

# **An Approach to Text Mining using Information Extraction**

**Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis**

Centre for Research in Information Management  
Department of Computation, UMIST, P.O. Box 88,  
Manchester, M60 1QD, UK

Email: {karanik, christos and babis}@co.umist.ac.uk <http://www.crim.org.uk>

## **Abstract**

In this paper we describe our approach to Text Mining by introducing TextMiner. We perform term and event extraction on each document to find features that are likely to have meaning in the domain, and then apply mining on the extracted features labelling each document. The system consists of two major components, the Text Analysis component and the Data Mining component. The Text Analysis component converts semi structured data such as documents into structured data stored in a database. The second component applies data mining techniques on the output of the first component. We apply our approach in the financial domain (financial documents collection) and our main targets are: a) To manage all the available information, for example classify documents in appropriate categories and b) To “mine” the data in order to “discover” useful knowledge. This work is designed to primarily support two languages, i.e. English and Greek.

## **1.Introduction**

The explosive growth of databases in almost every area of human activity has created a great demand for new, powerful tools for turning data into useful knowledge. To satisfy this need researchers from various technological areas, such as machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, econometrics, information retrieval, information extraction etc., have been exploring ideas and methods. All these efforts have led to a new research area often called data mining (DM) or Knowledge Discovery in Databases (KDD). We can define it as the nontrivial extraction of implicit, previously unknown, and potentially useful information from given data (Piatetsky-Shapiro and Frawley [1991]).

Until now, most of the work in Knowledge Discovery in Databases (KDD) has been concerned with structured data. However, a lot of information nowadays is available in the form of text, for example in documents, manuals, email, presentations on the Web, and so forth. Unlike the tabular information typically stored in Traditional Databases, this form of information has only limited internal structure. In order to take advantage of all this information we apply Text Mining (TM) technology.

## **2.The innovation of our approach**

Text Mining uses unstructured textual information and examines it in attempt to discover structure and implicit meanings “hidden” within the text. Most approaches to TM apply mining algorithms on labels associated with each document. These labels may be keywords extracted from the document or just a list of words within the document of interest.

In our approach the main contribution is that we apply mining algorithms on terms (meaningful sequence of words e.g. Department of Computation) combined with events (meaningful set of terms e.g. take-over, company1, company2) extracted from the documents. We believe that the most characteristic factors, which describe a document, are the terms and events mentioned within the document. It is very important to extract the “right” features in order to have accurate and useful results. Information Extraction (IE) is the technology [6], which involves in this pre-processing step. The resulting document representations are used as input to a clustering algorithm. We have developed a clustering algorithm appropriate for categorical data. Using this algorithm we are in a position to discover structure within the document collection. In addition classification techniques are used to further validate the results from clustering.

