International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 613-622

# Text Mining Process, Techniques and Tools : an Overview

Vidhya. K. A[1] & G. Aghila[2]

Text Mining has become an important research area, which refers to the application of machine learning (or data mining) techniques in the study of Information Retrieval and Natural Language Processing. In sense, it is defined as the way of discovering knowledge from ubiquitous text data which are easily accessible over the Internet or the Intranet. The survey of Text Mining techniques, Text Mining Applications, literature survey of various applications and tools has been presented. Text Mining techniques like Document clustering and Document Classification have been presented. Text mining based framework for applications like Summarization, Topic Discovery, Information Extraction, Information Retrieval terms and techniques in each method has been discussed. Various text mining and data visualization tools for application to patent information like their working mode, capabilities, data sources and result output have been presented. In depth analysis of algorithm related to classification techniques its advantages and disadvantages and the working mode has been presented.
Keywords: Text Mining, Information Extraction, Information Retrieval, Document Classification.

## 1. Introduction

Mining is the process of inferring for patterns with in a structured or unstructured data. There are various mining methods out of which they differ in the context and type of dataset that is applied. The process of extracting information and knowledge from unstructured text led to the need for various mining techniques for useful pattern discovery. “Data Mining (DM) and Text Mining (TM) [1] is similar in that both techniques “mine” large amounts of data, looking for meaningful patterns [19].”Some of the mining types are data, text, web, business Process and service mining.

DM looks for patterns within structured data, that is, databases. The underlying technologies are based on statistics and artificial intelligence, littering the field with buzzwords such as classification and regression trees (CART), chi-squared automatic induction (CHAID) [2], [14] neural networks and genetic algorithms. TM looks for patterns in unstructured data - memos and documents, pdf and text files. Consequently, it often uses language-based techniques, such as semantic analysis and intelligence.

Web Mining deals with the extraction of specific knowledge from the World Wide Web. More precisely [13], Web Content Mining is that part of Web Mining which focuses on the raw information available in Web pages [9]; source data mainly consist of textual data in Web pages (e.g., words, but also tags); typical applications are content-based categorization and content-based ranking of Web pages.

[1,2]Department of Computer Science, Pondicherry University, Pondicherry, India

Email: [1]avidhya06@gmail.com, [2]aghilaa@yahoo.com

The Web is transforming from a Web of data to a Web of both Semantic data and services [6] and this trend is providing us with increasing opportunities to compose potentially interesting and useful services from existing services. While the user may not sometimes have the specific queries needed in top-down service composition approaches to identify them, Service Mining the early and proactive exposure of these opportunities will be key to harvest the great potential of the large body of Web services.

Contemporary information systems record business events in so-called event logs [6] from which business process mining [10] takes these logs to discover process, control, data, organizational, and social structures. Although many researchers are developing new and more powerful process mining techniques [16] and software vendors are incorporating these in their software, few of the more advanced process mining techniques have been tested on real-life processes. However there are many mining techniques the orientation of this document is towards Text Mining techniques and its application

The paper is organized in the following way: Section II for Text Mining Process, Section III Text Mining Vs Information Extraction, Section IV Text Ming Vs Information Retrieval, and Section V Text Ming Vs Natural Language Processing, Section VI for Text Mining Vs Document Clustering, Section VII for Text Mining Vs Document Classification Section VIII for Text Mining Tools finally followed by Conclusion and References.

## 2. Text Mining Process

Text Mining (TM) refers some informational content included in any of the items such as: newspaper; articles; books; reports; stories; manuals; blogs; email, and articles in the WWW. The quantum of text of the present day is pretty

vast with ever-growing incremental power [19].

The prime aim of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. TM is an empirical tool that has a capacity of identifying new information that is not apparent from a document collection.

Figure 1 depicts the TM process that uses Information retrieval and Natural Language Processing to mine large dataset and infer the knowledge available in the dataset. The Process of TM includes searching, extracting, categorization where the themes are readable and the meaning is obvious. Typically text mining tasks include text categorization [15], text clustering, Information extraction, information retrieval [14], sentiment analysis, document summarization [4], [5] and entity relation modeling. TM, also known as Knowledge Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) refers generally to the process of extracting information and knowledge from unstructured text.

