IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.7, Special Issue.4, Feb. 2019 E-ISSN: 2347-2693

High Utility Text and Data Mining Methods

R. Kanimozhi

Dept. of Computer Application, Idhaya College for Women, Kumbakonam, Tamilnadu, India

Corresponding Author: Kanimvenkat@gmail.com

Available online at: www.ijcseonline.org

***Abstract*— Text data has continuous growth of volumes of data, automate extract ion of implicit, previously unknown, and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining corresponds to the extension of the data mining approach to textual data and is concerned with various tasks, such as extraction of information implicitly contained in collection of documents, or similarity-based structuring. This paper provides the reader with a very brief introduction to some of the theory and methods of text data mining. The intent of this paper is to introduce some of the current text mining methods that are employed within this discipline area. In this paper we provide some of methods of text datamining.**

***Keywords*— Text Mining, Text Mining Text Processing, Methods Text, Document clustering.**

I. INTRODUCTION

Besidestheusualquantitativeorqualitativedatafamiliartostatisti cians,thedatathatservesasinputto the information extraction algorithms can take any form, including imagery, video, audio or text. Here we will focus on text-based data sources. The textual data sources for information extraction can range from free from text to semi formatted text(html,xmland others) and includes those sources that are encoded in open source document formats (openDocument) and other proprietaryformats(MicrosoftWordandMicrosoftPowerPoint). Ifeelthattheproblem of extracting information from these datasources is one that offers great challenges to the statistical community. Theplanforthisarticleistodiscuss some of these challenges and to generate interest in the community in this particular topic area. Data mining on text has been designated at various times as statistical textprocessing. Knowledgediscoveryintext,intelligenttextanalysis,ornaturalla nguage processing, depending on the application and the methodology that is used[1]. Examples of text mining tasks include classifying document ssuch that each member of each group has similar meaning(clustering or unsupervised learning), and finding documents that satisfy some search criteria (information retrieval). In the interest of space, I will not provide discussions on information retrieval. The reader is referred to the recent work by Michael Berry that provides discussions on some of the recent work in these areas [2]

Text collection, in general, lacks the imposed structure of a traditional database. The text expresses a vast range of information, but encodes the information in a form that is difficult to decipher automatically. The data mining techniques are essentially designed to operate on structured databases. When the data is structured it is easy to define the set of items and hence, it becomes easy to employ the traditional mining techniques. Identifying individual items or terms is not so obvious in a textual database. Thus, unstructured data, particularly free-running text, places a new demand on data mining methodology. Specific techniques, called text mining techniques, have to be developed to process the unstructured textual data to aid in knowledge discovery.

The inherent nature of textual data, namely unstructured characteristics, motivates the development of separate text mining techniques. One way is to impose a structure on the textual database and use any of the known data mining techniques meant for structured databases. The other approach would be to develop a very specific technique for mining that exploits the inherent characteristics of textual databases. Irrespective of the approach chosen for text mining, one cannot ignore the close interactions of other related subjects, such as computational linguistics, natural languages processing, and information retrieval.

II. RELATEDWORK

The application of data mining to non-structured or less-structured text files. Text mining helps the organizations to find the hidden content of documents, including additional

© 2019, IJCSE All Rights Reserved

useful relationship and group documents by common themes. The use of Text Mining is the data mining a decision support methods assume that the data is stored in one or more tables, organized in a number of fields with a predefined range of possible values the application the previous studies of data mining have focused on structured data, such as relational, transactional, and data warehouse data. the available information is stored in text databases(ordocument databases),which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic forms, such as electronic publications, e- mail, CD-ROMs, and the www.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research. Information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. There is a relationship between areas like Information retrieval (IR) , Information Extraction(IE), and computational linguistics with text datamining.

Information Retrieval

IR is concerned with finding and ranking documents that match the user's information needs. The way of dealing with textual information by the IR community is a keyword-based document representation. A body of text is analyzed by its constituent word, and various techniques are used to build the core words for a document. The goals are to find documents that are similar, based on some specification of the user and to find the right index terms in a collection, so that querying will return the appropriatedocument.

2.1.1 Measures for Text Retrieval

A text retrieval system has just retrieved a number of documents based on input in the form of a query. we can assess accurate or correct system as let the set of documents relevant to a query be denoted as

{Relevant} , and the set of documents retrieved be denoted as {Retrieved}. The set of documents that are both relevant and retrieved is denoted as

{Relevant} п {Retrieved}. There are two basic measures for assessing the quality of text retrieval precision and Recall.

Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., correct responses). It is defined as

$$\text{Precision}= \frac{|Relevant \cap \{Retrieved\}|}{|Retrieved|}$$

Recall: This is the percentage of documents that are relevant to the query and retrieved. it is formally defined as

$$\text{Recall}= \frac{|Relevant \cap \{Retrieved\}|}{|Retrieved|}$$

Information Extraction

IE has the goal of transforming a collection of documents, usually with the help of in IR system, into information that is more readily digested and analyzed. IE extracts relevant facts from the documents, while IR selects relevant documents. In general, IE works at a finer granularity level than IR does on the documents. Most IE systems use machine learning or data mining techniques to learn the extraction patterns or rules for documents semi-automatically or automatically, text mining is part of the IEprocess.

The results of the IE process could be in the form of a structured database, or could be a compression or summary of the original text or documents. One could view for the former that IE is a kind of pre- processing stage in the text mining process, which is the step after the IR process and before data mining techniques are performed. in a similar view, IE can also be used to improve the indexing process, which is part of the IR process. In another viewpoint, IE is an instance of textmining.

Computationallinguistics

Corpus-based computational linguistics computes statistics over large text collections in order to discover useful patterns. These patterns are used to inform algorithms for various sub problems within natural language processing, such as part-of-speech tagging, word-sense disambiguation, etc. The aim of text data mining also rather similar to this. However, within the computational linguistics framework, patterns are discovered to aid other problems within the same domain, whereas text data mining is aimed at discovering unknown information for different applications.

III. TEXT DATA MINING METHODS

Levels of textprocessing

Word Level

The bag of words or vector representation takes single words found in the training corpus as features ignoring the sequence in which the words occur. This representation is based on the statistic about single words in isolation. Such feature is said to be Boolean, if we consider whether a word either occurs or does not occur in a document. The feature is said to be frequency based if the frequency of the w ord in a document is taken into consideration.

Words Properties

Relations among words:

© 2019, IJCSE All Rights Reserved

Homonomy : Same form, but different meaning (e.g. bank: river bank, financial institution)

Polysemy : same form, related meaning (e.g. bank: blood bank, financial institution)

Synonyms: different form, same meaning (e.g. singer,vocalist)

Hypolymy: One word denotes a subclass of an another (e.g. breakfast, meal)

Word frequencies in texts have power distribution small number of very frequent words , big number of low frequency words.

Stop-Words

Stop words are those common words that do not add meaningful content to the document. Some examples are: the, and, but, or . Stop words can be a pre-specified list of words or they can be dependent on the content orcorpus.

The feature selection includes removing the case, punctuation, infrequent words, and stop words. A good site for the set of stop-words for the English language is

www.dcs.gla.ac.uk/idom/ir_resources/linguistic_until /stop_words

Stop-words are word that from non-linguistic view do not Cary information they have mainly functional role , usually we remove them to help the methods to perform better. Many of the most frequently used words in English are worthless in IR and text mining

– these words are called stop words.

Latent SemanticIndexing

Latent Semantic Indexing (LSI) transforms the original document vectors to a lower dimensional space by analyzing the correlation structure of terms in the document collection, such that similar documents that do not share terms are placed in the same topic.

Stemming

Stemming is a process which reduces words to their morphological roots. For example the words "informing", "information", "informer", and informed" would be stemmed to their common root "inform", and only the latter word is used as the feature instead of the former four.

Steaming is often applied in the area of information retrieval, where the goal is to enhance system performance and to reduce the number of unique words. Stemming is the process of removing suffixes and prefixes, leaving the root or stem of the word. For example, the words protecting, protected, protects, and protection could be reduced to the root word protect. This makes sense, since the words have similar meaning. However, some stemmers would reduce the word probate to probe and the word relativity to relate, which convey different meanings. This does not seem to affect results in information retrieval, but it could have some undesirable consequences in classification and clustering. Stemming and the removal of stop words will reduce the size of the lexicon, thus saving on computation resources. The porter stemming algorithm is a common employed stemming procedure.[3]

Different forms of the same word usually problematic for text data analysis, because they have different spelling and similar meaning (e.g. user, using, used,..) (e.g. engineer, engineered, they are by most approaches treated as completely unrelated words.

Stemming is a process of transforming a word into its stem. Cutting off a suffix (e.g., computing ->compute)

Lemmatization is a process of transforming a word into its normalized form. Replacing the word,, most often replacing a suffix (e.g computing -> compute)

Basic stemming methods

Remove ending

(i) If a word ends with a consonant other than s, followed by an s, then deletes.

