A Systematic study of Text Mining Techniques

Pravin Shinde & Sharvari Govilkar

Dept. of Information Technology, Mumbai University

ABSTRACT

Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from machine learning, natural language processing (NLP), data mining, information retrieval (IR), and knowledge management. Text mining involves the pre-processing of document collections such as information extraction, term extraction, text categorization, and storage of intermediate representations. The techniques that are used to analyse these intermediate representations such as clustering, distribution analysis, association rules and visualisation of the results.

KEYWORDS

Text categorization, IR, clustering, visualisation.

1. INTRODUCTION

Text mining can be referred as a knowledge intensive process in which using a various suites of analysis tools, user interacts with a document collection. The text mining also extracts the useful information from data sources through the explorations and identifications of interesting patterns, which are similar or analogous to data mining. In this case of text mining, the data sources are document collections, and patterns are not found among formalised database records but in the unstructured textual data in the documents in these collections.

Certainly, from seminal research on data mining the text mining derives much of its direction and inspiration. So, it is not surprising to find that data mining and text mining systems have many high-level architectural similarities. For instance, both types of systems rely or based on pattern-discovery algorithms, presentation-layer elements and pre-processing routines such as visualisation tools to enhance the output data. Further, text mining adopts many of the specific types of patterns in its core knowledge discovery operations that were first introduced and vetted in data mining research.

2. TEXT ENCODING

It is necessary to pre-process the text documents and store the information in a data structure for mining large document collections, which is more suitable for further processing than a plain text file. Various methods exist that try to exploit also the syntactic structure and semantics of text document, most text mining approaches are based on the idea that a text document can be represented by a set of words, which means a text document is described based on the set of words contained in it.

2.1. Text Mining Pre-processing Techniques

There are two ways of categorizing the structuring techniques of document are according to their task, algorithms and formal frameworks that they use.

DOI: 10.5121/ijnlc.2015.4405

Task oriented pre processing approaches envision the process of creating a structured document representation in terms of tasks and subtasks and usually involve some sort of preparatory goal or problem that needs to be solved such as extracting titles and authors from a PDF. In pre processing approaches are rely on techniques such that classification schemes, probabilistic models, and rule-based systems approaches for analysing complex phenomena that can be also applied to natural language texts.

2.1.1. Task Oriented Approach

A document has a variety of possible representations tree. The task of the document parsing process is to take the most raw representation and convert it to the representation through which the meaning of the document surfaces.

A divide and conquer strategy is typically selected to face with this extremely difficult problem and the problem is divided into a set of subtasks, each of which is solved separately. The subtasks can be divided broadly into three classes preparatory processing, general purpose NLP tasks, and problem dependent tasks.

The task of the preparatory processing is to convert the raw input into a stream of text, possibly labelling the internal text zones such as paragraphs, or tables, columns. Sometimes it is possible to extract some document level fields such as <Author> or <Title> in cases in which the visual position of the fields allows their identification.