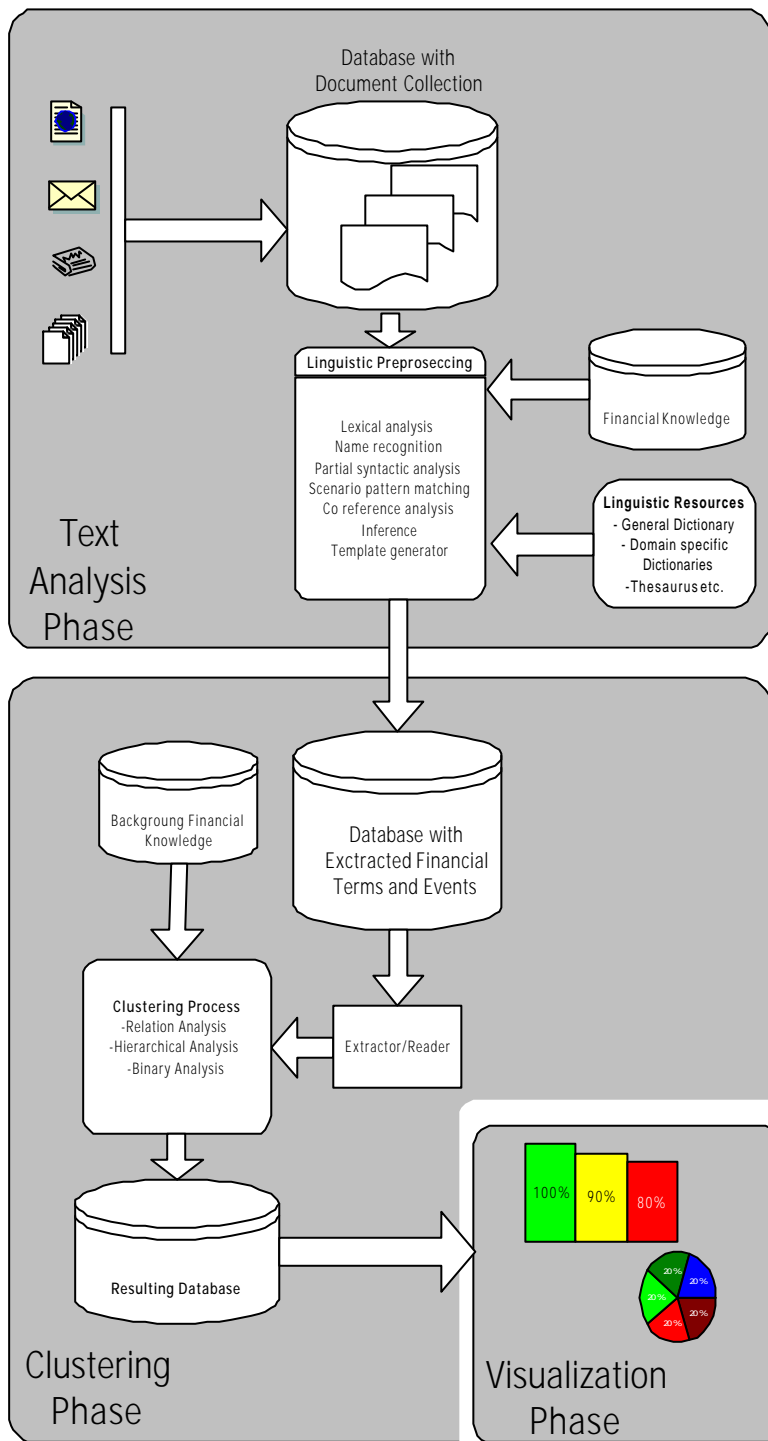


Figure 1 TextMiner Architecture

### **3. Case study**

As we have mentioned above the domain in which we are going to apply the TextMiner approach is the stock market. The domestic capital markets are of the most valuable financial institutions for the global economy. In micro-economic level the capital market offers to companies the opportunity to gather low priced investment capital. In macro-economic level the domestic capital market accumulates significant valuable income, as a result of the daily operations, which through the taxation assists the reduction of the state deficits. In addition in macro-economic level the investments contribute to the improvement of the economy's competitiveness, create new jobs and forward the exportation of products and services. As a consequence they increase the inflow of exchange for the country.

Users of financial news (Financial newspapers, news agencies, etc.), financial analysts, employees of Securities Companies and individual' users with financial knowledge, have to deal daily with vast amounts of information. All these users know that the progress of the capital market depends a lot on relevant news. That's why they have to analyse all the news that appear on newspapers magazines and other textual resources. Until now the processing of the unstructured textual databases has to be done manually by the users.

Our target is to automate much of the users' work. Firstly we want to manage information stored in textual databases (collection of documents) and secondly to extract useful knowledge.

### **4. Information Extraction**

Information Extraction is the mapping of natural language texts (such as newswire reports, newspaper and journal articles, electronic mail, World Wide Web pages, any textual database, etc.) into predefined, structured representation, or templates, which, when filled, represent an extract of key information from the original text [14]. The information concerns entities of interest in the application domain (e.g. companies or persons), or relations between such entities, usually in the form of events in which the entities take part (e.g. company takeovers, management successions etc.). Once extracted, the information can then be stored in databases to be queried, data mined, summarised in natural language, etc.

The first necessary step is the linguistic pre-processing. It consists of a number of linguistic techniques such as tokenization, part of speech tagging, lemmatization etc in order to feed the information extraction system.