(ii) if a word ends in es, drop thes

(iii) if a word ends in ing, delete the ing unless the remaining word consists only of one letter or ofth.

(iv) if a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a singleletter.

Transform words

If a word ends with "ies" but not "eies" or "aies" then "ies y."

N-grams ofwords

Other feature representations are also possible, such as using information about word positions in the document, or using n-grams representation (word sequences of length up to n). Simple way for generating phrases based on statistics-frequent n-grms;

N-gram is a sequence of n consecutive words (e.g data mining" is a2-gram, "word for Windows" is a 3- gram). "Frequent n-grms" are the n-grams which appear in the document collection at least minifreq (eg., 4) times.

Part-of-Speech(POS)

One important feature is the POS. There can be 25 possible values for POS tags. Most common tags are noun, verb, adjective and adverb. thus, we can assign a number 1,2,3,4 or 5, depending on whether the word is a noun, verb, adjective, adverb or any other, respectively.

© 2019, IJCSE All Rights Reserved

Thesaurus(WordNet)
WordNet is the most well developed and widely used lexical database for English. It consist from 4

Databases (nouns, verbs, adjectives, and adverbs).Each database consists from sense entries each sense consists from a set of synonyms, e.g. musician, instrumentalist, player. Each wordNet entry is connected with other entries in the graph through relations.

PositionalCollocations
The values of this type of feature are the words that occur one or two position to the right or left of the given word.

Higher OrderFeatures
Other features include phrases, document concept categories, terms, hypernyms, named entities, dates, email addresses, locations, organizations, or URLs. These features could be reduced further by applying some other feature selection techniques, such as information gain, mutual information, cross entropy, or odds ratio.
Once the features are extracted, the text is represented as structured data, and traditional data mining techniques can be used. The techniques include discovering frequent sets, frequent sequences and episode rules.

DocumentLevel
Summarization: Text summarization can be applied on a single document or a set of documents written in one or several languages e.g. capturing on- line news from several countries and providing a summary of them involves multi-document and multi-languagesummarization.

There are several ways to provide text summary by providing keywords that help capturing the main topic(S) of the text either for human understanding or for further processing such as indexing and grouping of documents, books, pictures. Highlighting or extracting the most important sentences. Generating new sentences based on the whole text, as for instance used by a new human in writing book reviews.

Summarization by keywords
Automatic extraction of key phrases from text. the approach is the document is treated as a set of phrases, where each phrase is either keyphrase (positive example) or non-keyphrase (negative example). Machine learning is used to generate a model for classifying phrases (decision tree, linear model, ..). Each phrase is described by features, e.g. the number of words in a phrase, the position in the text where the phrase first occurs, the frequency of the phrase. Generate model is used to classify each phrase from documents as keyphrase or non-keyphrase.Theproblemisdifficultasonlyasmallproportion of examples are positive (usually less than 1%).

Summarization by keywords
Automatic extraction of key phrases from text.
Approach is
(i). The document is treated as a set of phrases, where each phrase is either keyphrase (positive example) or non-keyphrase (negativeexample).
(ii). Machine learning is used to generate a model for classifying phrases (decision tree, linear model.,). (iii). Each phrase is described by features, e.g. the number of words in a phrase, the position in the text where the phrase first occurs, the frequency of the phrase..
(iv). Generate model is used to classify each phrase from document as keyphrase or non-keyphrase.
(v) The problem is difficult as only a small proportion of examples is positive (usually less than 1%).

SentenceLevel
Summarization by highlighting sentences Automatic extraction of sentences from text. Approach is
(i) . Document is treated as a set of sentences where each sentence is either important , classify sentences using machine learning.
(ii). Each sentence described by features, eg. position in the document (beginning of the document, beginning of a section, length presence of thematic words (most frequent words), presence of cue phrases (fixed-phrases such as "in conclusion" or "to summarize").

Single Document Visualization
Visualization of a single document, visualizing of big text corpora is easier task because of the big amount of information. Visualization of a single (possibly short) document is much harder task because we cannot count of statistical properties of the text (lack of data) . we must rely on syntactical and logical structure of the document..

