Scientific Investigation of Text Mining

Sedigheh Damavandi

MSc in Information Technology, software design, Shiraz University, E-learning School

Abstract: The increasing growth of database almost in any area of human activity has caused that increases need for new powerful tools for changing the data into useful knowledge. To meet this need, researchers in areas such as machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, econometrics, information retrieval, information extraction and .... have explored methods and ideas. Unstructured nature of these texts, applying the same methods that we use about no text database, makes it impossible. Therefore methods and specific algorithms are required for processing (pre-processing) for extracting useful patterns. Text mining uses unstructured textural information and examines it to explore the structure and hidden implicit meanings in the text. In this article, we examine one of the newest fields studied in data mining; text mining. In this paper are described existing methods for pre-processing, classification, information extraction, methods of finding relations. At the end of any part, evaluation and comparison carry out on expressed methods in that part and in the end; number of text mining applications is expressed.
[Sedigheh Damavandi. **Scientific Investigation of Text Mining.** *N Y Sci J* 2015;8(7):24-35]. (ISSN: 1554-0200). http://www.sciencepub.net/newyork. 5

Keywords: Text Mining, Information Technology, Database

1. Introduction

Text mining is discovery of previous unknown information by new computers that automatically extract information from different written sources. Key factor is relevance of extracted information with each other to create new data and or new hypotheses to more than this look for them with more common experimental methods. In search the user usually look for what is already known and has been written by someone else. The problem is that all the information that is currently irrelevant is putting aside to you find the information that you need and in addition your relevant information. In text mining aims to discover the unknown information, things that are not yet known, as well as it can not still be registered.

1-1 Text Mining

Text mining means automatic extraction of new information and unknown from various written sources and the first time it was raised by Feldman et al. Text mining is different with Internet search. So that the internet search users are looking for things that already known and have been written by other people but in text mining aims to discover information that previously are unknown and subjects that anyone them still has not written and another is that in process of internet search user will find also cases unrelated to the needed subject but user in text mining focuses only on the his desired subject. Most active applied range of text mining is in biological sciences and medicine, for example its application is in the discovery of patterns and relationships from MEDLINE textual databases. The stored data in further the textual database is semi-structured data and because not completely are unstructured and nor completely structured. For example, a document is, including a number of structured fields such as title, authors, date of publication, and the category from other hand including some of the unstructured textual components, such as abstract and content. Information retrieval techniques such as (methods indexing text) are created to handle non-structured documents. Retrieval techniques of old information for a large amount of textual data that is increasingly rising are inefficient. Without knowing the contents of documents, formulating the appropriate Query to analyze and extract useful information from data is difficult. Users need tools to compare different documents, organize documents based on their relevance and finding patterns. Therefore, one of the newest research fields in data mining: text mining expanded for this purpose. Text mining means searching patterns in unstructured text. Text mining is used for discovering automatically favorite knowledge or useful from semi-structured text. Several techniques have been proposed for the text mining, including conceptual structure, exploring association rule, the decision tree, methods of rules inference, also information retrieval techniques for works such as matching the documents, organizing, clustering.

Among the problems that exist in the field of text mining is, discovering useful knowledge from unstructured text or semi structured, which has attracted much attention. Traditional data mining methods assume that the information is in the form of relational databases, so for many applications, such as electronic information is not available in helpful semi-structured or unstructured form. Without text mining, processing unstructured textual databases must be

done manually by the users, which is very cumbersome. So we can say the aim of text mining is, automating the large amount of users work. Sometimes, instead of text mining word is used word of "text data mining" and also famous name "knowledge discovery in text" or KDT. Text mining, its emphasis is on finding new knowledge of the text, (usually knowledge that implicitly is in the documents) while retrieving information finds documents that are the more relevant. This paper studies the text mining field. Text mining can be considered as an interdisciplinary method for information retrieval, machine learning, statistical, mathematical linguists and especially data mining. Because text-mining has root in many technologies, so there are different definitions for it. People who had a work history in the field of data mining, was wanting to apply the same concepts and methods available at data mining on texts and their definitions was according on the same grounds. But those who had came from the community of computational linguists, wanted to give the ability to the computer to be able to understand the text, and this is end of what is expected of text mining. {Tehran University}

1.2 knowledge discoveries and its relationship with text mining

Knowledge discovery in the databases is formed (KDD) in the early 80s, in the reference to general concept, high level and following the search for knowledge in information. KDD word for describing all the stages of information extraction from databases and also explain the early works objectives of the decision rules application. Sometimes, data mining is used as a synonym for KDD. This means that all aspects of the process of knowledge discovery are data mining. Data mining sometimes is considered as part of the KDD process and describes modeling phase. In the total, KDD, process of finding useful information and patterns from data be said, and data mining is using algorithms to find useful information in the KDD process. KDD is process of identifying patterns understandable, useful, new and valid in data. In fact the goal is to find patterns and hidden relationships in this data. Among the properties that can be used to measure the quality of the found patterns in the data include the ability of human understanding, validation by statistical criterion, novelty and usefulness.