TM starts with a collection of documents; which would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system [21], yielding an abundant amount of knowledge for the user of that system. The following figure explores the detail processing methods in Text Mining.

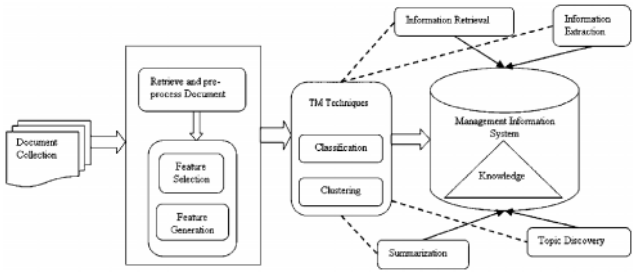



Fig. 1: Text Mining Process

The document collection from figure 1 is set of files might be with any extension like PDF, txt or even flat file extension which are normally collected and named as noisy unstructured text data found in informal settings such as online chat, SMS, emails, message boards, newsgroups, blogs, wikis and web pages. Also, text data set is created by processing spontaneous speech, printed text and handwritten text contains processing noise.

The dataset is an unstructured dataset of documents which are pre-processed using the following three rules:

- Tokenize the file into individual tokens using space as the delimiter.
- Removing the stop word which does not convey any meaning.
- Use porter stemmer algorithm to stem the words with common root word.

Feature Generation: The process of TM involves generating features in a spread sheet format which is simpler and more restrictive than open- ended data mining. TM is unstructured because it is very far the spread sheet model that we need to process data for prediction. Even then, this type of transforming data from text to spread sheet model can be highly methodical and there is need for an organized procedure to fill in the cells of spread sheet.

Feature Selection: Feature Selection algorithms are adhoc which in turn is the process of selecting the important features which requires an exhaustive search of all subsets of features of chosen cardinality. If the large numbers are available this is impractical for supervised learning algorithms the search is for satisfactory set of features instead of optimal set.

After the appropriate selection of features the text mining techniques are incorporated for the applications like Information retrieval, Information Extraction, Summarization and Topic Discovery for necessary knowledge discovery process. The process of using Knowledge Discovery in Database (KDD), which is the fundamental step in Text Mining, knowledge experts can obtain important strategic information for their business. KDD has more intensive transformation methods to cross-examine traditional databases, where data are in structured form, by automatically finding new and unknown patterns in huge quantity of data. Mostly, structured data represent only a little part of the overall organization knowledge and the knowledge is incorporated in textual documents.

Figure 1 depicts the knowledge stored in the management information system where the knowledge is stored and retrieved when the system gets a new training set the system goes for incremental learning rather than the initial learning process available.

For Example, Text Data mining [18], [10] in customer relationship management applications can contribute significantly to the business rather than randomly contacting a customer through a call center or sending mail, a company can concentrate its efforts on customer that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across the dataset so that one may predict which channel and which offer the customer is most likely to respond to across all potential offers that are made out of it.

### 3. TEXT MINING VS INFORMATION EXTRACTION

Information Extraction (IE) is the process of automatic extraction of structured information such as entities, relationship between entities and attributes describing entities from unstructured texts. Mostly useful information such as names of people, places or organization mentioned in the text is extracted without a proper understanding of the text. Traditional data mining systems assumes that the information to be mined is already in the form of relational database [17].

Information Retrieval deals with the problem of finding relevant documents in a collection. Information Extraction identifies useful relevant text in a document. Useful information is defined as text segment and its associated attributes.

