Proposed Architecture for Automatic Conversion of Unstructured Text Data into Structured Text Data on the Web

[1]CH.Madhusudhan, [2]K.Mrithyunjaya Rao..
1 Associate Professor, Dept of MCA
2 HOD , Dept of CSE, St. Johns Institute of Science and TechnologyVaagdevi College of Engineering and Technology, R.R.Dist, Andhrapradesh, India Warangal, Andhrapradesh, India

Abstract
Data mining, and in particular text mining, has attracted much attention in recent years due to the vast amounts of data available, and the rate of growth. Data mining tools can be used to uncover patterns or hidden relations in the available data, and can potentially contribute greatly to business strategy decisions, knowledge bases, and scientific and medical research. In contrast to data mining, where one looks for patterns and knowledge in structured databases, text mining deals with unstructured, or semi structured, text data such as reports, e-mails or web-pages.

Key Words
Data Mining, Text Mining, Text Classification, Text Mining Methods.

1. Introduction

The Management of unstructured data is acknowledge as one of the most critical unsolved problems in data management and business intelligence fields in current times. The major reason for this unresolved problem is primarily because of the actuality that the methods, systems and related tools that have established themselves so successfully converting structured information into business intelligence, simply are ineffective when we try to implement the same on unstructured information.New methods and approaches are very much necessary. It is a known realism that huge amount of information is shared by the organization explosion across the globe has resulted in opening a lot of new avenues to create tools for data management and business intelligence, simply are ineffective when we try to implement the same on unstructured information. New methods and approaches are very much necessary. It is a known realism that huge amount of information is shared by the organizations across the world over the web. Data stored in most text databases are semistructured 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 semistructured 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 data mining.

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.

Text Mining is a special case of data mining that aims at extracting of useful information from textual data. Besides the usual quantitative or qualitative data familiar to statisticians, the data that serves as input to 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 semiformatted text (html, xml and others) and includes those sources that are encoded in open source document formats (open Document) and other proprietary formats (Microsoft Word and Microsoft Powerpoint). I feel that the problem of extracting information from these data sources is one that offers great challenges to the statistical community. The plan for this article is to discuss 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 text processing. Knowledge discovery in text, intelligent text analysis, or natural language processing, depending on the application and the methodology that is used [1]. Examples of text mining tasks include classifying documents such 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

Manuscript received December 5, 2013
Manuscript revised December 20, 2013

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.
Classification is a data mining technique used to predict group membership for data instances [1]. Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave.
Classification is the most popular data mining technique. Text classification tasks can be divided into two sorts: supervised document classification where some external mechanism provides information on the correct classification for documents or to define classes for the classifier, and unsupervised document classification. Where the classification must be done without any external reference, this system does not have predefined classes. There is also another task called semi supervised document classification, where some documents are labeled by the external mechanism.
Text (or Document) classification is an active research area of text mining, where the documents are classified into predefined classes. Mostly-text documents include letters, newspapers, articles, blogs, technical reports, proceedings, and journal papers, etc. Document Filtering, also based on the classification algorithm to extract the relevant documents related to specific topic from the set of documents [1].
Data classification is a two step process in first step; a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple assumed to belong to a predefined class, as determined by one of the attributed called the class label attributes.In the second step the model is used for classification. First the predictive accuracy of the model (or classifier) is estimated.
Section 2 describes about the issues in Text Mining, section 3 defines Proposed Architecture.

2. Related Work

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 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(or document 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 semistructured 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 semistructured 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 data mining.

2.1 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 appropriate document.

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} ⊓ {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

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

Recall: This is the percentage of document that is relevant to the query and retrieved. It is formally defined as

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

2.2 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 IE process.

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 an another viewpoint, IE is an instance of text mining.

2.3 Computational linguistics:

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 subproblems 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

2.1 Missing Data

Missing data values cause problems during both the training phase and to the classification process itself. Missing values in the training data must be handled and may produce an inaccurate result. Missing data in a tuple to be classified must be able to be handled by the resulting classification scheme. There are many approaches to handling missing data.

