to lose concepts and should not be too large to keep noise. Dumais (Dumais, 1991) performed information retrieval and found that performance increased over the first 100 dimensions, hitting the maximum, and then falling off slowly. Also, the higher number had been tested in this experiment, but yielded no difference in results. Thus, 100 seemed to be a good start for the maximum number of SVD dimensions. Moreover, since the vocabulary of the collection was not too general and not too controlled, it fit in between Binary and None frequency weights. Therefore, Logarithm was an appropriate frequency weight applying here. Furthermore, Entropy, GF-IDF, and None were three term weighting techniques which had been selected for this experiment.

Table 3.4. Parameter Settings for Term-Document Frequency Matrix Conversion Stage

Property	Value
Compute SVD	Yes
SVD Resolution	High
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	Log
Term Weight	Entropy

3.1.2. Core Mining Processing. Clustering technique was applied to cluster comments from students into clusters. Table 3.5 shows some key parameter settings for this process. Automatically cluster was enabled to allow clustering on the data set. The number of clusters was unknown; thus, it was not possible to define the exact number of clusters. The maximum number was set to 10 since the document collection was small and 10 clusters should be sufficient to cover all ideas. Expectation-maximization (EM) clustering technique was being used. The number of descriptive terms was set to 7. This number is reasonable for the size of data. Clustering worked on the term-frequency matrix after dimensional reduction (i.e., SVD) had been applied.

Table 3.5. Parameter Settings for Clustering of Core Mining Processing

Property	Value
Automatically Cluster	Yes
Exact or Maximum Number	Maximum
Number of Clusters	10
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Descriptive Terms	7
What to Cluster	SVD Dimensions

3.2. RESULTS

After all required parameters were set appropriately, the Text Miner node was run. Different data sets had been tested. They were responses from open-ended questions of the surveys performed on different modules in different lab sessions. In this research, survey comments from three modules, Environmental, Surveying, and Geotechnical, have been fed in the text miner. The complete results from all data sets were summarized in Appendix B. In this section, only results from one data set were discussed. The resulted clusters from text mining, with three different term weighting algorithms, were shown in Tables 3.6, 3.7, and 3.8. The survey comments were collected from 28 students who responded to the question "Please list ways in which the lab activity that covered air pollution sources and transport could be improved." Note that the plus sign (+) in front of each term means the term is the root term which were stemmed from different variation terms. For example, "+ good" can be "good," "better," or "worse."

Table 3.6. Clusters from Text Mining with Entropy Term Weighting (Log/Entropy)

#	Descriptive Terms	Freq	Percentage	RMS std.
1	+ make, + step, + short, + do, on, in	18	0.642857142…	0.2152035…
2	answer, + software, no, + good	9	0.321428571…	0.1746291…

Table 3.7. Clusters from Text Mining with GF-IDF Term Weighting (Log/GF-IDF)

#	Descriptive Terms	Freq	Percentage	RMS std.
1	+ step, + show, on	7	0.25	0.1457834…
2	+ software, + good	4	0.14285714…	0.1665495…
3	depth, in	3	0.107142857…	0.2114772…
4	answer, + do, no	5	0.178571428…	0.2013215…
5	+ short, do, better, much, + make, + not, + easy	8	0.285714285…	0.1650697…

Table 3.8. Clusters from Text Mining with IDF Term Weighting (Log/IDF)

#	Descriptive Terms	Freq	Percentage	RMS std.
1	+ make, + short, do, better, air, much, + direction	17	0.607142857…	0.1953363…
2	no, more, explanation, answer, + software, in, + do	10	0.357142857…	0.2106838…

From the results, descriptions for each cluster were constructed from its descriptive terms and shown in Table 3.9. The analysts have to use their domain knowledge in the set of documents to construct the descriptions appropriately. In addition, the constructed descriptions had been checked and corrected based on the actual comments in the collection to ensure that descriptions were relevant to the raw data. Consequently, the descriptions for each cluster in Table 3.9 should cover all ideas across each cluster.

From Table 3.9, different weighting methods yielded different number of clusters. Log/Entropy, Log/GF-IDF, and Log/IDF resulted in 2, 5, and 2 clusters, respectively. Moreover, each cluster was labeled with different sets of descriptive terms. Because the number of descriptive terms was set to 7, descriptive terms count for each cluster did not exceed the number. In this case, Log/GF-IDF generated 5 clusters, the highest number among the others, while each cluster contained the least count of descriptive terms (i.e., only 2-3 terms for cluster #1-4). Thus, it seemed that this method pointed out the most ideas as well as performed best on distinguishing an idea apart from the others due to the number of clusters and the small numbers of descriptive terms. By contrast, it would be

more difficult to group terms and construct sentences from the results of the other methods since each cluster contained more than one idea.

Table 3.9. Descriptions of Each Cluster Constructing from Descriptive Terms

Cluster No.	Descriptive Terms	Descriptions
Log / Entropy		
1	+make, +step, more, +short, +do, on, in	Make steps shorter. Need or do more on “something.”
2	answer, +software, no, +good	No answer. Need better software.
Log / GF-IDF		
1	+step, +show, on	Show steps on “something.”
2	+software, +good	Need better software.
3	depth, in	Labs/instructions are too much in depth.
4	answer, +do, no	No answer
5	+short, do, better, much, + make, not, +easy	Shorten “something.” (e.g., lab/steps/descriptions) Make “something” (e.g., lab/steps/descriptions) easier. Do “something” (e.g., lab/steps/descriptions) better.
Log / IDF		
1	+make, +short, do, better, air, much, +direction	Make better and shorter directions to do the air pollution lab.
2	No, more, explanation, answer, +software, in, +do	No answer. Need more explanation of doing something in the software/application.

Grouping terms incorrectly may lead to confusion and misunderstanding. For example, “no,” “more,” “explanation,” and “answer” could be interpreted to “no explanation and more answers”, or “no answer and more explanations.” The two

sentences conveyed different meanings and lead to different conclusion. Nevertheless, using multiple techniques and combining the results would be able to help resolve confusion. For instance, Log/GF-IDF proposed cluster #4 which carried the idea “No answer.” When “no” and “answer” were found in cluster #2 from Log/Entropy and Log/IDF, an assumption could be made that the two terms should be grouped together. Moreover, the common ideas repeating in the clustering results from all methods could be assumed as important ideas which actually presented in the collection. For example, “make” and “short” appeared in cluster #1, 5, and 1, respectively, from Log/Entropy, Log/GF-IDF, and Log/IDF, that was, both of them appeared in the results from all weighting techniques being used in this experiment. Therefore, “make ‘something’ shorter” could be confirmed as one opinion from the students.