We aim to extract from the documents, financial terms (Company names, actions between companies, people, places, exchanges etc.) and financial events of interests. We are interested in specific events between specific companies/persons. In other words, a specific event (e.g. take-over) between company A and company B will be used as a feature to label the document as well as the same event (i.e. Take-over) between different companies.

In our case we define a number of events within the financial domain. We keep this information in a table called Event Type. Each event has a different number of attributes to be described.

For example:

<b><u>Event type:</u></b>	Take-over.
<b>Date:</b>	March 15
<b>Company Predator:</b>	Neurosoft SA.
<b>Company Target:</b>	IBM.
<b>Type of takeover:</b>	Friendly
<b>Value:</b>	240,000,000 pounds
<b>Document ID:</b>	Doc1
...	

We have described this particular event (Takeover) with these attributes (Date, Company Predator, Company Target, Type of takeover, Value, Document ID...); another event could be described with different attributes.

After extracting from the document collection all the events, which concerns us, we fill a table, which has the following structure.

<b>Event ID</b>	Event Type	Date	Company (increased capital)	Company Predator	Company Target	Type of takeover	Value	Doc ID
Event 1	Takeover	March 15		Neurosoft SA.	Microsoft	Friendly	240,000,000	Doc1
Event 2	Share capital increase	April 10	Microsoft				250,000,000	Doc2

Table 1 Extracted Events

After extracting the financial terms and events we are going to construct the table, which will be used as input for the clustering algorithm. Documents in the database can be viewed as records, and the corresponding terms/events (that belong to each document) as attributes of the record. This is demonstrated better in the following table.

Documents	Extracted financial terms and event
Doc1	{Microsoft, Neurosoft, Take-over, Department of computation, ... Event1...}
Doc2	{Microsoft, IBM, Share capital increase, ... Event2...}
...	...

Table 2 Data input for the Clustering algorithm.

## 5. Clustering algorithm

We apply clustering algorithms in the resulting database in order to discover structure within the document collection. Clustering in TM means the creation of subsets from a collection of documents. A cluster is here defined as a group of documents having features, which are more similar to each other than to the features of any other group. In other words, documents from one cluster share some common features, which distinguish them from the other documents.

Clustering does not need any predefined categories in order to group the documents. Thus, the aim of this analysis is to produce a set of clusters in which the internal similarity of documents (inside the cluster) is maximised and the external similarity is minimised.

Many algorithms have been applied depending on the data collection and the task to be accomplished. Hierarchical Clustering and the Binary Relational Clustering are of the well known ones. In the former approach, clusters are arranged in cluster trees, where related clusters appear in the same branch of the tree, while the latter creates a flat cluster structure.

This automatic analysis can be useful in many tasks. First of all, it can provide an overview of the contents of a large document collection. Another task is identification of hidden structures within groups of objects. The process of finding related information can then become easy and accurate. In addition, new trends (or new policies, depending on the content of the document), which have not been mentioned in other documents, can be discovered in documents. Finally duplicate documents in a collection can be detected. These are just few of the tasks, which can be supported by clustering.

Cluster analysis is a traditional statistical technique. Objects can be partitioned into groups, which are easy to discover by algorithms making use of similarity, as measured by the numeric distance between objects. But, not every kind of similarity can be estimated numerically. For example, the distance between the concepts, UMIST and LSE. Even though these distances can be transformed into numbers, any such transformation would be difficult and subjective. So traditional distance-based cluster analysis is not possible to produce meaningful results when processing concepts.

In Conceptual clustering clusters are not just collection of entities with numerical similarity. Clusters are understood as groups of objects that together represent a concept. It does not produces only clusters, but also descriptions of the related concepts.

For Conceptual clustering we need a set of attribute descriptions of some entities, a description language for characterising clusters of such entities, and a classification quality criterion. The

target is to partition entities into clusters in a way that maximises the quality criterion and in the same time to determine general descriptions of these clusters.

It is important to mention that in conceptual clustering the properties of cluster descriptions are taken into consideration in the process of determining the clusters. This is a major difference between conceptual clustering and conventional clustering. In conventional clustering method we determine clusters according to a similarity measure. This similarity measure is a function of only the properties of the objects being compared and of no other factor. In contrast conceptual clustering takes into account not only the properties of the objects but also two other factors: The language that the system uses to describe the clusters and the environment, which is the set of neighbouring examples.

