ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

A tutorial review on Text Mining Algorithms

Mrs. Sayantani Ghosh[1], Mr. Sudipta Roy[2], and Prof. Samir K. Bandyopadhyay[3]
Department of Computer Science and Engineering[1,2,3]
University of Calcutta, 92 A.P.C. Road,
Kolkata-700009, India.

ABSTRACT- As we enter the third decade of the World Wide Web (WWW), the textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic—just a keystroke or mouse click away .It can be viewed as one of a class of non traditional Information Retrieval (IR) strategies which attempt to treat entire text collections holistically, avoid the bias of human queries, objectify the IR process with principled algorithms, and "let the data speak for itself." These strategies share many techniques such as semantic parsing and statistical clustering, and the boundaries between them are fuzzy. In this paper different existing Text Mining Algorithms i.e Classification Algorithm, Association Algorithm, Clustering Algorithm is briefly reviewed, stating the merits / demerits of the algorithms. In addition some alternate implementation of the algorithms is proposed. Finally the logic of these algorithms are , merged to generate an algorithm which will perform the task of Classification of a data set into some predefined classes, establish relationship between the classified date and finally cluster the data based on the association between them into groups.

Keywords: Data Mining, Text Mining, Classification, Clustering, Association, Agglomerative, Divisive, Information Retrieval, Information Extraction.

1. INTRODUCTION

Labour-intensive manual text-mining approaches first surfaced in the mid-1980s, but technological advances have enabled the field to advance swiftly during the past decade. Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. As most information (common estimates say over 80%) is currently stored as text, text mining is believed to have a high commercial potential value. Increasing interest is being paid to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning. Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values

2. STAGES OF TEXT MINING PROCESS

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined together into a single workflow.

2.1. Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

2.2. Natural Language Processing (NLP) is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. Often this is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools.

2.3. Data Mining (DM) is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful knowledge. When used in text mining, DM is applied to the facts generated by the information extraction phase. We put the results of our DM process into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually.

2.4. Information Extraction (IE) is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generated by NLP systems.

3. PROBLEMS OF TEXT MINING

One main reason for applying data mining methods to text document collections is to structure them. A structure can significantly simplify the access to a document collection for a user. Well known access structures are library catalogues or book indexes. However, the problem of manual designed indexes is the time required to maintain them. Therefore, they are very often not up-to-date and thus not usable for recent publications or frequently changing information sources like the World Wide Web. The existing methods for structuring collections either try to assign keywords to documents based on a given keyword set (classification or categorization methods) or automatically structure document collections to find groups of similar documents (clustering methods). The problem of Text Mining is therefore Classification of data set and Discovery of Associations among data. In order to overcome from the problems of Data Mining the following algorithms have been designed.

4. TASKS OF TEXT MINING ALGORITHMS

- Text categorization: assigning the documents with pre-defined categories (e.g decision trees induction).
- Text clustering: descriptive activity, which groups similar documents together (e.g. self-organizing maps).
- Concept mining: modelling and discovering of concepts, sometimes combines categorization and clustering approaches with concept/ logic based ideas in order to find concepts and their relations from text collections (e.g. formal concept analysis approach for building of concept hierarchy).
- Information retrieval: retrieving the documents relevant to the user's query.
- Information extraction: question answering.

5. TYPE OF TEXT MINING ALGORITHM

5.1 Classification Algorithm

The **Classification problem** can be stated as a training data set consisting of records. Each record is identified by an unique record id, and consist of fields corresponding to the attributes. An attribute with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical attribute. One of the categorical attribute is the classifying attribute or class and the value in its domain are called class labels.

5.1.1 Objective:

Classification is the process of discovering a model for the class in terms of the remaining attributes. The objective is to use the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set.

5.1.2 Classification Models:

The different type of classification models are as follows:

1. Decision Tree
2. Neural Network
3. Genetic Algorithm

5.1.2.1. Classification Using Decision Tree:

- Sequential Decision Tree based Classification
- Parallel Formulation of Decision Tree based Classification.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