Not only listing the resulted clusters and their descriptive terms, the program also assigned each comment to each cluster. The complete list of cluster assignments of this data set was included in the Appendix C. From the list, some comments were assigned to “.” cluster. It means that the comments were outliers which contain ideas different from majority of the whole collection. Hence, it could not be categorized into a particular cluster. This unclassified comment made the total number of frequencies falling in clusters (i.e., 27 for this example) less than the total number of all comments in the collection (i.e., 28 in this case). Text mining could only capture the repeated ideas which were commonly issued by many students. Outliers would be ignored. For instance, a few students expressed that videos were helpful, but this idea did not appear in any cluster. Moreover, some students not only commented that they needed more explanations, but also specified that they wished to see explanations of what they were doing in the program and how it pertained to air pollution. However, these clarifications were not captured by text mining process because text mining collects “major” concerns, not outliers.

The fewer number of clusters, the more opportunity that a comment would be clustered to the correct group. Using clusters from Log/IDF, “Make it easier to understand.” could be clustered reasonably to either cluster #1, or 2. Moreover, there was one comment “More descriptive instructions. More description of what is being done.” It was classified as cluster #1 from Log/Entropy which made sense, whereas was classified

as cluster #4 which did not relate to the comment at all. In this data set, Log/Entropy and Log/GF-IDF generated 67.86% while Log/IDF yielded 60.71% correctness of clustering. The correctness rate was described in more detail in section 4.

4. RESULT ANALYSIS AND CONCLUSION

4.1. COMPARISON OF RESULTS FROM DIFFERENT DATA SETS

Each module may carry jargons and terms specific to the lab instruction in that module. For example, fill site and borrow sites are the names of sites mentioned in the instruction of the Geotechnical lab. Thus, the terms "fill" and "borrow" appeared in comments from Geotechnical lab, but not in the other comments from then other modules. Evaluators needed to be aware of the jargons and did not confuse the specific term with the actual ideas hidden in the comments. Moreover, most of feedback contained "no answer" so that "no" and "answer" are supposed to be paired if they appear in the same set of descriptive terms.

Similar questions gave similar feedback. There are common ideas extracted from all modules. Students thought that the learning system was practical and helpful to complete labs. The labs were applicable to the real world, but they were too long and needed to be shortened. Besides, the labs needed more explanations and better instructions as well as needed to be easier to follow. Videos were useful, but confusing and hard to follow. Table 4.1 lists these ideas as well as shows the key terms leading to the ideas.

Table 4.1. Common Ideas Attached with the Descriptive Terms

Ideas	Module / Question	Referenced Results				
Students thought that the learning system was practical and helpful to complete labs.	Env. SP09 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	answer, + software, no, + good	9	0.32143…	0.17463…
	Env. SP09 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	No, answer, + learn, gis, + strength, lab	9	0.32143…	0.18465…
	Env. SP10 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	on, lab, helpful, + do, + software, + time, would	32	0.47761…	0.11902…

Table 4.1. (Continued) Common Ideas Attached with the Descriptive Terms

Ideas	Module / Question	Referenced Results				
	Geotech. F09 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	+ make, soil, would, + do, with, + help, + good	37	0.69811…	0.13828…
The labs were applicable to the real world, but they were too long and needed to be shortened.	Env. SP09 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	+ make, + step, more, + short, + do, on, in	18	0.64286…	0.21520…
	Env. SP09 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	World, real, real world, in, air, + do, air pollution, applicable, can, could	15	0.53571…	0.18154…
	Geotech. F09 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	class, applicable, gis, + learn, new, + software, real world	13	0.24528…	0.13647…
	Sur. SP10 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	real, in, + software, world, lab, will, + plan	29	0.74359…	0.19040…
Besides, the labs needed more explanations and better instructions as well as needed to be easier to follow.	Env. SP10 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	But, long, + continue, + do, hard, but, + not	21	0.31343…	0.12448…
	Env. SP10 / Q6	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	+ teacher, + explanation, more, + instruction, + good, little, on	28	0.41791…	0.13375…
	Geotech. F09 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	+ instruction, + open, follow, explain, zip, + easy, + file	14	0.26415…	0.14488…

Table 4.1. (Continued) Common Ideas Attached with the Descriptive Terms

Ideas	Module / Question	Referenced Results				
	Sur. SP10 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		1	in, + not, more, + software, + instruction, could, understand	27	0.69231…	0.18313…
Videos were useful, but confused and hard to follow.	Env. SP10 / Q2	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	Video, + video, hard, follow, + confuse, + do, + weakness	35	0.52239…	0.12779…
	Sur. SP10 / Q1	#	Descriptive Terms	Freq	Percentage	RMS std.
		2	Step, on, information, can, tool, + video, + help	8	0.20513…	0.18761…

4.2. COMPARISON OF RESULTS FROM TEXT MINING AND THE PRIOR EVALUATION

The past evaluation of the civil engineering learning system was discussed in section 2. There were many consistencies in the overall results from those past researches and from text mining in this study, described in the previous subsection. Both studies resulted in the same conclusions that the lab was applicable to the real world and suggested that sufficient and good explanations were important for completing the task.

The past studies were agreed that the lab was strongly related to “real world” engineering. The students rated the lab session as more applicable to the real world learning than the class lecture and text books. In addition, the studies claimed that the students required additional guidance and context as well as a more elaborate introduction. A sufficient explanation was needed for completing each task since the students wanted to understand the big picture of what and why they were doing.

Subsection 4.1 presents the common conclusion of multiple results from text mining in general. Students thought that the learning system was practical and helpful to complete labs. The labs were applicable to the real world, but they were too long and needed to be shortened. Besides, the labs needed more explanations and better instructions as well as needed to be easier to follow.

Other than these consistencies, there were a few ideas which both analyses supplemented each other with different knowledge. Text mining pointed out that the videos were useful, but confusing and hard to follow. This point was response from different types of questions applied in the past studies; thus, it did not appear in their outcome. Moreover, the past studies indicated that the learning system was motivational and engaging. This was not the raw input from students, but it was a result interpreted in higher level based on the raw opinions. Text mining straightforwardly provided basic understanding of the documents without applying any complicated algorithm to interpret data in high level; therefore, the point did not appear in the text mining outcome. Furthermore, in the past studies, they also included suggestions from the students that additional features could be included in the learning system as well as understanding the big picture of the lab by adding explanations of what and why they were doing would help their learning. These points were in-depth clarifications which were not expressed by majority of the students. Thus, text mining did not present the outliers or rare cases such these expressions since the outcome from text mining is a set of major concerns of the document collection.