Visualization can be used in order to evaluate feature extraction techniques within an exploratory data analysis framework. It can similarly be used to help discern data structure after the application of dimensionality reduction schemes. It can be used to visually identify cluster structure or to help in the identification. of outliers. Finally we note that visualization can be used to aid in discovery by suggesting interesting associations between documents or terms. This type of approach is fruitful when comparing disparate corpora. Let"s take a moment to highlight some of the interesting applications of visualization to document collections.Morris and Yen have developed an interesting visualization scheme to identify relationships among disparate entity types[6]. Their schema uses agglomerative clustering to solve for entity groups among the two data types. For example one data type might be research fronts while the other data type is author collaboration groups. They have also used their technique to study the temporal evolution of topic areas as a function of citationcount.

© 2019, IJCSE All Rights Reserved

Simple approach is:

(i). The text is split into thesentences.

(ii). Each sentence is deep-parsed into its logical form.

(iii). Resolution is performed on all sentences e.gall „he‟,„she‟,„they‟,„him‟,„his‟,„her‟,etc.references to the objects are replaced by its propername.

(iv). From all the sentences we extract [Subject – Predicate- Object triples](SOP)

(v"). SOPs form links in the graph

(vi). finally, we draw a graph.

Text Segmentation

Text segmentation is the process of dividing the text in to sentences. Divide text that has no given structure into segments with similar content. example applications are topic tracking in news (spoken news). identification of topics in large, unstructured text databases. Text Segmentationapproach:

(i). Divide text intosentences.

(ii). Represent each sentence with words and phrases itcontains.

(iii). calculate similarity between the pairs of sentences.

(iv). Find a segmentation 9sequence of delimiters), so that the similarity between the sentences inside the same segment is maximized and minimized between the segments.

This approach can be defined wither as optimization problem or as sliding window.

Document-Collection Level

Text Representation

Set-of-words document representation. word of weighting. In the set-of-words representation each word is represented as a separate variable having numeric weight (importance").The most popular weighting schema is normalized word frequency. TFxIDF:

Tf(w)- term frequency (number of word occurrences in a document).

Df(w) – document frequency(number of documents containing the word).

N- number of all documents.

TfIdf(w)- relative importance of the word in the document.

Dimensionality Reduction:

Dimensionality reduction in machine learning is usually performed to improve the prediction performance, improve learning efficiency to provide faster predictors possibly requesting less information on the original data. to reduce complexity of the learned results, enable better understanding of the underlying process.

The space in which the document resides istypically thousands of dimensions or more. Given thecollection of documents, along with the associatedinterposing distance matrix, one is often interested infinding a convenient lower-dimensional space toperform subsequent analysis. This might be chosen tofacilitate visualization, clustering, or classification.One hopes that by applying dimensionality reduction,one can remove noise from the data and better applyour statistical data mining methods to discover subtlerelationships that might exist between the documents.Let‟s first consider a particularly interestingprojection (a way to reduce the number ofdimensions0 that can be calculated directly from theterm-document matrix. it turns out that one can makeuse of a well-known theorem from linear algebra toobtain a set of useful projections via singular valuedecomposition. this has come to be known in text tdata mining and natural language processing as latent semantic indexing (analysis).[4].

Approaches to dimensionality reduction: Map the original features onto the reduced dimensionality space by:

(i). selecting a subset of the original features , no feature transformation, just select a feature subset. (ii). Constructing features to replace the original features by using methods fromstatistics.

(iii). by using background knowledge for constructing new features to be used in addition/ instead of the original features (can be followed by feature subset selection). general background knowledge 9sum or product of features) , domain specific background knowledge (parser for ted data to get noun phrases, clustering of words, user-specified function).

Dimensionality Reduction:

Feature subset selection. Approaches to feature subset selection:

(i). Filters: Evaluation function independent ofthe learningalgorithm

(ii). Wrappers : Evaluation using modelselection based on the machine learningalgorithm.

(iii). Embedded approaches: Feature selectionduring learning.

(iv). Simple Filters: Assume feature independence(problems with large number of features,e.g. textclassification).

Text Categorization

Document Categorization, Automatic Document Categorization Task . given a set of documents labeled with content categories, the goal is to build a model which would automatically assign right content categories to new unlabeled documents. Content categories can be unstructured (e.g., Reuters) and Structured (e.g. yahoo). Algorithms for learning document classifiers, thee are popular algorithms for text categorization like Support vector Machines, Logistic Regression, Perception algorithm, Naïve Bayesian classifier, Nearest Neighbor algorithmetc.,

© 2019, IJCSE All Rights Reserved

Text DocumentClustering

In document clustering the documents are grouped together. this is done to partition or segment the documents into components that then give the user a more general view of the data. For document clustering we can use the normal clustering techniques, e.g. partition and hierarchical methods. Documents can be represented using Vector space model. for distance function cosine similarity measure is commonlyused.

