Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2006 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2006

Information Mining: Integrating Data Mining and Text Mining for Business Intelligence

Quanzhi Li
New Jersey Institute of Technology

Yi-fang Brook
New Jersey Institute of Technology

Follow this and additional works at: http://aisel.aisnet.org/amcis2006

Recommended Citation
Li, Quanzhi and Brook, Yi-fang, "Information Mining: Integrating Data Mining and Text Mining for Business Intelligence" (2006). *AMCIS 2006 Proceedings*. 182.
http://aisel.aisnet.org/amcis2006/182

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Information Mining: Integrating Data Mining and Text Mining for Business Intelligence

Quanzhi Li
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102
quanzhi.li@njit.edu

Yi-fang Brook Wu
Information Systems Department
New Jersey Institute of Technology
Newark, NJ 07102
wu@njit.edu

ABSTRACT:

Data mining and text mining can help decision makers obtain business intelligence and make informed decisions, but using one of them gives us only a partial picture. The application of data mining can lead to questions that cannot be answered with only numbers. Therefore, decision makers will need text mining to drill the textual data to find explanations for numbers. On the other hand, the application of text mining will also raise questions that cannot be answered with only text. We need to examine and utilize findings from both. However, most of the current text mining applications and data mining applications are not integrated. In this paper, a framework for combining these two technologies is described. In this framework, a taxonomy complemented by feature indexing and full-text indexing will bridge data mining and text mining. The technical challenges of the integration are also discussed.

Keywords:

Data mining, text mining, information mining, taxonomy, business intelligence, feature indexing.

INTRODUCTION

Data mining and text mining are becoming more and more popular in recent years. These two technologies can help people obtain business intelligence. In a broader point of view, data mining may consist of the mining of structured data and also unstructured data. Structured data refers to the data stored in the transactional database, data mart, or data warehouse, such as the numeric data. Unstructured data mainly refers to the textual data. Examples of unstructured data are project proposals, emails, business memos, reports, and news articles. They are considered unstructured, because they cannot be easily processed or managed like the numerical data stored in the relational database, and their attributes, such as the meanings of words and the relationships between words, are not easily understood or manipulated by computer programs. In this broader sense, data mining includes text mining. But more often, researchers take a narrower point of view on these two terms, considering data mining and text mining as two different concepts - data mining refers to discovering knowledge from only the structured data and text mining refers to finding knowledge from the unstructured textual data. In this paper, we consider data mining and text mining as two different concepts. Other terms bearing similar meaning with data mining include knowledge extraction, knowledge mining, data analysis and pattern analysis (Chang et al, 2001; Fayyad and Uthurusamy, 1996).

Data mining became popular long before text mining, but text mining is also becoming popular in recent years and more and more researchers have focused on this area. It has begun to work together with data mining in discovering business intelligence. The growing interest in integrating unstructured text (or text mining technology) in business intelligence is due to a number of factors. According to independent studies by Delphi Group, approximately 80 percent of all enterprise data are unstructured. Using unstructured data along with structured data can improve business analysis (Inmon 2004; Creese 2004; Sullivan 2001; Unitas 2002). Analysis in structured data can lead to questions that cannot be answered by just working on numbers. This will lead the decision makers to research on the unstructured data, which may in turn raise questions that need to be checked against the structured data again. Therefore, these two technologies can work together to help the users make wiser decisions. For instance, they have been used together to improve fraud detection in the property casualty insurance industry (Sullivan, 2001). The term "information mining" is used to refer to the integration of data mining and text mining technologies to provide decision makers business intelligence. Figure 1 shows that, to obtain business intelligence, decision

makers need to work with both data mining on structured data and text mining on unstructured data (Sullivan, 2001). In this paper, the phrases "integrating data mining and text mining" and "integrating structured data and unstructured data" are used interchangeably, since data mining works on structured data and text mining works on unstructured data.

The problem with current development of data mining and text mining techniques is that they are not semantically integrated together to provide business intelligence. If there a question is raised during the data mining process, users need to leave the data mining system and enter the text mining system to do further research to find an answer, and vice versa. In this paper we try to propose a framework, which can integrate the two systems together to provide a unified platform for users. When users encounter a question in one system and want to further explore it using the other kind of technology, the unified system can automatically bring users the relevant information. For example, if a user is analyzing a product's market using data mining techniques and needs textual data related to this product to do further analysis, then the unified system can automatically bring relevant documents or document metadata to the user, and the user can further drill down the textual data using text mining tools. As mentioned before, we call this unified mining process as information mining. Besides the proposed framework, in this paper we also discuss some challenges of implementing this integration.