4.3. EVALUATION OF EFFECTIVENESS

In order to evaluate effectiveness of the text mining process, the aspects need to be considered are the quality of the clusters and the ability to assign comments into the most appropriate clusters.

4.3.1. Quality of the Clusters. In this research, good clusters are clusters which summarize the entire document collection as effectively as the qualitative approach, as well as which are attached with understandable descriptive terms, and easy to construct sentences from the descriptive terms. Based on these characteristics, two analyses were performed.

First, since the goal of this research was to bring text mining to assist the evaluation process of the civil engineering learning system, it is critical to ensure that text mining was able to provide insights to the system evaluators as effectively as the non-text mining method, or even higher. Thus, the results from text mining and the past qualitative

studies had been compared and contrasted, as what explained in subsection 4.2. Text mining could capture only main concerns issuing by majority of the students while the qualitative method was able to provide outliers or rare cases since it required human efforts to read the entire comment collection line-by-line. Nevertheless, the main ideas which text mining extracted from the data were consistent with the evaluation from the past studies using the qualitative approach.

Second, descriptive terms, resulted from the text miner, needed to be meaningful which the analysts could easily form sentences from. In order to estimate ability to construct sentences from the resulted descriptive terms, s survey had been conducted among two groups: a non-experienced group comprising ordinary people who never experienced GIS or the learning system before, and an experienced group consisting of people who are familiar with GIS and the learning system.

The survey asked the participants to construct sentences from resulted key terms from text mining. Outcome from the Environmental module, spring'10 was used. There were four open-ended questions; hence, four different data sets of outcome. The questions were "Do you feel that the activities (steps) in the GIS lab were redundant (repeated)? If so, was redundancy helpful? Why?", "Please list the strengths and weaknesses of the web-based learning system you used for the GIS lab activity.", "Please list the strengths and weaknesses of the GIS lab in terms of its applicability to "real world" activities.", and "Please suggest ways in which the lab activity could be improved." Each data set consisted of three results from three different weighting methods, Log/Entropy, Log/GF-IDF, and Log/IDF. The same survey was sent to participants from the two groups. As examples, feedback from one data set is depicted in Table 4.2. The question is "Please list the strengths and weaknesses of the web-based learning system you used for the GIS lab activity." from the Environmental module, spring 2010.

According to the feedback, the experienced group performed better since he constructed sentences which were much more relevant to the original student comments than the other group. For instance, in fact, the students complained that the lab took too long and they needed more time to work on the lab. The experienced group got the point, but the non-experienced group interpreted it that the software is helpful to do the lab on

time. Also, the experienced group was able to know that step by step method was helpful which fit the fact; whereas, the non-experienced group misinterpreted it to the software would helpful to do the lab on time.

Table 4.2. Examples of the Sentences Constructed by the Two Groups

Weighting Method (Frequency Weight / Term Weight)	**Clustering Result from Text Mining**		**Constructed Sentences**	
	#	**Descriptive Terms**	**Experienced**	**Non-Experienced**
Log / Entropy	1	on, lab, helpful, + do, + software, + time, would	The lab was helpful, more time with the software	The software was helpful to do the lab on time
	2	video, hard, follow, + confuse, + do, but, + not	The video was hard to follow and confusing	Video was hard to follow but not confusing to do
Log / GF-IDF	1	but, long, + confuse, + do, hard, but, + not	Not helpful and was long and confusing	It was long and confusing to do but not hard
	2	on, + time, step, + strength, + do, + step, + software	Step by step method was helpful to do software	The strength of software was to do the steps on time
Log / IDF	1	on, lab, would, helpful, + do, + software, + time	Lab was helpful with the software, more time	The software would helpful to do the lab on time
	2	video, + video, hard, follow, + confuse, + do, + weakness	A weakness was that the video was hard to follow and confusing	The weakness of the video was it was hard to follow

Not only this data set, the overall result also reflected that ordinary people who did not know and access the system before found difficulty to group key terms and made some sentences which were not relevant to the actual student comments. By contrast, people who were familiar with the system (i.e., lab teaching assistants, and lab designers) performed better in forming sentences closely relating to the ideas actually presented in

the feedback collection. This was consistent with what Strauss and Corbin (Strauss & Corbin, 1998) had defined, "…Experience and knowledge are what sensitizes the researcher to significant problems and issues in the data and allows him or her to see alternative explanations and to recognize properties and dimensions of emergent concepts…"

4.3.2. Correctness of the Cluster Assignments. There is no common, well-defined technique to evaluate the cluster assignments. Different methods were applied, mostly based on statistics, in different researches. Bartlett and Albright (2008) used a misclassification rate, ranging from 0 to 1, to compare performance of different pre-processing techniques applying on sentiment analysis. Pang and Lee (2004) defined percentage of accuracy to compare results from two machine learning methods: support vector machines (SVMs) and Naive Bayes (NB), across different variables. These two researches performed sentiment analysis, classifying documents based on sentiments into two pre-defined clusters: positive and negative groups. Thus, the nature of the resulted clusters was different from this research in a certain degree. In sentiment analysis, the clusters are known before text mining process is started. By contrast, clustering has no cluster defined before text mining processing. The number of clusters as well as their labels can be varied depending on the input data. This makes it is much more difficult to define correctness of the clustering assignments since the set of the clusters are not confirmed correct.

Therefore, in this research, the assumption that the clusters resulted from text mining are best constructed to summarize the document collection was made, in order to evaluate correctness of the clustering assignments. Then, a statistical technique similar to the ones used by Bartlett and Albright (2008), and Pang and Lee (2004) was employed. This research defined a correctness rate as a quantitative variable which was used to evaluate correctness of the clustering assignments. A correctness rate is a ratio of the number of comments clustered appropriately to the total number of all comments, where a comment clustered appropriately is a comment containing at least one idea relevant to the descriptions of the cluster which the comment was assigned to. Equation (16) shows the formula for the correctness rate.

Correctness Rate = #Comments clustered appropriately / #All comments (16)

Figure 4.1 illustrated correctness rates calculated from different data sets across different modules. These data sets were feedback from different types of questions; thus, the dissimilarity of the questions affected the differences in the correctness rates. Each weighting scheme did not make much difference in term of correctness. The average ranged from 60% to 70%.