1-3 Related search areas

This lack of information in the world today is not the problem, but is the lack of knowledge that can be obtained from this information. Millions of web pages, millions of words in the digital library and thousands pages of information in each company are only a few of these information sources. But can not be introduced a source of knowledge in this between specifically. Knowledge is a summary of the information and also conclusions and the product of think and analyze data mining information is a much efficient method for discover information from structured data that is stored in tables. Data mining extracts patterns from the transactions, groups data and also classifies it. Using data mining we can find out some association between data items that have filled a database. However, we have a problem with data mining and it is the lack of generality in its application. Most of our knowledge if do not become non-digital, are completely unstructured. Digital libraries, news, e-books, many of financial documents, scientific articles, and almost everything that you can find in the Web are not structured. As a result, we can not use data mining teachings about them directly. However, there are three main approaches for dealing with this large volume of non-structured data, includes information retrieval, information extraction and natural language processing.

2- Methods for pre-processing of texts

To explore, a large set of documents, it is essential that the documents be preprocessed and information be stored in a suitable data structure for next processing. In this field, there are several methods that try to use the syntactic structure and semantic of text. In most of methods documents is displayed as a set of words (Showing bag of words). Most of text mining methods, applies search algorithms on the labels attributed to the each document. These labels may become extracted keywords from the document or only a list of words in the document. To show the least important a word in a document commonly is used vector showing, for every word a numerical importance amount is stored. Main and important methods that are based on these ideas are: vector space model, probability model and logical model.

To extract all the words of a text, a tokenization process is required in which a text by removing all punctuation marks and replacing the tabs and other non-textual characters to a empty space character, converted into a stream of division words, and then, this show is used for next processing. To set of words obtained from merging all documents in a set, dictionary of that documents set is called. To reduce the size of the dictionary and descriptive aspects of a document set can be used methods of filtering, root finding to reduce words that describe the documents. Filtering methods deletes words from the dictionary and therefore from documents. Among these methods can be mentioned stop word filtering and delete of stop word. For reducing the number of the dictionary words can use methods of term selection. In methods of term selection only selected words is used for

describing the documents. In methods of term selection only selected words for describing the documents are used. One method for selection of etymology Karsh tries to make primary forms of words. For example, deleting the sign of addition (s) from nouns or ing of verbs and ....after the etymology process, each word by its root is displayed. In fact, here words that have the common root are converted to their roots, words extraction is on the basis of their purity degree.

2-1 vector space model

This model is able to analyze a large set of documents affectively. This method was introduced for information retrieval and indexing but now, it is used in some methods text mining. In this model, documents and query as vectors in the m-dimensional space are displayed. That in this space each dimension is a term. The order of term is the basic concept such as a word or phrase. Vector elements are corresponded with a term weight. D document will be displayed in form d = (x_1, x_2……..x_n) that each xi shows the importance of i term in the d document. Here, we determine similarity based on distance between vectors (document with document or document with query). Whatever vectors be closer, are considered more like to each other. That this similarity defines based on angle of two vectors or vector multiplication or cosine Similarity. For assigning weights to the terms can be used estimation Tf and IDF. In one term Tf that number of repetitions in the document is more, it is more important and more weight

2-2 linguistics preprocessing

Sometimes, it is possible, linguistics pre-processes also be used for increase the available information about the terms. For this purpose, the following methods are applied:

A) Syntactic labeling: for each term specifies that is noun, verb, adjective and..... Although large number does not believe that this is part of the text mining, for example, a system named GATE at the University of Sheffield, in a digital library have been inserted for this intention. GATE includes tools for labeling sentences. For example, this system can find within a text, name of geographical locations, name of persons and something like this. The system further includes a data extraction to knowledge extraction. POS often plays an important role in Natural Language Processing. In fact, this is the first step in natural language processing and natural language processing is one of the bases of text mining.

B) Chunking text: The purpose is grouping adjacent words in a sentence.

C (word meaning disambiguation: Here we try to remove the ambiguities in the meaning of a word or phrase. For example, a bank may mean a financial institution or beach edge or the river. So instead of terms, meanings can be stored in the display of vector space. This causes getting bigger dictionary. But it also makes the term meaning be considered in display.

3 -Text mining process

Text mining is a process that includes many technological fields. Information retrieval, data mining and artificial intelligence and computational linguistics are all fields that in this field have a role. But there are two main phase generally in the text mining process.

The first phase is, preprocessing documentation. The output of the first phase can be having two different formats, based on the document and based on the concept. In the first format, what is important for us, how to better display for documentation. This can be its convert to an intermediate format and semi-structured, or applying an index on it or any type other display that makes working with documents, more efficient. In the case any existence in the display finally again will be a document. In the second type improving to the display of documentary, concepts and meanings contained in the document, as well as the relationship between them and any kind of another conceptual information that can be extracted, is extracted from text. In this type of display, another we are not facing with the documentation as an existence but with concepts are facing that has been extracted from this document. The next step is knowledge extraction from these middle forms of documentation display. Depending on the method of displaying a document, process of Knowledge extraction for a document is different. Display based on document are used, for grouping, classification, visualization and so on, while the display based on the concept is used for finding relationships between concepts, automatic making of thesaurus and ontology and so on.

