

Date: January 7, 2006.

**Using Association-Rule Mining Techniques to
Discover Theme Patterns from Text
(for Text Summarization).**

**Student Name: Catherine Inibhunu
Course: 60-510 Fall 2005
Instructor: Dr. Richard Frost.
Supervisor: Dr. Christie Ezeife.**

TABLE OF CONTENTS

1. Introduction.....	4
2. Basic Issues.....	6
2.1 Extraction of Sentence Segments.....	6
2.2 Evaluation of Extracting Systems.....	6
2.3 Weighting and Ranking Segments.....	7
2.4 Memory Management.....	7
2.5 Lexical and Semantic Analysis.....	7
2.6 Representation of Summaries.....	8
3. Approaches.....	9
3.1 Feature Extraction.....	9
3.1.1 Cue Markers.....	9
3.1.2 Declarative Information Analysis Language.....	9
3.1.3 Parallelism.....	10
3.1.4 Converting unstructured text to structured format.....	11
Table 3.1 Papers that have addressed the Extraction Problem.....	12
3.2 Evaluation of Extraction Mechanisms.....	13
3.2.1 Challenges of Automatic Summarization.....	13
3.2.2 Empirical Study of Feature Selection Metrics.....	13
Table 3.2 Papers that have addressed the Evaluation of Extraction Algorithms...	14
3.3 Weighting and Ranking Mechanism.....	15
3.3.1 Summarization with relevance Measure.	15
3.3.2 Summarization as Feature Selection for Text Categorization.....	16
3.3.3 Sentence-Selection Heuristic.....	16
Table 3.3 Papers that have used Weighting and Ranking Mechanism.....	17
3.4 Memory Management and Computation.....	18
3.4.1 Association Rules	18
3.4.2 Association Terms with Text Categories.	18
3.4.3 Rare-but-Important Associations.....	19
Table 3.4 Papers that have addressed the Memory Management and Computation Problem.....	20
3.5 Lexical and Semantic Analysis.....	21
3.5.1 Text Segments and Text Themes.	21
3.5.2 Using Lexical Chains.	21
3.5.3 Semantic Orientation	22
3.5.4 Using Predefined Text Classes instead of WorldNet.....	23

3.5.5 Mining partial features	24
Table 3.5 Papers that have addressed the Extraction Problem.....	25
3. 6 Representation of Summaries.....	26
3.6.1 The Babylon Project	26
3.6.2 What is this Text About?	27
3.6.3 Using Shrinkage.	27
3.6.4 Discovering Technological Intelligence	27
3.6.5 Semantic Thumbnails	27
Table 3.6 Papers that have addressed the Summary Representation Problem.....	28
4. Conclusion.....	30
5. Acknowledgement	32
6. Annotation Bibliography.....	33
7. Bibliography.....	46
8. Appendix 1.....	52

List of tables and corresponding page numbers

- (a) **Table 3.1 papers that have addressed the Extraction Problem...(page 12)**
- (b) **Table 3.1 Papers that have addressed the Evaluation of
Extraction Algorithms(page 14)**
- (c) **Table 3.3 Papers that have used Weighting and Ranking
Mechanism(page17)**
- (d) **Table 3.4 Papers that addressed the Memory Management
And Computation problem(page 20)**
- (e) **Table 3.5 Papers that addressed Lexical and Semantic Analysis..(page 25)**
- (f) **Table 3.6 Papers that addressed the Summary
Representation Problem.....(page29)**
- (g) **Table 4. Papers that have used Association-Rule Mining Techniques.**

1. INTRODUCTION

The world has accepted computers as the best means for storing information. This is due to the fact that it is very easy to save data, it is convenient, any one with access to a computer can do it, and most importantly, information stored can be shared among many users, or transferred to other locations. However, as more text documents are stored in large databases, it becomes a huge challenge to understand hidden patterns or relations in the data. Since text data is not in numerical format, it cannot be analyzed with statistical methods.

Various mechanisms have been proposed for analyzing textual data. These include, clustering algorithms that classify documents into a constant number (k) of distinct clusters (Krishna et al., 2001). This becomes a problem when the text documents themselves do not fit into these k clusters. Categorization is another approach that has been used (Bekkerman and Allan, 2003) where predefined classes are given. A scan performed on source documents assigns each document to the class that best represents it. This approach fits only domain-specific environments, thus documents that do not have predefined categories are not analyzed. Probabilistic models assign various weights to different words in a document (Meir and Zhang, 2003), but some core key words with low occurrence or frequency end up getting the lowest probabilistic measure leading to poor analysis. Association rules have also been used in creating text summaries. However the algorithms used are based on the traditional Apriori-like structure that normally performs recursive scans on the entire database to get frequent items. This was proved to be slow and inefficient in (Zaiane and Antonie, 2002).

In this report, a survey has been conducted on mechanisms for understanding text. An investigation of various mechanisms that use association-rule mining (ARM) techniques for creating text summaries has been carried out.

Association-rule mining finds interesting relationships among large sets of data items stored in a given database. Let D be such a database, an association rule r , is denoted as $A \rightarrow B$ where A and B are disjoint sets in D . The confidence of r is the conditional probability $P(B|A)$ and the support of r is determined by the prior probability of A and B , $P(A \text{ and } B)$. Given a minimum support threshold and a minimum confidence threshold, ARM generates all association rules that are above the given thresholds.

A table containing the papers that have used ARM techniques is included in Appendix 1.

The rest of the survey is organized as follows: Section 2 presents the issues in text summarization. Section 3 presents approaches by various researches to text summarization. At the end of each subsection in section 3, there is a table that gives a brief note on techniques taken by various researchers is given. Section 4 has the conclusion, section 5 has the acknowledgements, section 6 presents the annotations of the 23 most important references, and section 7 presents the complete bibliography. Tables included in the report are also available in appendix 1 for easy access.

2. BASIC ISSUES

2.1 Extraction of Sentence Segments

There is a huge challenge as to how sentence segments should be extracted from text for them to yield important information about the original document. At the same time, when a segment is extracted, can it be combined with other extracted segment to form a document summary? “A summary will not be as good as an abstract”, (Chuang et. al., 2000). Since documents are not structured in a standard way, does structuring the data before extraction of features make the procedure more feasible? These are some of the important questions that are addressed by (Mooney and Buenscu, 2005).

2.2 Evaluation of Extraction Systems

Accuracy in feature selection is regarded as one criteria for measuring a text-summarization mechanism (Forman, 2003). If a user looks at extracted document segments, they should be able to infer what would be the real context of the original text document. Any system that provides such knowledge would be ideal for text summarization. However, present systems are not able to handle documents from multiple sources (Hahn and Mani, 2000).

2.3 Weighting and Ranking Segments

Redundancy is a huge problem for document summaries. Various researchers have introduced weighting measures and ranking mechanisms to avoid this problem (Gong and Liu, 2001; Kotcz et. al., 2001; McDonald and Chen, 2002).

2.4 Memory Management

In traditional databases, the Apriori approach was the foundation of many rule-generation algorithms. It was originally designed for identifying frequent patterns in structured data. Apriori requires several scans of the entire database thereby taking up much memory. This leads to much inefficiency; however improvements to the Apriori approach could lead to generation of frequent patterns in text documents (Holt and Chung, 1991; Zaiane and Antonia, 2002; Phan et. al., 2005).

2.5 Lexical and Semantic Analysis

For a document summary to represent an original document, the semantics in the summary should be a component of the original document. Terms that are closer in meaning should be grouped in the same summary. Various studies have been done on summarization using lexical analysis by (Silber and McCoy, 2000), using semantic orientation of document segments in (Turney and Littman, 2003), and discovery of rules by understanding the lexical knowledge in the document (Sakurai and Suyama, 2000).