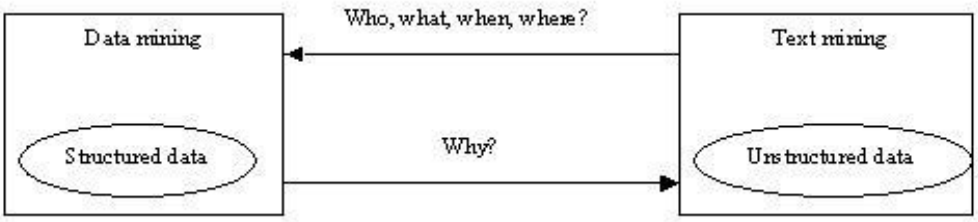

Figure 1. Data mining and text mining work together to get business intelligence

PREVIOUS STUDIES

Some researchers consider data mining a synonym of knowledge discovery in database, but others consider it only a key step of the process of knowledge discovery in databases. A typical knowledge discovery in databases includes two main steps. The first one is data preprocessing, which includes data cleaning, data transformation and data reduction. The other one is data mining, which includes pattern discovery from data, pattern evaluation and knowledge presentation (Chang et al, 2001). Some of the major pattern discovery processes of data mining are: concept description (Cleveland, 1993), association rule mining (Agrawal et al, 1993), data classification, data clustering and predication. Many previous studies on data mining have focused on association rule mining.

The goal of text mining is the discovery, recognition or derivation of new information from large collection of text (Hearst, 1999). Text mining goes far beyond merely using statistical models, which are often done with text files. The typical tasks of text mining include association discovery, trend discovery and event discovery (Chang, 2001). Sometimes people also consider document classification and clustering as one type of text mining techniques. Text mining exploits some techniques and notations of data mining, but there are differences between these two kinds of knowledge discovery processes. The biggest difference occurs during the data preparation and feature selection step (Dorre 1999). As text mining deals with unstructured textual data, there are no explicit features that can be directly used for mining. Therefore, a text mining system has to prepare the text by adding a very complex feature extraction function.

Several previous studies have mentioned the possibility of linking data mining with text mining (or linking structured data with unstructured data). Sullivan (2001) points out that one possible way of combining them is to link the terms in data warehouses with the terms in document warehouses, making a connection between the structured and unstructured data. Inman (2004) mentioned that one possible approach to bridge the two kinds of data environments (structured and unstructured) is by means of a search engine. With a search engine, data can be sought in one environment and brought to the other environment. In Unitas's published white papers (Unitas, 2001, 2002), three kinds of integration strategies for combining structured and unstructured data are mentioned: front end integration, middle tier integration and back end integration. In the front end integration, the user requests are translated into queries (structured and unstructured) and submitted to the underlying search engines. In the middle tier integration, search engines are built for structured data and unstructured data separately. These two kinds of data are handled separately and not integrated together in this tier. In the back end integration approach, they propose to migrate the unstructured data into a format that can be easily indexed and queried, for example, putting them in content warehouses. Previous studies did not describe the details on how to combine

these two kinds of technologies, and they did not provide a framework or a clear integration method to follow. In this paper, we will present a framework on how to integrate these two kinds of mining technologies, and describe what kinds of approaches to use to seamlessly link these two kinds of data.

A FRAMEWORK FOR INTEGRATING DATA MINING AND TEXT MINING

Figure 2 shows the high-level integration architecture. For a typical integration, there are three levels. All these three levels work together to provide a unified information mining solution. The lowest level consists of a data mining system and a text mining system. They provide the basic functions of data mining and text mining. These two systems may reside in two separate places. For example, each of them may have its own operating system, database, server, and other supporting services. On the other hand, they may also be on the same machine, sharing the same operating system and other supporting services, but not semantically integrated. The middle level is the integration level. It semantically integrates the two low-level systems, building links between them. The highest level is the information mining user interface. It provides a coherent, unified interface for decision makers. In the following subsections, we will describe the three levels in details, with emphasis on the middle level, which is the core part of the integration.