4-Methods of text mining

The main reason of appalling data mining methods for textual documents is, structuring them. Famous database structures are including library catalogs, or books indexes. Popular database structures include library catalogs, or books indexes. Problem of designed indexes manually are the time required to maintain them. Therefore, information sources that are not much appropriate change web Problem indexes designed manually are the time required to maintain them. Therefore, information sources that much change, like web are not appropriate. Existing methods for structuring set is including: classification methods and clustering methods. Following initially general phase of the text mining process is described. Then some of the important methods for texts classification along with

evaluation are presented. Then methods for automatically extraction of useful patterns from textual documents along with evaluate them are described. Combining these methods with structuring methods (clustering and classification) provide powerful tools to explore useful patterns in textual collections that at the end of this section, some of the methods that these two methods for exploring useful patterns from texts have combined, are expressed.

4-1 Information Extraction

The starting point for computers to analyze unstructured text is, using information extraction. Information extraction Software identifies key terms (phrases) and their relationship with the text. That does this work by looking for fields defined in text, a process that calls similar samples. This software acquires the relationship between all the people, places and times until provide meaningful information (targeted) for the user. This technology can be very useful when facing with a large amount of text. Common data mining assumes that information for extracting now puts in the form of relational databases. Unfortunately, a number of applications, electronic information are available just in the form free documentation of natural language instead of a structured database, during proposing the IE, issue is conversions of textual documents set in form of more structured database. That structured database by IE model can be applied for KDD model for the additional search of information.

Search algorithm that is an algorithm for improvement of calling, by using extraction rules as follows. Be note that the final decision whether or not is for extraction predicted layers that whether or not exist layer (or any of the synonyms) in a substring of documentation. If layer find in the text, extractor studied this definite prediction and extracted layer.

Input: RB is the set of prediction rules

D is the set of documents

Outpute: F is the set of slot fillers extracted

Function Information Extraction (RB,D)

F:=ø

For each example D Є D do

Extract fillers from D using extraction rules and add them to F

For each rule R in the prediction rules base RB do

If R fires on the current extracted fillers

If the predicted filler is a substring of D

Extract the predicted filler and add it to F return F

4-2 Classification

The purpose of the texts classification is, attributing the classes predefined into textual documents. For example, a new news that are entered we say belongs to the athleticclass or political or artistic. Various methods are used for the classification of documents. In the classification there are a training set of documents that are known for these classes. By using this set, the classification model is characterized, and then by using that class the new document that enters is characterized. To measure the performance of the classification model, a test set is considered which is independent of the training set. And labels that for these documents are estimated by the model are compared with the real label of documents. The proportion of documents that has been properly classified to total number the documents called accuracy. Three criterions are used for comparison of classifier.

1) Precision: fraction of retrieved documents that is relevant.

2) Recall: indicates the fraction of relevant documents retrieved. These two criterion are defined as follows:

There is a compromise between precision and recall. That's why, another criterion called the F-score that makes a compromise between both, is used for measuring the total performance of classifier.

4-1-2 Selection of index term

For a set of document may exist more than 100 thousand different word, but words must be selected that are more informative. With decrease the number of words, complexity of classification reduces. Common criterion used for ranking is information gain. The criterion for term tj is defined as follows:

P (c_l) Equal to fraction of the training documents that have t_j class.

P(tj =0P) and (tj =1) are not equal to the number documents that have the term t_j IG(tj) are calculated for all term and terms that their IG is too low are removed from the dictionary. The following some common methods of data mining for text classification are described.

4-2-2 Classifier Na¨ıve Bayes

A method of Classification is probability. Class of one document is according to the words that have appeared in a document. It is noteworthy that each document precisely belongs to a class. Na¨ıve Bayes classification have a learning step in which the probabilities $P(t_i|L_c)$ (number of documents that in set of training includes term tj and their class label is Lc divided by the total documents of training set) are estimated. In step of classification, estimated probabilities for classification of a new sample in accordance with the law Bayes are used. To reduce the number of probabilities $P(t_i|L_c)$ that must be estimated can be used methods of index term

selection. Although this method due to its independence assumption is somewhat possible unrealistic, but in practice good results are obtained from it:

Labeling manually documentations of training set is tough work. Some papers use the not labeled documents for the training set. Assume that from a small training set have obtained that ti word has Strong correlations with the class Lc. May from the Not labeled documents obtain that tj has Strong correlations to ti thus could concluded that tj is a good predictors for the Lc class. In this method not labeled documents, but in practice improve the results of classification. In from combination Native Bayes and EM1 are used and classification error is reduced to 30%.