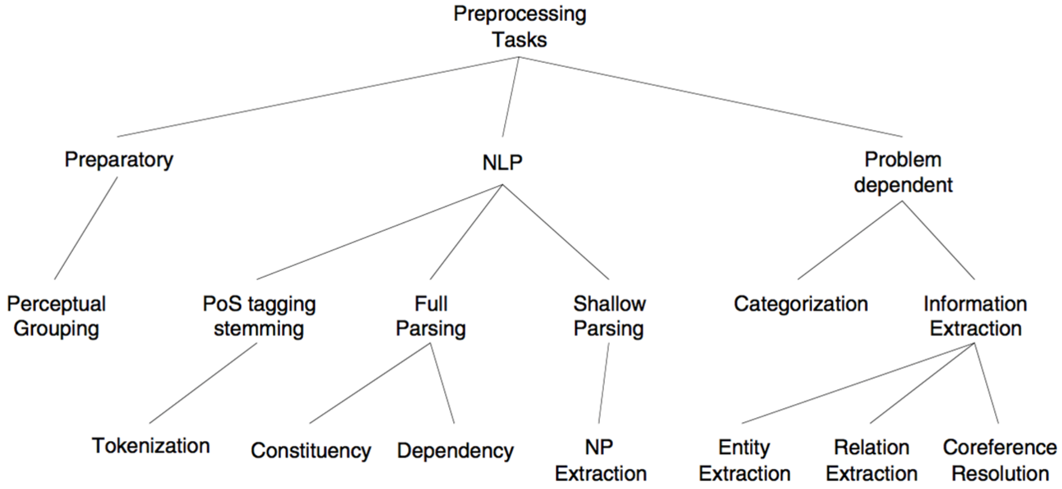

Fig.1. A taxonomy of text preprocessing tasks. [3]

2.1.1.1. General Purpose NLP Tasks

It is currently an orthodox opinion that language processing in humans cannot be separated into independent components. Various experiments in psycholinguistics clearly demonstrate that the different stages of analysis like phonetic, morphological, syntactical, semantical, and pragmatically occur simultaneously and depend on each other.

2.2. Problem-Dependent Tasks: Text Categorization and Information Extraction

The final stages of document structuring create representations that are meaningful for either later processing phases or direct interaction of the text mining system user. The nature of the features sharply distinguishes between the two main techniques: text categorisation and information extraction (IE). Text categorisation and IE enable users to move from a "machine readable" representation of the documents to a "machine understandable" form of the documents.

3. Categorization

Probably the most common portion in analysing complex data is the categorization or classification of elements. Described abstractly, the task is to classify a given data instance into a pre-specified set of categories. Applied to the domain of document management, the task is known as text categorization, given a set of categories (subjects, topics) and a collection of text documents.

3.1. Machine Learning Approach to TC

In this approach, by learning the properties of categories from a set of pre classified training documents, the classifier is built automatically. In this case the learning process is an instance of supervised learning because the process is guided by applying the known true category assignment function on the training set. The clustering is also called as unsupervised version of the classification task. For classifier learning there are many approaches available some of them are variants of more general ML algorithms and others have been created specifically for categorization.

3.1.1. Probabilistic Classifiers

Probabilistic classifiers show the categorization status value CSV (d, c) with the probability P(c | d) where document d belongs to the category c and compute this probability by an application of Bayes' theorem:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

The marginal probability P(d) need not be computed because it is constant for all categories. To calculate P(d | c), we need to make some assumptions about the structure of the document d. With the document representation as a feature vector d = (w1, w2 , . . .), the most common assumption is that all coordinates are independent, and thus the classifiers resulting from this assumption are called Naive Bayes (NB) classifiers. They are called "naive" because the assumption is never verified and often is quite obviously false. However, the attempts to relax the naive assumption and to use the probabilistic models with dependence so far have not produced any significant improvement in performance.

$$P(d|c) = \Pi_i P(W_i|c)$$

3.1.2. Decision Tree Classifiers

A decision tree (DT) classifier is a tree in which the internal nodes are labelled by the features, the edges leaving a node are labelled by tests on the feature's weight, and the leaves are labelled by categories. A DT categorises a document by starting at the root of the tree and moving successively downward via the branches whose conditions are satisfied by the document until a leaf node is reached. The document is then assigned to the category that labels the leaf node.