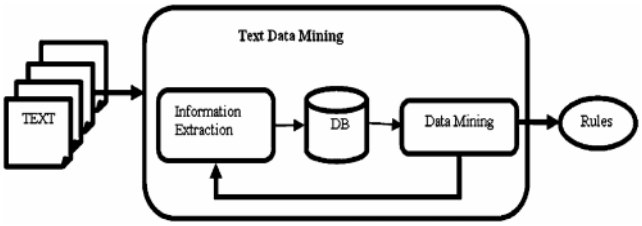



Fig. 2: IE based TM Framework

Table I
IE System Available

| IE Systems | Work Method | Stages |
|---|---|---|
| SRI International developed Fastus– Finite state Automaton for Text Understanding System | Regex –Regular Expression actually represents the string pattern. | Five –Complex words, Basic Phrases, complex phrases, structures and Merged Structures. |
| Rapier – Robust Automated Production of Information Extraction Rules | Set of Rules (patterns) that are applied to text to locate relevant information | Three parts – PreFiller, Filler, Post Filler. |
| The LaSIE System | IE will benefit from, the best computational models of human language processing. One system which is in this tradition is the LaSIE (LArge Scale Information Extraction) system developed at Sheffield. | Four Stages- text pre-processing, lexical and terminological processing, syntact analysis and semantic interpretation, discourse interpretation |

### 4. TEXT MINING VS INFORMATION RETRIEVAL

Information Retrieval (IR) is finding a document of an unstructured nature usually text that satisfies an information need from within large collections usually stored on computers. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching and IR can also cover other kinds of data and information problems beyond that specified in the core definition above.

These range from fully linguistic (based on parsing the sentences) to fully statistical (e.g., based on counting word co-occurrences). While doing an IR research it is proved that phrases are valuable indexing units and yield improved search effectiveness however, the style of phrase generation used is not that critical. Studies comparing linguistic phrases to statistical phrases have failed to show a difference in their retrieval performance.

Some IR systems also use multi-word phrases (information retrieval) as index terms. Since phrases are considered more meaningful than individual words, a phrase match in the document is considered more informative than single word matches. Several techniques to generate a list of phrases have been explored. Automatic extraction of metadata is an important application of TM techniques. However, existing automatic document retrieval techniques bypass the metadata creation stage and work on the full text of the documents directly [Salton and McGill, 1983] in which the basic idea is to index every individual word in the document collection.

The documents are represented as a “bag of words” that is, the set of words that they contain, along with a count of how often each one appears in the document which makes it easier for retrieval. There are various IR systems available in which two systems AMORE and MAPBOT has been listed with its working mode as follows:

AMORE: Advanced Multimedia Oriented Retrieval Engine [7] The Harvest Information Discovery and Access System for text indexing and searching, and using the content oriented image retrieval (COIR) library for image retrieval.

Advantage: Its an automatic indexing of both text and image from one or more Web sites.

MAPBOT: An interactive Web based map information retrieval system [11] in which Web users can easily and efficiently search geographical information with the assistance of a user interface agent (UIA). Each kind of map feature such as a building or a motorway works as an agent called a Maplet.

MAPBOT, an active map system using software agent technology is presented to solve these problems.

Advantage: Maplets to communicate with information agents outside the system to retrieve more information for the user.

There are many models available for IR process which can be broadly classified as:

- Classical models of IR based on mathematical knowledge that was easily recognized and well understood simple, efficient and easy to implement. The three classical information retrieval models are: Boolean, Vector and Probabilistic models.
- Non-Classical models of IR are based on principles other than similarity, probability, Boolean operations etc on which classical retrieval models are based on information logic model, situation theory model and Interaction model.
- Alternative models of IR .Alternative models are enhancements of classical models making use of specific techniques from other fields. Example:" Cluster model, fuzzy model and latent semantic indexing (LSI) models."

Table II depicts various IR models available their working mode, advantages and their disadvantages.

Table 2
IR Models

| Model | Working Mode | Advantages | Disadvantages |
|---|---|---|---|
| Boolean Model | A document is represented by a set of key terms: chosen from a fixed set of key terms, or, possibly automatically, from the documents themselves. | Easy to implement and has low computational cost. | Formulation of good query is not feasible |
| | | Its query language is more expressive than that of other models. | Relative importance for the keywords cannot be specified |
| | | The model is fit for users who know exactly what they are looking for. | Arranging the retrieved document in order of relevance is difficult |
| Vector Model | In a Vector-model IR system containing n key terms, an n-dimensional space is defined such that each axis is associated with a different key term. | It is possible to assign weights to key terms in a query. | Key terms are supposed to be independent. |
| | | The similarity measure can be used to present the results in order of relevance. | In a query no logic relations (such as AND, |
| | | Many researchers assume that the retrieval results obtained with the vector model are better than those obtained with the Boolean Model. | OR and NOT) between key terms can be used. |
| Probabilistic Model | The basis for the probabilistic model is the probability ranking principle best possible retrieval results are achieved when documents are shown in the order of their probable relevance to the Query. | The effectiveness of the probabilistic model is clearly better than the Boolean model, and slightly worse than the vector model. | Key terms are supposed to be independent of each other (as in the vector model). |
| | | | There is no method for estimating term relevance at the beginning, when no relevant documents are known. |
| Connectionist Model | Neural networks create a form of connectionism this is also applicable in IR. IR purposes each key term can be associated with an input neuron and each document with an output neuron. A query is presented to the network by activating the neurons which are associated with the desired key terms. | Learning is a part of the model. | Hidden layers make the constraint of independent key terms redundant. |

