Text Mining and Its Applications

Bharati N Kharade

Abstract: **As computer networks become the backbones of science and economy, enormous quantities of machine readable documents become available. Computerization and automated data gathering has resulted in extremely large data repositories e.g. Walmart: 2000 stores, 20 M transactions/day. Unfortunately, the usual logic-based programming paradigm has great difficulties in capturing the fuzzy and often ambiguous relations in text documents. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text.**

In this paper, text mining is described as a method for information retrieval, machine learning, statistical analysis and especially data mining. First these methods are given and then defined text mining in relation to them. Later sections give different approaches for the main analysis tasks preprocessing, classification, clustering, information extraction and visualization. The last section explains number of successful applications of text mining.

Keywords: **data mining, machine learning, text mining, text categorization, clustering, text visualization**

1. INTRODUCTION

Text mining is a new area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. [2]

The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, etc., or you could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. In the most general terms, text mining will "turn text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects,
the application of unsupervised learning methods (clustering), etc.

G.H. Raisoni College of Engineering and Management,Pune.
bharati27@gmail.com

2. KNOWLEDGE DISCOVERY IN DATABASES

Fayyad has defined Knowledge Discovery in Databases (KDD) as follows [1]:

”Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”

Knowledge discovery in databases is a process that is defined by several processing steps that have to be applied to a data set of interest in order to extract useful patterns. These steps have to be performed iteratively and several steps usually require interactive feedback from a user.

2.1 Data Mining, Machine Learning and Statistical Learning

There is *data mining as synonym for KDD*, meaning that data mining contains all aspects of the knowledge discovery process. *Data mining is considered as part of the KDD-Processes* and describes the modeling phase, i.e. the application of algorithms and methods for the calculation of the searched patterns or models.

Databases are necessary in order to analyze large quantities of data efficiently. Since the analysis of the data with data mining algorithms can be supported by databases and thus the use of database technology in the data mining process might be useful.

Machine Learning (ML) is an area of artificial intelligence concerned with the development of techniques which allow computers to learn by the analysis of data sets. ML is also concerned with the algorithmic complexity of computational implementations.

Statistics deals with the science and practice for the analysis of empirical data. It is based on statistical theory which is a branch of applied mathematics. Within statistical theory, randomness and uncertainty are modeled by probability theory. Today many methods of statistics are used in the field of KDD [2].

2.2 Text Mining

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. There are different definitions of text mining, according to specific perspective of different research areas:

Text Mining = Information Extraction. The first approach assumes that text mining essentially corresponds to information extraction the extraction of facts from texts.

Text Mining = Text Data Mining. Text mining can be also defined similar to data mining as the application of algorithms and methods from the fields of machine learning and statistics to texts with the goal of finding useful patterns. For this purpose it is necessary to pre-process the texts accordingly. Many authors use information extraction methods, natural language processing or some simple preprocessing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied.

Text Mining = KDD Process. Following the knowledge discovery process model, we frequently find in literature text mining as a process with a series of partial steps to extract information as well as the use of data mining or statistical procedures for the extraction of not yet discovered information in large collection of text [2].

3. DATA MINING METHODS FOR TEXT

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. There are following methods of Text Mining [1][2].

3.1 Classification and Clustering Methods

Classification methods are used to assign data to predefined categories. A variety of techniques are available (e.g., decision trees, naïve Bayesian classifiers, and nearest neighbor classifiers).