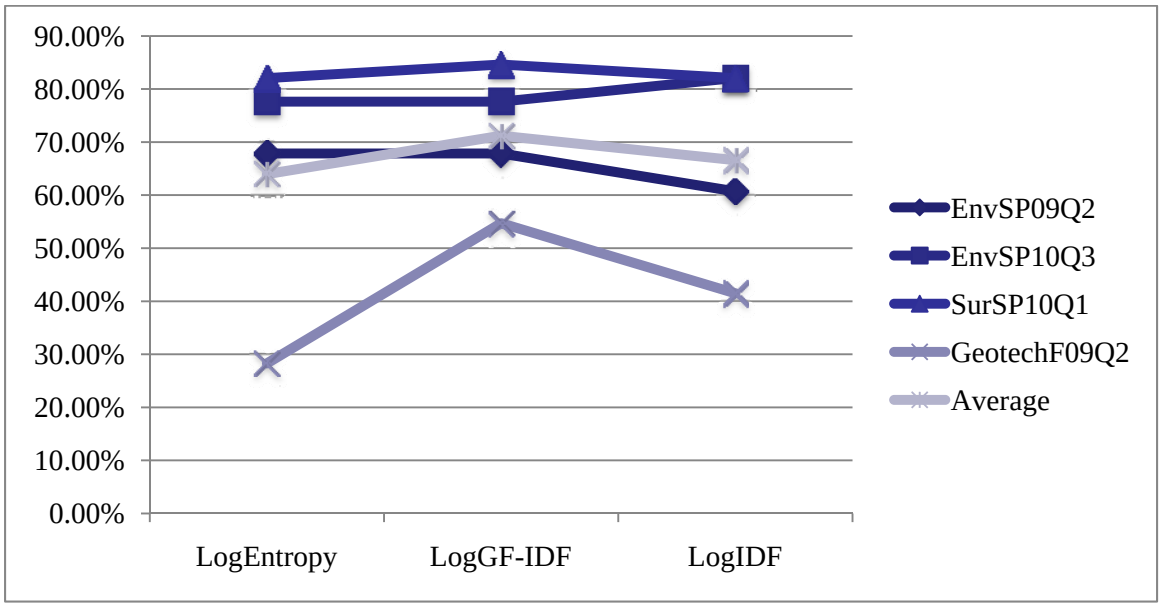

Figure 4.1. Correctness Rates of Clustering Assignments from Different Data Sets

EnvSP09Q2 asked “Please list ways in which the lab activity that covered air pollution sources and transport could be improved.” EnvSP10Q3 asked “Please list the strengths and weaknesses of the GIS lab in terms of its applicability to “real world” activities.” SurSP10Q1 asked “Please list the strengths of the lab activity that covered planning utility a route (fiber optic cable), in terms of its effect on learning and motivation, and it’s applicability to “real world” engineering.” GeotechF09Q2 asked “Please list ways in which the lab activity that covered soil borrow sites could be improved.” Note that EnvSP10Q3 and SurSP10Q1 asked questions more directed than those in EnvSP09Q2 and GeotechF09Q2; therefore, the feedback was more focused. EnvSP10Q3 and SurSP10Q1 specifically queried about strengths and weaknesses, so the answers were in the sense of strengths and weaknesses of the learning system. By

contrast, EnvSP09Q2 and GeotechF09Q2 asked for suggestions of the way in which the lab could be improved. Answers could be anything and more varied. Since text mining generally captures main concerns which are repeated a lot of times in the document collection, it performs better when texts are less in variation. Hence, EnvSP10Q3 and SurSP10Q1 yielded higher values of correctness rate due to their focused response from students.

To make it easier to understand the calculation of correctness rate, the computation of the EnvSP09Q2 data set is discussed in more detail as a sample data. For example, there are two clusters named “Make it understandable” and “Make it easier and shorter.” The comment is “Please make the steps easier.” If the comment is clustered into “Make it easier and shorter,” it will be considered clustered appropriately although the comment does not say “shorter.” On the other hand, if it is clustered into “Make it understandable,” it is not acceptable since none of the idea in the comment relates to “understandable.” The calculation of this data set is displayed in Table 4.3. The value 1 was assigned to the column “Correct?” of the comment which was appropriately clustered; otherwise, 0 was assigned. Then, the correctness rate was calculated from this column and shown at the end of the table.

Table 4.3. Correctness Rate Calculation

Name	Log/Entropy (2 clusters)		Log/GF-IDF (5 clusters)		Log/IDF (2 clusters)	
	Cluster	Correct?	Cluster	Correct?	Cluster	Correct?
A	1	1	3	1	1	1
A1	2	0	1	0	1	0
A2	1	1	3	1	2	1
A3	1	1	5	1	1	1
A4	1	1	5	1	1	1
A5	1	1	5	1	1	1
A6	1	1	5	1	1	1
A7	2	0	2	0	2	1
A8	1	1	1	1	1	0
A9	1	1	5	1	1	1
A10	1	0	3	1	1	0
A11	1	1	1	1	1	0
A12	2	1	2	1	2	1

Table 4.3. (Continued) Correctness Rate Calculation

Name	Log/Entropy (2 clusters)		Log/GF-IDF (5 clusters)		Log/IDF (2 clusters)	
	Cluster	Correct?	Cluster	Correct?	Cluster	Correct?
A12	2	1	2	1	2	1
A13	.	0	4	0	.	0
A14	2	1	1	0	1	0
A15	1	1	5	1	1	1
A16	1	1	1	1	1	1
A17	2	1	4	1	2	1
A18	2	0	2	0	2	0
A19	2	1	4	1	2	1
A20	2	1	2	1	2	0
A21	1	1	5	1	1	1
A22	1	1	4	0	2	1
A23	1	1	1	1	2	1
A24	1	0	5	1	1	0
A25	1	0	.	0	1	0
A26	2	0	1	0	1	1
A27	1	0	4	0	2	0
		67.86%		67.86%		60.71%

4.4. CONCLUSION

Text mining has been used in evaluating surveys of the civil engineering learning system and consistent outcome was obtained. The outcome from text mining addressed the similar ideas in primary stage of the qualitative analysis. This same set of knowledge has been used to improve lab instructions and the learning system itself.

In text mining, the pre-processing stage is very important and dominates the entire process. There are several techniques available to be applied in text mining. Text miners have a chance to adopt or ignore techniques based on the nature of their data sets. From the experiment in applying text mining in survey comments, text mining is able to cluster comments into clusters without pre-defined labels. Attached with each cluster is a set of descriptive terms which summarize the idea of each one. Analysts are able to read only these descriptive terms, instead of the entire documents, to obtain the ideas of the entire collection.