2.6 Summary Representation

The quality of a text summarization is mostly measured by how informing it is to a potential user (Merrill, 2003). However, most summarizers are domain specific and end up not being very useful to users in different domains. For example, a summarizer for medical documents might be very useful to a researcher in the medical field. However, the same summarizer, if used in business communities, especially call centers, may not be helpful at all. Several techniques have been proposed for better representation of document summaries; this includes the Babylon project (Merrill, 2003), use of informative views by (Hernandez and Grau, 2003), and hierarchical summaries in (Kongthon, 2004) and (Iperiotis and Gravano, 2004). Finally the use of thumbnails by (Sengupta et. al., 2004) is a technique borrowed from image representation.

3. APPROACHES

3.1 Feature Extraction

3.1.1 Cue Markers

The problem of automatic text summarization was addressed in (Chuang and Yang, 2000). They proposed a method that uses cue markers in extracting segments from sentences, an adaptation of work done by (Marcu, 1996) in his Phd Dissertation. First, a sentence is identified as having different clauses. For a sentence having two clauses, the first clause is called the nucleus and the second clause the subordinate. The two clauses are joined by rhetoric relations referred to as cue markers, e.g. “but”, “because”, “if”, and “however”. The nucleus is considered important and used to create a summary. They took rhetoric relations of nearby segments instead of considering all segments generated. Features of each segment identified as important are then put in a vector that is later used to create a summary. The authors claim that experiments conducted showed that their algorithm performed reasonably well compared to the Microsoft Word Summarizer in terms of recall, precision, and accuracy of classification.

3.1.2 *Declarative Information Analysis Language*

The problem of knowledge discovery from text by extracting key words from within a text document was explored by (Yonatan et. al., 2001). The authors presented ClearStudio, a system that has modules for extracting information.

The system includes rules for defining important features to be extracted. Features include events, facts, and words with meaning given the document domain. The rules were developed with DIAL (Declarative Information Analysis Language) a language for writing rules in internet explorer developed in (Appelt et. al., 1993). The authors claim that the system was successfully used to create rules in diverse domains, e.g. analysis of patents and financial news.

3.1.3 *Parallelism*

The problem of categorizing documents was also addressed in (Castillo and Serrano, 2004). The authors begin by referring to the work of (Yang and Hanovar, 1998). They use parallelism to develop a multi-strategy classification for text documents. Different learners are explored and each carries out its own feature selection separately. Each learner receives a document, the learner first removes stop words, then stemming, frequency of word occurrence is recorded; words that appear in a title are given a higher frequency value. Outputs of each learner are then combined to form a classification model. The authors claim that the presence of different learners allowed the system to be domain independent thereby giving good classification results.

3.1.4 *Converting Unstructured Text to Structured Format*

Conversion of unstructured data to a structured format was adapted by (Mooney and Bunescu, 2005). They extended their earlier work, described in (Mooney and Califf, 2003), where regular expressions were used to simulate rules from training examples. In this paper they took a different approach; labeling words in the documents sequentially. A word is given three labels; a pre-filler (a tag before the word), a filler that matches the word to be extracted, and a post-filler which identifies a word after a filler. Based on a word's context, a word is determined to be either a key document feature or not. The authors did experiments on Biomedical journals; identifying human proteins. In order to model sequences appropriately, they compared the Hidden Markov Model (HMM) and the Conditional Random Field (CRF) model. They claim that the CRF model provided the best model for sequencing labels by capturing the dependence between labels in adjacent words.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Chuang and Yang, 2000)	Automatic Text Extraction.	Using Cue Markers as a guide to segment extraction from sentences.	(Marcu, 1996)
(Yonatan et. al., 2001)	Information Extraction from Text.	Declarative Information Analysis system that provided rules for text extraction.	(Appelt et al., 1993)
(Castillo and Serrano, 2004)	Multistrategy Classifier System.	Parallelism as a strategy for text document classification.	(Yang and Hanovar, 1998)
(Mooney and Califf, 2005)	Natural language Information Extraction.	Converting unstructured text to structured text with aid of sequential word labeling.	Extends earlier work described in (Mooney and Cliff, 2003)

Table 3.1 Papers that have addressed the Extraction Problem

3.2 Evaluation of Extraction Mechanisms

3.2.1 Challenges of Automatic Summarization

The problem of generation of summaries from online sources was addressed by (Hahn and Mani, 2000) who provide a comprehensive review document summarization techniques. They pointed out the limitations of available summarization tools as mere extraction tools; not able to create a document abstract from extracted text features. They also argued that the tools cannot handle multiple document sources and in particular non-textual data. They suggested that evaluation techniques are needed for evaluating document summarizers; a summary must be a complement of the source.

3.2.2 Empirical Study of Feature-Selection Metrics

The problem of evaluation of different feature-selection methods performed on data was also addressed by (Forman, 2003). He presented a comprehensive report on evaluation of twelve feature-selection methods performed on data from various sources e.g. Reuters was one source. An analysis was done on each technique based on its accuracy, F-Measure, precision, and recall using the WEKA system. The experiments main focus was to obtain the classification with best performance without taking into account the number of features selected to get the performance. The author noted that it was surprisingly hard to pick the best metric measure techniques could perform differently in different domains. However the author did not do any evaluation using association rules.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Hahn and Mani, 2000)	Evaluating extraction procedures.	-This is a comprehensive review of various issues that affect most of the tools available in the industry for text summarization.	-This is an independent work and did not refer to any work. -It's a good paper to understand text summarization problem.
(Forman, 2003)	Evaluation of different feature-selection method.	-Presented a comprehensive study of experiment done to 12 feature selection methods.	-They used the WEKA system from University of Waikato to perform most experiments.

Table 3.2 Papers that have addressed the Evaluation of Extraction Algorithms

3.3 Weighting and Ranking Mechanisms

3.3.1 Summarization with relevance Measure

The problem of text extraction with ranking was addressed by (Gong and Liu, 2000). They proposed two text-summarization methods that ranked sentences extracted from original documents. Some of those ranked sentences were then used to create summaries. The first method uses information-retrieval methods for measuring the relevance of a sentence. A relevance score of a sentence in a document is computed by calculating the weight of terms in the sentence. All sentences with higher scores are used to create a summary. The second method uses latent semantic analysis where a document is represented by an $m \times n$ matrix; m is the number of terms and n is the number of sentences. A term frequency is then calculated by identifying how many sentences the term occurs in. Singular vectors are then created of all the terms that have similar meanings. These singular vectors are then used to create summaries. The author's main goal was to be able to extract the highest ranked and disjoint sentences for creating a document summary. The authors claim that the results of experiments, conducted on CNN Worldview news program, using the two methods, and compared to summaries developed by three human evaluators, were quite comparable.

3.3.2 Summarization as Feature Selection for Text Categorization

The same problem was also addressed by (Kotcz et. al., 2001). They proposed an algorithm for creating document summaries using only words extracted from the original document. They referred to work described in (Mahesh, 1997). First the algorithm assigns a weighing measure for each feature extracted, then unique terms are evaluated depending on their relevance weights. The words with the highest score are then used to form the document summary. The authors claim that experiments conducted on Reuters-corpus demonstrated informative summaries when matched with original documents.

3.3.3 Sentence-Selection Heuristic

A different approach to the summary-creation problem was taken in (McDonald and Chen, 2002) where summaries were created using predefined keywords and sentence heuristics. They presented TXTRACTOR, a tool for ranking text segments that is strongly related to work described in by (Carbonell and Goldstein, 1998). The segments are later used to create document summaries. They describe three steps that are used in creating the summaries; (1) evaluation of sentence based on position, length and format of words in a sentence (capitalized or not), (2) sentence segmentation by identifying topic boundaries, (3) segment ranking using some heuristics and finally, summary creation.

Low-ranking segments are eliminated during summary creation.

