Role of Text Mining in Business Intelligence

Palak Gupta[1], Barkha Narang[2]

Abstract

This paper includes the combined study of business intelligence and text mining of uncertain data. The data that is used in current business domains is not precise, accurate and complete. Instead, data is considered uncertain and therefore this uncertainty is propagated to the results produced by Business Intelligence (BI). Till now, websites most often used text-based searches, which only found documents containing specific user-defined words or phrases. Through use of a semantic web, text mining can find content based on meaning and context rather than just by a specific word. This improves our search and BI results and facilitates social networks analysis or counter-intelligence. The text mining software acts similar to an intelligence analyst or research librarian, focusing on more limited scope of analysis. The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to extract automatically specially from unstructured data. May be for this reason, there has been little work in text data mining to date, and most people talk about it as information access or many have not used text directly to discover unknown information. Through this paper we wish to uncover such areas and define Business Intelligence, structured or unstructured data, text mining, and then discuss the potential applications and limitations of text mining. The idea behind discussing this is to draw attention to exciting new kinds of problems and BI trends like green computing, social networking, data visualization, mobile BI, predictive analytics, composite applications, cloud computing, multi-touch and Software-as-a-Service (SaaS). This paper would outline recent ideas about how to pursue exploratory data analysis over text and what we consider to be real text data mining efforts.

Keywords- Business Intelligence, Text mining, structured data, unstructured data

[1] Assistant Professor, Jagannath International Management School, Kalkaji, New Delhi
E-mail:vaishpalak@rediffmail.com Mobile: 9650012542

[2] Assistant Professor, Jagannath International Management School, Kalkaji, New Delhi
E-mail:barkhaanarang@gmail.com Mobile: 9971253940

Introduction

Business Intelligence (BI) refers to technologies and applications for collecting, storing and analyzing business data that ultimately helps the enterprise to make better decisions (De Ville, 2001). BI techniques are generally computer based that provide past, current and future trends of the enterprise. Its applications include activities of decision support system, query and reporting, data mining, complex event processing, online analytical processing, process mining, business performance management, text mining, statistical and predictive analysis.

Business Intelligence helps in good decision making and ensures competitive intelligence by analyzing the company's internal data along with the information of the competitors (Bergeron and Hiller, 2002). BI applications in an enterprise are diverse providing enterprise reporting to serve strategic management of business, executive information system, collaboration platform to allow sharing of inside and outside business data and electronic data interchange. An important aspect of BI is knowledge management that helps companies in making good strategies through proper insight and experiences. BI also provides pro-active alerts in form of colour change of reports if some mishap occurs.

BI helps companies to analyze their loads of data for decision making but not all data is structured and simple to understand as some data exists in unstructured or semi-structured form in which searching and interpretation consumes a lot of time. Thus, making decisions in such situations becomes complex.

Metadata- Handling Unstructured and Semi-structured data

Businesses store volumes of data in the form of web pages, emails, video and image files, news and reports which are called semi structured or unstructured data. In practise, such data leads to waistage of time in searching and leads to poor decisions as volumes of unstructured data are stored in variety of formats and referred by different technologies. To solve this problem, metadata which is data about data, should be kept with unstructured or semi-structured data. By the techniques of information extraction and automatic categorization, metadata can be generated in the form of summaries or topics.

Text Mining- Innovation in BI to gain competitive advantage

It is a process of automatic extraction of information from large unstructured text (Hart, nd). In text mining, patterns are identified from natural language text rather than databases as in data mining. It also differs from web mining as in contrast to structured data of web, text mining is usually done on unstructured data to derive pieces of genuine information for business. It is different from web search as in web, searching is of something that is already known and has been written by someone else but in text mining, unknown information is discovered for future predictions.

The phrase "text mining" is used to denote any system that analyzes large quantities of natural language text by parsing it and then detects lexical or linguistic usage patterns in an attempt to extract correct information (Sebastian, 2002**).** It is about looking for patterns in text by automatic extraction. This linking of extracted information is very important as it leads to formation of new facts and hypotheses that can be explored further by experiments. Text mining is quite different from data mining as data mining implies extraction of implicit, previously unknown and useful information from data (Frank and Witten, 2000) but with text mining, the information to be extracted is not at all hidden and is explicit. Text mining explores data in a form that can directly be feeded into computers without any human intervention. It grew from related technologies based on statistics, probability theory and artificial intelligence for example performance of any search can be measured by counting success and failures.