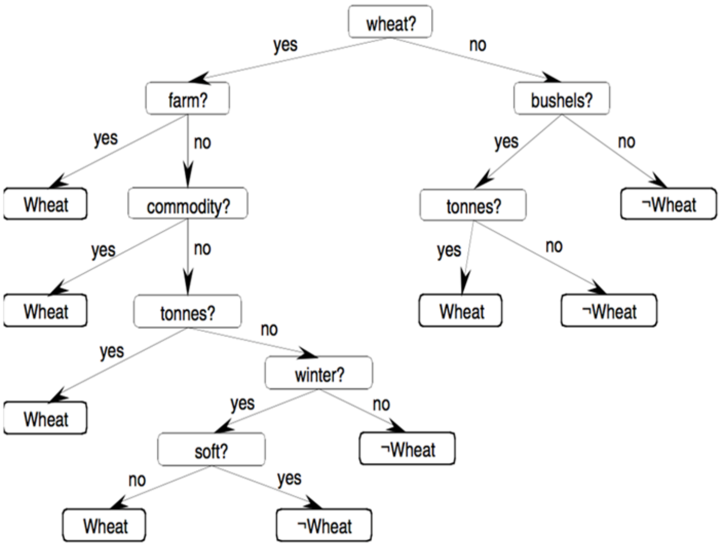

Fig. 2 A Decision Tree classifier. [3]

3.1.3. Neural Networks

Neural network (NN) can be built to perform text categorization. Normally, the input nodes of the network receive the feature values categorization status values produced by output nodes and the dependence relations represent by link weights. For classifying a document the feature weights are loaded into the input nodes, activation of the nodes is propagated forward through the network, and the final values on output nodes determine the categorization decisions.
The NN are trained by back propagation, where as the training documents are loaded into the input nodes. If a misclassification error occurs then it is propagated back through the network and modifying the link weights in order to minimise the error.

3.1.4. Support Vector Machines

The support vector machine (SVM) algorithm is very effective and fast for text classification problems.

A binary SVM classifier in geometrical terms can be seen as a hyperplane in the feature space separating the points that represent the positive instances of the category from the points that represent the negative instances. The classifying hyperplane is chosen during training as the unique hyperplane that separates the known positive instances from the known negative instances with the maximal margin. The margin is the distance from the hyperplane to the nearest point from the positive and negative sets. The Figure 3 is an example of a maximal margin hyperplane in two dimensions.

SVM hyperplane are determined by a relatively small subset of the training instances which are called the support vectors. The SVM classifier has an important advantage in its theoretically justified approach to the over fitting problem, which allows it to perform well irrespective of the dimensionality of the feature space. Also, it needs no parameter adjustment.

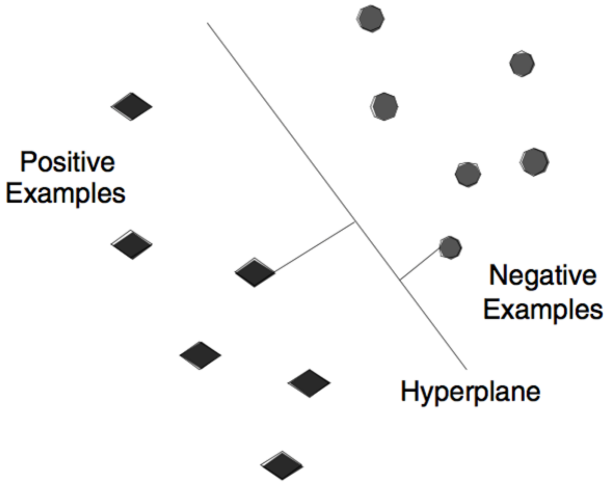

Fig. 3. Diagram of a 2-D Linear SVM. [3]

4. Clustering

Clustering method can be used in order to make groups of documents with similar content or information. The result of clustering is typically a partition P which is a set of clusters P. Every cluster consists of a number of documents they should be similar and dissimilar to clusters of other documents. Clustering algorithms compute the clusters based on the attributes of the data and measures of similarity or dissimilarity.

4.1. Clustering Algorithms

Several different variants of an abstract clustering problem exist. A flat (or partitioned) clustering produces a single partition of a set of objects into disjoint groups, whereas a hierarchical clustering results in a nested series of partitions.

The most commonly used algorithms are the K-means (hard, flat, shuffling), the EM-based mixture resolving (soft, flat, probabilistic), and the HAC (hierarchical, agglomerative).

4.1.1. K-Means Algorithm

The K-means algorithm partitions a collection of vectors {x1, x2,..,xn} into the set of clusters {C1, C2, . . . Ck}. The algorithm needs k cluster seeds for initialization. They can be externally supplied or picked up randomly among the vectors.