They compared their approach to that of a tool created using the TextTiling algorithm and claim that their approach gave better summaries.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Gong and Liu, 2001)	Text extraction with ranking.	-Two summarization methods are proposed. -1 st method use information retrieval to measure sentence relevance. -2 nd method use latent semantic analysis.	This was an independent work and did not refer to any previous work.
(Kotcz, 2001)	Feature extraction with ranking.	-A technique that assigns weights to each feature extracted. -High scoring features are used to create summaries.	(Mahesh, 1997)
(McDonald and Chen, 2002)	Ranking text segments.	-Creation of heuristics for ranking segments using Maximal Marginal Relevance.	(Carbonell and Goldstein, 1998)

Table 3.3 Papers that have used Weighting and Ranking Mechanisms

3.4 Memory Management and Computation

3.4.1 Association Rules

The problem of mining association rules from words in text databases was addressed by (Holt and Chung, 1999). They proposed two algorithms for mining association rules between words in a text database by modifying the Apriori Algorithm in (Agrawal and Srikant, 1994) and the Direct Hashing and Pruning Algorithm (DHP) in (Park, 1997). Firstly, all items in the document are ordered in a lexical manner then the first frequent items into smaller sets. Each partition is examined separately to identify new frequent patterns. This is done to reduce the amount of memory required in counting the number of frequent items unlike the original Apriori, or the (DHP) approach, that requires much memory due to repetitive scan of the entire database in rule generation. The authors claim experiments done using their improved algorithms out-performed both the original Apriori and the DHP methods in identifying frequent patterns on large text databases and in memory usage.

3.4.2 Associating Terms with Text Categories

An extension of the Apriori algorithm in (Agrawal and Srikant, 1994) was also made by (Zaiane and Osmar, 2002) in the creation of the document-categories problem. Two algorithms are proposed. Association-Rule-based Categorizer By Category (ARC-BC) is the first algorithm, it discovers rules for one category at a time.

After discovering the frequent items, rules are then generated by making the rule the consequence of the frequent item sets in the category. The second algorithm, Association-Rule-based Categorizer for All Categories (ARC-AC) handles the entire training set as one entity. A dominance factor is introduced to enhance overlapping categories; the idea is to pick the most-dominant categories to represent a document. The authors claim that their algorithm presented an efficient training phase in addition to having rules that were understandable compared to most text classifiers.

3.4.3 *Rare-but-Important Associations*

A different approach to the mining of association rules was taken in (Phan et. al., 2005). A focus was done on discovering rare but important rules from huge collections of documents with optimal computation time. An FP-Tree technique developed in

(Han et al., 2000) was used for discovering instances that would not be discovered with common training models in aid of classifying irregular instances. They defined two integers, l_{sup} ; the lowest support permitted and u_{sup} the upper limit support. These two integers were much less than T , the total transactions being considered.

A rule that has an antecedent value of l_{sup} and a confidence greater or equal to the permitted confidence is termed as “rare-but important”. Due to the huge amount of combinations of items that are discovered in extracting these rare-but important rules, the FP- tree was used to improve computation time.

The authors claim that experiments done on English language chunking and in recognition of name entities showed accurate results. They also claim that there is a relationship between association rules and statistical learning in discovering patterns from document sources.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Holt and Chung, 1999)	Mining association rules from text.	-Two algorithms presented: (1) Improved Apriori. (2) Improved Direct Hashing and Pruning.	(Agrawal and Srikant, 1994) and (Park et. al., 1997)
(Zaiane and Osmar, 2002)	Efficient mining of association rules from text.	-Two algorithms presented: (1) Discover rules from each category at a time. (2) Discover rules from all categories at one scan.	(Agrawal and Srikant, 1994)
(Phan et, al., 2005)	Discover rare but important rules from text.	-They introduce two integers an upper and lower bound. -A rule that falls in the lower bound but with above average confidence is rare but important.	(Han et al., 2000)

Table 3.4 Papers that have addressed the Memory Management and Computation Problem

3.5 Lexical and Semantic Analysis

3.5.1 Text Segments and Text Themes

Two techniques for text segmentation were presented in (Salton et. Al., 1996). The authors extend their earlier work described in (Salton et. al., 1994). The first technique is a chronological text-segment decomposition algorithm and the second is semantic decomposition of text into themes. The segments are considered as parts of the document that normally stand out as independent e.g. an abstract, a conclusion, or an introduction. The semantic decomposition involves gathering all pieces of a document that contain related information; text themes. They used the segments created and the text themes identified to match a theme to a specified segment/segments. A threshold is introduced; the segments and themes that pass the threshold test are used to characterize the text structure thereby forming summaries. The authors claim that the two algorithms gave readable outputs especially when tested on complex text structures; however they suggested improvements which could be made by adding relevant material during the segmentation stage.

3.5.2 Using Lexical Chains

A method for summarizing large documents using lexical chains was taken in (Silber and McCoy, 2000). They refer to work described by (Barzilay and Elhadad, 1997).

A linear-time algorithm is presented for extracting lexical chains in large documents using WordNET as the word dictionary. The algorithm first creates chains, then after a scan on the document, a word is inserted into the right chain depending on its number sense.

A score for a chain is calculated after a word is inserted in it. The resulting chains are used to form the final document summary. The authors claim that experiments conducted showed similar results with those done by (Barzilay and Elhadad, 1997) but their technique was more efficient as it used a linear time complexity.

3.5.3 Semantic Orientation

A method for inferring the semantics of a word based on its statistical association was introduced by (Turney and Littman, 2003). Their focus was on identifying positive or negative measure of words; distinguishing antonyms from synonyms of a given word. Given a word, its semantic orientation is calculated by measuring the difference between its association with positive and negative words. The greater of the two measures is taken as the dominant factor, thereby giving the word its semantic orientation.

A suggested application for this orientation is in marketing strategies, particularly in creating movie and automobile reviews. The authors claim that experiments tested on 3596 words that were manually labeled as either positive or negative produced an accuracy of about 82.8%.

3.5.4 Using Predefined Text Classes instead of WorldNet

A different approach was taken by (Sakurai and Suyama, 2004). Text data was put into sets of words then, using lexical analysis, they generated key phrases. They refer to work done by (Ichimura et. al., 2000). Instead of using WordNET, a list of text classes are provided by a user and then used together with the extracted key phrases to identify hidden rules. Fuzzy decision trees are used in matching a key phrase to its right text-class node in the tree. The authors claim that experiments done on email analysis using key phrases gave more valid rules compared to using just mere words in a document.

3.5.5 Mining Partial Features

A technique for mining reviews commented on by customers was introduced by (Hu and Liu, 2004). They proposed a method for creating summaries on product reviews. They refer to work described by (Turney and Littman, 2003). Their main focus was on the features that customers had placed opinions on.

They did not rewrite summaries rather they extracted the features commented on by customers, then determined if an opinion was positive or negative and then a summary was created.

First they extract features that were viewed as important depending on the count of the features; they then use a pruning mechanism to reduce the volume of output given and also to eliminate redundancy. The semantic orientation proposed by (Turney and Littman, 2003) is also applied to understand the semantic orientation of the reviews. The authors claim that experiments conducted on products generally sold or bought online gave effective feedback.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Salton et. al., 1996)	Replacing large text for easy information retrieval.	-Chronological decomposition of text into text segments. -semantic decomposition of text segments to text themes. -use text themes to characterize original text structure.	(Salton et. al., 1994)
(Silber and McCoy, 2000)	Summarizing large documents with efficiency.	-Capturing document content and forming lexical chains using WorldNet. -High scoring chains form summaries	(Barzilay and Elhadad, 1997)
(Turney and Littman, 2003)	Evaluating semantic orientation of a word.	-Identifying how a word measures by taking the difference between its association to positive and negative words. -The greater of the two is taken as the dominant factor.	(Hatzivassiloglou and McKeown, 1997).
(Sakurai and Suyama, 2004)	Summarization of Large documents with efficiency.	-Breaking documents into lexical chains and using predefined text classes instead of key concept dictionaries like WorldNet.	(Ichimura et. al., 2000)
(Hu and Liu, 2004)	Developing a technique for mining customer reviews.	Extraction of features commented on by customers then identifies positive and negative reviews.	(Turney and Littman, 2003)

Table 3.5 Papers that have addressed the Extraction Problem

3. 6 Representation of Summaries

3.6.1 The Babylon Project

An approach for understanding knowledge underlying GlaxoSmithKline (Pharmaceutical Company) databases was proposed by (Merrill, 2003). He presented the Babylon Project, a text-mining framework that was intended to be extensible over a variety of domains. An ontology in the system enables users to view concept hierarchies, tools for querying the knowledge base and mechanisms for exploring results. The first prototype use of the system was to mine drug reports to identify reactions, events, or interaction between drug to drug by using knowledge present in the documents. The future of this project will be extended for other use within the company.