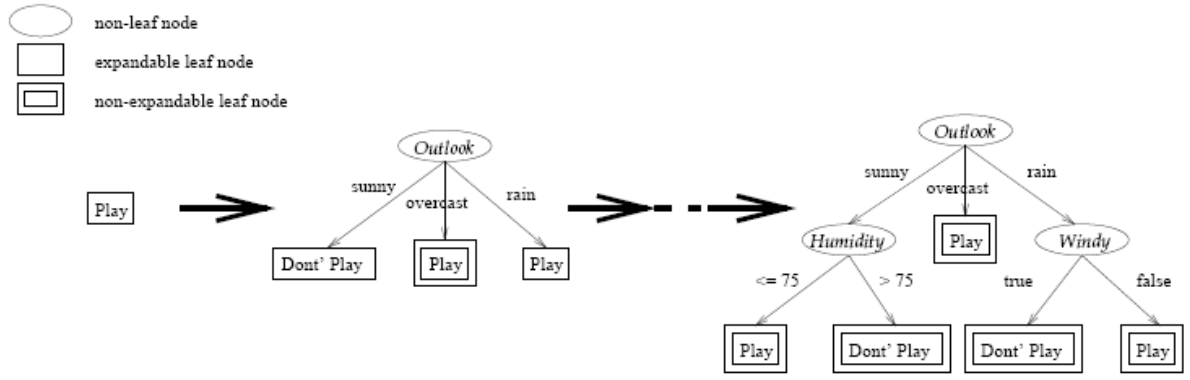
5.1.2.1.1. Sequential Decision Tree based Classification:

A decision tree model consists of internal node and leaves. Each of the internal node has a decision associated with it and each of the leaves has a class label attached to it. A decision tree based classification consists of two steps.

1. Tree induction – A tree is induced from the given training set.
2. Tree pruning – The induced tree is made more concise and robust by removing any statistical dependencies on the specific training data set.

5.1.2.1.1.1. Hunt's method: The following gives the recursive description of Hunt's method for constructing a decision tree from a set T of training cases with classes denoted fC1,C2, . . . ,Ckg.

Case 1 T contains cases all belonging to a single class Cj. The decision tree for T is a leaf identifying class Cj .



(a) Initial Classification Tree (b) Intermediate Classification Tree (c) Final Classification Tree

Figure1: Hunt's method

Figure 1 shows how Hunt's method works with the training data set. In case 2 of Hunt's method, a test based on a single attribute is chosen for expanding the current node. The choice of an attribute is normally based on the entropy gains of the attributes. The entropy of an attribute is calculated from class distribution information. For a discrete attribute, class distribution information of each value of the attribute is required. Outlook at the root of the decision tree shown in Figure 1. Once the class distribution information of all the attributes are gathered, each attribute is evaluated in terms of either entropy [Qui93] or Gini Index [BFOS84]. The best attribute is selected as a test for the node expansion.

5.1.2.1.1.2. C4.5 Algorithm: The C4.5 algorithm generates a classification–decision tree for the given training data set by recursively partitioning the data. The decision tree is grown using depth–first strategy. The algorithm considers all the

Case 2 T contains cases that belong to a mixture of classes. A test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes fO1, O2, . . . , Ong. Note that in many implementations, n is chosen to be 2 and this leads to a binary decision tree. T is partitioned into subsets T1, T2, . . . , Tn, where Ti contains all the cases in T that have outcome Oi of the chosen test. The decision tree for T consists of a decision node identifying the test, and one branch for each possible outcome. The same tree building machinery is applied recursively to each subset of training cases.

Case 3 T contains no cases. The decision tree for T is a leaf, but the class to be associated with the leaf must be determined from information other than T. For example, C4.5 chooses this to be the most frequent class at the parent of this node.

possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct value of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted

for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attribute.

Recently proposed classification algorithms SLIQ[MAR96] and SPRINT [SAM96] avoid costly sorting at each node by pre-sorting continuous attributes once in the beginning.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

5.1.2.1.1.3. SPRINT Algorithm: In SPRINT, each continuous attribute is maintained in a sorted attribute list. In this list, each entry contains a value of the attribute and its corresponding record id. Once the best attribute to split a node in a classification tree is determined, each attribute list has to be split according to the split decision. A hash table, of the same order as the number of training cases, has the mapping between record ids and where each record belongs according to the split decision. Each entry in the attribute list is moved to a classification tree node according to the information retrieved by probing the hash table. The sorted order is maintained as the entries are moved in pre-sorted order.

Decision trees are usually built in two steps. First, an initial tree is built till the leaf nodes belong to a single class only. Second, pruning is done to remove any over fitting to the training data. Typically, the time spent on pruning for a large dataset is a small fraction, less than 1% of the initial tree generation.

Advantages are they are inexpensive to construct, easy to interpret, and easy to integrate with the commercial database and they yield better accuracy. Disadvantages are it cannot handle larger data sets that are it suffers from memory limitations and it has low computational speed.