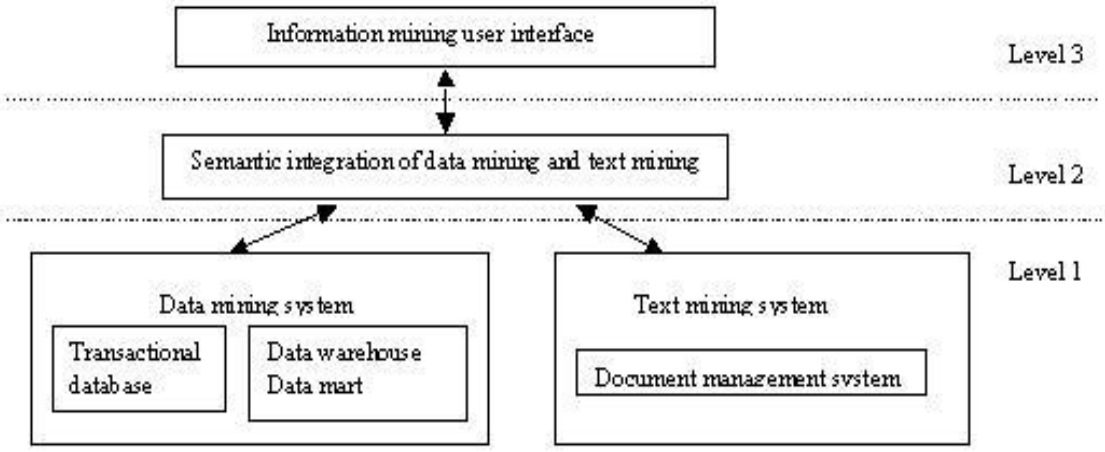

Figure 2. The framework of combining data mining and text mining technologies

Level 1: The Data Mining System and the Text Mining System

In this section, we describe the main components of level 1. Level 1 is the foundation of the whole information mining system. As we mentioned in last section, data mining mainly deals with the structured data, while text mining focuses on the unstructured data. On the data mining side, there are usually the following components: data warehouses, data marts, transactional databases, text mining tools for mining the structured data, etc. The typical mining processes include association rule mining, concept description mining, clustering, classification and prediction. Compared to text mining, the structured data and data mining techniques have been well explored by previous studies; therefore, in this section we only describe the components on the text mining part in details.

The text mining side usually includes a document management system (or content management system or document warehouse) and some text mining tools for extracting knowledge form the textual documents. In this paper, the terms *document management system, content management system* and *document warehouse* are used interchangeably, but usually a document management system or content management system has more functions than a document warehouse. A document management system integrates internal documents as well as documents from external sources. Only internal or only external documents can only provide limited analysis or business intelligence. When the textual data from different sources are analyzed and integrated together, more values will be obtained. For example, if the internal documents on marketing and sales are integrated with outside industry analysis, market reports or competitor's advertising information, we may do more analysis and discover more valuable information. Some organizations may already have a functioning document management system; some others may just have some internal documents, such as project reports and business contracts, organized into different file folders. Figure 3 shows the typical components on the text mining side. They are also described below.

Document Repository: both internal and external documents should be retrieved, preprocessed and stored in the repository. For some external documents, if it is impossible to store the full-text documents, their metadata should be extracted to represent them. A retrieval service can be used to collect useful documents from outside.

Metadata Repository: depending on the organization and the domain it belongs to, there may need different kinds of document metadata. Some typical ones are document title, source, format, creation date, authors, keywords and summary. Some metadata are easy to extract, such as titles and author information, but some other metadata, such as document keywords and summary, may need complicated natural language processing and text mining techniques. For example, most documents do not have author-assigned keywords, and so some text analysis techniques should be applied to automatically extract keywords for these documents. Several document keyword extraction algorithms have been proposed in previous studies (Wu et al, 2005; Frank et al, 1999; Turney, 2000).

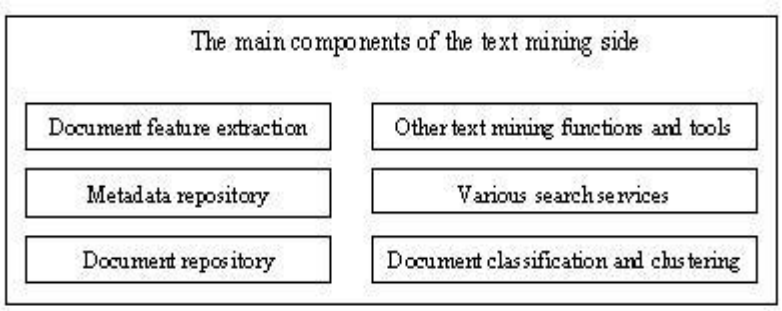

Figure 3. Key components on the text mining side

Document Feature Extraction: for different organizations and domains, different types of entities need to be extracted from the documents. Some examples are person name, location (state, city), time, company name, and product name. Information extraction, pattern recognition, and machine learning techniques can be applied to extract special features from text.