3.6.2 What is this Text About?

Presenting informative views of text in aid of visualization and navigation of documents was an approach taken by (Hernandez and Grau, 2003). They proposed a technique for providing a user with different levels of abstraction of an original document. The authors refer to work done by (Hearst, 1999). The first level is the global level, that contains the document's main topic, and the second level contains sub topics extracted from the document. The authors' future intention is to develop an evaluation framework to test the significance of the proposed system to a user during information retrieval.

3.6.3 Using Shrinkage

A mechanism for improving the coverage of approximating content summaries was developed by (Iperiotis and Gravano, 2004). This was an advancement of their earlier work described in (Iperiotis and Gravano, 2003). Due to the fact that similar databases tend to have similar vocabularies, the authors proposed a mechanism for having one database complement another thereby creating a hierarchy of related vocabularies. The authors claim that experiments done on 315 real web databases, as well as on TREC data, showed shrinkage-based content summaries developed were more complete than non-shrunk summaries. They also claim that document sample size was never increased during classification.

3.6.4 Discovering Technological Intelligence

In her dissertation (Kongthon, 2004), the management of information for Research and Development (R&D), was the main focus. The author extended the work done on Technology Opportunities Analysis (TOA) developed in the Technology Policy and Assessment Center, at the Georgia Institute of Technology, USA.

Two Text-Mining algorithms using association rules were proposed for gathering related terms in text data. The first is a tree-like network for capturing important themes of a hierarchical structure, and the second group concepts together to form a thesaurus for data preprocessing. The author claims that experiments conducted on abstracts from Thai Science and technology publications showed good results on supporting decision making on Science and Technology in Thailand.

3.6.5 Semantic Thumbnails

The use of semantic thumbnails to represent document collections was proposed by (Sengupta et. al., 2004). Just like in image processing, a thumbnail is a compressed representation of an original image. This technique is similar to work done in BioKnot at university of Indiana, USA. In this paper, semantic content of a document is used to create a document summary; a semantic thumbnail of the document. The authors' main focus was on processing documents in XML format. They claim that experiments done showed good accuracy of recall on keyword based searches on the web. Future on this work intends to look into accommodating time on document content when creating semantic thumbnails.

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Merrill, 2003)	Developing application for knowledge exploration and management for GlaxoSmithKline.	-Babylon project. -An ontology based system that allow users to view concept hierarchies, tools for querying the knowledge base and exploring results.	Independent work done at GlaxoSmithKline Data Exploration Sciences Division
(Hernandez and Grau, 2003)	Providing informative views for text visualization and navigation.	-Different levels of abstraction. -Global Level presents the main topic. -Second Level has subtopics extracted from documents. -Third Level has details of chosen segment.	(Hearst, 1999).
(Iperiotis and Gravano, 2004)	Mechanism for improving coverage of approximating content summaries.	-Creating hierarchies of related vocabularies.	(Iperiotis and Gravano, 2003)
(Kongthon, 2004)	Management of information for Research and Development.	-Using association rule mining techniques to gather rules from text. -Group rules to form hierarchies.	Extension of work done on Technology Opportunities Analysis (TOA) at Georgia Institute of Technology, USA.
(Sengupta, 2004)	Document Representation.	-Using semantic thumbnails to represent document contents. -Concept taken from image representation.	Refer to work done on BioKnot, a Bioinformatics setting at University of Indiana, USA

Table 3.6 Papers that have addressed the Summary Representation Problem

4. CONCLUSION

This report contains a review of research on text summarization using association-rule mining techniques. Most of the algorithms presented deal with various aspects of feature extraction and text-summary generation from key features. Six main issues were identified as a concern for any good summarization algorithm.

This includes issues with extraction of features from a document; efficient tools are needed for information extraction from large text databases (Yonatan et. al., 2001).

Evaluating how informative extracted features are to a user of the original document is also a major issue and various researchers are looking towards having informative views (Hahn and Mani, 2000).

While ranking could be a useful method for evaluating features extracted, the question still remains, how should features be assigned weights and how should these features be ranked in order to eliminate non-important words. An attempt to have summaries with a wider coverage of the original document is done in (Gong and Liu, 2001).

New techniques that deal with memory management and computation time problems are needed. Most of the proposed algorithms still use Apriori like approach, (Zaiane and Osmar, 2002; Holt and Chung, 1999). This takes up a lot of memory space and a lot of computation time. An improvement was done with the use of an FP- Tree in (Phan et. al., 2005), however, the original FP-Tree in (Han et. al., 2000) was intended to be used in structured relational databases and does not work well with unstructured data.

Understanding the semantic and lexical orientation of words so as to group similar words in the same category is another issue. Some researchers like (Silber and McCoy, 2000) use the WorldNet Ontology, but other researchers like (Sakurai and Suyama, 2004) create their own text classes they see WorldNet as a limited to English text only.

The final product from any text-summarization system should be informative and easy to use (Hearst, 1999). Hierarchical structures have been proposed by (Iperiotis and Gravano, 2004), however performance of information retrieval in a these systems is quite inefficient and summaries need to be compressed even more (Iperiotis and Gravano, 2004). Semantic thumb nails proposed in (Sengupta et. al., 2004) are a compressed representation of the original document, however this system is meant to be used for only constant documents, when the contents of original documents changes, the semantic thumbnails have to be changed too. A term frequency is used in (Sengupta et. al. , 2004), semantic orientation is the future work for these researchers.

5. ACKNOWLEDGEMENT

I would like to thank Dr. Richard Frost, the instructor for course 60-510 in fall 2005 for all the help in making this a successful report. Thank you for all the good remarks you gave every time I sent you my horrible reports, you never said it was bad, but always gave suggestions of what needed to be modified or included in the report. I would recommend any one to take this course with Dr. Frost, he was very understanding and always available for any help anyone needed.

The next person to thank is my supervisor Dr. Christie Ezeife. She is the most amazing person and professor I have met. She has guided me through many endeavors in the university and especially in creating an excellent research path to follow. She is always there to help even with no appointments; she never turns anyone down. Thank you Dr. Ezeife for being such a good person and professor, dedicated to students welfare and academic success.

6. ANNOTATIONS FOR 23 IMPORTANT REFERENCES

(Used a double asterisk for milestone papers)

1. ** Castillo, M. D. and Serrano, J. I. (2004). A Multistrategy Approach for Digital Text Categorization from Imbalanced Documents. *ACM SIGKDD Exploration Newsletter*. 6(1), 70-79.

The problem addressed by the authors of this paper is a mechanism for categorizing documents. The authors begin by referring to work of (Yang and Hanovar, 1998). They used parallelism to develop a multi-strategy classification of text documents. Different learners were explored and each carried out its own feature selection separately. Outputs of each learner were then combined to form a classification model. The authors claim that experiments done yielded good classification performance.

2. Chuang, W.T. & Yang, J (2000). Extracting Sentences for Text Summarization: A Machine Learning Approach. *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens Greece. pp. 125-159.

The problem addressed in this paper is the development of an automatic text summarizer. They refer to the work done by (Marcus, 1997). The authors proposed a method that used cue markers in extracting segment from sentences. Segments were represented by predefined features. A supervised learning approach was then used to extract sentence segments that were viewed as important and then used in creating a document summary. The authors claim that experiments conducted showed that their algorithm performed reasonably well compared to Microsoft Word Summarizer in terms of recall, precision and accuracy of classification.