The three weighting methods produced different clusters with different descriptive terms. Also, the numbers of comments falling into each cluster (i.e., frequencies) may be different among results from different weighting methods, even when the methods generated the same set of clusters labeled with the same set of key terms. Notice that the total number of comments was sometimes not equal to the sums of frequencies from all clusters. It means that some comments were not able to be classified in any cluster. Those comments were ones containing, other than stop words, rare terms (i.e., outliers) which did not occur in the other comments in the collection. Also, there is a cluster carrying no descriptive term. Not only blank comments, comments containing few key terms were classified in this type of clusters.

From the results, some limitations and issues of text mining were found. Since text mining is a semiautomatic process, it requires human efforts with domain knowledge to be involved in some degree. First of all, human effort is required in order to construct sentences from the resulted key terms to interpret meanings of each cluster. For example, “easy, over, air, less, lab, make, short” may convey the ideas “Make the Air lab easier, shorter with fewer (less) steps.” Also, words are stemmed, so analysts have to guess which form should be used to construct the sentence. For instance, “easy” can be “easy” or “easier”. The student can either think that the lab is easy, or the lab should be made easier. This variation can change the meaning of the cluster. Moreover, some words which inverse the sentiment of the statement such as ‘not’, ‘never’, and ‘rarely’ are not identified which term is their pair. Thus, interpretation might be misled if the text miners or analysts are not familiar with the domain and document collection. In addition, outliers or extreme cases will be ignored. Text mining captures only important terms which represent the main focus or concerns of the document collection as a whole. Some terms which occur only in a few documents will not be included in any specific cluster. The ideas which differ from majority can hardly be captured by text mining. Sometimes, those ideas are important and might be useful since they capture issues which others fail to concern. Furthermore, it is hard to define all equivalent terms such as synonyms, jargons, and especially misspellings. Besides, the ability to recognize the equivalent terms also depends on writing. There are several ways to express the same idea. A phrase can be equivalent to a word. For instance, “make it easy” means simplify, but text mining

cannot detect that both of them are equal. If the writers wrote documents using totally different words and writing styles, text mining might consider those equivalent comments as outliers and fail to include the important idea as a cluster.

Regardless of these limitations, text mining makes it easier to summarize the ideas from student comments from a potential large collection of comments. However, to yield best performance, the analysts need to work with people who are familiar with the system such as instructors and teaching assistants of the classes to construct sentences for major ideas. Moreover, combining results from different methods (i.e. Log/Entropy, Log/IDF, etc.) will help constructing sentences and benefit analysis of the data. Also, the open-ended questions should be designed to fit the usage of text mining and reduce the effects from limitations. For instance, text mining cannot detect words inversing sentiment such as “not”; thus, a survey should not ask for strengths and weaknesses in the same question. Since both aspects are opposite in ideas, asking for them in the same question would lead to confusion, mismatch, and wrong conclusion when the analysts try to construct sentences from the resulted key terms. Due to the misinterpretation that might have, it is better to include only one query into a question. The question “do you feel that the activities (steps) in the GIS lab were redundant (repeated)? If so, was redundancy helpful? Why?” is not appropriate to analysis by text mining since it contains three queries in the question. It should be broken into three separate questions. The outcome from this preliminary study could be used to help develop appropriate questions.

4.5. FUTURE WORKS

Future works may include effective approaches to evaluate the clustering results. Also, sentiment analysis might be helpful for evaluators as well. Classifying the comments by sentiment, not by category, predicts whether each comment from students is positive or negative. This will assist system evaluators and developers to understand students’ satisfaction towards the learning system.

Furthermore, there is potential to incorporate text mining in the qualitative analysis in order to yield the most productive results. It is possible to employ text mining in microanalysis, which is the primary stage of the qualitative analysis. Text mining can be used to extract initial notes in order to reduce human efforts required to read all

comments at the beginning stage of the analysis. Then, human efforts could be put in place after that to derive high-level categories from the initial sets. In this case, the outliers could be ignored since the main categories were generally broad and were more likely to cover all rare cases. Also, the cluster assignments from text mining could be linked to the high-level categories. For each comment, it was possible to replace the initial cluster, generated by the text mining, with its relevant high-level category, constructed by humans.

In addition, multi-stage clustering can also be applied. The first clustering is for extracting initial notes. Then, the initial notes are inputs to the next clustering to obtain subcategories. Finally, the final clustering clusters subcategories into major categorie

APPENDIX A.
LAB SURVEY: OPEN-ENDED PART

Laboratory Questionnaire

Questionnaire Description:

On the questionnaire that follows, you will be asked to answer questions that will help us to evaluate and improve activities in this class. Your responses will be completely confidential. Neither your instructor, nor those scoring the data will see your name associated with your responses.

12. Please list the strengths of the lab activity that covered air pollution sources and transport, in terms of its effect on learning and motivation, and it's applicability to "real world" engineering.

13. Please list ways in which the lab activity that covered air pollution sources and transport could be improved.

APPENDIX B.
ALL RESULTS FROM TEXT MINING

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please list ways in which the lab activity that covered air pollution sources and transport could be improved. (Asked 28 students in, Spring'09 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ make, + step, + short, + do, on, in	18	0.64286…	0.21520…
		2	answer, + software, no, + good	9	0.32143…	0.17463…
	Log / GF-IDF	**#**	**Descriptive Term**	**Freq**	**Percentage**	**RMS std.**
		1	+ step, + show, on	7	0.25	0.14578…
		2	+ software, + good	4	0.14286…	0.16655…
		3	depth, in	3	0.10714…	0.21148…
		4	answer, + do, no	5	0.17857…	0.20132…
		5	+ short, do, better, much, + make, + not, + easy	8	0.28571…	0.16507…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ make, + short, do, better, air, much, + direction	17	0.60714…	0.19534…
		2	no, more, explanation, answer, + software, in, + do	10	0.35714…	0.21068…

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please list the strengths of the lab activity that covered air pollution sources and transport, in terms of its effect on learning and motivation, and it's applicability to "real world" engineering. (Asked 28 students in, Spring'09 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	no, answer, + learn, gis, + strength, lab	9	0.32143…	0.18465…
		2	+ software, + have, in, air, + step, + do, but, air pollution, applicable, can	18	0.64286…	0.20714…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	no, answer, + learn, + strength, pollution, lab, + software, + have	12	0.42857…	0.21104…
		2	world, real, real world, in, air, + do, air pollution, applicable, can, could	15	0.53571…	0.18154…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	air, gis, air pollution, applicable, can, engineering, help, may, people, with	13	0.46429…	0.19145…
		2	no, but, answer, + learn, + have, lab, + not, world, real, real world	14	0.5	0.21222…