5.1.2.1.2. Parallel Formulation of Decision Tree based Classification

The goal of parallel formulation of decision tree based classification algorithms are scalability in both runtime and memory requirements. The parallel formulation overcome the memory limitation faced by the sequential algorithms, that is it should make it possible to handle larger data sets without requiring redundant disk I/O. Also parallel formulation offer good speedup over serial algorithm.

Type of parallel formulations for the classification decision tree construction is

- Synchronous Tree Construction Approach
- Partitioned Tree Construction Approach
- Hybrid Parallel Formulation

5.1.2.1.2.1. Synchronous Tree Construction Approach

In this approach, all processors construct a decision tree synchronously by sending and receiving class distribution information of local data. Major steps for the approach are shown below:

1. Select a node to expand according to a decision tree expansion strategy (eg. Depth-First or Breadth-First), and call that node as the current node. At the beginning, root node is selected as the current node.
2. For each data attribute, collect class distribution information of the local data at the current node.
3. Exchange the local class distribution information using global reduction [KGGK94] among processors.
4. Simultaneously compute the entropy gains of each attribute at each processor and select the best attribute for child node expansion.
5. Depending on the branching factor of the tree desired, create child nodes for the same number of partitions of attribute values, and split training cases accordingly.
6. Repeat above steps (1–5) until no more nodes are available for the expansion.

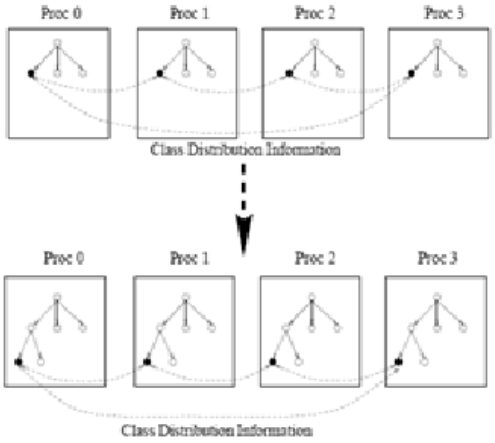


Figure 2: Synchronous Tree Construction Approach with Depth-First Expansion Strategy

In Figure 2 the root node has already been expanded and the current node is the leftmost child of the root (as shown in the top part of the figure). All the four processors cooperate to expand this node to have two child nodes. Next, the leftmost node of these child nodes is selected as the current node (in the bottom of the figure) and all four processors again cooperate to expand the node. The advantage of this approach is that it does not require any movement of the training data items. Disadvantages are this algorithm suffers from high communication cost and load imbalance. For each node in the decision tree, after collecting the class distribution information, all the processors need to synchronize and exchange the distribution information. Hence, as the tree deepens, the communication overhead dominates the overall processing time.The other problem is due to load imbalance. Even though each processor started out with the same number of the training data items, the number of items belonging to the same node of the decision tree can vary substantially among processors.

5.1.2.1.2.2. Partitioned Tree Construction Approach

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

In this approach, whenever feasible, different processors work on different parts of the classification tree. In particular, if more than one processors cooperate to expand a node, then these processors are partitioned to expand the successors of this node. Consider the case in which a group of processors Pn cooperate to expand node n.
The algorithm consists of following steps:
Step 1 Processors in Pn cooperate to expand node n using the method described above. **Step 2** Once the node n is expanded in to successor nodes, n1, n2, . . . , nk , then the processor group Pn is also partitioned, and the successor nodes are assigned to processors as follows:
Case 1: If the number of successor nodes is greater than |Pn|,
1. Partition the successor nodes into |Pn| groups such that the total number of training cases corresponding to each node group is roughly equal. Assign each processor to one node group.
2. Shuffle the training data such that each processor has data items that belong to the nodes it is responsible for.
3. Now the expansion of the sub trees rooted at a node group proceeds completely independently at each processor as in the serial algorithm.
Case 2: Otherwise (if the number of successor nodes is less than |Pn|),
1. Assign a subset of processors to each node such that number of processors assigned to a node is proportional to the number of the training cases corresponding to the node.
2. Shuffle the training cases such that each subset of processors has training cases that belong to the nodes it is responsible for.
3. Processor subsets assigned to different nodes develop subtrees independently. Processor subsets that contain only one processor use the sequential algorithm to expand the part of the classification tree rooted at the node assigned to them. Processor subsets that contain more than one processor proceed by following the above steps recursively.