## 5. TEXT MINING VS NATURAL LANGUAGE PROCESSING

Text and data mining approach has been used in Natural language processing to overcome the difficulties with the codes, keywords and search techniques involved in knowledge or pattern discovery. The output of this mining process helps analyst to discover the trends and new occurrences in the data available. Many TM algorithms have been developed to maintain the text sources of bilingual corpus or multi lingual corpus.

Natural Language processing, Text mining and Machine learning methods can be applied for mining of interesting data available online for example gene ontology or protein-protein function mapping using certain tools. For example Rapier and Tagging process is done through for certain biomedical corpus available.

Rapier is a machine learning tool that learns information extraction rules from a set of documents and associated templates. This form of representation is grammar rule induction. The Rapier works on tagged documents directly instead of templates. We ran this instance of Rapier on our manually tagged training corpus to produce a set of grammar rules. There are various kinds of rules available for working in which interaction rules outplay most of the linguistic applications:

POS: Noun phrase; Semantic: Gene Ontology

Word: 'is'

POS: Verb past participle

Word: 'by'

POS: Noun phrase; Semantic: Gene Ontology.

Kernel-based learning methods have been applied to various information extraction tasks. For simple data representations (e.g., "bag-of-words") in which features can be easily extracted, some basic kernel functions such as linear kernel, polynomial kernel, and Gaussian kernel are often used. For data in structured representation, convolution kernels are frequently used (Collins and Duffy 2002).[22]

## 6. TEXT MINING VS DOCUMENT CLUSTERING

Document clustering (also referred to as text clustering) is one of the most important text mining methods that are developed to help users effectively navigate, summarize, and organize text documents. Figure 3 depicts the overview of document clustering process which is done by organizing a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or organize the results returned by a search engine in response to a user's query [3].

Clustering can significantly improve the precision and recall of data in information retrieval systems [7], which is an efficient way to find the nearest neighbors of a document available [12]. The problem of document clustering is generally defined as follows: given a set of documents, it should be partitioned into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters.

The important aspect in clustering is the similarity measure which enhances the optimization of clustering problems, good choice of similarity measure will lead to improvements in clustering performance [8], [12]. The similarity between two documents is computed with one of several similarity measures based on two corresponding feature vectors, e.g. cosine measure, Jaccard measure and Euclidean distance measure.

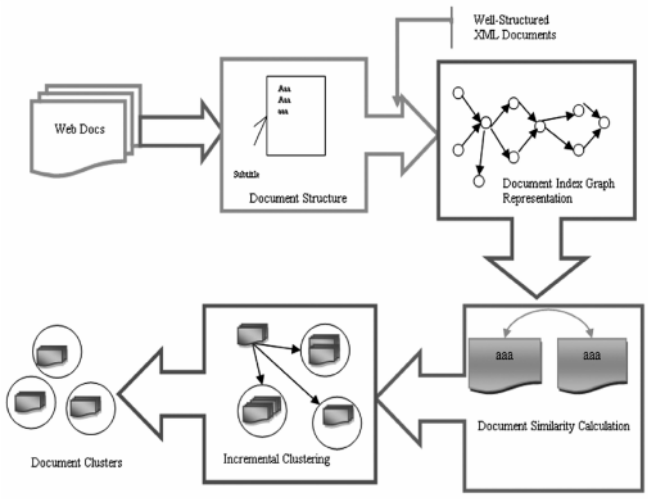



Figure 3: Document Clustering

The topic detection problem (Allan, Carbonell, Doddington, Yamron, & Yang, 1998) [19] consists of determining for each incoming document, whether it reports on a new topic, or it belongs to some previously detected topic. A topic detection system forms topic-based clusters of documents. The clusters in our top level of the hierarchy can be seen as detected topics in a stream of news. Also, the lower levels of the hierarchy can present topics with different detail levels.