Document Classification and Clustering: all the documents should be classified into predefined categories or be clustered into clusters based on how similar they are with each other. Documents can be classified or clustered based on their full text or their metadata, such as document keywords or abstracts.

Search Services: Search services should provide enough flexibility to the decision makers. The necessary search services may include document full-text search, metadata search, search by example and search by browse. "Search by example" means the system will find objects similar to the example provided by the user. "Search by browse" refer to the mechanism that users can find what they want by browsing a directory or other kinds of hierarchical structures, like Yahoo's directory.

Other Text Mining Functions or Tools: besides the common statistical methods that can be applied to textual documents, the system should also have the ability to fulfill the typical text mining processes, such as association rule discovery and trend discovery. Association rule discovery may be used to find relationship between objects from the text collection. Knowledge bases may be built based on the relations of various objects. For example, based on the relationships among different concepts (terms) discovered by association rule mining technique or other text analysis algorithms, we may build a concept hierarchy for a research domain, an organization or a document collection (Lin et al, 1998; Feldman et al, 1996; Bot et al 2005). For some organizations, event discovery may also be necessary. For instance, based on the analysis of competitors' advertisements, we may find new events, such as a new product launch or strategy changes of competitors. Usually a shift in vocabulary and changes in term frequency distribution indicate a new event.

In this section we have described the fundamental components of information mining: data mining, text mining and other support components that facilitate the data mining and text mining processes. Next section will present the methods used to integrate these two kinds of mining.

Level 2: Integrating Data Mining and Text Mining

To semantically link the structured data and unstructured data to provide a unified information mining platform, we propose to use the following methods to integrate them together: a taxonomy to cover and index all the important entities from both sides, some feature indexes to index special objects not covered by the taxonomy (e.g. customer names), and a full-text index to index all the terms that are not covered by the taxonomy and feature indexes. These three methods complement each other and work together to build a semantic link between the structured data and the unstructured data. They are further explained below.

Taxonomy

Taxonomy sometimes is also called ontology, thesaurus or knowledge base. It is a knowledge representation scheme for organizing relationships between objects or concepts (terms). A term in a taxonomy usually has parent terms, which are broader concepts, and child terms, which are narrower concepts. It also has synonyms and related terms.

Taxonomy is one of the main methods we propose to integrate the two kinds of mining systems. It will link the data mining side and the text mining side. It is a bridge between the structured data and the unstructured data. The taxonomy used to integrate the two sides is usually domain-specific (or organization-specific), and contains the important domain-specific terms. It will try to cover all the meaningful entities from both sides. For each term in the taxonomy, its parent terms, child terms, synonyms and related terms will also be included in this taxonomy, if they exist and can be identified automatically or manually. This will make it possible to provide more options (functions) when connecting the structured data with the unstructured data, and vise versa. For example, suppose a sales person is using data mining technique to analyze the trend of sale of certain kind of car, e.g. "Honda Civic," and she/he also wants to analyze the related textual data to get a better idea. A rich taxonomy will make it possible to link this product to not only the textual data directly related to it (e.g. documents about "Honda Civic") but also other related textual data (e.g. documents about "Honda," documents about "Civic LX" or "Civic EX,"or documents about "Toyota Corolla").

Each term in the taxonomy has a data structure attached. Among other information, this data structure also holds two kinds of pointers, one of which points to the data on the data mining side (e.g. the name of a database table field), and the other points to the textual data on the text mining side. For instance, the pointer to the text mining side may point to a folder where documents related to this term are located; it may also points to a single document, summary of a document or other kinds of metadata. It is also possible that there are several small-scale taxonomies, instead of one, for the whole structured data and unstructured data, and each of them focuses on a sub domain or division. For example, if an organization sells both computer software and hardware products, two taxonomies may be built, one for software products and concepts, and another one for hardware.

The suggested possible approaches to build a taxonomy are:

1. Building the taxonomy manually.

   The builders should try to place all the meaningful entities from the structured data side (table names, table field names, meaningful string data in the tables, etc) and the unstructured data side in the taxonomy. They also need to manually identify terms' synonyms, parent terms and child terms, and other related terms. This manual process is time-consuming and costly, but the created taxonomy will have a clear structure and high precision.