4-3-2 nearest neighbor classification

Instead of making explicit models for different classes, the other way is that be selected documents from a training set that are similar to current document and current document class is equal to the a class that the majority have similar documents. In method of classification K to nearer neighbor, k to document from training set that has highest similarity (based on a similarity criterion defined) to the current document as neighbor of that document is selected.

There are a lot of similarity criterions in text mining. A simple method is, counting the number of common words in the two documents. This method should be normalized for documents with the different length. Also words may be more reflective for the content of a document. As a standard method for measuring the similarity can be cited cosine similarity. For determining the class of di document similarity S(di,dj) for all documents in the training set is calculated. Then k most similar documents of the training set as neighbors of current document are elected and di document class is equal to a class that most documents of its neighbor have the class. In this method k optimal amount can be estimated from another training set by cross-validation.

Classification of nearest neighbor is a non-parametric method. According to done method of k to near neighbor has good performance in practice. The problem is that during classification, many calculations are necessary (calculating the similarity of a document with all the documents in training set). Some expansions of Nearest Neighbor method are expressed.

4-2-4 Decision trees

For making decision trees is used from a decision strategy and overcome. In fact, we have a training set that it should be divided. For an M training set with labeled documents, ti word should be selected for the classification of training set. For the selection of ti word can be used criterion of information Gain that previously is described. Then M based on the selected ti term is divided into two subsets. + Mi subset is, including ti documents, and - Mi subset is, including documents that are not ti term in them. we repeat the same process for + Mi and –Mi and will continue this work until all documents in a subset are belonging to a class (for example, all belong to the class Lc) in which case, label of that node become equal with the class corresponding with documents and that node convert to leaf and other we do not divides it, or this process will continue until no longer do not exist a term for subsets classification. In this case, a label put on it that the majority documents of that section have labels. Therefore, in accordance with this process a tree are obtained from the laws that leaves are corresponding with classes. Decision trees are a standard tool in data mining. The rapid and scalable algorithms are in both variables and the size of training set. One of the decision trees problems for text mining is that depends only a small number of terms. A better method is the use of boosting decision tree in which a set of supplements decision trees are made. In this method reduces errors.

4-2-4 the core methods and SVM

SVM is a classification algorithm supervised that has been used successfully to classify text. Usually d document by vector (td1, ..., tdN) from the number of words will be displayed. Text mining is that only a small number of terms of a SVM can separate two classes: a positive class L1 [58] [2] (by y = 1 is shown) and L2 negative class (which is shown with y = -1). In space of input vectors, one hyperplane by setting y = 0 in the below linear equation is defined.

SVM specifies one hyperplane that is putted between positive and negative samples of training set. Bj parameters are sat up such that the distance (which is called the margin) between hype lane and the closest positive or negative sample become maximum. Documents which have distance from hyperplan, support vectors are called, and specify actual location of hyper plan. Usually only a small fraction of the documents are support vector. A new document with term vector if the amount be classified in L1 otherwise in L2 is classified. SVM can be used with non-linear predictor.

The most important feature SVMs is that learning is almost independent from the dimensions of the feature space. This method does not need to select character because inherently the points of data (support vector) for a good classification will be choice. Therefore, this method is particularly suitable for texts classification. In the case of textual data, the choice of kernel functions has little effect on the accuracy of classification.

4-2-4 Classifier evaluation

In this section efficiency of different classification methods that have been described in previous sections, we compare with each other. The efficiency of classification different methods on newsgroups set of 20 pathfinders is shown in Table 1. In Table 1, the efficiency of 5 methods native Bayes, C4.5 decision tree, k to the nearest neighbor, SVM and boosted tree have been compared with each other

Tables 1: compare the efficiency of classification different methods

F-value	Methods
0,795	na¨ıve Bayes
0.794	decision tree C4.5
0.856	k to the nearest neighbor
0.870	SVM
0.878	boosted tree

4-3 Classification

Classification includes identifying the main topics of documentation with substituting documentation in the set of pre-defined subjects. At the time of classification of a documentary, computer program often will meet with documentation as "package of the words". This does not attempt to process information like information extraction. In more exact statement, classification only accounts words that appear, and identifies from these cases, major issues which cover the documentation. Classification Often on a Semantic culture in which subjects have been defined already and relationships are quoting with searching extensive information, the more limited terms, synonyms, and related terms. Classification tools normally are having methods for documentation classification, as well as documentation that often have a concept in special subject.