Text mining represents flexible approaches to information management, research and analysis on textual materials. It also helps in natural language processing as it automates the analysis of links of citations in text and hyperlinks in web. Natural language depends solely on "common-sense" knowledge which is exceptionally difficult to encode in algorithmic form (Lenat, 1995) so text mining came up as an outgrowth of "real-text" mindset which provides shallow processing of unrestricted text and deep processing of domain specific material.

Data Warehouse Vs Document Warehouse- Repository for Text Mining

Data Warehouse is actually a repository of business or enterprise databases which gives a picture of historical and current organization's operations (Date, 2003). It focuses on internal sources as quality control, sales, inventory and production. Data warehouse is designed to enable efficient management decision making as it presents a coherent picture of business conditions at a single point of time. It involves development of systems that enable extraction of data from operating system and installation of a warehouse database system that helps managers to access data in flexible ways.

However, Document warehouse is a repository for Business Intelligence keeping variety of document types from different sources to automatically extract and store the salient features of documents. It provides semantic integration of related documents of an enterprise. Document warehouse is a software framework for analysis, sharing and reuse of unstructured or semi-structured data as multimedia or text documents which is different from data warehouse where structured data as tabular sales reports are analyzed.

Data warehouse integrates multiple related data stores and provides managers volumes of information for analytical and decision support purposes. Its environment includes relational database, Extraction, Transportation, Transformation and Loading (ETL) solutions, client analysis tools and Online Analytical Processing (OLAP) engine. But Document warehouse leads to Enterprise Document Repository (EDR) of multiple unstructured information bases that helps in efficient, rapid and global access to information of various types in various forms.

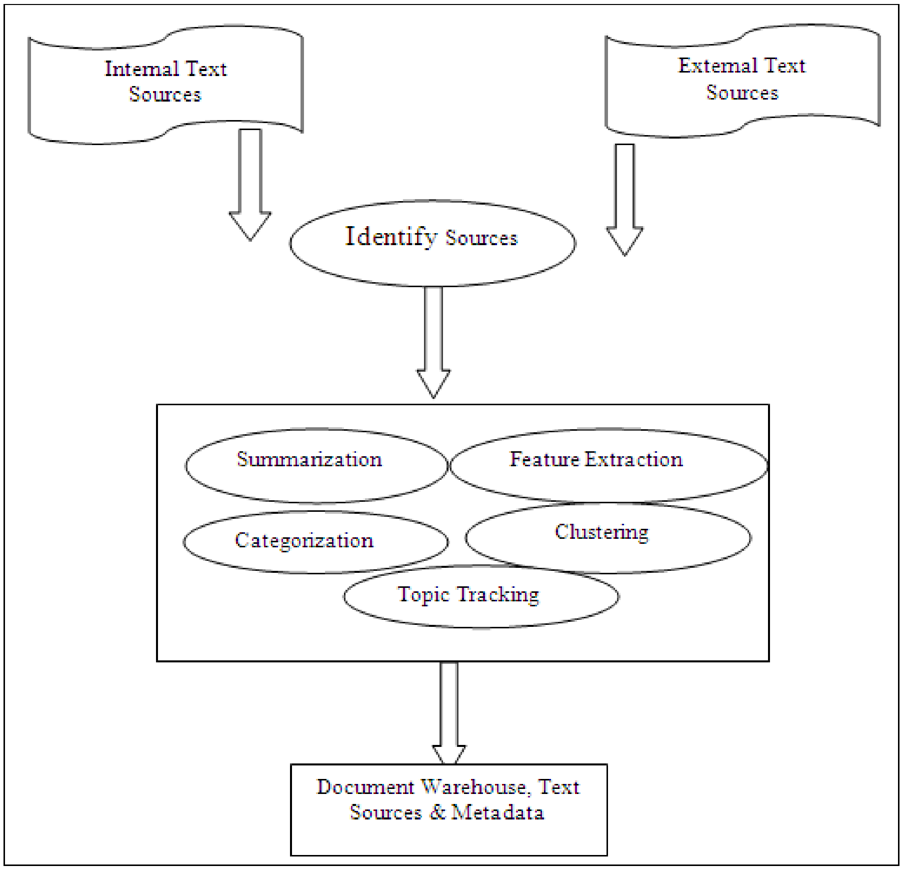

Figure 1 Steps to Construct Document Warehouse
(Source Gao Li, Chang Elizabeth, and Han Song, "Powerful Tool to Expand Business Intelligence: Text Mining", World Academy of Science, Engineering and Technology, 2005)