2. Building the taxonomy automatically

   Taxonomy building tools may be used to automatically extract meaningful entities from the structured data side and the unstructured data side. For some special table field names (e.g. a field named "F_Name," which refers to "First name"), some natural language processing techniques need to be applied (e.g. analyzing the functional descriptions of some table fields) to find their real meanings and represents them appropriately. The edit distance technique can also be used to match table filed name with taxonomy entries. One good thing with the automatic method is that, after a period of time, when new structured data or unstructured data are added to the system, the taxonomy can be updated automatically. The taxonomy building tools should be able to handle single words and also phrases, since many concepts or entities are represented by phrases instead of single words. Usually, automatic methods can easily find a term's related terms, but it is difficult to find synonyms, parent terms and child terms accurately. And the built taxonomy will have a lower quality than the one manually built.

3. Building the taxonomy semi-automatically

   A. Manually first, then automatically

      The experts manually build a simple high-level taxonomy first, especially for the entities of the structured data side, which are difficult to process by automatic methods. Then the automatic methods are used to find the related terms from the textual documents or structured data for the manually identified taxonomy entries.

   B. Automatically first, then manually.

      A taxonomy is built using automatic methods first. Then experts will correct the errors and add other necessary terms.

Feature indexing

In this paper, feature indexing refers to the indexing of some special entities. Some special entities may not be identified and organized in a meaningful way by the automatic taxonomy building systems. Examples are customer names (person names or company names), product names or location names. Whether or not these kinds of entities can be organized in the taxonomy depends on a few factors, such as the taxonomy building methods, the characteristics of the documents, and the domain of the organization. In the integration of data mining and text mining, feature indexing is a complement to the taxonomy. This kind of index connects the special entities on the data mining side with the corresponding entities on the text mining side. Each index item also has two kinds of pointers, one of which points to the structured data side, and the other points to the corresponding objects on the textual data side. Feature indexing is especially useful for certain organizations. The following scenario is an example. Suppose we have a feature index of all the customer names, and after looking at a customer's structured payment history, we have some questions needed to be answered by looking at the textual documents related to this customer (e.g. contracts). From the information mining system interface, by just clicking on some contextual menus, the textual information related to this customer will be displayed to us automatically. In this example, the underlying connection between the structured data and the textual data is the feature index. It makes this customer's structure data (the payment history) dynamically liked to the related textual information (e.g. contracts).

Full-text indexing

Full-text indexing is a complement to the taxonomy and feature indexing. It is called full-text indexing because it will index all the terms appearing in the textual documents. The terms may be only single words or it may also include multiple-word phrases. The full-text indexing will also index all the terms on the structured data side that are not covered by the taxonomy or the feature indexes. Because this full-text index will cover all the terms on the unstructured data side, this index can be combined with the full-text index of the unstructured data side if there already exists one. Usually before integrating both sides together, there is already one full-text index for all the documents on the unstructured data side to serve the search services and other functions of the document management system. Therefore, very often what we need to do is to update the existing full-text index with the terms from the structured data side. The updated index can serve as both a bridge of the two sides and as well as a full-text index for the general search services on the text mining side.

Level 3: Information Mining Interface for Business Intelligence

Level 3 is the highest level of the integration. It is the interface between the underlying systems and the decision makers. When a user is working with the information mining system, the system can automatically, dynamically generate link or menu information (e.g. dynamic links, popup windows, menu in context, etc) based on what the user is working on. It fulfills this by detecting the objects (context) the user is working with and then accessing system level 2, the integration layer, to see if there is any related information from the structured data side or the unstructured data side that should be provided to the user. The interface can be web-based or otherwise.

THE CHALLENGES

As discussed in previous sections, integrating data mining system with text mining system can bring us benefits, but nowadays most of organizations still run their structured data system and unstructured data systems separately. They are not semantically integrated together. The following challenges, among others, are the reasons behind the scene

1. Lack of the underlying fundamental systems. Some organizations may have only data mining system or data warehouse system, but not text mining system or document management system. Building a text mining/document management system is a more challenging task than building a data mining/data warehouse system, because text mining is a relatively new concept compared to data mining and data warehouse, and processing textual documents from different sources is more difficult. Some organizations only have some internal documents, without any processing of these documents and metadata. On the other hand, although it is easier to build a data warehouse than a document warehouse, to integrate different database schemas from multiple departments or applications into one data warehouse is not an easy job. Without a good data warehouse or document warehouse system as the underlying support, it is not easy to integrate text mining system and data mining system.
2. Building taxonomy is a challenging job. As described before, although there are many approaches to build a taxonomy, none of them is a perfect solution. Without a complete, accurate, easy-updated taxonomy, we cannot effectively integrate the structured data and unstructured data. In some organizations, each department has their own taxonomy. How to integrate these taxonomies into one organization-wide taxonomy is also a question.