The algorithm proceeds as follows:

Initialization

K seeds, either given or selected randomly, form the core of k clusters. Every other vector is assigned to the cluster of the closest seed.

Iteration:

The centroid Mi of the current cluster is computed:

$$M_i = |C_i|^{-1} \sum_{x \in C_i}^{l} x$$

Each vector is reassigned to the cluster with the closest centroid.
Stopping condition:
At convergence – when no more changes occur.
The K-means algorithm maximises the clustering quality function Q:

$$Q(C_1, C_2, \dots, C_k) = \sum_{C_i} \sum_{x \in C_i} Sim(x - M_i)$$

If the distance metric (inverse of the similarity function) behaves well with respect to the centroids computation, then each iteration of the algorithm increases the value of Q. A sufficient condition is that the centroid of a set of vectors be the vector that maximises the sum of similarities to all the vectors in the set. This condition is true for all "natural" metrics. It follows that the K-means algorithm always converges to a local maximum.

The K-means algorithm is popular because of its simplicity and efficiency. The complexity of each iteration is O(kn) similarity comparisons, and the number of necessary iterations is usually quite small.

4.2 Hierarchical Agglomerative Clustering (HAC)

The HAC algorithm begins its work with each object in particular cluster and proceeds, according to some chosen criterion it is repeatedly merge pairs of clusters that are most similar. The HAC algorithm finishes when everything is merged into a single cluster. Binary tree of the clusters hierarchy is provided by history of merging.
The algorithm proceeds as follows:

Initialization:
Each and every object is put into a separate cluster.

Iteration:
Find the pair of most similar clusters and merge them.

Stopping condition:
Repeat step 2 till single cluster is formed.

When everything is merged into single cluster different versions of the algorithm can be produced, then it is calculated the similarity between clusters. The complexity of this algorithm is O(n2s), where n is the number of objects and s the complexity of calculating similarity between clusters. Measuring the Quality of an algorithm needs human judgment, which introduces a high degree of subjectivity.

Given a set of categorised (manually classified) documents, it is possible to use this benchmark labelling for evaluation of clustering's. The most common measure is purity. Assume {L1, L2,..., Ln} are the manually labelled classes of documents, and {C1, C2, . . . , Cm} are the clusters returned by the clustering process. Then,

$$Purity(C_i) = max_j |L_j \cap C_i| / |C_i|$$

5. Information Extraction

The Natural language texts have information, which is not suitable for computers for analysis purpose. Where as computers uses large amount of text and extract useful information from passages, phrases or single words. So Information Extraction can be considered as restricted form of natural language understanding and here we know about the semantic information, we are seeking for. The task of information Extraction is to extract parts of text and assign specific attribute to it.

5.1. Hidden Markov Models

One of the main problem of standard classification approaches they are not considered the predicted labels of the surrounding words and it can be done using probabilistic models of sequences of labels and features. The Hidden Markov model (HMM) based on conditional distributions of current labels L(j) given the previous label L(j−1) and the distribution of the current word t(j) given the current and the previous labels L(j), L(j−1).

$$p(L|t) = \frac{1}{const} exp\left(\sum_{j=1}^{n}\sum_{r=1}^{k_j} \lambda_{jr} f_{jr}(L_j, t) + \sum_{j=1}^{n-1}\sum_{r=1}^{m_j} \mu_{jr} g_{jr}(L_j, L_{j-1}, t)\right)$$

The algorithm is required the training set and their correct label for computing their frequency. The Viterbi algorithm is an efficient learning method which exploit the sequential structure. The HMM were successfully used for named entity extraction.

6. Visualization Methods

The Information provided by graphical visualization is more better, comphrensive and faster understandable than pure text based description so it is best for mining the large document collection. Most of the approaches of text mining are motivated by the methods which had been proposed in the area of visual data mining, information visualizations and explorative data mining.

This method can improve the discovery or extraction of relevant patterns or information for text mining and information retrieval systems. Information that allow a visual representation comprises aspects of result set, keyword relations or ontology are considered the aspects of the search process itself.