Text classification aims at assigning pre-defined classes to text documents. An example would be to automatically label each incoming news story with a topic such as ”sports”, ”politics”, or ”art”. To measure the performance of a classification model a random fraction of the labeled documents is set aside and not used for training. We may classify the documents of this *test set* with the classification model and compare the estimated labels with the true labels. The fraction of correctly classified documents in relation to the total number of documents is called *accuracy* and is a first performance measure. *Precision* quantifies the fraction of retrieved documents that are in fact relevant, i.e. belong to the target class. *Recall* indicates which fraction of the relevant documents is retrieved.

precision = #{relevant∩ retrieved} / #retrieved
recall = #{relevant ∩ retrieved} / #relevant

Clustering seeks to identify a finite set of abstract categories that describe the data by determining natural affinities in the data set based upon a pre-defined distance or similarity measure. Clustering can employ categories of different types (e.g., a flat partition, a hierarchy of increasingly fine-grained partitions, or a set of possibly overlapping clusters). Clustering can proceed by agglomeration, where instances are initially merged to form small clusters and small clusters are merged to form larger ones; or by successive division of larger clusters into smaller ones. Some clustering algorithms produce explicit cluster descriptions; others produce only implicit descriptions [1][2].

3.1.1. Nearest Neighbor Methods

Nearest Neighbor algorithms support clustering and classification by matching cases internally to each other. A simple example of a nearest neighbor method would be as follows: given a set $\boldsymbol{X} = \{x_1\ x_2\ x_3\ldots\ x_n\}$ of vectors composed of *n* features with binary values, for each pair (*xi*, x_j), $x_i^1\ x_j$,, create a vector *vi* of length *n* by comparing the values of each corresponding feature n_i *of* each pair (x_i , x_j), entering a 1 for each n_i feature with matching values match and 0 otherwise. Then sum the v_i values to compute the degree of match. Those pairs (x_i, x_j) with the largest result are the nearest neighbors. In more complex nearest neighbor methods, features can be weighted to reflect degree of importance. Domain expertise is needed to select salient features, compute weights for those features, and select a distance or similarity measure. Nearest neighbor approaches have been used for text classification [2].

3.1.2. Relational Learning Models

Relational learning models are inductive logic programming applications. Their foundation is logic programming using Horn clauses, a restricted form of first-order predicate logic. Logic programming describes relations on objects using declarative subject-predicate representations and uses classical deductive logic to draw conclusions. Data mining has been conducted by using inductive logic programming to generate database queries with predicate logic query syntax [2].

3.1.3. Genetic Algorithms

Genetic algorithms can be used both for classification and for discovery of decision rules. Named for their Darwinist methodology, genetic algorithms use processing that is analogous to DNA recombination. A population of "individuals," each representing a possible solution to a problem, is initially created at random or drawn randomly from a larger population. Pairs of individuals combine to produce "offspring" for the next generation, and mutation processes are used to randomly modify the genetic structure of some members of each new generation. Genetic algorithms perform categorization using supervised learning, training with a set of data and then using the known correct answers to guide the evolution of the algorithm using techniques akin to natural selection. Genetic methods have advantages over neural networks, because they provide more insight into the decision process [2].

3.2 Statistical Methods

3.2.1 Linear Regression and Decision Trees

Linear regression (or *correlation*) methods are used to determine the relationships between variables to support classification, association and clustering. Variations include univariate and multivariate regression. One common use of linear regression is to support generation of a decision tree. Decision trees are typically induced using a recursive algorithm that exhaustively partitions the data starting from an initial state in which all training instances are in a single partition, represented by the root node, and progressively creates sub-partitions that are represented by internal or leaf nodes. Each node will correspond to a rule characterizing some explicit property of the data, so generation of a decision tree is a restricted form of rule induction in which the resulting rules are mutually exclusive and exhaustive. Decision tree induction is fairly straightforward, but the results will only be useful if the available features provide sufficient basis meaningful categorization. To reduce computational complexity, heuristics are often applied to the selection of linear properties that implicitly omit from consideration the vast majority of potential rules. Rule extraction from decision trees can be used in data mining to support hypothesis validation [1] [5].

3.2.2. Nonlinear Regression and Neural Networks

Nonlinear Regression algorithms are used to support classification, association and clustering. Neural networks determine implicit rules where the classes invoked are not defined classically. One objection to the use of neural networks is that "the results often depend on the individual who built the model". This is because the model, the network topology and initial weights, may differ from one implementation to another for the same data. Unsupervised learning methods require no feedback from a domain expert; instead, the network is used to discover categories based on correlations within the data. The alternative is supervised (or reinforcement) learning, in which expert feedback is given as part of the training set to indicate whether a solution is correct or incorrect.