3. How to match the taxonomy with the database schemas. The taxonomy and database schemas have different structures and logics, no matter it is a relational database schema or a star schema for a data warehouse. How to do the mapping and matching between these different structures is a research question and worth further exploring.

CONCLUSION AND FUTURE RESEARCH

In this paper, we propose a framework to integrate text mining and data mining techniques to form a unified information mining system for decision makers to obtain business intelligence. One of the main components in this integration framework is taxonomy, which is used to semantically link the structured data in the data mining side and the unstructured data in the text mining side. One of our future research topics is to explore the feasibility and effectiveness of some automatic methods of building taxonomies. Another future research direction is to implement the framework proposed in this paper in one or two domains and evaluate its effectiveness.

ACKNOWLEDGMENT

Partial support for this research was provided by the United Parcel Service Foundation, the National Science Foundation under grants DUE-0226075, DUE-0434581 and DUE-0434998, and the Institute for Museum and Library Services under grant LG-02-04-0002-04.

REFERENCES

1. Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of data*, May, Washington D.C., 207-216.
2. Chang, G., Healey, M. J., McHugh, J. A.M. and Wang, J. T. L. (2001) Mining the World Wide Web: An information search approach. Kluwer Academic Publishers. Norwell, MA.
3. Bot, R.S, Wu, Y.B., Chen, X., Li, Q. (2005) Generating Better Concept Hierarchies Using Automatic Document Classification, in Proceedings of ACM Fourteenth Conference on Information and Knowledge Management (CIKM 2005), Bremen, Germany.
4. Creese, G. (2004) Duo-Mining: Combining Data and Text Mining, http://www.DMReview.com.
5. Cleveland, W. (1993) Visualizing data. Hobart Press. Summit, New Jersey.
6. Dorre, J., Gerstl, P., and Seiffert, R. (1999) Text mining: Finding nuggets in mountains of textual data, In *Proceedings of the 5th ACM SIGKDD*, 398-401.
7. Fayyad, U. and Uthurusamy, R. (1996) Data mining and knowledge discovery in database. *Communication of the ACM*, 39(11):24-26.
8. Feldma, R., and Hirsh, H. (1996) Mining associations in test in presence of back-ground knowledge. In *Proceeding of the 2nd International Conference on knowledge discovery and Data mining*. Portlan, Oregon, 343-346.
9. Frank, E., Paynter, G., Witten, I., Gutwin, C., and Nevill-Manning, C. (1999) Domain-specific keyphrase extraction. *Proceeding of the sixteenth international joint conference on artificial intelligence*, San Mateo, CA, 668-673.
10. Hearst, M. A. (1999) untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, August, College park, Maryland.
11. Inman, B. (2004) managing unstructured data, Information Lifecycle Mangement (ILM) newsletter
12. Lin, S.-H., Shih C.-S., Chen, M.C, Ho, J.M. Ko, M.T., and Huang, Y.M. (1998) Extraction classification knowledge of Internet documents with mining term association. *In proceedings of the 21st Annual International ACM SIGIR conference*, August, Melbonrne, Australia, 241-248.
13. Sullivan, D. (2001) Document warehousing and text mining. Wiley Computer Publishing. Canada.
14. Turney, P. D. (2000) Learning algorithm for keyphrase extraction. *Information Retrieval*, 2(4), 303-336.
15. Unitas (2001). Managing Unstructured Content in Enterprise Information Portals (EIPs), http://www.unitas.com, 2001.
16. Unitas (2002). A Single View: Integrating Structured and Unstructured Data/Information within the Enterprise, http://www.unitas.com, 2002.
17. Wu, Y.B, Li, Q. Bot,R. and Chen, X. (2006) Finding nuggets in documents: A machine learning approach, Journal of the American society for information and technology (JASIST). Volume 57, Issue 6, 740-752.