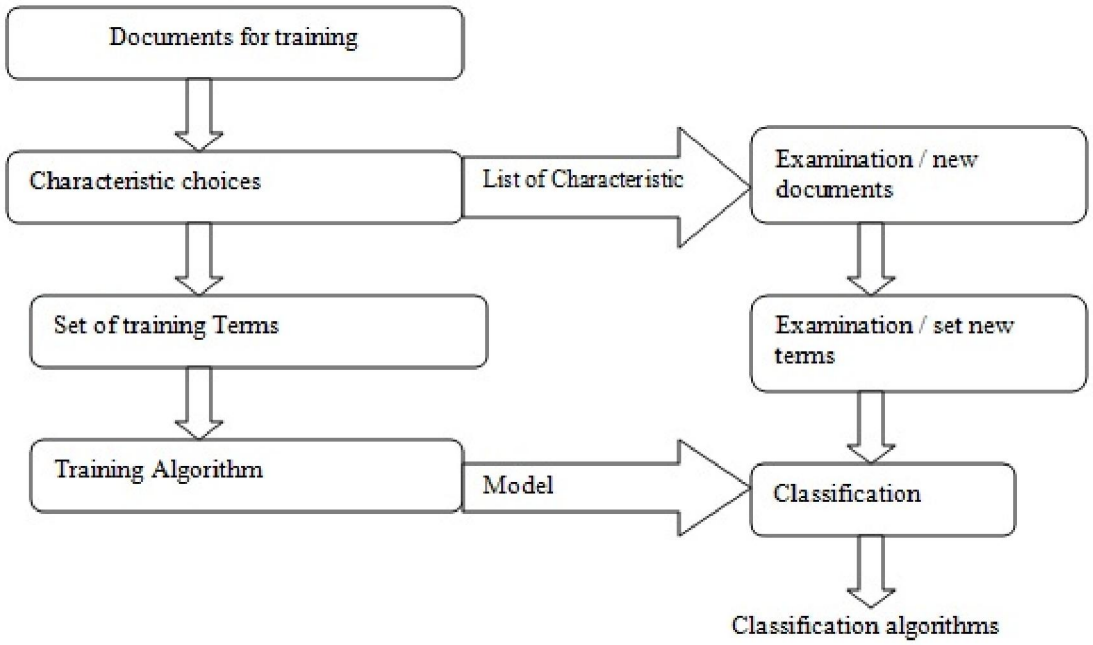

Figure 1 - current Diagram from textual classification

Similarly, for summarizing, classification can be used with tracking subject for other characteristic of documentation s for personal seeking information personally about a subject. Returned documentation from tracked subject can be categorized with content valuable, So that people can offer an important issue for first relevant documentation. Classification can be used in a number of applied areas. A number of jobs and industry stipulates that financial support of customers or respond to customers questions about various subjects. If they can use the general framework of classification for the documentation classification from on the subject, then the customers or final users will be able to access the information that they search with so many speed. The aim of the textual classification, classification of a documentation set is in form of a number of pre-defined the fixed categories. Each documentary may belong to more than one class. Aim of using supervised learning algorithms is, training classifiers

from on known samples (not labeled documentation) and automatic implementation of unknown case classification (not labeled documentation). Figure 8 shows the current general diagram from work of textual classification. A set of labeled documentation from on the source o = (d1,d2,…, dn) belong to the categories set C = [c1, c2, ... cp] studies. Textual classification Work is, training classifiers to use this documentation, and assign categories to new documentation. At this stage training, n documents are putted in p separate folders (are arranged), while the folders are consistent with the type of classes. The next stage, the set of training data through another selection process are prepared.

4-4Pluralization

Pluralization is a technique which is used for documentation of similar groups, but this is different with the classification of documents that be pluralization on the edges instead it and is used for subjects that already were defined. Another advantage of pluralization is in documentation that can be appear in multiple sub-subjects, Therefore, the pluralization of useful documentation, from search results will not be removed. A basic algorithm of pluralization creates a path of subjects for each document and size of values from how good documentation in an each pluralization. Pluralization Technology can be useful in the information classification of management system, which may include thousands of documents.

In the k-means pluralization algorithm, while the same calculation is between textual documentation, which only not investigate eigen vector based on algorithm of statistical successive terms, but combine the relevant degree among words together, Then the relations between keywords in form of calculated is accepted, that hereby sensitivity of consecutive input fields, for the specific text is less, that the semantic understanding examine accuracy of similar additive effects from the short text and simple sentences and moreover the right and speed of calling the results of textual pluralization. Algorithm model with co-mining design was shown in Figure 2.