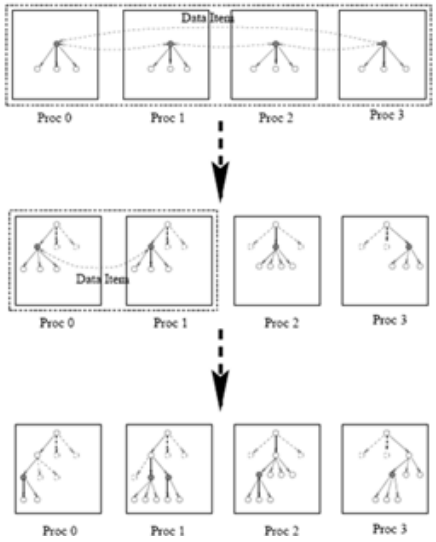


Figure 3: Partioned Tree Construction Approach

At the beginning, all processors work together to expand the root node of the classification tree. At the end, the whole classification tree is constructed by combining subtrees of each processor.
Figure 3 shows an example. First (at the top of the figure), all four processors cooperate to expand the root node just like they do in the synchronous tree construction approach. Next (in the middle of the figure), the set of four processors
is partitioned in three parts. The leftmost child is assigned to processors 0 and 1, while the other nodes are assigned to processors 2 and 3, respectively. Now these sets of processors proceed independently to expand these assigned nodes.
In particular, processors 2 and processor 3 proceed to expand their part of the tree using the serial algorithm. The group containing processors 0 and 1 splits the leftmost child node into three nodes. These three new nodes are partitioned in two parts (shown in the bottom of the figure); the leftmost node is assigned to processor 0, while the other two are assigned to processor 1. From now on, processors 0 and 1 also independently work on their respective sub trees.
Advantages:

- The advantage of this approach is that once a processor becomes solely responsible for a node, it can develop a subtree of the classification tree independently without any communication overhead.

Disadvantages:

- The first disadvantage is that it requires data movement after each node expansion until one processor becomes responsible for an entire subtree. The communication cost is expensive in the expansion of the upper part of the classification tree.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

- The second disadvantage is poor load balancing inherent in the algorithm. Assignment of nodes to processors is done based on the number of training cases in the successor nodes. However, the number of training cases associated with a node does not necessarily correspond to the amount of work needed to process the subtree rooted at the node.

5.1.2.1.2.3. Hybrid Parallel Formulation

The hybrid parallel formulation has elements of both schemes. The Synchronous Tree Construction Approach incurs high communication overhead as the frontier gets larger. The Partitioned Tree Construction Approach incurs cost of load balancing after each step. The hybrid scheme keeps continuing with the first approach as long as the communication cost incurred by the first formulation is not too high. Once this cost becomes high, the processors as well as the current frontier of the classification tree are partitioned into two parts. The description assumes that the number of processors is a power of 2, and that these processors are connected in a hypercube configuration. The algorithm can be appropriately modified if P is not a power of 2. Also this algorithm can
be mapped on to any parallel architecture by simply embedding a virtual hypercube in the architecture.

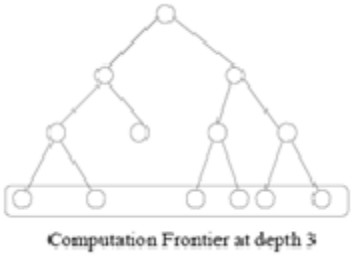


Figure 4: The computation front during computation phase

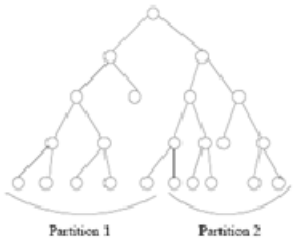


Figure 5: Binary partioning of the tree to communication costs

5.1.2.2. Classification Using Neural Network:

In Supervised learning, we are given a set of example pairs (x,y), x ϵ X, y ϵ Y and the aim is to find a function f in the allowed class of functions that matches the examples. In other words, we wish to infer the mapping implied by the data. The cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain. Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation).The supervised learning paradigm is also applicable to sequential data (e.g., for speech and gesture recognition).

With respect to the above specification the following assumptions have been considered.

(1)Multi-Layer Perceptions is the simple feed forward neural network is actually called a Multilayer perception (MLP). An MLP is a network of perceptions. The neurons are placed in layers with outputs always flowing toward the output layer. If only one layer exists, it is called a perception. If multiple layers exist, it is an MLP.

(2) Back Propagation algorithm is a learning technique that adjusts weights in neural network by propagating weight changes backward from the sink to the source nodes.