In our case we consider a document database (textual database) containing financial terms and events per document. These data can be used to cluster the documents such that documents with similar referring patterns (to terms/events) are in a single cluster. For example, one cluster may consist of documents referring to the local market, while another may consist of documents referring to the international, European etc. market. The clusters then can be used to characterize the different documents groups, and these characterization can be used in other data mining tasks, such as relevance analysis, classification etc. In addition it is very important that the user can have a quick overview of a large document collection.

In our database the attributes of the data points are non-numeric. Documents in the database can be viewed as records with Boolean attributes, each attribute corresponding to a single financial term/event. The value of the attribute is True if and only if the document contains the corresponding term/event; otherwise it is False. Boolean attributes are a special case of categorical attributes. Of course categorical attributes are not always true or false, but could have any finite set of values.

Traditional clustering algorithms that use distances or vector products methods between points for clustering are not appropriate for Boolean and categorical attributes. This observation forces us to apply the general framework of the ROCK algorithm [17] and the concept of links in order

to overcome the limitation of the traditional algorithms according to categorical data. As we have mentioned above we are going to use categorical attributes in order to characterize the documents.

An additional important factor is the concepts of Relative Closeness (RC) and Relative Interconnectivity (RI) presented in Chameleon algorithm [16]. Existing clustering algorithms find clusters that fit some static model. Although effective in some cases, these algorithms can break down; that is, cluster the data incorrectly, if the user does not select appropriate static-model parameters. The problem is that they use this static model of the clusters and they do not use information about the nature of individual clusters as they are merged. Rock belongs in the scheme of algorithms that ignores information about the closeness of two clusters as defined by the similarity of the closest items across two clusters. This last observation leads us to inherit these concepts (RI, RC) and to try combining them with the Rock algorithm, which will be the general framework.

In our approach we apply the clustering algorithm Rock [17]. Rock introduces a novel concept of clustering that is based on links in order to measure the similarity between a pair of data points. Clustering points based on only the closeness or similarity between them is not strong enough to distinguish two “not well-separated” clusters because it is possible for points in different clusters to be neighbors. As **link(pi, pj)** is defined the number of common neighbours between pi and pj. From this definition follows that if link(pi, pj) is large, then it is more probable that pi and pj belong to the same cluster.

The algorithm (Figure 1) accepts as input the set S of n sampled points to be clustered, and the number of desired clusters k.

**Procedure** Cluster (S, k)

**Begin**

1. link := compute\_links(S)
2. for each  $s \in S$  do
3.     q[s] := build\_local\_heap(link, s)
4.     Q := build\_global\_heap(S, q)

```

5.   while size(Q) > k do {
6.     u := extract_max(Q)
7.     v := max(q[u])
8.     delete(Q,v)
9.     w := merge(u,v)
10.    for each x ∈ q[u] ∪ q[v] do {
11.      link[x, w] := link[x, w] + link[x, v]
12.      delete(q[x] , u ); delete(q[x],v)
13.      insert(q[x], w, g(x, w));insert(q[w], x, g(x, w))
14.      update(Q, x, q[x])
15.    }
16.    insert(Q, w, q[w])
17.    deallocate(q[u]); deallocate(q[v])
18.  }

```

**End**

Figure 2

## **6.Classification algorithm**

Our approach makes further use of KDD techniques to exploit the knowledge discovered after applying clustering. More specifically classification is applied using the descriptions derived by clustering as the class attribute. This gives further insight of the document collection and possibly verifies the previous findings.

By applying decision tree classification algorithms on the templates under examination, we can retrieve hierarchies of concepts and test the validity of the discovered descriptions. This can possibly iterate a further clustering step following refinement of the database by feature subset selection.

## **7.Greek, a particular language**

Modern Greek is a language difficult to describe computationally. One reason is that several different inflectional endings, prefixes, infixes and a stress marker participate to the inflection or

conjugation of Modern Greek words, resulting to 4-7 word-forms for a Noun and up to 250 word-forms for a Verb. Another reason is that, the sentence constituents can be found at various positions (e.g. subject, verb, and objects). The syntactic role of word-forms can, most of the time, be extracted from the syntactic information attached to their inflectional endings, without considering their position in the sentence. One last reason, which makes Greek language difficult to describe computationally, is the existence of numerous characteristics carried over from Ancient Greek. Many old forms are in use interchangeably to new forms, along with old syntactic structures and archaic expressions.