Data Warehouse provides information related to what, when, who, where and how aspects but document warehouse answers the main query of users that is "why". It gathers information from both internal and external sources to enable long-term strategic management. All data collected from various sources undergo summarization, categorization, feature extraction, clustering and topic tracking to give textual data for storage in document warehouse. This textual data can be in the form of normal documents, summaries, language translations or metadata. This process of constructing the document warehouse is explained in the Figure 1.

Text Mining Process

Whether text data is used for descriptive purposes, or for predictive purposes, or both, the following steps in the diagram are executed to fulfil text mining-

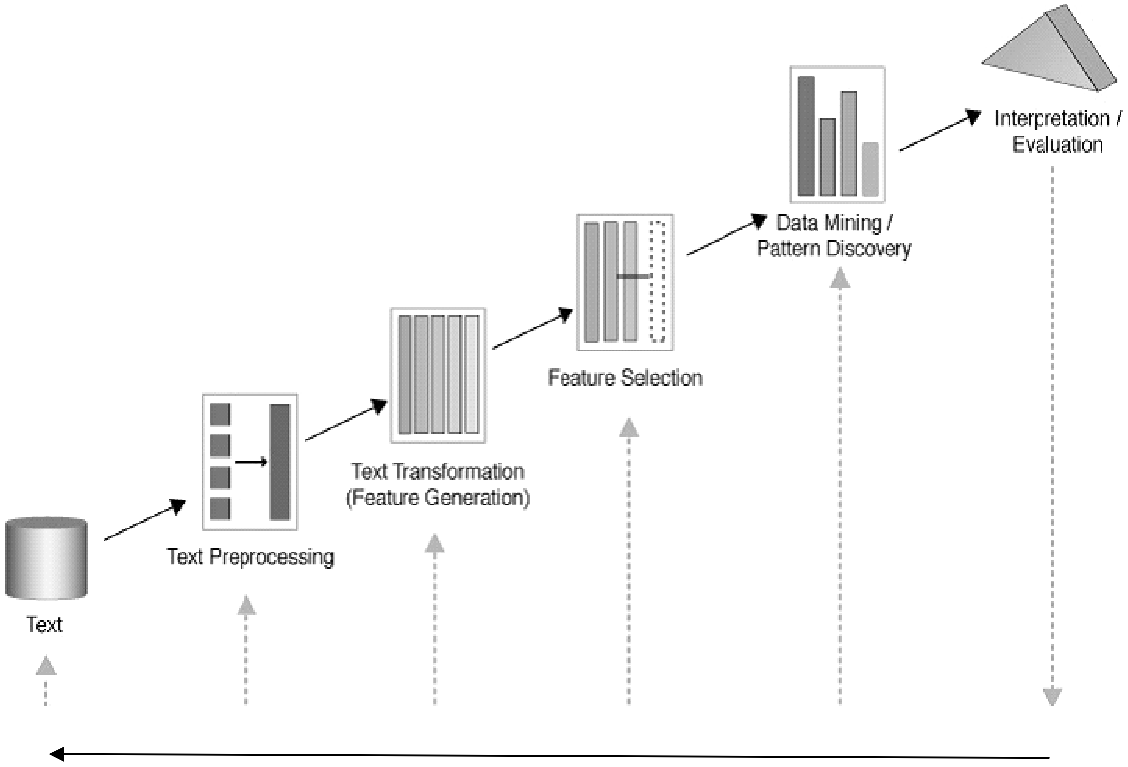
Interpretation / Evaluation
Data Mining / Pattern Discovery
Feature Selection
Text Transformation (Feature Generation)
Text Preprocessing
Text

Figure 2 Process of Text Mining

(Source Wasilewska Anita, "CSE-634 Data Mining: Text Mining", 2007 http://www.cs.sunysb.edu /~cse634/ presentations/TextMining.pdf)

- **Text-** It involves following steps-
  a. **Document Clustering**

The database textual data, whether unstructured or structured, first undergoes process of clustering which decomposes textual data and generates quantitative representation suitable for further analysis and decision making.

b. **Text Characteristics**

The text used for mining should allow dependency of words and phrases and allow different input modes for human or automated consumers. Text mining tries to find ambiguous, erroneous, misleading and unstructured text.