- Ignore the missing data.
- Assume a value for the missing data. This may be determined by using some method to predict what the value could be.
- Assume a special value for the missing data. This means that the value of missing data is taken to be a specific value all of its own.

2.2 Measuring Performance

The performance of classification methods is usually examined by evaluating the accuracy of the classification. Since classification is often a fuzzy problem, the correct answer may depend on the user.

Classification accuracy is usually calculated by determining the percentage of tuples placed into the correct class. This ignores the fact that there also may be a cost associated with an incorrect assignment to the wrong class.

2.3 Preparing the data for classification and prediction

The steps may be applied to the data in order to help improve the accuracy, efficiency, and scalability of the classification or prediction process.

- Data Cleaning
- Relevance analysis
- Data transformation

2.4 Comparing Classification Methods:

Classification and prediction methods can be compared and evaluated according to the following criteria:

Predictive accuracy: This refers to the ability of the model to correctly predict the class label of new or previously unseen data.

Speed:this refers to the computation costs involved in generating and using the model.

Robustness: This is the ability of the model to make construct the model efficiently given large amounts of data.

Scalability: this refers to the ability to construct the model efficiently given large amounts of data.

Interpretability:this refers to the level of understanding and insight that is provided by the model.

Classification methods categorization

Statistical distancedecision treeneural networkrule based

Text classification is a fundamental task in document processing. The goal of text classification is to classify a set of documents into a fixed number of predefined categories/classes. A document may belong to more than one class. When classifying a document, a document is represented as a "bag of words".It does not attempt to process the actual information as information extraction does. Rather, in simple text classification task, it only counts words (term frequency) that the document covers e.g. if in the document, cricket word comes frequently then "cricket" is assigned as its topic (or class) [3][4]. Classification is a two step process. First step is Model construction and the second step is Model Usage.

Model Construction: Also called Training Phase or Learning Phase; the set of documents used for model construction is called training set. It describes a set of predetermined classes. Each document/sample in the training set is assumed to belong to a predefined class (labeled documents). The model is represented as classification rules, decision trees, or mathematical formulae [4] [5].

Model Usage: this is the 2nd step in classification. Also called Testing phase or classification Phase, it is used for classifying future or unlabeled documents. Then known label of test document/sample is compared with the classified result to estimate the accuracy of the classifier. For e.g. the labeled documents of the training set, is used further to classify unlabeled documents. Test set is independent of training set [3] [4] [5].

The text classification task is to train the classifier using these labeled documents, and assign categories/classes to the new, unlabeled documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process [4] [5].

Any classifier is unable to understand a document in its raw format; a document has to be converted into a standard representation. it is observed from previous research that words work well as feature for many text classification techniques[4][6].

In feature space representation, the text documents are represented as sequence of words called "Bag of Words". Bag of Words (Bow) is one of the basic methods of representing a document. The Bow is used to form a vector representing a document, with one component in the vector for each word in the document. These components are computed using term frequency (TF). Such representation of a set of documents as a vector is known as Vector Space Model [7]. Text can also be tokenized using inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation [3].

Text classification also represents many challenges and difficulties. First, it is difficult to capture high level semantics from a few key words. Second, high dimensionality (thousands of features) and variable length of the documents, place both efficiency and accuracy demands on classification systems [7].

3. Proposed Architecture

Our Automatic Conversion of Unstructured Text Data into Structured Text Data on the Web Architecture, shown in below figure, facilitatesmining of large databases and data warehouses. The architecture consists of four layers; the lowest layer is the data repository layer, which consists of the supporting databases and data warehouses. On top of it is the database (DB) layer, which provides a multi-dimensional view of data for online analytical processing and mining. The essential layer of data mining is the OLAP/OLAM layer, which consists of two engines, one for online analytical processing and one for mining. Finally, on top of the OLAP/OLAM layer, which provides easy-to-use interfaces? These interfaces let users construct data warehouses, create multidimensional databases, select the desired set of data, perform interactive OLAP and mining, and visualize and explore the results.