Text summarization [4], [5] is significantly helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning.

## 7. TEXT MINING VS DOCUMENT CLASSIFICATION

Text classification tasks like supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents. Categorization [15] often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms.

Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic. Using supervised learning algorithms, the objective is to learn classifiers from known labeled documents and perform the classification automatically on unlabeled documents. Figure.8 shows the overall flow diagram of the text categorization task. Consider D = [d1, d2,.. ., dn] belonging to a set of classes C = [c1,c2,…,cp].

The Text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class.

### A. Classification Algorithms a Comparative Study

There are various algorithms for Document classification out which few algorithms has been listed in the above table along with their merits and demerits. From the table IV inference is that certain algorithm work well in small datasets while certain other algorithms work well in large dataset. The most intriguing fact is that hybrid algorithms yield better classification accuracy rather than individual techniques.

a. Naive Bayes Classifier: The Naýve Bayesian classifier (NB) is a simple but effective text classification algorithm which has been shown to perform very well in practice (McCallum & Nigam, 1998; Mitchell, 1997). The basic idea in NB is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document.

Working Mode: The supervised learning method in which it follows the multinomial distribution for document classification, where $P(t_k|c)$ is the conditional probability. It works with the bayes theorem and the following equation infers the conditional probability of a document d belonging to the labeled class c as,

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c)$$

$$c_{map} = \arg\max_{c \in C} \dot{P}(c \mid d) = \arg\max_{c \in C} \dot{P}(c) \prod_{1 \le k \le n_d} \dot{P}(t_k \mid c)$$

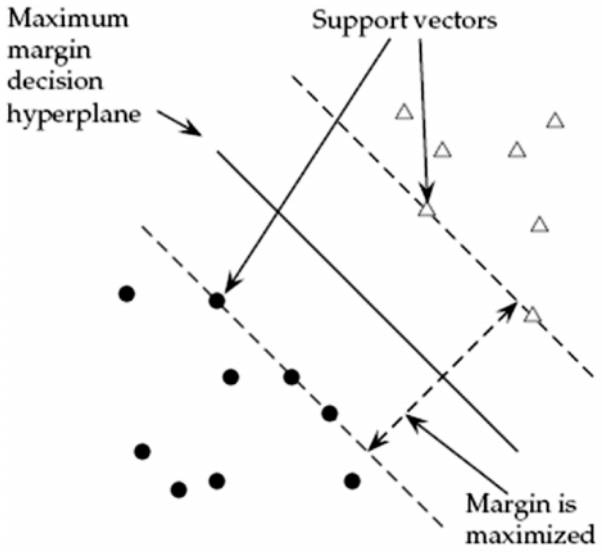
Merits: Good Classification Accuracy especially with Large Datasets and it is easy to Implement.

Demerits: Naive Bayes selects poor weights for the decision boundary.

The Systematic problem with Naive Bayes is that features are assumed to be independent and the Magnitude for the weights for classes with strong word dependencies is larger than for classes with weak word dependencies.

### B. Support Vector Machines(SVM)

SVM operates by finding a Hyper-surface in the space of possible inputs. The hyper-surface attempts to split the positive examples from the negative examples by maximizing the distance between the nearest of the positive and negative examples to the hyper-surface. Intuitively, this makes the classification correct for testing data that is near but not identical to the training data.




Working Mode: A decision hyper plane can be defined by an intercept term b and a decision hyper plane normal vector $\vec{w}$ which is perpendicular to the hyperplane. This vector is commonly referred to as weight vector. To choose among all the hyperplanes that is perpendicular to the normal vector. The linear equation of the SVM classifier is given as follows,

$$f(\vec{x}) = \text{sing}(\vec{w}^T \vec{x} + b)$$

$$\vec{x}' = \vec{x} - yr \frac{\vec{w}}{|\vec{w}|}$$

and so satisfies $\vec{w}^T \vec{x}' + b = 0$. Hence:

$$\vec{w}^T (\vec{x} - yr \frac{\vec{w}}{|\vec{w}|} + b = 0$$

Solving for r gives:[2]

$$r = y\frac{\vec{w}^T\vec{x} + b}{|\vec{w}|}$$

if $f(\vec{x})$ yields +1 then it belongs to the same class, if -1 then belongs to some other class. The geometric margin of the classifier is the maximum width of the band that can be drawn separating the support vectors of the two classes.