Modern Greek inflection system exhibits a narrower range of morph syntactic ambiguity, compared to the corresponding ambiguity revealing in low inflected languages (e.g. in English almost every Noun can be a verb), and is mainly attributed to the frequent occurrence of functional words, such as Articles and Pronouns.

## **8. Conclusion and future work**

In this paper, we introduced our approach for data mining textual data (Text Mining). We presented TextMiner, a text-mining tool designed and developed at UMIST. We are currently in the stage of optimising the performance of the tool and also currying out extensive testing. At the same time we carry out tests in other domains.

Our innovation is that we label the documents, in order to cluster them, by using the terms (sequence of words with particular meaning e.g. Department of Computation) and events (set of terms with a particular meaning) mentioned within the documents. Additionally we have developed a clustering algorithm appropriate for categorical data. We apply further mining algorithms in the extracted information.

In the future, our target is to extract valuable knowledge by correlating information from the structured and the unstructured databases (see overall architecture in figure 3). For example in the financial case study we will investigate how the share price behave according to the announcement of the event “split”. We believe that the need for analysing data from diverse types of data sources is significant.

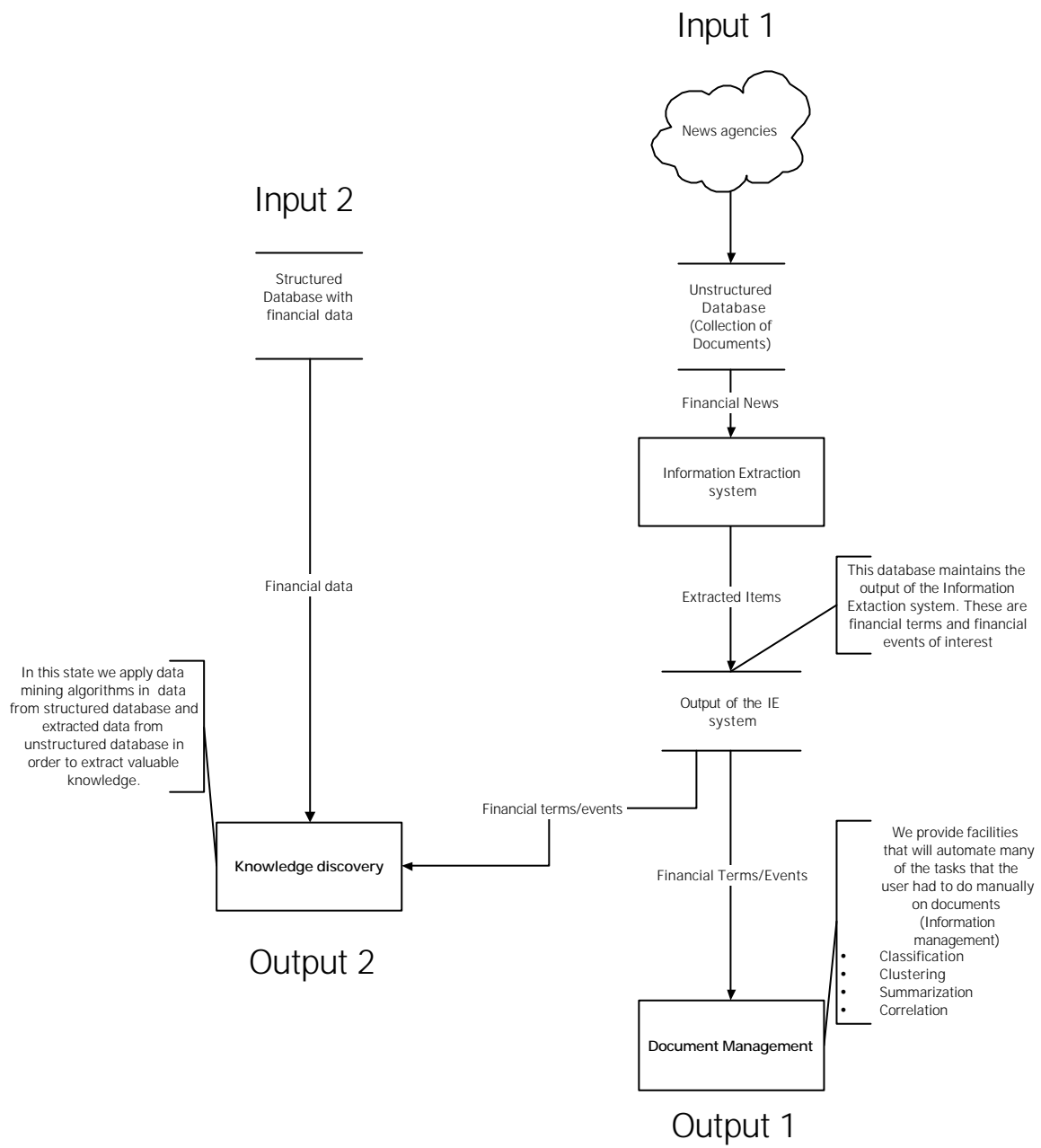


Figure 3

## References

- [1] Cutting D., Karger D., Pedersen J., and Tukey J. (1992), "*Scatter/gather: A cluster-based approach to browsing large document collections*", In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- [2] El-Hamdouchi A. and Willet P. (1986), "*Hierarchic document clustering using ward's method*", In proceedings of the Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- [3] El-Hamdouchi A. and Willet P. (1989), "*Comparison of hierarchic agglomerative clustering methods for document retrieval*", The Computer Journal, 32(3)