The OLAM architecture provides a modularized and systematic design for a data mining system and provides several benefits. First, it takes advantage of widely available, comprehensive information processing infrastructure. These systems have been systematically constructed around relational database management

system and data warehouses which can store huge amounts of data. Data cleaning, integration, and consolidation have been largely performed in the construction of data warehouses. An efficient OLAM architecture should use existing infrastructure in this way rather than constructing everything from scratch.

Another benefit of the OLAM architecture is that it provides an OLAP based exploratory data analysis environment. The integration of database and data warehouse at one end and online analytical processing and mining at the other facilitates two things. First, it becomes possible to mine different subsets of data and at different levels of abstraction by drilling, pivoting, slicing, and dicing a multidimensional database and the intermediate data-mining results.

Secondary, it facilitates online, interactive selection of data-mining functions and interestingness thresholds. Performing, these functions interactively and viewing the results with data/knowledge visualization tools will greatly enhance the power and flexibility of exploratory data mining.

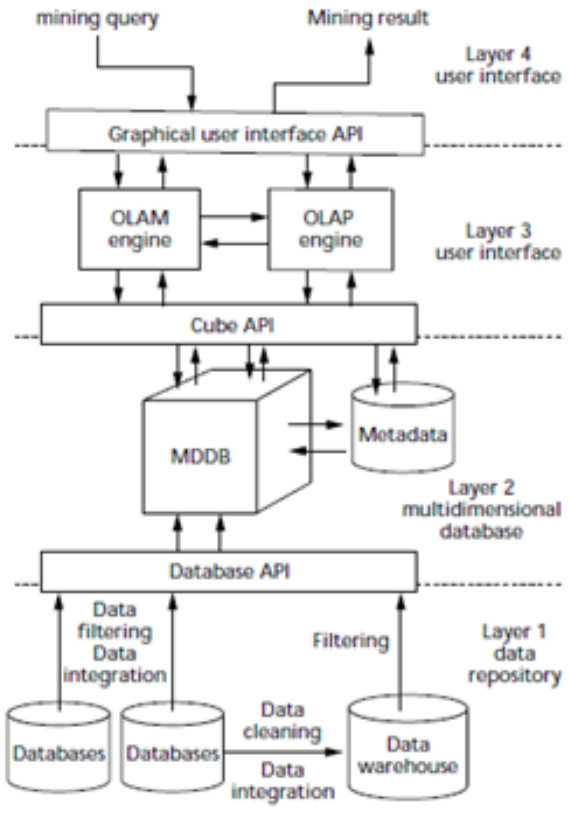

Figure: Represents MR Text data on the Web Architecture

3.2.1 The Method: In this paper we have proposed a web representation of multi-relational context dependent text data system for unstructured text data. First of all, our system will extract the string of unstructured text data, typed by user. After extracting the question string, we will do the text pre-processing and do the text transformation i.e. feature generation and then we perform feature selection, after that we apply different pattern discovery and data interpretation and evaluation is performed then the data is represented in multidimensional data format and applied multi-relational patterns applied then the data is represented on the web format.

Text data representation on web

1. Text Pre processing
2. Initialization process
3. structured representation process

The below figure represents the unstructured text data is converted into text file format.

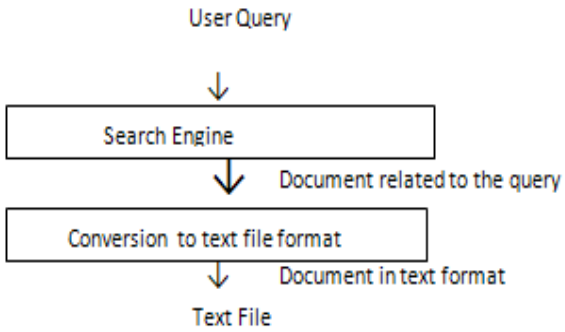

Figure:Represents Document Collection

The following figure represents the removing the irrelevant data from the text file and prepares the processed text data.