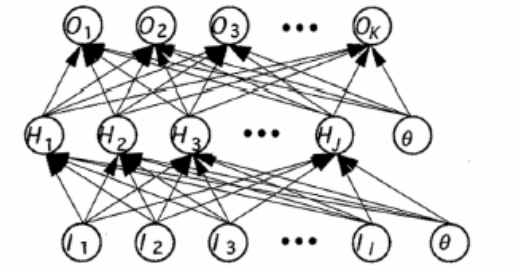


Figure 6: The typical structure of a back propagation network

Advantages of Neural Network:

- Artificial neural networks make no assumptions about the nature of the distribution of the data and are not therefore, biased in their analysis. Instead of making assumptions about the underlying population, neural networks with at least one middle layer use the data to develop an internal representation of the relationship between the variables.
- Since time-series data are dynamic in nature, it is necessary to have non-linear tools in order to discern relationships among time-series data. Neural networks are best at discovering nonlinear relationships.
- Neural networks perform well with missing or incomplete data. Whereas traditional regression analysis is not adaptive, typically processing all older data together with new data, neural networks adapt their weights as new input data

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

becomes available.

Disadvantages of Neural Network:

- No estimation or prediction errors are calculated with an artificial neural network
- Artificial neural networks are "black boxes," for it is impossible to figure out how relations in hidden layers are estimated.

Tasks of Neural Network:

The tasks to which artificial neural networks are applied tend to fall within the following broad categories:

- Function approximation, or regression analysis, Including time series prediction and modeling.
- Classification, including pattern and Sequence.
- Recognition, novelty detection and sequential decision making.
- Data processing, including filtering, clustering, Blind source separation and compression.

5.1.2.3. Classification using Genetic Algorithm:

Genetic algorithms are heuristic optimization methods whose mechanisms are analogous to biological evolution .In Genetic Algorithm, the solutions are called individuals or chromosomes. After the initial population is generated randomly, selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next generation. The quality of an individual is measured by a fitness function.

5.1.2.3.1. Genetic Operators

The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Before genetic operators are applied, parents have been selected for evolution to the next generation. The crossover and mutation algorithm is used to produce next generation. The probability of deploying crossover and mutation operators can be changed by user. In all of next generation, WTSD has used as the fitness function.

5.1.2.3.2. End Condition

GA needs an End Condition to end the generation process. If there is no sufficient improvement in two or more consecutive generations; stop the GA process. In other cases, time limitation can be used as a criterion for ending the process.

5.2 Algorithm for Discovering Associations:

5.2.1 Objective: In order to discover associations present in the data. The problem was formulated originally in the context of the transaction data at supermarket. This market basket data, as it is popularly known, consists of transactions made by each customer. Each transaction contains items bought by the customer. The goal is to see if occurrence of certain items in a transaction can be used to deduce occurrence of other items, or in other words, to find associative relationships between items. Traditionally, association models are used to discover business trends by analyzing customer transactions. However, they can also be used effectively to predict Web page accesses for personalization. For example, assume that after mining the Web access log, Company X discovered an association rule "A and B implies C," with 80% confidence, where A, B, and C are Web page accesses. If a user has visited pages A and B, there is an 80% chance that he/she will visit page C in the same session. Page C may or may not have a direct link from A or B. This information can be used to create a dynamic link to page C from pages A or B so that the user can "click-through" to page C directly. This kind of information is particularly valuable for a Web server supporting an e-commerce site to link the different product pages dynamically, based on the customer interaction.

5.2.2 Type:

5.2.2.1. Parallel Algorithm for Discovering Associations: The problem can be stated as given a set of items, association rules predict the occurrence of some other set of items with certain degree of confidence. The goal is to discover all such interesting rules. There are several properties of association models that can be calculated.

5.2.2.2. Sequential Algorithm for finding Association: The concept of association rules can be generalized and made more useful by observing another fact about transactions. All transactions have a timestamp associated with them; i.e. the time at which the transaction occurred. If this information can be put to use, one can find relationships such as if a customer bought book today, then he/she is likely to buy a book in a few days time. The usefulness of this kind of rules gave birth to the problem of discovering sequential patterns or sequential associations. In general, a sequential pattern is a sequence of item-sets with various timing constraints imposed on the occurrences of items appearing in the pattern. Example Consider the instance that A, B, C, D are the set of transactions such that (A) (C,B) (D) encodes a relationship that event D occurs after an event-set (C,B), which in turn occurs after event A. Prediction of events or identification of sequential rules that characterize different parts of the data, are some example applications of sequential patterns. Such patterns are not only important because they represent more powerful and predictive relationships, but they are also important

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

from the algorithmic point of view. Bringing in the sequential relationships increases the combinatorial complexity of the problem enormously. The reason is that, the maximum number of sequences having k events is O(mk2k1), where m is the total number of distinct events in the input data. In contrast, there are only (m C k) size-k item-sets possible while discovering non-sequential associations from m distinct items.