3. Feldman, R., Aumann, Y., Libetzon, Y., Ankori, K., Schler, J., and Rosenfeld, B. (2001) A Domain Independent Environment for Creating Information Extraction Modules. *Proceedings of the 10th International Conference on Information and Knowledge Management*. Atlanta, GA, USA. pp. 586-588.

The problem addressed by the authors of this paper is information extraction for knowledge discovery from text documents. First, the algorithm extracted all information on a text document that was termed as important. Using a declarative language, the authors defined rules for extracting important information from text. The authors claim that their system provided efficient tools that were used to develop rules and interpret them.

4. Forman, G. (2003). An Extensible Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*. 3, 1289-1305.

The problem addressed by the author of this paper is an evaluation of different feature selection methods performed on data. The author gave a comprehensive report on evaluation done to twelve feature selection methods performed on data from various sources e.g. Reuters. An analysis was done on each technique based on their accuracy, F-Measure, Precision, and recall using the WEKA system. The experiments main focus was to obtain the classification with best performance without taking into account the number of features selected to get the performance. The author claim that in terms of F-measure especially recall, the Bi-Normal Separation outperformed all other methods. However the author did not do any evaluation using association rules.

5. Gong, Y. and Liu, X. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *Proceedings of the ACM SIGIR*, New Orleans, USA. pp. 19-25.

The problem addressed by the authors of this paper is extraction of text then creating summaries for those texts. The authors presented two text summarization methods that ranked sentences extracted from original documents then used some of those sentences to create summaries. The first method used information retrieval methods for measuring the relevance of a sentence. The second method used latent semantic analysis. The author's main goal was to be able to extract the highest ranked and disjoint sentences for creating a document summary. The authors claim that the results produced on experiments conducted on CNN Worldview news program using the two methods and compared to summaries developed by three human evaluators were quite comparable.

6. Hahn, U. and Mani, I. (2000) The Challenges of Automatic Summarization. *IEEE Journal in Computing*. 33(11), 29-36.

The problem addressed by the authors of this paper is generation of summaries from online resources in a timely manner. The authors gave a comprehensive review on document summarization. They pointed out on the limitations of available summarization tools as mere extraction tools; not able to create a document abstract from extracted text features. They also argue that the tools cannot handle multiple document sources and in particular non-textual data. The authors suggested that evaluation techniques are needed for evaluating document summarizers; a summary must be a complement of the source.

7. **Hernandez, N. & Grau, B. (2003). What is this Text About?.
Proceedings of the 21st Annual Conference on Documentation, San Fransisco, California. pp 117-124.

The problem addressed by the authors of this paper is providing informative views of text in aid of visualization and navigation of documents. The authors refer to (Hearst, 1999). They proposed a technique for providing a user with different levels of abstraction of an original document. The first level was the global level, that contained the document main topic, and the second level contained sub topics extracted from the documents. The authors' future intention is to developing an evaluation framework to test the significance of the proposed system to a user during information retrieval.

8. Holt, J. D and Chung S.M. (1999) Efficient Mining of Association Rules in Text Databases. *Proceedings of the eighth international conference on Information and knowledge management* Kansas City, MO USA. pp. 234-242.

The problem addressed by the authors of this paper is mining association rules from words in text databases. Two algorithms for mining association rules between words in a text database were proposed in this paper. The authors modified the Apriori Algorithm in (Agrawal and Srikant, 1994) and the Direct Hashing and Pruning Algorithm (DHP) in (Park et. al., 1997). In the proposed algorithms, the authors partitioned the frequent items into smaller sets and then processed each partition separately thereby reducing the amount of memory required in the original Apriori or the (DHP).

The authors claim experiments done using their improved algorithms out performed both the original Apriori and the DHP in identifying frequent patterns on large text databases and on memory usage.

9. Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Seattle, Washington, USA. pp 168-177.

The problem addressed by the authors of this paper is dealing with thousands of product reviews from customers. The authors proposed a method for creating summarization for customer reviews and refer to work done by (Turney and Littman, 2003). They focused on the features that customers had placed opinions on. They do not rewrite summaries rather they extracted the features commented on by customers, then identified if an opinion was positive or negative and then a summary was created. The authors claim that experiments conducted on products generally sold or bought online gave effective feedbacks.

10. **Ipeirotis, P. G. and Gravano, L. (2004). When one Sample is not Enough: Improving Text Database Selection Using Shrinkage. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*. Pp 767-778.

The problem addressed in this paper is a mechanism for improving the coverage of approximating content summaries. The authors refer to their earlier work done in (Ipeirotis and Gravano, 2004). Due to the fact that similar databases tend to have similar vocabularies, they proposed a mechanism for having one database complement one another thereby creating hierarchies of related vocabularies.

The authors claim that experiments done on 315 real web databases as well as on TREC data showed shrinkage based content summaries developed were complete than non-shrunk summaries. They also claim that document sample size was never increased during classification.

11. **Kongthon, A. (2004). A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management. *PhD Dissertation, Georgia Institute of Technology*. Georgia, USA. 212 pages.

The problem addressed in this dissertation is the management of information for Research and Development (R&D). The author extends the work done on the Technology Opportunities Analysis (TOA) developed in the Technology Policy and Assessment Center, Georgia Institute of Technology, USA. Two Text Mining algorithms using association rules were proposed for gathering related terms in text data. The first was a tree-like network for capturing important themes of a hierarchical structure, and the second grouped concept together to form a thesaurus for data preprocessing. The author claims that experiments conducted on abstracts from Thai Science and technology publications showed good results on supporting decision making on Science and Technology in Thailand.

12. Kotcz, A., Prabhakar, V., Kalita, J. (2001). Summarization as Feature Selection for Text Categorization. *Proceedings of the CIKM*, Atlanta GA, USA. pp 365-370.

The problem addressed by the authors of this paper is evaluating effectiveness of text categories created with summarization techniques. The authors proposed an algorithm for creating document summaries using only words extracted from the original document. They refer to work done in (Mahesh, 1997). First the algorithm assigns a weighing measure for each feature extracted, then unique terms are evaluated depending on their relevance weights. The words with the highest score are then used to form the document summary. The authors claim that experiments conducted on Reuters-corpus demonstrated good performance on the relevance of summaries when matched with original documents.

13. **McDonald, D and Chen, H. (2002). Using Sentence-Selection Heuristic to Rank Text Segments in TXTRACTOR. *Proceedings of the 2nd ACM/IEEE CS-Joint Conference on Digital Libraries*, Portland, Oregon, USA. pp. 28-35.

The problem addressed by the authors of this paper is creating summaries from a user defined number of sentence. The authors refer to work done in (Carbonell and Goldstein, 1998). They presented TXTRACTOR, a tool for ranking text segments. The segments were later used to create document summaries. They presented three steps that they used in creating the summaries; (1) evaluation of sentence based on position, length and format of words in a sentence (capitalized or not), (2) sentence segmentation by identifying topic boundaries, (3) segment ranking using some heuristics and finally, summary creation. Low ranking segments were eliminated during summary creation.

They compared their approach to that of a tool created using the TextTiling algorithm and claim that their approach gave better summaries.

14. Merrill, G.M (2003). The Babylon Project: Toward an Extensible Text-Mining Platform. *IT Professional* 5(2), 23-30.

The problem addressed in this paper is enhancing applications for knowledge exploration and management. The author focused on ways for understanding knowledge underlying in GlaxoSimthKline (Pharmaceutical Company) databases.

He presented the Babylon Project, a text mining framework that was intended to be extensible over a variety of domains. The first prototype in the system was to mine drug reports to identify reactions, events, or interaction between drug to drug by using knowledge present in the documents. The future on this project will be extended for other use within the company.

15. **Mooney, R. J and Bunescu R. (2005). Mining Knowledge from Text Using Information Extraction. *ACM SIGKDD Exploration Newsletter*. 7(1), 3-10.