- **Text Pre-processing-** It is a extensive step involving following mechanisms-

  a. **Text Cleanup-** this process involves removal of advertisements from web pages and normalization of data so that the resultant text is free from redundancy and spurious text generation.
  b. **Tokenization-** it is parsing of unstructured text in which the text is split into sets of tokens on the basis of hyphens, apostrophes, white space and punctuations found in the text.
  c. **Parts of Speech Tagging-** in this process, grammar rules are applied on the text with their corresponding parts of speech.
  d. **Word-sense Disambiguation-** it allows finding different meanings or senses of a word implied in different situations.
  e. **Semantic structures-** it allows studying the semantics of text stored in the document warehouse and either performs full parsing and produces a parse tree for each sentence or does chunk parsing for nouns and verbs in each sentence. However, full parsing leads to grammatical mistakes and wrong sentence splits so partial parsing is preferred.

- **Text Transformation/ Attribute Generation-** Attribute generation generates labels/ attributes from the text based on their features. It involves following steps-

a. **Text Representation-** It allows representation of text by features and their occurrences using approaches of "Bag of words" and "Vector space" where each word is represented as an individual variable having numeric weight.

b. **Feature Generation-** this process selects features of a document so as to improve its representation which may be earlier misleading or redundant. It uses approaches of selection before use or selection based on use.

- **Feature Selection-** It involves reduction in text dimensionality and irrelevant attributes to deal with issues of scarcity of resources and feasibility (Liu and Motoda, 2003). It is used to improve text representation by selecting a subset of features either before using them in a classifier or how well they perform in classifier. Though the approach of selection before use is independent of many classifiers and has low costs in computation but it is less effective in comparison to the approach of selection based on use as it evaluates on performance.

- **Data Mining-** This is a pure application-dependent stage also called as Knowledge Discovery and Data Mining (KDD). It provides extraction of useful, valid, understandable patterns from databases, texts and web. It provides ways to make best use of data through rapid computerization (Pyle, 2003 and Dunham, 2005). Data Mining provides a number of tasks to get work done simpler which are discussed below-

Classification- To classify future data into known classes
Sequential Pattern Matching- it is based on the sequential rule A->B which implies that event B will always be followed by event A.
Association- It is a rule X->Y such that X and Y are data item sets.
Clustering- it involves finding clusters of related or similar traits in groups
Deviation Detection- to analyze and find the significant changes in data
Data Visualization- enables usage of graphical ways to show hidden patterns in data.
Data Mining basically analyzes large volumes of business numeric data to derive knowledge that can be used for competitive advantage.

- **Interpretation/ Evaluation-** It is the final step in text mining process which could lead to either termination or iteration. Text mining process could be finally terminated if well-suited results are achieved for business intelligence or the process could be finally iterated if the results are not up to the mark or are used as a part of further inputs.

Techniques in Text Mining to support BI

Text mining is the technology to extract knowledge from structured repositories and document warehouse enabling business intelligence operations. It is build on a number of techniques as discussed below-

1. **Clustering**

It is an automatic process that divides volumes of documents into groups that are related to each other on the basis of common themes or properties. It is a method to find out what a collection actually contains. An example of clustering in text mining is to analyze customers' emails and find the one's overlooked bearing a common theme. The goal of this analysis is to find a set of clusters having lesser intra-cluster similarity than inter-cluster similarity and finally try to maximize it. Thus, clustering eases the browsing process to find related information by identifying hidden similarities and over viewing contents of large documents.

Clustering is applied to identify properties of a set of data as date, cost etc. and divide them into clusters. These clusters or subsets can be used to find hidden similarities, provide brief on large database collection and simplify browsing process to find related or linking information. Clustering is used by Text Miner search engine which is supported by a robust algorithm to find groups that are more similar than other members of same or other groups. Text mining uses Hierarchical Clustering for textual data where it merges two similar clusters into one. This process continues until the final root cluster is obtained. So, in this process of hierarchy both inter and intra-cluster properties are easily revealed.