5.3. Clustering Algorithm:

5.3.1. Objective: Clustering is a division of data into groups of similar objects Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters.

Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.

5.3.2. Clustering Algorithms:

Clustering Algorithms are classified into following two methods:

5.3.2.1. Hierarchical Methods: Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity.

Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) .

An **agglomerative** clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters.

A **divisive** clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

Advantages are 1) Embedded flexibility regarding the level of granularity 2) Ease of handling of any forms of similarity or distance 3) Consequently, applicability to any attribute types. Disadvantages are 1) Vagueness of termination criteria 2 The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement

5.3.2.2 Partitioning Methods: In data partitioning algorithms, which divide data into several subsets. Since checking all possible subset systems is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters. One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found. More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find. One advantage of probabilistic methods is the interpretability of the constructed clusters. Having concise cluster representation also allows inexpensive computation of intra-clusters measures of it that give rise to a global objective function.

6. PROPOSALS

In this section we have made certain proposals that can be implemented as a modification to the existing Text Mining Algorithms as defined in the previous sections.

Association Algorithm: In Sequential Algorithm, the Sequential pattern between the data elements can be determined by associating a timestamp with each data, this time is assigned based on the arrival time of each data . If this information can be put to use, one can find relationships such as if a customer bought book today, then he/she is likely to buy a book in a few days time.

However this technique of finding the sequential pattern between the data item is more applicable if we are considering the dynamic data set where the number of data in a dataset varies dynamically with time.

Advantages:

- Every data item can be uniquely identified. There is less possibility of overlapping.
- The mechanism is simple to implement and incurs less overhead than assigning a timestamp to each data item.
- This technique is suitable when considering static data set.

Disadvantages:

- When using a very large dataset, it is required to generate a unique sequence number for each item; this may not always be feasible.
- There lies a chance that more than one same item may be assigned a same sequence number belonging to a different dataset.

Clustering Algorithm: In the Hierarchical Clustering Algorithm, there are two approaches

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

namely (a) Top- Down Approach in which a single cluster is divided into smaller clusters. (b) Bottom – Up Approach in which several smaller clusters are merged into a single cluster.

This two approach can however be combined into a single approach which is a sandwich of both the Top- Down and the Bottom- Up Approach. In this approach, if we start from a particular cluster say C1, the cluster C1 at that level along with the other clusters in the same level can be combined to form a single cluster, similarly the cluster C1 can be divided into smaller clusters.

Example: Each level consists of clusters. The top-down approach can be explained as, at the topmost level, there is the cluster country, India. This cluster can be further classified into several clusters containing metropolitan cities . The cities are further divided into clusters called districts. The bottom-up approach will finally generate the cluster INDIA

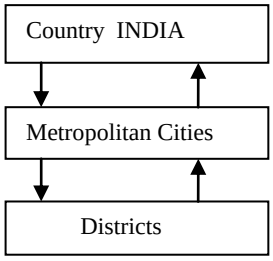


Figure 6: Hierarchical Clustering

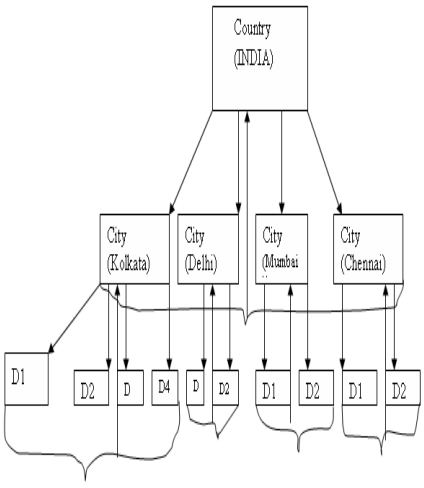


Figure 7: Sandwich Approach

Advantage of this approach is a combination of both the top-down and the bottom- up approach, it will utilize the advantage of both the approach and disadvantage of it is more complicated than the other two approaches.

7. WORKFLOW WITH SET OF SAMPLE TEST CASES

1. Implementation of Classification Algorithm:

Input: Newspaper Document and the location.