The problem addressed by the authors in this paper is the use of information extraction for identifying relations on text documents. The authors' extends their earlier work in (Mooney and Califf, 2003) that used regular expression to stimulate rules from training examples. In this paper they took a different approach; labeling words in the documents, then using techniques for modeling sequences from the resulting labels to identify hidden patterns. The authors did experiments on Biomedical journals; identifying human proteins.

In order to model sequences appropriately, they compared the Hidden Markov Model (HMM) and the Conditional Random Field (CRF) model. They claim that the CRF provided the best model for sequencing labels.

16. **Phan, X. H., Ho, T.B., Nguyen, L.M. and Horiguchi, S. (2005). Improving Discriminative Sequential Learning with Rare-but-Important Associations. *Proceedings of the ACM Symposium on Applied Computing*. Chicago, Illinois, USA. pp. 304-313.

The problem addressed by the authors of this paper is discovering of rare but important rules from huge amount of documents. The authors used an FP-Tree technique in (Han et al., 2000) for discovering instances that would not be discovered with common training models in aid of classifying irregular instances. The authors' claim that experiments done on English language chunking and in recognition of name entities showed accurate results. They also claim that there is a relationship between association rules and statistical learning in discovering patterns from document sources.

17. **Sakurai, S. S. and Suyama, A. (2004). Rule Discovery from Textual Data based on Key Phrase Patterns. *Proceedings of the ACM Symposium on Applied Computing March 2004*, Nicosia Cyprus. pp. 606-612.

The problem addressed by the authors in this paper is a technique for discovering rules from text without using concept dictionary. The authors begin by referring to work done in (Ichimura et. al., 2000). They proposed an algorithm that decomposed text data into sets of words using lexical analysis there by generating key phrases.

A list of text classes were then provided by a user and used together with the extracted key phrases to identify hidden rules from the underlying text by use of fuzzy decision trees; a key phrase was matched to the right text class in the tree. The authors claim that experiments done on email analysis using key phrases gave valid rules compared to using just mere words in a document.

18. **Salton, G., Singhal, A., Buckley, C. and Mitra, M. (1996). Automatic Text Decomposition Using Text Segments and Text Themes. *Proceedings of the Seventh ACM Conference on Hypertext*. Washington DC, USA. pp. 53-65.

The problem addressed by the authors of this paper is creation of summaries from text that acted as a representation of the original document for simplifying information retrieval. The authors refer to their earlier work in (Salton et. al., 1994). In this paper they presented two techniques; a chronological text segments decomposition algorithm and a semantic decomposition of text into themes. An interaction between the outputs of the two algorithms was then used to characterize text structure thereby forming text summaries. The authors claim that the two algorithms gave readable outputs especially when tested on complex text structures; however they suggested improvements could be done by adding relevant materials during segmentation stage.

19. Sengupta, A. Dalkilic, M and Costello, J. (2004). Semantic Thumbnails - A Novel Method for Summarization Document Collections. *Proceedings of the 22nd International Conference on Design of Communication: The Engineering Quality Documentation*, Memphis, Tennessee, USA. pp.45-51.

The problem addressed by the authors of this paper is the use of semantic thumbnails to represent document collections. The authors adapt the technique used in BioKnot a Bioinformatics System at University of Indiana. Just like in image processing, a thumbnail is a compressed representation of an original image. In this paper, semantic content of a document was used to create a document summary; semantic thumbnail of the document. The authors' main focus was on processing documents in XML format. They claim that experiments done showed accuracy of recall on keyword based searches on the web. Future on this work intends to look into accommodating time on document content when creating semantic thumbnails.

20. Silber, H. G. and McCoy, K.F. (2000). Efficient Text Summarization Using Lexical Chains. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, New Orleans, USA. pp. 252-255.

The problem addressed by the authors of this paper is a methodology for summarizing large documents using lexical chains. The authors refer to work done by (Barzilay and Elhadad, 1997). A linear time algorithm was presented for extracting lexical chains in large documents using WorldNET as the word dictionary.

The algorithm first created chains then after a scan on the document, a word was inserted to the right chain depending on its number sense, a score for a chain was calculated after a word was inserted in it. The resulting chains were used to form the final document summary. The authors claim that experiments conducted showed similar results with Barzilay and Elhadad algorithm.

21. Turney., P. D. and Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*. 21(4), 315-346.

The problem addressed by the authors of this paper is on evaluation of semantic orientation of words in text documents. The authors presented a method for inferring the semantics of a word based on its statistical association with sets of positive or negative words; distinguishing antonyms from synonyms. They used techniques proposed in (Hatzivassiloglou and McKeown, 1997) for semantic word orientation. They claim that experiments tested on 3596 words that were manually labeled as either positive or negative produced an accuracy of about 82.8%.

22. Yonatan, R. F., Libetzon, L., Ankori, K. Schler, J. and Rosenfeld, B. (2001). A Domain Independent Environment for Creating Information Extraction Modules. *Proceedings of the 10th International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA. pp 586-588.

The problem addressed in this paper is the discovery of knowledge from text based on key words extracted from the text itself. The authors presented ClearStudio, a system that gave modules for extracting information that is based on work done in (Appelt et. al., 1993). The system included rules for defining important features to be extracted. Features included events, facts, and words with meaning given the document domain. The rules were developed with DIAL (Declarative Information Analysis Language). The authors claim that the system was successfully used to create rules in diverse domains e.g. analysis of patents and financial news.

23. ** Zaiane, O. and Antonie, M.L. (2002). Classifying Text Documents by Association Terms with Text Categories. *Proceedings of the 13th AustralAsian Conference on Database Technologies*. pp. 215-222.

The problem addressed by the authors of this paper is exploring the use of association rule mining techniques for creating document categories. The authors refer to work done in (Agrawal and Srikant, 1994). They presented two algorithms for building a classifier; the first took into consideration only about a single category while the second one handled the entire training set as one entity. A dominance factor was introduced to enhance overlapping categories. The authors claim that their algorithm presented efficient training phase in addition to having rules that were understandable compared to most text classifiers.

7. BIBLIOGRAPHY

1. Abiteboul, S., Aggrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., DeWitt, D., Franklin, M., Molina, H. G., Gawlick, D., Gray, J., Haas, L., Halevy, A., Hellerstein, J., Ioannidis, Y., Kersten, M., Pazzani, M., Lesk, M., Maier, D., Naughton, J., Schek, H., Sellis, T., Silberschatz, A., Stodgrass, R., Ullman, J., Weikum, G., Wisdom, J. and Zdonik, S. (2005). The Lowell Database Research Self-Assessment. *Communications of the ACM*, 8(5) 111-118.
2. Agichtein, E. and Ganti, V. (2004). Mining Reference Tables for Automatic Text Segmentation. *Proceedings of the 10th ACM SIGIDD International Conference on Knowledge Discovery and Data Mining*. (pp. 20-29).
3. Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference*. pp.487-499.
4. Agrawal, R. Bayardo, R. J. and Srikant, R. (2000). Athena: Mining-based Interactive Management of Text Databases. *Proceedings of the 7th International Conference on Extending Database Technology (EDBT)*, Konstanz, Germany.
5. Alisa, K. (2004). A Text Mining Framework for Discovering Technological Intelligence to Support Science & Technology Management. *PhD Dissertation, Georgia Institute of Technology*, Georgia, USA. pp.105-112.
6. Appelt, D. E., Jerry, R. H., Bear, J., Israel, D. Kameyama, M. and Tyson, M. (1993) The SRI MUS-5 JV-FAVSTUS Information Extraction System. *Proceedings of the 5th Message Understanding Conference*, Baltimore, MD.
7. Baralis, E. and Chiusano, S. (2004). Essential Classification Rule Sets. *ACM Transactions on Database Systems*, 29(4), 635-674.
8. Basu, S., Mooney, R. J., Pasupuleti, K.V. & Ghosh, J. (2001). Evaluating the Novelty of Text-Mined Rules Using Lexical Knowledge. *Journal of KDD*, pp. 233-238.
9. Barzilay, R. and Elhadad, M. (1997). Using Lexical Chains for Text Summarization. *In Proceedings of the Intelligent Scalable Text Summarization Workshop*, Madrid, Spain. pp 10-17.
10. Califf, M.E. and Mooney, R.J. (2003) Bottom relational learning of pattern matching for information extraction. *Journal of Machine Learning Research*, 4 pp122-210.