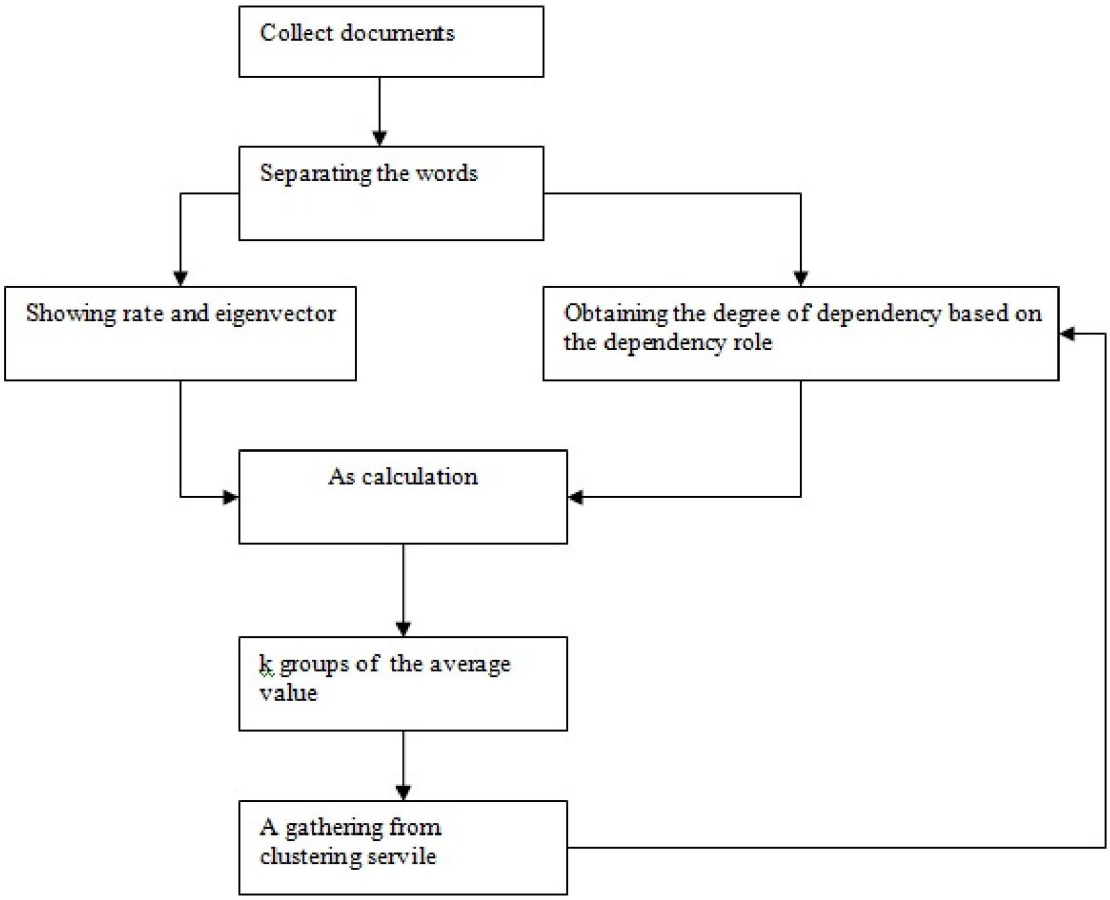

Figure 2 - current diagram from k clustering, a means based on text mining

In pluralization method based on words relationships (WRBC), pluralization process of the text consists four main parts: preprocess text, calculation of relationship between words, words pluralization and text classification. In Figure 3 we see.

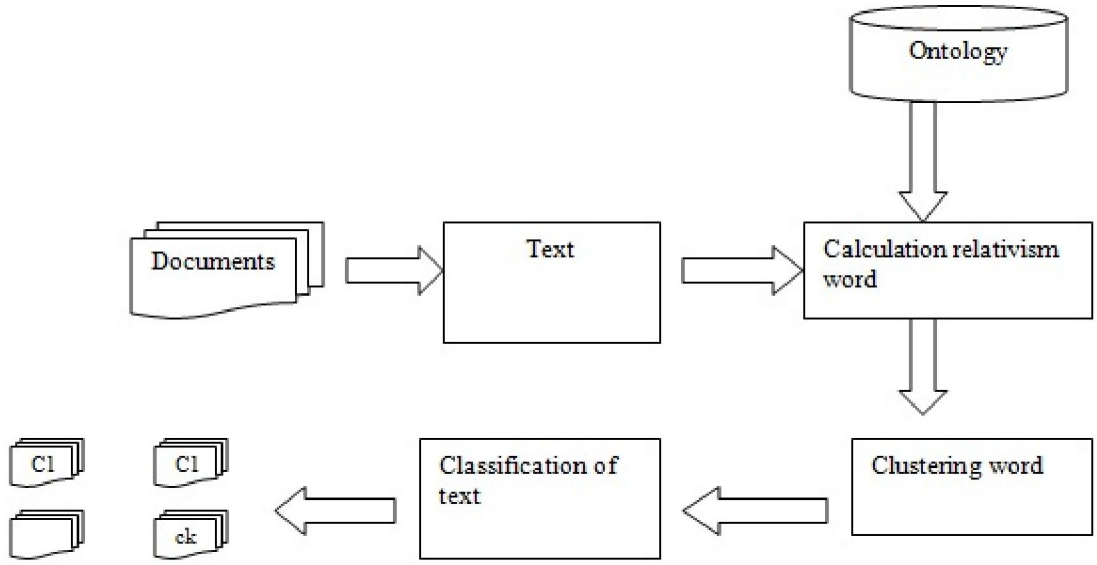

Figure 3: relativity of word based on clustering methods

The first step in the text pluralization is for converting the documentation, which usually are string of characters in the form of suitable representations for classification work:

1) Delete the interdiction words: interdiction words, words are more common that not transmit information (for example, pronouns, prepositions, conjunctions, etc ..). Cleaning interdiction words can improve the results of pluralization.