To convert the text to quantitative data, we can use any of the clustering methodologies familiar to statisticians. For example, we could apply k-means clustering, agglomerative clustering, or model-based clustering (based on estimating finite mixture probability densities)[5]. Rather than go into detail on these well-known methods, we offer the following graph-based approach.

Searching for groups:

(i). clustering is unsupervised or undirected.

(ii). Unlike classification, in clustering, no pre- classified data.

(iii). Search for groups or clusters of data points (records) that are similar to oneanother.

(iv). Similar points may mean: similar customers, products, that will behave in similarways.

Group similar points together

(i). Group points into classes using some distance measures, within-cluster distance, and between cluster distance.

Major clustering Techniques:

(i). Partitioning algorithms: Construct various partitions and then evaluate them by some criterion (ii). Hierarchy algorithms; create a hierarchical decomposition of the set of data (or objects) using somecriterion

(iii). Density-based: based on connectivity and densityfunctions.

(iv). Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of the model to eachother.

IV. CONCLUSION

We have provided introduction to text data mining in this paper. There are numerous challenges to the text data mining. The identification of features that capture semantic content is one area of importance. The Identification of features that capture semantic content is one area of importance.

ACKNOWDEDGEMENT

My thanks to all the experts who have contributed towards the development of the template.

REFERENCES

[1]. BAEZA-YATES, R. AND RIBEIRO-NETO, B. (1990). Modern Information Retrieval. Addison Wesley.

[2].BERRY, M.W.(2003) . Survey of Text Mining: Clustering, classification and Retrieval (Hardcover). springer.

[3] PORTER, M.F. (1980). Algorithm for suffix striping, Program, 130-137.

[4]. DEERVESTER, S., DUMAIS, S.T., FURNAS, G. W., AND LANDAUER, T.K. (1990). Indexing by latent semantic analysis. Journal of the Am. Soc. for Information Science 41, 6, 391-407.

[5]. DUDA, R.O., HART, P.E., AND STORK, D.G. (2000). Pattern Classification, Second ed. Wiley Interscience. MR1802993.

[6]. MORRIS, S.A. AND YEN, G.G. (2004). Crossmaps: visualization of overlapping relationships in collections of journal papers. Proceedings of the National Academy of Sciences of the United states of America supplement 1 101, 5291-5296.

[7] Fayyad U, Piatetsky-Shapro G, and Smyth P 1996 from data mining to knowledge discovery: an overview. In Fayyad U, Piatetsky Shapiro G, Smyth P, and Uthurusamy R (eds) Advances in Knowledge Discovery and Data Mining. Cambridge, MA, AAAI/MIT press:1-34.

[8] Dunham, M.H. (2003). Data Mining- Interdictory and Advanced Topics. Prentice-Hall, NewJersey.

[9] Witten, I.A. and Frank, E. (2000). Data Mining-Practicall Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, SanFrancisco.

[10] Groth, R. (2000). Data Mining – Building competitive Advantage. Prentice-Hall, NewJersey.

[11] G. Giuffrida, W.W.Chu, D.M.Hassens, NOAH: An Algorithm for Mining Classification Rules from Datasets with Large Attribute Space. In Proceedings of 12th International Conference on Extending Database(EDBT), Konsta, Germenay, March2000.

[12] Q. Zou, W.W. Chu, D. Johnson, H.Chiu, A Pattern Decomposition Algorithm for Finding All frequent Patterns in Large Datasets. ICDM2001:673-674.

[13] W.W. Chu, K.Ching, C.C.Hsu, H.Yau, An Error based Conceptual Clustering Method for Providing Approximate Query Answers. Communications of the ACM, 39(13), December,1996.

[14] J.Han, J. Pei, Y.Yin, Mining Frequent Patterns without Candidate Generation. 2000 ACM SIGMOD Intl. Conference on Management ofData.

[15] C.M.Ho, P.H.Huang, J.lew, J.D.Mai, V.Lee, Y.C.Tai, Intelligent System Capable of Sensing-Computing-Actuating, Keynote Address, 4th Intl. Conference on Intelligent Materials, Society of Non-Traditional Technology. Tokyo, Japan, October 1998.

© 2019, IJCSE All Rights Reserved