7. Applications and merits/demerits

Classification of news as a Text: In the daily newspaper the users would like to see stories of people at different places and organizations etc. such task are tedious when we do it manually. So in this case text mining approach like information extraction can be used to do this kind of task which would retrieve the template having different entity and their relationship with each other in the structured format. Which can be putted into the database, then we can applied for retrieving the interesting patterns.

Analysis of the Market trends: Everybody knows that corporate market around us is how much growing fast, in order to know about our competitors and the growth of an organizations and their number of the employees. To get such information, manual work is a tedious task or impossible task. But by using text mining approaches like classifications or information extractions it is easy to simplify the task.

Analysis of the junk Emails: This is a common application for text mining is in automatic analysis of the junk E-mails that are undesirable. The classification technique of text mining can be used to classify such mails on the basis of pre-defined frequently occurring terms.

Merits of Text mining:

i) As database can store less amount of information, this problem has been solved through Text Mining.

ii) Using the technique such as information extraction, the names of different entities, relationship between them can easily be found from the corpus of documents set.

iii) Text mining has solved the problem of managing such a great amount of unstructured information for extracting patterns easily; otherwise it would have been a great challenge.

Demerits of Text mining:

i) No programs can be made in order to analyse the unstructured text directly, to mine the text for information or knowledge.

ii) The information which is initially needed is nowhere written.

8. Conclusion

In this paper the introduction of text mining and its methods has been tried to cover. Because of this we motivated this field of research, and gave more formal definition to the terms, which are used herein and presented the brief overview of text mining and its methods, their properties and their applications.

Now days there has been lot of work did on the document using text mining methods. The improvement for text mining is still an interesting, open issue and as in current world scenario time is the prime constraint of any application. So as to do fast work with highest performance one can think to implement the existing methods on parallel platform.

REFERENCES

[1] M. Nagy and M. Vargas-Vera, “Multiagent ontology mapping framework for these mantic web,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 693–704, Jul. 2011.

[2] C. Lu, X. Hu, and J. R. Park, “Exploiting the social tagging network for web clustering,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no.5, pp. 840–852, Sep. 2011.

[3] R. Feldman and J. Sanger, the Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data. New York: Cambridge Univ. Press, 2007.

[4] M. Konchady, Text Mining Application Programming. Boston, MA: Charles River Media, 2006.

[5] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,“Effective Pattern Discovery for Text Mining,” IEEE Trans. on knowledge and Data engineering, vol. 24, no. 1, Jan.2012

[6] Calvillo, E.A.Padilla, A. Munoz, J. Ponce, J. “Searching Research Papers Using Clustering and Text Mining”, IEEE Conference Publication ,11-13 March 2013.

[7] Rodrigo Miranda Feitosa, Nilson Santos, “Social Recommendation in Location- Based Social Network using Text Mining,” 2013 4th International Conference on Intelligent Systems, Modelling and Simulation.

[8] Shaidah Jusoh and Hejab M. Alfawareh,“Techniques, Applications and Challenging Issue in Text Mining,” IJCSI, Vol. 9, Issue 6, No 2, November 2012

[9] Mrs. Sayantani Ghosh, Mr. Sudipta Roy,“A tutorial review on Text Mining Algorithms”, IJARCCE, Vol. 1, Issue 4, June 2011

[10] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/text.htm#CIHFDAAB "Oracle® Data Mining Concepts of Text mining"

[11] http://www.ijarcce.com/upload/june/6-A%20tutorial%20review%20on%20Text%20Mining%20Algorithms.pdf "A tutorial review on Text Mining Algorithms"

[12] http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf. "A Brief Survey of Text Mining"

Authors

Mr Pravin Shinde, Pursuing ME (Artificial Intelligence and Robotics) from Pillai Institute of Information Technology, New Panvel.

Sharvari Govilkar is working as Associate professor in Department of Information Technology at Pillai Institute of Information Technology, New Panvel. Her qualifications are M.E. Computer Science, Ph.D. (pursuing) and having more than 14 years of teaching experience.