4. APPLICATIONS OF TEXT MINING

4.1 Analyzing open-ended survey responses:

In survey research (e.g., marketing), it is not uncommon to include various open-ended questions pertaining to the topic under investigation. The idea is to permit respondents to express their "views" or opinions without constraining them to particular dimensions or a particular response format. This may yield insights into customers' views and opinions that might otherwise not be discovered when relying solely on structured questionnaires designed by "experts". For example, you may discover a certain set of words or terms that are commonly used by respondents to describe the pros and cons of a product or service (under investigation), suggesting common misconceptions or confusion regarding the items in the study[6].

4.2 Automatic processing of messages and emails:

Another common application for text mining is to aid in the automatic classification of texts. For example, it is possible to "filter" out automatically most undesirable "junk email" based on certain terms or words that are not likely to appear in legitimate messages, but instead identify undesirable electronic mail. In this manner, such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed (automatically) to the most appropriate department or agency; e.g., email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments; at the same time, the emails are screened for inappropriate or obscene messages, which are automatically returned to the sender with a request to remove the offending words or content[6].

4.3 Analyzing warranty or insurance claims, diagnostic interviews, etc.

In some business domains, the majority of information is collected in open-ended, textual form. For example, warranty claims or initial medical (patient) interviews can be summarized in brief narratives, or when you take your automobile to a service station for repairs, typically, the attendant will write some notes about the problems that you report and what you believe needs to be fixed. Increasingly, those notes are collected electronically, so those types of narratives are readily available for input into text mining algorithms. This information can be usefully exploited to, for example, identify common clusters of problems and complaints on certain automobiles, etc. Likewise, in the medical field, open-ended descriptions by patients of their own symptoms might yield useful clues for the actual medical diagnosis.

4.4 Investigating competitors by crawling their web sites.

Another type of potentially very useful application is to automatically process the contents of Web pages in a particular domain. For example, you could go to a Web page, and begin "crawling" the links you find there to process all Web pages that are referenced. In this manner, you could automatically derive a list of terms and documents available at that site, and hence quickly determine the most important terms and features that are described. It is easy to see how these capabilities could efficiently deliver valuable business intelligence about the activities of competitors.

4.5 Enhancing Web Search

One way to enhance users' efficiency and experience of Web search is by means of *meta-search engines*. Traditionally, meta-search engines were conceived to address different issues concerning general-purpose search engines, including Web coverage, search result relevance, and their presentation to the user. A common approach to alternative presentation of results is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered "maps" (kartoo.com). However, topics generated by clustering may not prove satisfactory for every query [5].

4.6 Patent Analysis

In recent years the analysis of patents developed to a large application area. The reasons for this are on the one hand the increased number of patent applications and on the other hand the progress that had been made in text classification, which allows to use these techniques in this due to the commercial impact quite sensitive area. Meanwhile, supervised and unsupervised techniques are applied to analyze patent documents and to support companies and also the European patent office in their work. The challenges in patent analysis consist of the length of the documents, which are larger then documents usually used in text classification. Usually every document consists of 5000 words in average. More than 140000 documents have to be handled by the European patent office (EPO) per year. They are processed by 2500 patent examiners in three locations. In several studies the classification quality of state-of-the-art methods was analyzed. Text clustering techniques for patent analysis are often applied to support the analysis of patents in large companies by structuring and visualizing the investigated corpus. Thus, these methods find their way in a lot of commercial products but are still also of interest for research, since there is still a need for improved performance [2].