Merits: The model is relatively easy to understand and the model can be applied without modification on new collections. The SVM model gives a very good retrieval quality.

Demerits: Translating the training set into a higher-dimensional space incurs both computational and learning-theoretic costs.

### C. Rocchio's Algorithm

Rocchio's Algorithm is based on the relevancy Feedback Algorithms for Document Relevancy. Relevancy Feedback Models are an effective way of modifying and expanding user queries (such as search engines). Rocchio's Algorithm is one of the earliest methods used for queries.

Working Mode: Rochhio's method working with set of document vector $\vec{d}$ so that document with similar content have similar vector where $c_j$ represent the respective class.

$$H_{TFIDF}(d^i) = \arg\max_{C_j \in C} \frac{\vec{C}_j}{\left\|\vec{C}_j\right\|} \cdot \frac{d^i}{\left\|\vec{d}^i\right\|}$$

which was further enhanced as

$$H_{TFIDF}(d^i) = \arg\max_{C_j \in C} \frac{\overrightarrow{\sum_{i=1}^{|F|} C_j^{\,i} \cdot d^{r(i)}}}{\sqrt{\sum_{i=1}^{|F|}\left(C_j^{(i)}\right)^2}}$$

Rochhio shows that each prototype vector maximizes that mean similarity of the positive training examples with the prototype vector Cj minus the mean similarity of the negative training examples with the prototype vector Cj where |F| is the cardinality of documents.

Merits: Its is easy to implement and a very fast Learner for training a system with relevance Feedback Mechanism

Demerits: This method yields low classification accuracy with linear combination which is too simple for classification and the Constant $\alpha$ and $\beta$ are empirical.

### D. K-Nearest Neighbor (K-NN)

K-NN is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. k nearest neighbors of a training data are computed first. Similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class.

Working Mode:

$$y(d_i) = \arg\max_m \frac{\sum_{x_j \in top\_n\_kNN(c_m)} Sim(d_i, x_j)\, y(x_j, c_m)}{\sum_{x_j \in top\_n\_kNN(c_m)} Sim(d_i, x_j)}$$

$$x_i = \frac{(1 + \log(TF(w_i, d))) \cdot \log(\frac{|D|}{DF(w_i)})}{\sqrt{\sum_j [(1 + \log(TF(w_j, d))) \cdot \log(\frac{|D|}{DF(w_j)})]^2}}.$$

where $d_i$ is a test document, $x_j$ is one of the neighbors in the training set, $y(x_j, c_k) \in \{0, 1\}$ indicates whether $x_j$ belongs to class $c_k$, and $Sim(d_i, x_j)$ is the similarity function for $d_i$ and $x_j$. The original k nearest neighbors are calculated which compute the probability that one document belongs to a class by using only the top n nearest neighbors for that class.

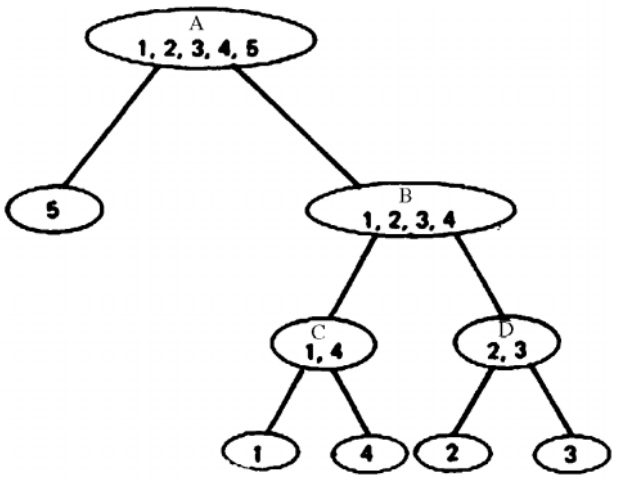
Merits: This method yields good classification accuracy.