Design and Output:

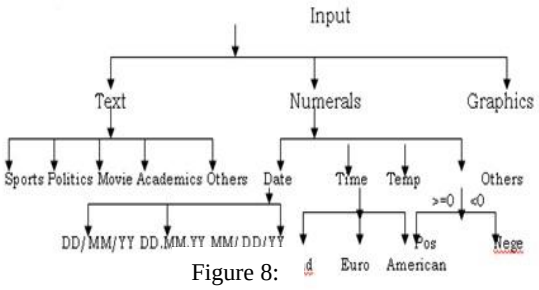


Figure 8:

(a) Input: Newspaper Document (Specified topic)

Query: Determine whether the news is a follow up or fresh news.

Modification Req: Yes

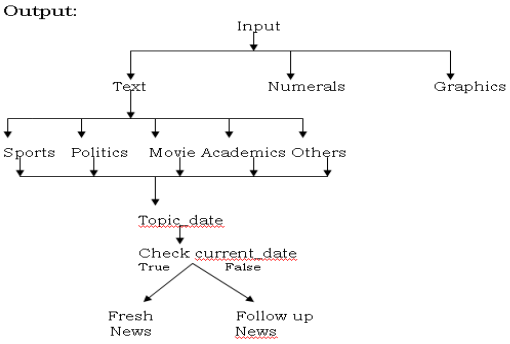


Figure 9:

(b) Input: Newspaper Document

Query: List of IT jobs specialization in Oracle with workplace preferably in Kolkata or otherwise

Modification Req: Yes

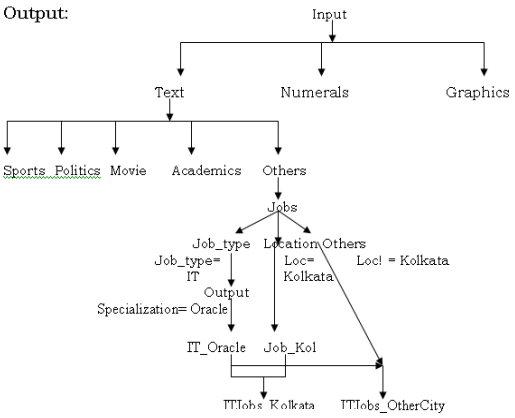


Figure 10:

3. Implementation of Clustering Algorithm

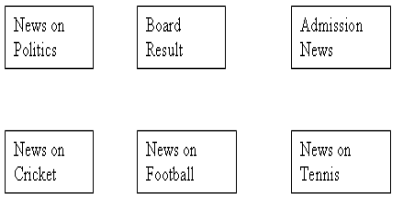
ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

User Query:

Case1:

(a) Input: Number of Cluster = 6.



(b) Query: State the Political Information involved in Sports and Academics.

Figure 11:

(c) Output:

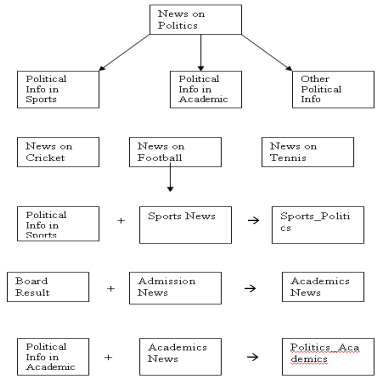


Figure 12:

4. Depiction of Combined Approach of Classification, Association and Clustering Algorithm:

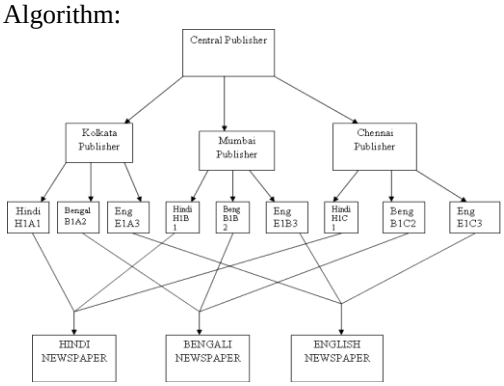


Figure 13:

- Consider the Central Publisher as the predefined base class. The base class is classified into sub classes Kolkata Publisher, Mumbai Publisher and Chennai Publisher based on the name of the city name as the classification condition. (Here I have considered only three cities)
- The association between the subclasses Bengali, English and Hindi Newspaper is established based on unique sequence number.
- The agglomerative clustering is used to club the smaller clusters i.e the Bengali, English and Hindi newspapers of different cities into appropriate clusters Bengali, English and Hindi Newspaper , containing the news of different cities in the specified language.