4.7 Text Classification for News Agencies

In publishing houses, a large number of news stories arrive each day. The users like to have these stories tagged with categories and the names of important persons, organizations and places. To automate this process the Deutsche Presse-Agentur (DPA) and a group of leading German broadcasters (PAN) wanted to select a commercial text classification system to support the annotation of news articles. Seven systems were tested with a two given test corporation of about half a million news stories and different categorical hierarchies of about 800 and 2300 categories. The Deutsche Presse-Agentur now is routinely using a text mining system in its news production workflow [4]. Due to confidentiality, the results can be published only in anonymized form. For the corpus with 2300 categories the best system achieved at an F1-value of 39%, while for the corpus with 800 categories an F1-value of 79% was reached. In the latter case, a partially automatic assignment based on the reliability score was possible for about half the documents, while otherwise the systems could only deliver proposals for human categorizers. Especially good are the results for recovering persons and geographic locations with about 80% F1-value. In general there were great variations between the performances of the systems. In usability experiment with human annotators the formal evaluation results were confirmed leading to faster and more consistent annotation. It turned out, that with respect to categories the human annotators exhibit a relative large disagreement and a lower consistency than text mining systems. Hence the support of human annotators by text mining systems offers more consistent annotations in addition to faster annotation [2].

4.8 Bioinformatics

Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that corresponds to instances of concepts that are of interest to biologists. Examples of such entities include the names of proteins, genes and their locations of activity such as cells or organism names. Entity recognition is becoming increasingly important with the massive increase in reported results due to high throughput experimental methods. It can be used in several higher level information access tasks such as relation extraction, summarization and question answering. For practical applications the current accuracy levels are not yet satisfactory and research currently aims at including a sophisticated mix of external resources such as keyword lists and ontologies which provide terminological resources [2].

4.9 Anti-Spam Filtering of Emails

The explosive growth of unsolicited e-mail, more commonly known as spam, over the last years has been undermining constantly the usability of e-mail. One solution is offered by anti-spam filters. Most commercially available filters use black-lists and hand-crafted rules. On the other hand, the success of machine learning methods in text classification offers the possibility to arrive at anti-spam filters that quickly may be adapted to new types of spam. There is a growing number of learning spam filters mostly using naive Bayes classifiers. A prominent example is Mozilla's e-mail client. Michelakis et al. [MAP+04] compare different classifier methods and investigate different costs of classifying a proper mail as spam. They find that for their benchmark corpora the SVM nearly always yields best results. To explore how well a learning-based filter performs in real life, they used an SVMbased procedure for seven months without retraining. They achieved a precision of 96.5% and a recall of 89.3%. They conclude that these good results may be improved by careful preprocessing and the extension of filtering to different languages [2].

4.10 Mining Bibliographic Data

Vojvodina, the northern province of Serbia, is home to many educational and research institutions. In 2004, the Provincial Secretariat for Science and Technological Development of Vojvodina started collecting data from researchers employed at institutions within its jurisdiction. Every researcher was asked to fill in a form, provided as an MS Word document, with bibliographic references of all authored publications, among other data. Notable properties of the collection are its incompleteness and diversity of approaches to giving references, permitted by information being entered in free text format [5][6].

5. CONCLUSION

In this paper, a brief introduction is given to the broad field of text mining. A more formal definition of the term is used herein and presented a brief overview of currently available text mining methods, their properties and their applications to specific problems. Text mining concept, its approaches and few applications in different areas are described.

6. REFERENCES

1. Han.J, Kamber. M. Data Mining Concepts and Techniques.
2. Andreas N¨urnberger, Gerhard Paaß, Fraunhofer AiS. A Brief Survey of Text Mining.
3. Feldman, R., Sanger, J., The Text Mining Handbook. Cambridge University Press, 2007.
4. C.H.A. Koster, M. Seutter, and J. Beney. Classifying patent applications with window. In *Proceedings Benelearn,* Antwerpen, 2001.
5. G. Paaß and H. deVries. Evaluating the performance of text mining systems on real-world press archives. In *Proc. 29th Annual Conference of the German Classification Society (GfKl 2005).* Springer, 2005.
6. Miloš Radovanovic, Mirjana Ivanovic. Text Mining: Approaches And Applications