Demerits: K-NN uses all features equally in computing similarities and leads to poor similarity measures and classification errors when only a small subset of the features is useful for classification.

### E. Decision Tree Classifier

Decision Tree Classifier uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Ex: classification trees or regression trees.

In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.



Working Mode: The above figure depicts a decision tree where the individual nodes have been presented, either the tree can work up using the bottom-up approach or top down approach in which the following equation depicts the hybrid approach which works well for both the approach,

$$J_i = (x - \hat{M}_i)^T \hat{\Sigma}_i^{-1}(x - \hat{M}_i) + \ln\left|\hat{\Sigma}_i\right| \; i = n_L \text{ or } n_R$$
$$x \in n_L \text{ if } J_{n_L} < J_{n_R} \; x \in n_L \text{ if } J_{n_L} < J_{n_R} \qquad (1)$$

where x is the data sample of N dimensions

$\hat{M}_i$ is the sample mean of node i

$\hat{\Sigma}_i$ is the sample covariance of node i

$n_L$ and $n_R$ represent left and right node, respectively.

Merits: Yields good classification results primarily on low dimensional datasets and easy to generate rules which reduce Complexity.

Demerits: Training time is relatively expensive and suffers from over fitting by which it is not able to handle continuous variable well.

## 8. TEXT MINING TOOLS

A high-level overview of some key text mining and visualization tools [19] is presented to provide a comparison of text mining capabilities, perceived strengths, potential limitations, applicable data sources, and output of results, as applied to chemical, biological and patent information. Examples of tools to be discussed include sophisticated text mining software packages, some simpler full-text searching tools, and a few data visualization tools that could be integrated with the more sophisticated software packages and full-text searching tools have been discussed.

Like Clear Forest Analysis certain other data visualization tools that are designed and used by certain vendors has been listed in [2].The following table depicts few tools their capabilities and their output result form and the website available.

Table 5
Text Mining Tools

| Vendor Tools | Working Mode | Capabilities | Data Sources | Results Output | Website |
|---|---|---|---|---|---|
| Cheshire3 | Fast XML search engine, written in Python for extensibility and using C libraries for speed. | search, retrieve, browse and sort | The British Library ISTC<br>The Archives Hub<br>EU co-funded Digital Preservation Project | Extract data into one or more indexes after processing with configurable workflows to add extra normalization and processing | http://www.cheshire3.org/ |
| CFG parser | The parser is currently tested only on linux and gcc. | information extraction system, | parse a huge collection of documents such as a Web corpus, or to build an interactive (real-time) information extraction system, | offers a reasonable performance (an f-score of 85%) with high-speed parsing (71 sentences/sec) | http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/chunkparser/ |
| Part-of-speech tagger | The tagger is tested only on linux and gcc. | large-scale information extraction and real-time NLP applications | web document corpus | Part-of-speech (POS) tagger offers fast tagging (2400 tokens/sec) with a state-of-the-art accuracy (97.10%). | http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger/ |
| Text Mining with Rapid Miner | Developed in Java | Clustering, classification, sentimental analysis and Information Extraction and Named Entity Recognition | Example e-mail spam detection, automatic e-mail routing, adaptive personal news filtering, sentiment analysis of text documents like news, web pages, blogs, e-mail, or PDF document | Form of structured data set | http://rapid-i.com |
| ClearForest Text Analytics | Text mining | Semantic analysis/ Natural Language Processing | Structured and unstructured text from web, internal documents patents, etc. | Structured data entities, lists, visualization tools – trend graphs, category maps | http://www.clearforest.com/Technology/TechnologyOverview.asp |

Generally, Patent documents contain important research results that are valuable to the industry, business, law, and policy-making communities. When carefully analyzed, they might show technological details and relations, reveal business trends, inspire novel industrial solutions, or help make investment policy (Campbell, 1983 and Jung, 2003).

Patent analysis or mapping requires considerable effort and expertise. As can been seen, these processes require the analysts to have a certain degree of expertise in information retrieval, domain-specific technologies, and business intelligence.