2. **NLP/ Computational Linguistics**

It is also called as Natural Language Processing (NLP) which combines computer's ability to process volumes of text at high speed and human's ability to understand natural language, spellings and contextual meanings. Computational Linguistic is a technique of text mining in which computer gets the ability to process, analyze and deduce patterns from natural language so as to enable algorithms for problems including context-sensitive grammar, part-speech tagging and bilingual dictionary creation and word-meaning disambiguation. Generally, a computer is speedy and accurate in its tasks but it understands only machine language whereas humans communicate in natural language. So, to bridge the gap between the two, computational linguistic generates statistics over large text collections in natural language using technologies as question-answering, topic tracking, concept linkage, summarization and information visualization. New tools are now designed in NLP to reach to more targeted information.

3. Categorization

This tool of text mining categorizes volumes of documents into "topics" or "themes" and provides a cheap alternative to cataloguing (Yang and Pederson, 1997). It is a kind of "supervised" learning in which categories are determined in advance for each document. Text categorization or text classification is the assignment of natural language documents to predefined categories according to their content (Sebastiani, 2002). It utilizes many machine learning techniques to provide automatic metadata extraction, document indexing and maintenance of large categories of web resources.

4. Information Retrieval

It is the basic and core technique of text mining enabling users to find documents that satisfies their information needs (Baeza-Yates and Ribeiro-Neto, 1999). Information retrieval is a broad field with many subject areas and has developed models for representing large collections of text such that users can find documents in particular topics of their interest (Gao, Chang and Song, 2005). Information retrieval utilizes techniques of vector space model to minimize cost of document and query representation by calculating Euclidian distance between vectors representing documents and query. Latent Semantic Indexing is another Information retrieval technique dealing with synonymy and polysemy problems in text.

IR is actually used as a first step for searching data in document warehouse by researchers or users to reach to a solution of a particular problem. Though it retrieves loads of information from warehouse, it may not highlight the exact topic of interest to the user which he wanted to extract. So, IR poses problems in representing and identifying documents related to a particular set of topics.

5. **Pattern Recognition**

It is a text mining technique providing pattern searching with words, morphological and syntactic properties. At one end, it provides word/ term matching which is easier to implement but requires a lot of manual intervention while at the other end it allows pattern searching based on relevancy signatures (Riloff and Lehnet, 1994) provided by speech tagger. It is quite different from pattern matching which is to find regular expressions in programming languages as it mainly finds pre-defined patterns in text.

Applications of Text Mining in Business Intelligence

- Scientific Data Analysis
- Marketing, identifying potential customers and market segmentation
- Biomedical Sciences
- Document warehouse for SAP, a FileNet's commercial Software that enables SAP's business applications to access document images.
- Fraud detection