2) Basic words: with basic words, the method processes suffixes of transition for the production of these words. Do not do this work for words group that have the same conceptual meaning as work, workers, worked and working.

3) Filtering: v vocabulary domain in ontology for filtering is used. By filtering documentation for the domain of relevant, words are studied. This can reduce dimensions of the documentation. The main issue in pluralization of the statistical text is high dimensions of the characteristics space. Standard pluralization techniques can not face with such large pluralization of characteristics, until when processing highly in computer terms is harmful (expensive). We can present documentation with some fields of vocabulary through resolving the aspects of the issues. At the beginning of words pluralization, is selected a word randomly from internal rows. Other words are added to this row or new row until when all words be belong to the m row. This method emplacement of a word in other rows make possible and is corresponded with other data. This method realizes pluralization of the words with pluralization of relevant words and then runs textual classification.

4-5 Connection with concepts

Relationship tools, relevant documents concepts connects together with the identification of concepts usually divided, and helps to users to find information that they might not want find by using traditional search methods. This search for information creates instead of seeking. Relationship conceptual concepts in text are text mining, especially in the fields of biomedicine, while much research does on it this allows researchers to read all the information and they create relationships for other researchers. Arbitrarily, the software of the concepts communication can identify the relationship between diseases and ways of them treatment as the human can not do it. For example, the solution of a text mining software can identifies easily relations between the subjects or X, Y and Z, Y which are well known. But text mining tools can achieve potential relationship between x that sometimes humanity researchers still can not acquire its other side because a large amount of information that it for sorting in all the time of communication can have.

4-6 Information Visualization

Virtual text mining or information visualization covers large textual sources in virtual hierarchy or the representations and provides search capabilities and also simple search. .Docminer has been shown in Figure 4, which is the representation tool of large amounts of text, which allows to users that analyze

virtually the concepts. The user can communicate with the map of documentation, with zooming; grading and creating sub-map. Information visualization is useful when work require a wide range of documentation with total limitation and searches relevant subjects. Can use visualization of information to identify the terrorist networks or find information about the offenders that may be completely without communication, this can be provided for them with a design from possible relationships among suspicious relationships so that they can study the relationships that they can not design for themselves.

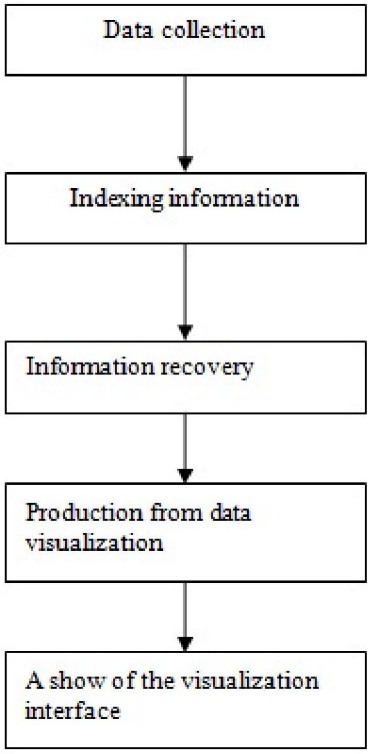

Figure 4- Information Visualization

Information visualization is an interpretation which may be empliment in three stages:

1) Data Provision: means to determining and obtaining the data from visualization of the original data space.

2) Extraction and the data analysis: analysis and extraction of required data from original data and space of data visualization.

3) Visualization of the project: means the application of special algorithms design for the designing data visualization space for considering visualization

Information Visualization model is divided into five structural stages:

1) Information set: for collecting information sources from the database or www is required

2) Index information: for indexing the information source for the case of the original data resources

3) Information retrieval: for searching lists of information in accordance with the results from on the main data sources based on needed recovery

4) Production of data visualization: for converting data in retrieval results in the form of data visualization.

5) Display of visualization relationship: for design of data visualization is until targets visualization and displays it on the visualization relationships. Inforismodel information visualization model is shown in Figure 4.

Responding to questions

Other application the parts of the natural language is, processing natural language questions, or answer to questions, which are faced with how find the best answer for the presented questions. Many Web sites that are equipped with the technology of answering questions, and allows the desired user that ask "question" and computer give an answer to it. Q & A can use a variety of text mining techniques. For example, it can use information extraction for entities extraction such as people, places, and events or considers questions classification in form of types of known questions (who, where, when, how, etc.)... Besides the web applications, companies can use Q & A techniques privately for applying those who are seeking answers for their frequently questions. Educational and medical sectors also may use Q & A in sectors that many questions are being asked and people want to respond it. Applied system of NL natural language asks a question from the user in English. The question then transfer into the idiomatic speech part (pos), which in the interrogative phrase and pos identified from each word questions, is included. The questions tagged, then with query generator that produces different types of questions, are used, which can be transferred into the search engine. The questions then with the search engine are implemented in orbit. Search engine provide documents that probably for our responses are searched. The documentation for these is investigated by extractor. Summaries extractor extracts summaries that include phrases / interrogative words from the documentation. These pieces to ranker that ranks based on the position of the algorithm.