11. Carbonell, J. and Goldstein, J (1998). The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *In SIGIR*, Melbourne, Australia, pp 335-336.
12. Castillo, M. D. D. and Serrano, J. I. (2004). Special issue on learning from imbalanced datasets: A Multistrategy Approach for Digital Text Categorization from Imbalanced Documents. *ACM SIGKIDD Exploration Newsletter*. 6(1), 70-79.
13. Chen, H.H., Tsai S.C. and Tsai, J.H. (2000). Mining Tables from Large Scale HTML Texts. *Proceedings of 18th Conference on Computational Linguistic*. pp. 166-172.
14. Chuang, W.T. & Yang, J (2000). Extracting Sentences for Text Summarization: A Machine Learning Approach. *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens Greece. pp. 125-159.
15. Feldman, R., Dagan, I. and Hirsh, H. (1998). Mining Text Using Keyword Distributions. *Journal of Intelligent Information Systems*, 10, 281-300.
16. Feldman, R., Aumann, Y., Libetson, Y., Ankori, K., Schler, J., and Rosenfeld, B. (2001) A Domain Independent Environment for Creating Information Extraction Modules. *Proceedings of the 10th International Conference on Information and Knowledge Management*. Atlanta, GA, USA. pp. 586-588.
17. Forman, G. (2003) An Extensible Empirical Study of Feature Selection Metrics for Text Categorization. *Journal of the Machine Learning Research*. 3, 1289-1305.
18. Fule, P. and Roddick, J. F. (2004). Experiences in Building a Tool for Navigating Association Rule Result Sets. *Proceedings of the 2nd Workshop on Australian Information Security, Data Mining, Web Intelligence and Software Internationalisation*, Dunedin, New Zealand. pp. 103-108.
19. Gong, Y. and Liu, X. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *Proceedings of the ACM SIGIR*, New Orleans, USA. pp. 19-25.
20. Hahn, U. and Mani, I. (2000). The challenges of Automatic Summarization. *Article in IT Professional*, 33(11), 29-36.

21. Hatzivassiloglou, V. and McKwoen, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL*. New Brunswick, NJ. pp 174 -181.
22. Hearst, M. (1999). User interfaces and Visualization. In R. Baeta-Yats and B. Ribeiro-Neto, editors, *Modern Information Retrieval*. pp. 257-322.
23. Hearst, M. (1999). Untangling Text Data Mining
Proceedings of ACL 37th Annual Meeting of the Association for Computational Linguistics. Marlyland, USA. pp. 3-10.
24. Hernandez, N. & Grau, B. (2003). What is this Text About?.
Proceedings of the 21st Annual Conference on Documentation, San Fransisco, California. pp.117 - 124.
25. Holt, J. D and Chung S.M. (1999) Efficient Mining of Association Rules in Text Databases. *Proceedings of the eighth international conference on Information and knowledge management* Kansas City, MO USA. pp. 234-242.
26. Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews.
Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery in Dat Mining, Seattle, Washington, USA. pp 168-177.
27. Huang C., Chuang, S. and Chien, L. (2004). LiveClassifier: Creating Hierarchical Text Classifiers through Web Corpora. *Proceedings of the 13th International Conference on World Wideweb*, New York, USA. pp. 184-192.
28. Ichimura, Y., Nakayama, Y., Miyoshi, M., Akahane, Y., Sekiguchi, T. and Fujiwara, Y. (2001). Text Mining System for Analysis of a Salesperson's Daily Reports. *Proc. Of Pacific Association for Computational Linguistics*. pp127-135.
29. Ipeirotis, P.G. and Gravano, L. (2004) When one Sample is not Enough: Improving Text Database Selection Using Shrinkage. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. Paris, France. pp. 767-778.
30. Kogut, P., Yen, J., Leung, Y., Sun, S., Wang, R., Mielczarek, T. and Hellar, B. (2004). Proactive Information GA Homeland Security Teams. *Communications of the ACM*. 47(3), 48-50.
31. Kotcz, A., Prabakarmurthi, V., Kalita, J. (2001). Summarization as Feature Selection for Text Cateorizatio. *Proceedings of the CIKM*, Atlanta GA, USA. pp. 365-370.

32. Krishna, K. and Krishnapuram, R. (2001). A Clustering Algorithm for Asymmetrically Related Data With Applications to Text Mining, *Proceedings of the CKIM*, Atlanta, GA, USA. pp. 571-573.
33. Lin, D. & Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA. pp.323-328.
34. Liu, B., Hu, M. and Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 11th International Conference on World Wide Web*, Chiba, Japan. pp. 342-351.
35. Mahesh, K. (1997) Hypertext Summary Extraction for Fast Document Browsing. *In working of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web* pp. 95-103.
36. Marcu, D. (1997). The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. *PhD Thesis, Department of Computer Science, University of Toronto*, Toronto, Canada. 374 pages.
37. McDonald, D. and Chen, H (2002). Using Sentence-Selection Heuristics to Rank Text Segments in TXTRACTOR. *Proceedings of the 2nd ACM/IEEE CS- Joint Conference on Digital Libraries*, Portland ,USA. (pp. 28-35).
38. Mei, Q. and Zhai, C. (2005) Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA. pp. 198-207.
39. Meir, R and Zhang, T (2003) Generalization Error Bounds for Bayesian Mixture Algorithm. *Journal of Machine Learning Research*. 4, 839-560.
40. Merrill, G.H. (2003). The Babylon Project: Toward an Extensible Text-Mining Platform. *IT Professional*, 5(2), 23-30.
41. Mooney, R. J and Bunescu R. (2005). Mining Knowledge from Text Using Information Extraction. *ACM SIGKDD Exploration Newsletter*. 7(1), 3-10.
42. Moniroga, S., Arimura, H., Ikeda, T., Sakao, Y. and Akamine, S. (2005) Key Semantic Extraction Dependency Tree Mining. *Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining KDD '05*. pp. 666-671.

43. Nahm, U. Y. & Mooney, R.J. (2002). Mining Soft-Matching Association Rules. *Proceedings of the 11th International Conference on Information & Knowledge Management*, McLean, Virginia, USA. pp. 681-683.
44. Park, J. S., Chen, M. S. And Yu, P. S. (1997). Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Trans. On Knowledge and Data Engineering*, (9)5, pp 813-825.
45. Phan, X.H., Ho, T.B., Nguyen, L. M., and Horiguchi, S. (2005). Improving Discriminative Sequential Learning with Rare-but-Important Associations. *Proceedings of the ACM Symposium on Applied Computing*, Chicago, Illinois, USA. pp. 304-313.
46. Pratt, W. and Yetisgen-Yildiz, M. (2003). LitLinker: Capturing Connections Across the Biomedical Literature. *Proceedings of the International Conference on Knowledge Capture*, Sanibel Island, FL. USA. pp.105-112.
47. Radev, D.R. & McKeown, K. (1998). Generating Natural Language Summaries from Multiple On-Line Sources. *Journal of Association for Computational Linguistics*, 24(4), 469-500.
48. Sakurai, S. and Ueno, K. (2004). Analysis of Daily Business Reports Based on Sequential Text Mining Method. *IEEE International Conference on Systems, Man & Cybernetics*, pp. 3279-3284.
49. Sakurai, S. S. and Suyama, A. (2004). Rule Discovery from Textual Data based on Key Phrase Patterns. *Proceedings of the ACM Symposium on Applied Computing March 2004*, Nicosia Cyprus. pp. 606-612.
50. Salton, G., Buckley, C. and Singhal, A. (1994). Automatic Analysis, Them Generation and Summarization of Machine readable Texts. *Sciences*, (3) 1421-1426.
51. Salton, G., Singhal, A., Buckley, C. and Mitra, M. (1996). Automatic Text Decomposition Using Text Segemnts and Text Themes. *Proceedings of the Seventh ACM Conference on Hypertext*. Washington DC, USA. pp. 53-65.
52. Sengupta, A. Dalkilic, M and Costello, J. (2004). Semantic Thumbnails - A Novel Method for Summarization Document Collections. *Proceedings of the 22nd International Conference on Design of Communication: The Engineering Quality Documentation*, Memphis, Tennessee, USA. pp. 45-51.
53. Silber, H. G. and McCoy, K.F. (2000). Efficient Text Summarization Using Lexical Chains. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, New Orleans, USA. pp. 252-255.