Developing Text Mining Solution for Business Intelligence

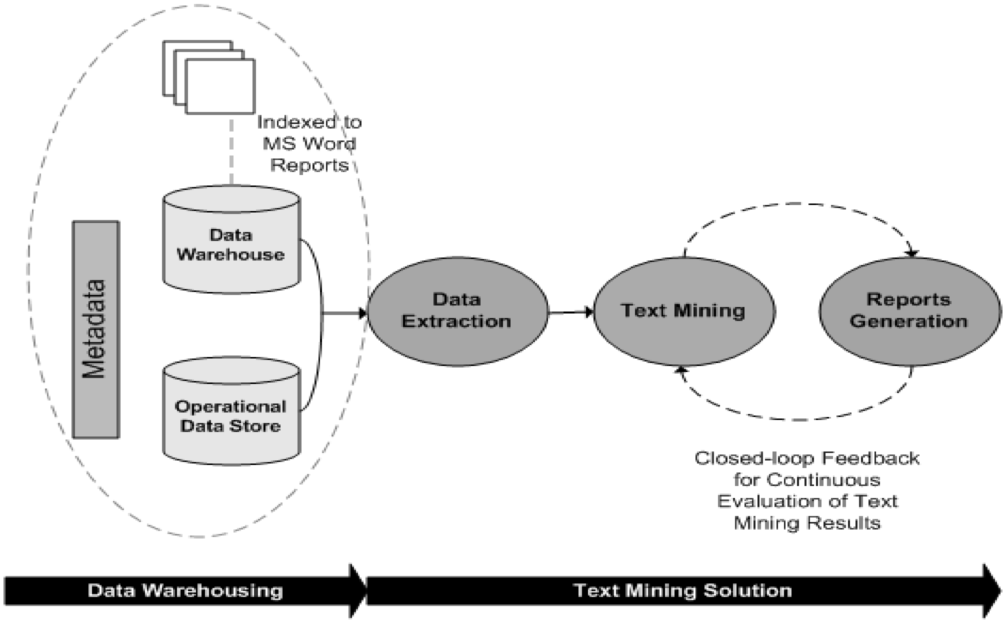

Figure 3 Text Mining Modules for Business Intelligence
(Source- Mailvaganam Hari, Text Mining for Fraud Detection- Creating cost effective data mining solutions for fraud analysis", 2007 http://www.dwreview.com/Data_mining/ Effective_ Text_Mining.html)

The customized solution as explained in figure above is developed into three modules-

1. **Metadata-** It manages enterprise databases in form of data warehouse and document warehouse which are indexed to MS Word reports.
2. **Data Extraction-** It pulls reports, manually or as a scheduled task from data warehouse repositories and stores them using the scripting language as Perl that enables changes by end users.

3. **Text Mining Solution-** this module contains data mining reports with likelihood of frauds and thus gives the client a competitive advantage and technical details that are kept as secret in the corporate for current and future trading. Text mining involves risk analysis and assessment which guides to quick, efficient, correct and fruitful decisions. It keeps an account of current and future risks, benefit of making one decision over other and the cost of taking one action over another.

Conclusion

Text Mining is an important aspect of Business Intelligence that helps users and enterprises in analyzing stored text in a better way so as to make better decisions, improve customer satisfaction and gain competitive advantage. It is better than data mining as it provides deeper insight into the expanding business domain and extracts more fruitful data for business intelligence. Its main objective is to derive new information from multi sources of raw text information which was never before thought of or faced. Text mining uses a number of software and technologies to help decision support system of an enterprise and keeps on generating alerts on market changeovers, mergers, poor performance and competition that in turn help the business to take corrective, measurable and preventive steps and be the leader.

References

Baeza-Yates, R. and Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Company.

Bergeron, P. and Hiller, C.A., 2008. “Competitive intelligence”, in B. Cronin, Annual Review of Information Science and Technology, Zedford, N.J.: Information Today, vol. 36, chapter 8.

Date C., 2003. Introduction to Database Systems. 8th ed., Upper Saddle River, N.J.: Pearson Addison Wesley, 2003.

De Ville, B., 2001. Microsoft Data Mining Integrated Business Intelligence for E-Commerce and Knowledge Management. Boston: Digital Press.

Dunham, M. H., 2005. Data Mining-Introductory and Advanced Topics. Prentice Hall.

Frank, E., Paynter, G., Witten, I.H., Gutwin, C. and Nevill-Manning, C. “Domain-specific keyphrase extraction.” Proc Int Joint Conf on Artificial Intelligence IJCAI-99. Stockholm, Sweden, pp. 668-673, 1999

Gao L., Chang, E. and Song, H., 2005. Powerful Tool to Expand Business Intelligence: Text Mining. World Academy of Science, Engineering and Technology.

Hart R.P., “The Text Analysis Program”, DICTION 5.0, Thousand Oaks, Calif.: Sage.

Lenat, D.B., 1995. CYC: A large-scale investment in knowledge infrastructure. Comm ACM, Vol. 38, No. 11, pp. 32–38.

Liu, H., Motoda, H. and Yu, L., 2003. “Feature Extraction, Selection, and Construction”, in N. Ye, editor, The Handbook of Data Mining, pp. 2- 41. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003.

Mailvaganam, H., 2007. Text Mining for Fraud Detection- Creating cost effective data mining solutions for fraud analysis. [Online] http://www.dwreview.com/Data_mining/Effective_ Text_Mining.html [Accessed on 27 November 2011]

Pyle, D., 2003. Business Modeling and Data Mining. Morgan Kaufmann, San Francisco, CA.

Riloff, E., 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, AAAI Press/The MIT Press, pp. 811-816.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47.

Wasilewska, A., 2007. CSE-634 Data Mining: Text Mining. [Online] http://www.cs.sunysb.edu /~cse634/ presentations/TextMining.pdf [Accessed on 26 November 2011]

Yang, Y., Pederson, J.O., 1997. A comparative study on feature selection in text categorization. Morgan Kaufmann, pp. 412-420