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Do you feel that the activities (steps) in the GIS lab were redundant (repeated)? If so, was redundancy helpful? Why? (Asked 67 students in, Spring'10 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	no, n/a	12	0.17910…	0.09254…
		2	yes, helpful, redundant, but, + learn, + help, + not	25	0.37313…	0.13324…
		3	+ time, + repeat, + do, + step, + have, redundancy, no	27	0.40299…	0.12666…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	no, n/a	12	0.17910…	0.09254…
		2	first time, first, difficult, yes, + do, + software, + remember	17	0.25373…	0.12764…
		3	+ make, but, redundant, + not	4	0.05970…	0.12484…
		4	+ have, in, + feel, step, + do, + not, redundant	31	0.46269…	0.11991…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	no, n/a	12	0.17910…	0.09254…
		2	+ not, + step, helpful, yes, redundant, but, + do	51	0.76119…	0.13558…

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please list the strengths and weaknesses of the web-based learning system you used for the GIS lab activity. (Asked 67 students in, Spring'10 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	on, lab, helpful, + do, + software, + time, would	32	0.47761…	0.11902…
		2	video, hard, follow, + confuse, + do, but, + not	34	0.50746…	0.12859…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	but, long, + confuse, + do, hard, but, + not	21	0.31343…	0.12448…
		2	on, + time, step, + strength, + do, + step, + software	40	0.59701…	0.11449…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	on, lab, would, helpful, + do, + software, + time	31	0.46269…	0.11813…
		2	video, + video, hard, follow, + confuse, + do, + weakness	35	0.52239…	0.12779…

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please list the strengths and weaknesses of the GIS lab in terms of its applicability to "real world" activities. (Asked 67 students in, Spring'10 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ allow, out, + activity, population, + problem, can, see	23	0.34328…	0.09979…
		2	+ relate, real world, but, will, gis, data, real	32	0.47761…	0.13618…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ would, world, real, + not, + do, + problem, real world	32	0.47761…	0.10529…
		2	+ area, lab, can, in, pollution, see, weakness	27	0.40299…	0.13204…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ activity, + learn, can, with, + weakness, + strength, + problem	29	0.43284…	0.10829…
		2	+ relate, real world, real, but, + do, data, world	28	0.41791…	0.13327…

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please suggest ways in which the lab activity could be improved. (Asked 67 students in, Spring'10 lab on Environmental module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ student, could, + have, would, + time, + lab, + software	33	0.49254…	0.11354…
		2	+ teacher, long, + good, on, more, +instruction, little	24	0.35821…	0.13467…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ explanation, long, + good, + instruction, more, ta, little	28	0.41791…	0.13324…
		2	could, + have, + software, in, + lab, + student, before	33	0.49254…	0.10701…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ student, could, + have, + software, + lab, + video, before	29	0.43284…	0.10980…
		2	+ teacher, + explanation, more, + instruction, + good, little, on	28	0.41791…	0.13374…

<table>
<tr><th>Question</th><th>Weighting Method (Frequency Weight / Term Weight)</th><th colspan="5">Clustering Result</th></tr>
<tr><td rowspan="8">Please list the strengths of the lab activity that covered soil borrow sites, in terms of its effect on learning and motivation, and it's applicability to "real world" engineering.

(Asked 53 students in, Fall'09 lab on Geotechnical module)</td><td>Log / Entropy</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td></td><td>1</td><td>class, applicable, gis, + learn, new, + software, real world</td><td>13</td><td>0.24528…</td><td>0.13647…</td></tr>
<tr><td></td><td>2</td><td>+ make, soil, would, + do, with, + help, + good</td><td>37</td><td>0.69811…</td><td>0.13828…</td></tr>
<tr><td>Log / GF-IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td></td><td>1</td><td>+ software, world, real world, + have, real, in, lab</td><td>33</td><td>0.62264…</td><td>0.12107…</td></tr>
<tr><td></td><td>2</td><td>gis, + learn</td><td>5</td><td>0.09434…</td><td>0.10339…</td></tr>
<tr><td></td><td>3</td><td>+ layer, able, + step, soil, + cost, see, + help</td><td>10</td><td>0.18868…</td><td>0.13920…</td></tr>
<tr><td></td><td>4</td><td>+ good</td><td>3</td><td>0.05660…</td><td>0.14939…</td></tr>
<tr><td>Log / IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td></td><td>1</td><td>world, real world, + have, real, + software, lab, + do</td><td>30</td><td>0.56604…</td><td>0.13073…</td></tr>
<tr><td></td><td>2</td><td>+ layer, + cost, + project, different, soil, see, gis</td><td>21</td><td>0.39622…</td><td>0.14518…</td></tr>
</table>

Question	Weighting Method (Frequency Weight / Term Weight)	Clustering Result				
Please list ways in which the lab activity that covered soil borrow sites could be improved. (Asked 53 students in, Fall'09 lab on Geotechnical module)	Log / Entropy	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	+ make, + open, follow, as, + work, class, + good	22	0.41509…	0.12883…
		2	may, + complete, require, more, + software, + site, + do	30	0.56603…	0.14952…
	Log / GF-IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	explanation, over, lecture, more, + software, + do, could	25	0.47170…	0.12186…
		2	+ instruction, + complete, require, instead of, + not, +site, may	23	0.43396…	0.14681…
	Log / IDF	**#**	**Descriptive Terms**	**Freq**	**Percentage**	**RMS std.**
		1	more, + work, time, may, require, lab, + site	37	0.69811…	0.13741…
		2	+ instruction, + open, follow, explain, zip, + easy, + file	14	0.26415…	0.14488…

<table>
<tr><th>Question</th><th>Weighting Method (Frequency Weight / Term Weight)</th><th colspan="5">Clustering Result</th></tr>
<tr><td rowspan="6">Please list the strengths of the lab activity that covered planning utility a route (fiber optic cable), in terms of its effect on learning and motivation, and it's applicability to "real world" engineering.

(Asked 39 students in, Spring'10 lab on Surveying module)</td><td rowspan="2">Log / Entropy</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td>1</td><td>answer, no</td><td>8</td><td>0.20513…</td><td>1.47393…</td></tr>
<tr><td>2</td><td>real, in, + software, world, lab, will, + plan</td><td>30</td><td>0.76923…</td><td>0.19206…</td></tr>
<tr><td rowspan="3">Log / GF-IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td>1</td><td>world, in, no, answer, + plan, + do, will</td><td>30</td><td>0.76923…</td><td>0.18303…</td></tr>
<tr><td>2</td><td>step, on, information, can, tool, + video, + help</td><td>8</td><td>0.20513…</td><td>0.18761…</td></tr>
<tr><td rowspan="3">Log / IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td>1</td><td>answer, no</td><td>8</td><td>0.20513…</td><td>5.73612…</td></tr>
<tr><td>2</td><td>real, in, + software, world, lab, will, + plan</td><td>29</td><td>0.74359…</td><td>0.19040…</td></tr>
</table>

<table>
<tr><th>Question</th><th>Weighting Method (Frequency Weight / Term Weight)</th><th colspan="5">Clustering Result</th></tr>
<tr><td rowspan="6">Please list ways in which the lab activity that covered planning a utility route (fiber optic cable) could be improved.