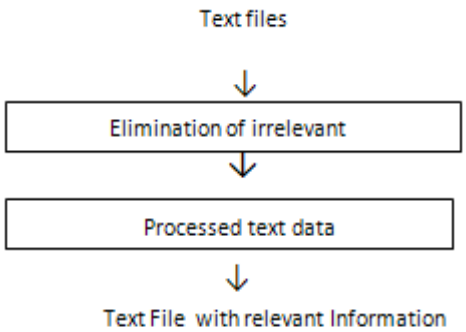

Figure: Represents elimination of irrelevant information

The following is the proposed algorithm for conversion of multi-relational text data to represent on the web format.

Algorithm:

Step 1: Gather the unstructured text data by using various gathering tools like search engines.

Step 2: Do the Text Pre-Processing

Step3: Do the Text Transformation (Feature Generation)

Step 4: Data Mining/ Pattern Discovery

Step 5: Data Interpretation/ Evaluation

Step 6: Create the Multi-dimensional Data

Step 7: Establish the relation among the data by using the Multi-Relational Data

Step 8: Finally convert the multi-relational data in to the required Web data format.

The following figure represents the flow chart for the multi relational context dependent unstructured text data to represent on the web.

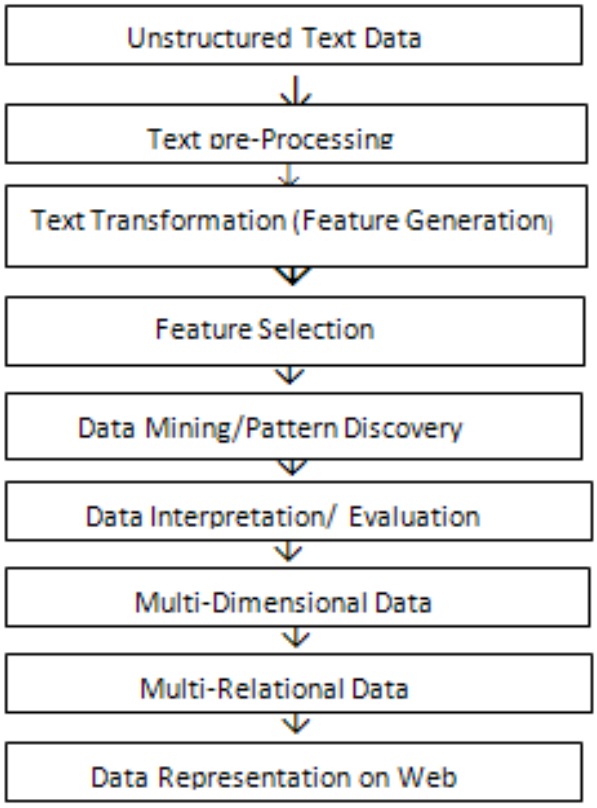

Figure:Represents MR Text data on the Web

4. CONCLUSION

In this paper we have proposed this architecture for representing unstructured text data in to structured data on the Web it helpful for people to view the structured data on the web and they can easily understand the data and they can use it for different purposes, Data Mining is inherently more powerful than Propositional Data Mining Our System will give the best way to present the data on the web.

5. 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, New Jersey.
[9] Witten, I.A. and Frank, E. (2000). Data Mining-Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.
[10] Groth, R. (2000). Data Mining – Building competitive Advantage. Prentice-Hall, New Jersey.
[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, Germany, March 2000.
[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 of Data.
[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.

[1]CH.Madhusudhan received his B.Sc from Osmania University in 1997, M.C.A. from Kakatiya University in 2002 and M.Phil from Madurai Kamaraj University in 2009. Currently he is pursuing Ph.D from Acharya Nagarjuna University. He is working as Associate Professor in M.C.A dept of St.Johns Institute of Science and Technology, R.R.Dist, Andhrapradesh, India. His research interests include Data Mining, Data Base Management Systems, Software Engineering and Object Oriented Programming,

[2]Prof. K.Mrithyunjaya.Rao HOD, Vaagdevi College of Engineering and Technology. Bollikunta, Warangal, Andhrapradesh, India. His areas of research include Data Warehousing and Data Mining, Database Management System, Software Engineering