## 9. CONCLUSION

TM also known as Text Data Mining or KDT refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics.

As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge.

The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process.

## REFERENCES


[1] Andreas Hotho, Andreas N¨urnberger, Gerhard Paaß, "A Brief Survey of Text Mining" May 2005.

[2] Aurora Pons-Porrata, Rafael Berlanga-Llavori, Jose´ Ruiz-Shulcloper "Topic discovery based on text mining techniques" Information Processing and Management 43 (2007).

[3] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze "An Introduction to Information Retrieval", Cambridge University Press Cambridge, England.

[4] Fang Chen, Kesong Han and Guilin Chen (2008), "An Approach to Sentence Selection based Text Summarization", Proceedings of IEEE TENCON02, 489-493.

[5] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami, Pooya Khosravyan Dehkordy and Asghar Tajoddin (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE Computer Society, 347-352.

[6] Federico Michele Facca, Pier Luca Lanzi "Mining Interesting Knowledge from Weblogs: a Survey" Data & Knowledge Engineering 53 (2005).

[7] F. Wiesman a, Arie Hasman a, H.J. van den Herik "Information Retrieval: an Overview of System Characteristics" International Journal of Medical Informatics, 47 (1997).

[8] Gang Kou· Chunwei Lou, G. Kou C. Lou "Multiple Factor Hierarchical Clustering Algorithm for Large Scale Web Page and Search Engine Clickstream Data" Annals of Operations Research, 2010.

[9] JIAN-SUO XU (2007), "TCBPLK: A New Method of Text Categorization", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, IEEE, 3889-3892.

[10] Li Gao, Elizabeth Chang, and Song Han (2005), "Powerful Tool to Expand Business Intelligence: Text Mining", Proceedings of World Academy of Science, Engineering and Technology, 8, 110-115.

[11] M. Lia,*, M. Qib "MAPBOT: a Web based Map Information Retrieval System". Information and Software Technology , 45 (2003) 691–698.

[12] Ming Zhao, Jianli Wang and Guanjun Fan (2008), "Research on Application of Improved Text Cluster Algorithm in Intelligent QA System", Proceedings of the Second International Conference on Genetic and Evolutionary Computing, China, IEEE Computer Society, 463-466.

[13] Rowena Chau*, Chung-Hsing Yeh. "A Multilingual Text Mining Approach to Web Cross-lingual Text Retrieval" Knowledge-Based Systems, 17 (2004) 219–227.

[14] Shu-Sheng Liaw, Hsiu-Mei Huang, "Information Retrieval from the World Wide Web: a User-focused Approach based on Individual Experience with Search Engines". Computers in Human Behavior, 22 (2006).

[15] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), "Experiments on Supervised Learning Algorithms for Text Categorization", International Conference, IEEE Computer Society, 1-8.

[16] Sougata Mukherjea, Kyoji Hirata, Yoshinori Hara' "Towards a Multimedia World-Wide Web Information Retrieval Engine". Computer Networks and ISDN Systems, 29 (I 997).

[17] Tamara Polajnar "Survey of Text Mining of Biomedical Corpora". August 2006.

[18] Tao Jiang, Ah-hwee Tan, Senior Member, IEEE, and Ke Wang "Mining Generalized Associations of Semantic Relations from Textual Web Content" IEEE Transactions on Knowledge and Data Engineering, 19, No. 2, February 2007.

[19] Vishal Gupta, Gurpreet S. Lehal. "A Survey of Text Mining Techniques and Applications". Journal of Emerging Technologies in Web Intelligence, 1, No. 1, August 2009".

[20] YunYun Yang *, Lucy Akers, Thomas Klose, Cynthia Barcelon "Text Mining and Visualization Tools – Impressions of Emerging Capabilities". World Patent Information, 30 (2008).

[21] Zhou Ning, Wu Jiaxin, Wang Bing and Zhang Shaolong (2008), "A Visualization Model for Information Resources Management", 12th International Conference Information Visualisation, China, IEEE, 57- 62.

[22] Jiexun Li, Harry Jiannan Wang, Zhu Zhang and J. Leon Zhao "A Policy-based Process Mining Framework: Mining Business Policy Texts for Discovering Process Models" Information System s and E-Business Management, 2010.