54. Tang, J. Li, H. Cao, Y. and Tang, Z (2005). Email Data Cleaning. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois. pp. 489-498.
55. Turney, P.D. (2002). Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. www.arxiv.org/.
56. Turney, P.D. and Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transaction on Information Systems*. 21(4), 315-346.
57. Wan, J. W.W. and Dobbie, G (2004). Mining Association Rules from XML Data Using XQuery. *Proceedings of the second workshop on AustralAsian Information Security, Data Mining & Web Intelligence and Software Internationalization*, Dunedin, NewZealand. pp. 169-174.
58. Yang, J. and Hanovar , V. (1998) Feature Subset Selection Using a Genetic Algorithm. *IEEE Intelligent Systems and their Applications*. 13(2) 44-49.
59. Yonatan, R. F., Libetzon, L., Ankori, K. Schler, J. and Rosenfeld, B. (2001). A Domain Independent Environment for Creating Information Extraction Modules. *Proceedings of the 10th International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA. pp. 586-588.
60. Zaiane, O. and Antonie, M.L. (2002). Classifying Text Documents by Association Terms with Text Categories. *Proceedings of the 13th AustralAsian Conference on Database Technologies*. pp. 215-222.
61. <http://www.research.ibm.com/journal/sj/404/nasukawa.html>
62. www.sims.berkeley.edu/~hears/text_mining.html
63. www.cs.uiowa.edu/~sehgal/
64. <http://biokdd.informatics.indiana.edu/jccostel/>

APPENDIX 1

(a)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Chuang and Yang, 2000)	Automatic Text Extraction.	Using Cue Markers as a guide to segment extraction from sentences.	(Marcu, 1996)
(Yonatan et. al., 2001)	Information Extraction from Text.	Declarative Information Analysis system that provided rules for text extraction.	(Appelt et al., 1993)
(Castillo and Serrano, 2004)	Multistrategy Classifier System.	Parallelism as a strategy for text document classification.	(Yang and Hanovar, 1998)
(Mooney and Califf, 2005)	Natural language Information Extraction.	Converting unstructured text to structured text with aid of sequential word labeling.	Extends earlier done in (Mooney and Cliff, 2003)

Table 3.1 Papers that have addressed the Extraction Problem....(page 12).

(b)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Hahn and Mani, 2000)	Evaluating extraction procedures.	.This is a comprehensive review of various issues that affect most of the tools available in the industry for text summarization.	-This is an independent work and did not refer to any work. -It's a good paper to understand text summarization problem.
(Forman, 2003)	Evaluation of different feature-selection method.	-Presented a comprehensive study of experiment done to 12 feature selection methods.	-They used the WEKA system from University of Waikato to perform most experiments.

Table 3.2 Papers that have addressed the Evaluation of Extraction Algorithms
...(page 14).

(c)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Gong and Liu, 2001)	Text extraction with ranking.	-Two summarization methods are proposed. -1 st method use information retrieval to measure sentence relevance. -2 nd method use latent semantic analysis.	This was an independent work and did not refer to any previous work.
(Kotcz, 2001)	Feature extraction with ranking.	-A technique that assigns weights to each feature extracted. -High scoring features are used to create summaries.	(Mahesh, 1997)
(McDonald and Chen, 2002)	Ranking text segments.	-Creation of heuristics for ranking segments using Maximal Marginal Relevance.	(Carbonell and Goldstein, 1998)

Table 3.3 Papers that have used Weighting and Ranking Mechanism...(page 17).

(d)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Holt and Chung, 1999)	Mining association rules from text.	-Two algorithms presented: (1) Improved Apriori (2) Improved Direct Hashing and Pruning	(Agrawal and Srikant, 1994) and (Park et. al., 1997)
(Zaiane and Osmar, 2002)	Efficient mining of association rules from text.	-Two algorithms presented: (1) Discover rules from each category at a time (2) Discover rules from all categories at one scan	(Agrawal and Srikant, 1994).
(Phan et, al., 2005)	Discover rare but important rules from text.	-They introduce two integers an upper and lower bound, a rule that falls in the lower bound but with above average confidence is rare but important	(Han et al., 2000)

Table 3.4 Papers that have addressed the Memory Management and Computation Problem ...(page 20).

(e)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Salton et. al., 1996)	Replacing large text for easy information retrieval.	-Chronological decomposition of text into text segments -semantic decomposition of text segments to text themes. -use text themes to characterize original text structure.	(Salton et. al., 1994)
(Silber and McCoy, 2000)	Summarizing large documents with efficiency.	-Capturing document content and forming lexical chains using WorldNet. -High scoring chains form summaries.	(Barzilay and Elhadad, 1997)
(Turney and Littman, 2003)	Evaluating semantic orientation of a word.	-Identifying how a word measures by taking the difference between its association to positive and negative words. -The greater of the two is taken as the dominant factor.	(Hatzivassiloglou and McKeown, 1997).
(Sakurai and Suyama, 2004)	Summarization of Large documents with efficiency.	-Breaking documents into lexical chains and using predefined text classes instead of key concept dictionaries like WorldNet.	(Ichimura et. al., 2000)
(Hu and Liu, 2004)	Developing a technique for mining customer reviews.	Extraction of features commented on by customers then identifies positive and negative reviews.	(Turney and Littman, 2003)

Table 3.5 Papers that have addressed the Extraction Problem...(page 25).

(f)

Authors	Problem Addressed	Proposed Solution	Referenced Work
(Merrill, 2003)	Developing application for knowledge exploration and management for GlaxoSmithKline	-Babylon project -An ontology based system that allow users to view concept hierarchies, tools for querying the knowledge base and exploring results.	Independent work done at GlaxoSmithKline Data Exploration Sciences Division
(Hernandez and Grau, 2003)	Providing informative views for text visualization and navigation.	-Different levels of abstraction. -Global Level presents the main topic -Second Level has subtopics extracted from documents. -Third Level has details of chosen segment	(Hearst, 1999).
(Iperiotis and Gravano, 2004)	Mechanism for improving coverage of approximating content summaries	-Creating hierarchies of related vocabularies	(Iperiotis and Gravano, 2003)
(Kongthon, 2004)	Management of information for Research and Development	Using association rule mining techniques to gather rules from text. -Group rules to form hierarchies	Extension of work done on Technology Opportunities Analysis (TOA) at Georgia Institute of Technology, USA.
(Sengupta, 2004)	Document Representation	-Using semantic thumbnails to represent document contents. -Concept taken from image representation	Refer to work done on BioKnot, a Bioinformatics setting at University of Indiana, USA

Table 3.6 Papers that have addressed the Summary Representation Problem ... (page 29).

(g)

Authors	Problem Addressed	Proposed Solution	ARM Technique
(Holt and Chung, 1999)	Mining association rules from text.	-Two algorithms presented: (1) Improved Apriori. (2) Improved Direct Hashing and Pruning.	Apriori
(Zaiane and Osmar, 2002)	Efficient mining of association rules from text.	-Two algorithms presented: (1) Discover rules from each category at a time. (2) Discover rules from all categories at one scan.	Apriori
(Phan et, al., 2005)	Discover rare but important rules from text.	-They introduce two integers an upper and lower bound. -A rule that falls in the lower bound but with above average confidence is rare but important.	Frequent Tree Patterns

Table 4. Papers that have used Association-Rule Mining Techniques.