(Asked 39 students in, Spring'10 lab on Environmental module)</td><td>Log / Entropy</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td></td><td>1</td><td>no, answer</td><td>9</td><td>0.23077…</td><td>0.01330…</td></tr>
<tr><td></td><td>2</td><td>lab, + have, + do, in, + not, more, + software</td><td>29</td><td>0.74359…</td><td>0.18784…</td></tr>
<tr><td>Log / GF-IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td></td><td>1</td><td>in, + not, more, + software, + instruction, could, understand</td><td>27</td><td>0.69231…</td><td>0.18313…</td></tr>
<tr><td></td><td>2</td><td>answer, no, + do, + have</td><td>11</td><td>0.28205…</td><td>0.10787…</td></tr>
<tr><td rowspan="3">Log / IDF</td><td>#</td><td>Descriptive Terms</td><td>Freq</td><td>Percentage</td><td>RMS std.</td></tr>
<tr><td>1</td><td>answer, no</td><td>8</td><td>0.20513…</td><td>0.00525…</td></tr>
<tr><td>2</td><td>lab, + have, + do, in, + not, more, + software</td><td>30</td><td>0.76923…</td><td>0.18770…</td></tr>
</table>

APPENDIX C.
CLUSTER ASSIGNMENTS

Question: Please list ways in which the lab activity that covered air pollution sources and transport could be improved.

Name	Comments	Log/Entropy (2 clusters)		Log/GF-IDF (5 clusters)		Log/IDF (2 clusters)	
		Cluster	Correct?	Cluster	Correct?	Cluster	Correct?
A	Better directions, Clear.	1	1	3	1	1	1
A1	Re-Do	2	0	1	0	1	0
A2	Go more in depth, spend more than one lab on the program.	1	1	3	1	2	1
A3	Make it shorter, you don't want much when you are bored out of your mind.	1	1	5	1	1	1
A4	Make it easier to understand	1	1	5	1	1	1
A5	Make it shorter, may be have it over a couple of days than just once. Make the learning curve more graduate.	1	1	5	1	1	1
A6	Make it shorter, make it slightly easier, make it more interesting.	1	1	5	1	1	1
A7	Introduction to the program, not just here and the link good luck.	2	0	2	0	2	1
A8	The movie's that showed each step were good and so were the test directions. I was unclear about what I was looking at on each step. We finish the whole lab, but was told one of our calc's were diff making everything wrong. Need checks.	1	1	1	1	1	0
A9	Cover air pollution sources and transport better, make the GIS part much simpler so it can actually be finishe on time.	1	1	5	1	1	1
A10	Too in depth for a starting tutorial.	1	0	3	1	1	0
A11	Liked step by step guidelines. Wanted to know why I was doing the steps, what was it showing me? What is the relevance?	1	1	1	1	1	0
A12	The program needs a better explanation as to why you are inputing certain information.	2	1	2	1	2	1
A13	Explain what we were doing to us.	.	0	4	0	.	0
A14	NA	2	1	1	0	1	0
A15	Don't do this lab in this class or make it shorter. Start out slow, don't try to make us learn the whole program at once.	1	1	5	1	1	1

Question: Please list ways in which the lab activity that covered air pollution sources and transport could be improved.

Name	Comments	Log/Entropy (2 clusters)		Log/GF-IDF (5 clusters)		Log/IDF (2 clusters)	
		Cluster	Correct?	Cluster	Correct?	Cluster	Correct?
A16	Less technical steps, more focus on air pollution.	1	1	1	1	1	1
A17	No answer	2	1	4	1	2	1
A18	Simplify- this is our first time working with this program and it was really confusing.	2	0	2	0	2	0
A19	No answer	2	1	4	1	2	1
A20	Use a program that is easier to understand.	2	1	2	1	2	0
A21	Better organized and set up. More qualitative info. Not as long.	1	1	5	1	1	1
A22	More descriptive instructions. More description of what is being done.	1	1	4	0	2	1
A23	Weakness: No explanation of the meaning of what each step means.	1	1	1	1	2	1
A24	Don’t make so complicated	1	0	5	1	1	0
A25	The video simulation could be correct. It would be helpful to see whats going on and the results we are looking for.	1	0	.	0	1	0
A26	Shorten it.	2	0	1	0	1	1
A27	Tell us what we are doing in the program and tell us how it protains to air pollution.	1	0	4	0	2	0
	% of Correctness		67.86%		67.86%		60.71%

BIBLIOGRAPHY

Albright, R. (2004). *Taming Text with the SVD.* Cary: SAS Institute Inc.

Ananiadou, S., & McNaught, J. (2006). *Text Mining for Biology and Biomedicine.* Norwood: Artech House, Inc.

Ananiadou, S., Kell, D., & Tsujii, J. (2006). Text Mining and Its Potential Applications in Systems Biology. *TRENDS in Biotechnology* , 571.

Archibugi, D., & Pianta, M. (1996). Measuring Technological Change through Patents and Innovation Surveys. *Technovation , 16* (9), 451-468.

Arora, R., & Bangalore, P. (2005). Text Mining: Classification and Clustering of Articles Related to Sports. *Proceedings of the 43rd annual Southeast regional conference* (pp. 153 - 154). Kennesaw, Georgia: ACM.

Atkinson, M., & Van der Goot, E. (2009). Near Real Time Information Mining in Multilingual News. *Proceedings of the 18th international conference on World wide web* (pp. 1153-1154). Madrid, Spain: ACM.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* New York: Addison Wesley.

Bartlett, J., & Albright, R. (2008). Coming to a Threatre Near You! Sentiment Classification Techniques Using SAS Text Miner. *SAS Global Forum.* San Antonio: SAS Institute Inc.

Becker, P., & Wallace, J. (2006). Eighty Ways to Spell Refrigerator. *SAS® Users Group International Conference.* San Francisco: SAS Institute Inc.

Berry, M. W., & Browne, M. (1999). *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* Philadelphia: Society for Industrial and Applied Mathematics.