- [4] Feldman R., and Hirsh h. (1996), "*Exploiting Background Information in Knowledge Discovery from Text*", Journal of Intelligent Information Systems.
- [5] Feldman R. and Dagan I. (1995), "*Knowledge Discovery in Texts*", In Proceedings of the First International Conference on Knowledge Discovery.
- [6] Grishman R. (1997), "*Information Extraction: Techniques and Challenges*", International Summer School, SCIE-97
- [7] Gaizauskas R. and Humphreys K. (1997), "*Concepticons vs. Lexicons: An Architecture for Multilingual Information Extraction*", International Summer School, SCIE-97
- [8] Hahn U., and Schnattinger K. (1997), "*Deep Knowledge Discovery from Natural Language Texts*", In Proceedings of the 3<sup>rd</sup> International Conference of Knowledge Discovery and Data Mining.
- [9] Lent B., Agrawal R. and Srikant R. (1997), "*Discovering Trends Text Databases*", In Proceedings of the 3<sup>rd</sup> International Conference on Knowledge Discovery.
- [10] Michalski R., Bratko I., and Kubat M. (1998), "*Machine Learning and Data Mining*", Wiley

- [11] Piatetsky-Shapiro G. and Frawley W. (Eds.) (1991), "*Knowledge Discovery in Databases*", AAAI Press, Menlo Park, CA.
- [12] Rajman M. and Besancon R. (1997), "*Text Mining: Natural Language Techniques and Text Mining Applications*", In Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapan & Hall IFIP Proceeding series. Leysin, Switzerland, Oct 7-10, 1997.
- [13] Raymond T.Ng and Jiawei Han. (1994), "*Efficient and effective Clustering methods for spatial data mining*", In Proc. Of the VLDB Conference, Santiago, Chile, Sept 1994.
- [14] Wilks Yorick (1997), "*Information Extraction as a Core Language Technology*", International Summer School, SCIE-97
- [15] Willet P. (1988), "*Recent trends in hierarchic document clustering: A critical review*", Information Processing and Management.
- [16] Karypis G., Eui-Hong (Sam) Han and Vipin Kumar, "*Chameleon: Hierarchical Clustering Algorithm Using Dynamic Modeling*", Computer Magazine, August 1999.
- [17] Guha Subipto, Rastogi Rajeev, and Kyuseok Shim, "*Rock: A Robust Clustering Algorithm for Categorical Attributes*".
- [18] G. Koundourakis, M.Saraee and B.Theodoulidis, "*Data Mining in Temporal Databases*"