8. CONCLUSIONS AND DISCUSSIONS

When a user gives a set of words as input for a search of specific information, Google perform the search on the existing documents available on the World Wide Web to find a match for the requisite information as per the user's query. While Data mining is typically concerned with the detection of patterns in numeric data, very often important (e.g., critical to business) information is stored in the form of text. Unlike numeric data, text is often amorphous, and difficult to deal with .Text mining generally consists of the analysis of (multiple) text documents by extracting key phrases, concepts, etc. and the preparation of the text processed in that manner for further analyses with numeric data mining techniques .A typical (first) goal in data mining is feature extraction, i.e., the identification of the terms and concepts most frequently used in the input documents; a second goal typically is to discover any associations between features (e.g., associations between symptoms as described by patients). Hence, a first step to text mining usually consists of "coding" the information in the input text; as a second step various methods such as Association Rules algorithms may be applied to determine relations between features.

9. REFERENCES

[1] N. Jovanovic, V. Milutinovic, and Z. Obradovic, Member, IEEE, "Foundations of Predictive Data Mining" (2002).
[2] Yochanan Shachmurove, Department of Economics, The City College of the City, University of New York and The University of Pennsylvania, Dorota Witkowska, Department of Management, Technical University of Lodz "CARESS Working Paper 00-11Utilizing Artificial Neural Network Model to Predict Stock Markets" September 2000.
[3] Margaret H.Dunham, "Data Mining- Introductory and Advanced Topics" Pearson Education, 2003, pages 106-112.
[4] Michael W. Berry and Malu Castellanos, Editors "Survey of Text Mining:Clustering, Classification, and Retrieval, Second Editio" Springe,September 30, 2007.
[5] http://en.wikipedia.org/wiki/Text_mining
[6] Ying Zhao and George Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. TR# 01-40, Department of Computer Science &Engineering, University of Minnesota, Minneapolis, 2000.
[7] Keno Buss, "Mining and Summarizing Customer Reviews" STRL, De Montfort University.

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2012

[8] Abdul-Baquee M. Shara "The Qur'an Annotation for Text Minin" School of Computin December 2009.
[9] Suman Chakraborty, Sudipta Roy, Prof. Samir K. Bandyopadhyay "Image Steganography Using DNA Sequence and Sudoku Solution Matrix", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 2, February 2012.
[10] Minqing Hu and Bing Liu "Mining and Summarizing Customer Reviews" Department of Computer Science ,University of Illinois at Chicago ,851 South Morgan Street,Chicago, IL 60607-7053.
[11] Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement , Ben Shneiderman and Catherine Plaisant "Discovering interesting usage patterns in text collections: Integrating text mining with visualizatio" http://hcil2.cs.umd.edu/trs/2007-08/2007-08.pdf
[12] http://store.elsevier.com/Practical-Text-Mining-and-Statistical-Analysis-for-Non-structured-Text-Data Applications /Gary-Miner/isbn-9780123869791/
[13] http://www.autonlab.org/tutorials/
[14] Un Yong Nahm "Text Mining with Information Extraction" National Science Foundation
[15] Louise Francis, FCAS, MAAA, and Matt Flynn "Text Mining Handboo" Casualty Actuarial Society E-Forum, Spring 2010.
[16] http://www.amazon.com/Principles-Adaptive-Computation-Machine-Learning/dp/026208290X
[17] Un Yong Nahm, Mikhail Bilenko and Raymond J. Moone " Two Approaches to Handling Noisy Variation in Text Minin" Proceedings of the ICML-2002 Workshop on Text Learning, pp. 18-27, Sydney, Australia, July 2002.

Mrs. Sayantani Ghosh
She has enrolled herself for Ph.D. in the Department of Computer Science and Engineering. Kolkata , India.
She received her B.Sc. degree in Computer Science from Bethune College in the year 2006, under the University of Calcutta. Sheranked 1st class 3rd in the same university. She completed her M.Sc. in Computer and Information Science from University College of Science and Technology, University of Calcutta in 2008. She did her M.Tech. in Computer Science and Engineering from the Dept. of Computer Science and Engineering, University of Calcutta in the year 2010. Currently she is working as an

Sudipta Roy
He is pursuing M.Tech in the Dept. Of Computer Science & Engineering , University of Calcutta, India. He received B.Sc(Phys Hons) from Burdwan University and B.Tech from Calcutta University. He is Author of more than five publications in National and International Journal. Field of interest is Biomedical Image Analysis, Image Processing, Steganography, Database Management System , Data Structure, Artificial Intelligence, Programming Languages etc.

Samir K Bandyopadhyay
He is Professor of Dept. Of Computer Science & Engineering, University of Calcutta, Kolkata, India.
He is Chairman, Science & Engineering Research Support Society(SERSC, Indian Part), Fellow of Computer Society of India, Sectional President of ICT of Indian Science Congress Association, 2008-