Bradley, P. S., Fayyad, U., & Reina, C. (1998). *Scaling EM (Expectation-Maximization) Clustering to Large Databases.* Redmond, WA: Microsoft Corporation.

Choudhary, A., Oluikpe, P., Harding, J., & Carrilo, P. (2009). The Needs and Benefits of Text Mining Applications on Post-Project Reviews. *Computers in Industry* , 728-740.

Cook, M., & Cook, C. W. (2000). *Competitive intelligence: create an intelligent organization and compete to win.* Dover, NH: Kogan Page.

Coussement, K. (2008). Employing SAS Text Miner Methodology to Become a Customer Genius in Customer Churn Prediction and Complaint E-mail Management. *SAS Global Forum.* San Antonio: SAS Institute Inc.

Dai HJ, C. Y. (2009). New Challenges for Biological Text-Mining in the Next Decade. *Journal of Computer Science and Technology* .

Dumais, S. T. (1991). Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers , 23* (2), 229-236.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Glossary. (n.d.). Retrieved February 11, 2010, from pathfinder solutions: http://www.pathfindersolutions.com.au/page/glossary.html

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. *ICWSM'2007.* Boulder: International Conference on Weblogs and Social Media.

Hall, R. H., Luna, R., Hilgers, G. M., Sullivan, J. M., Lawrence, W. T., & Buechler, M. R. (2005). Evaluation of a Prototype GIS Learning System to Teach Civil Engineering Concepts. *Proceedings of the World Conference on Education Multimedia, Hypermedia, & Telecommunications (EdMedia)* (pp. 3569-3574). Montreal, Canada: AACE.

Liddy, E. (2000). Text Mining. *Bulletin of the American Society for Information Science and Technology* , 13.

Luna, R. (2007). *Learn-Civil-GIS*. Retrieved March 2010, from Learn-Civil-GIS.org: http://learn-civil-gis.org

Macri, D. M., Tagliaventi, M. R., & Bertolotti, F. (2002). A Grounded Theory for Resistance to Change in a Small Organization. *Journal of Organizational Change Management* , 292 - 310.

McLachlan, G. J., & Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: John Wiley & Sons, Inc.

Mogotsi, I. C. (2007). News Analysis through Text Mining: A Case Study. *VINE , 37* (4), 516-531.

Montes-y-Gómez, M., Gelbukh, A., & López-López, A. (2001). Mining the News: Trends, Associations, and Deviations. *COMPUTACIÓN Y SISTEMAS , 5* (1), 14--24.

Mooney, R. J., & Nahm, U. Y. (2005). Text Mining with Information Extraction. *Proceedings of the 4th International MIDP Colloquium* (pp. 141-160). Bloemfontein: Van Schaik Pub.

Mullen, T., & Collier, N. (2004). Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 412-418). Barcelona, Spain: ACL Anthology.

Nasukawa, T., & Nagano, T. (2001). Text Analysis and Knowledge Mining System. *IBM Systems Journal* , 967.

Nasukawa, T., & Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77). Sanibel Island, FL: ACM.

Olson, D., & Shi, Y. (2007). *Introduction to Business Data Mining.* New York, NY: The McGraw-Hill Companies, Inc.

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the ACL,* (pp. 271-278).

Pham, Q.-K., Saint-Paul, R., Benatallah, B., Mouaddib, N., & Raschia, G. (2008). Mine your own business, mine others' news! *Proceedings of the 11th international conference on Extending database technology: Advances in database technology* (pp. 725-729). Nantes, France: ACM.

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

Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management , 24* (5), 513-523.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Scott, H. (2009). *What is Grounded Theory?* Retrieved May 18, 2010, from Grounded Theory Online: http://www.groundedtheoryonline.com/what-is-grounded-theory

Segall, R. S., Zhang, Q., & Cao, M. (2009). Web-Based Text Mining of Hotel Customer Comments Using SAS® Text Miner and Megaputer Polyanalyst®. *SWDSI 2009* (pp. 141-152). Oklahoma City: Decision Sciences Institute.

Shah, N. A., & ElBahesh, E. M. (2004). Topic-Based Clustering of News Articles. *Proceedings of the 42nd annual Southeast regional conference* (pp. 412 - 413). Huntsville, Alabama: ACM.

Sirmakessis, S. (2004). *Text Mining and its Application: Results of the NEMIS Launch Conference.* Heidelberg: Springer-Verlag.

Strauss, A., & Corbin, J. (1998). *Basic of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* Thousand Oaks, CA: Sage Publications, Inc.

Tandon, B. B., Hall, R. H., Luna, R., Sheng, H., & Boese, M. (2008). Integration of a GIS Learning Management System into Civil Engineering Curricula: An Evaluation. *Proceedings of the AACE E-Learn Conference* (pp. 497-505). Las Vegas, Nevada: AACE.

Tremblay, M., Berndt, D., Luther, S., Foulis, P., & French, D. (2009). Identifying Fall-Related Injuries: Text Mining the Electronic Medical Record. *Information Technology and Management* , 253-265.

Tseng, Y., Lin, C., & Lin, Y. (2007). Text Mining Techniques for Patent Analysis. *Information Processing and Management* , 1216-1247.

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2008). *Business Intelligence: A Managerial Approach.* Upper Saddle River, NJ: Pearson Education, Inc.

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., & Takeda, K. (2004). A Text-Mining System for Knowledge Discovery from Biomedical Documents. *IBM Systems Journal* .

Wagstaff, K., Cardie, C., Seth, R., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 577-584). Williamstown, MA: International Conference on Machine Learning.

West, C. (2001). *Competitive Intelligence*. Basingstoke: Palgrave.

Yu, W.-B., Lea, B.-R., & Guruswamy, B. (2007). A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting. *International Journal of Electronic Business Management , 5* (3), 211-224.

VITA

Nitsawan Katerattanakul was born in Bangkok, Thailand, on July 17, 1984. She received her bachelor's degree in Computer Engineering from Chulalongkorn University, Bangkok, Thailand. After her undergraduate graduation, she worked as a programmer for Imagimax Co., Ltd., and then worked as an IT engineer at Thai Flavor and Fragrance Co., Ltd., Bangkok, Thailand, from 2006 to 2008. In 2008, she started pursuing a Master's degree in Information Science and Technology at Missouri University of Science and Technology. From August 2009 until May 2010, she served as a Graduate Research Assistant in the Business and Information Technology Department at Missouri University of Science and Technology and was a part of "Introduction of GIS into Civil Engineering Curricula" project. She completed her degree and earned a Master's degree in Information Science and Technology from Missouri University of Science and Technology.