In the system QUA SAR (View Star) user mentions a question and this is given during it until analyzes the question and transfers recovery models. Then extracted question obtains an answer for the desired type of limitations and phrases with analytical questions and phrases recovery models.

The main objective of the questions analysis model is, creating a variety of expected questions from the interrogative text. This is a very important stage of the process until the extracted questions model use different strategies depending on the type of the desired questions. The other performed operations with this questions analysis model are in order to identify limits for using in the AE stage. These limitations with strings of extracted words from on pos labeled question by the pos patterns and rules have been created. For example, each string names (like “ozonhole" "ozone holes") as relevant patterns have been examined. Transient recovery model also help to answering users questions in the JIRS transitional recovery system and particularly with components of the N-grams search engine. Phrases with the relevant terms (namely without the interdiction words) are found by search engines using the IR’s original system. Sets of 1 g, 2 g, 3 g,.., n g are extracted from developed components and from users questions. In both conditions, n is the number of interrogative phrase. A comparison is done between the n gram set of components and user questions through obtaining weights per piece. The weight of each piece will be heavier if the components, including gram structures larger than the question, the extracted model of input questions with n returned piece have been created by the PR model and obtained limitations from the among the Q & A (including types of slightly questions).

Textual Crawler model in this moment is for each of the components with a set of templates for the kinds of desired questions and the preprocessing version of textual piece. The preprocessing Textual Part is based on separating all of the confirmative characters (punctuation) from words and opening descriptions of all the pieces. It is important that all the symbols and punctuation be maintained because we see that they usually suggest important sets for the unique responses: textual Crawler model starts this work with searching all the parts, subset systems of the desired patterns. Then each intended value for each S substring is found, depending on position limitations, if S is not contain any type of limited words. Filtering model achieve advantages of number of information resources, such as micro basic information or Web, as a result of putting aside the appropriate responses that do not match with followed patterns or is according unauthorized patterns. For example, a list of names of countries with four languages is including the basic information and putting aside the name of countries when seeking countries. When extracting the appropriate anathematized, Crawler textual model provide this, for next more appropriate value, if one of them exists. Finally, when using the all textual Crawler and text analysis, model of answer selection, choose appropriate answer for refer to the system.

Combination methods:

There are many approaches in knowledge extraction phase. However, all of these methods may be divided into two main categories. These two main categories, methods based on and methods performance based on knowledge. The first method, the designers are concerned about system performance and design the system so that they have the best performance and speed. The most common methods in this type of attitude are statistical methods and neural networks. Statistical methods are based on each type of statistical information that is extractable from the text. Such as repeating words alone, repeating words together and something like it, in the other hand methods are located based on knowledge that look at to this problem from another perspective. They try Firstly as much as possible extracts existing concepts from within set of texts and, secondly, establish relationships between these concepts. Using this method is very dependent on the NLP. In fact, it is a goal that NLP also will follow it and it is text understanding. Systems that use these methods currently are not high, but the DR-LINK from Syracus University is one of them.

Discotex Method

This method has been proposed in 2007 by Kanya. This method provides a new framework for text mining based on integrating information extraction system (IE) and the module of inference the standard rules (KDD). IE converts documents into more structured data. In fact searches certain pieces of data in documents to natural language, and a series of semi-structured textual documents converts into a more structured database. In this method, RAPIER and BWI have been used for making IE. Then built database by IE module is used by KDD module to explore more knowledge. In an improved version from this method derived law from KDD module for forecasting the missing information and improve the accuracy of IE module is used. For building KDD module has been used from APRIORI and RIPPER. The first task is, building a database by applying an information extraction system learned on the set of documents to natural language. In This way, standard data mining techniques on extracted data are applied. This knowledge discovery can be used for many tasks. In the proposed framework for text mining IE plays an important role. That preprocesses set of documents for transfer of extracted items in the data mining module. In this method, from two systems now in the market have been used for learning information extractor: 1) BWI 2) RAPIER

With training on a set of documents with their filled forms have been described, they acquire a basic

knowledge for the laws extraction and so can be tested on the new documents. BWI and RAPIER has been proved well on real applied programs seminar and USENET job posting example announcement is implemented. After creating an IE system that Solt desired set that has been extracts for applied program. By applying IE extraction patterns for each document for creating a collection of structured records can build a database of texts set. So standard KDD techniques on resulted database are applied to explore the relationships of interests.

To discover the predictions laws we assume any slot-value pairs in the extracted database as a separate binary features example graphic Earcel. Then will learn rules for prediction of any feature from all other features.

Text miner Method

In this method term and events of each document for finding the features that means in domain, is extracted, and then apply searching on the extracted features and labeled of each document. This system consists of two main components:

- Text analysis component
- Data mining component (Dt)

In this method, a semi-structured data changes for example documents cover to structured data stored in a data base. The second component applies data mining techniques on the output of the first component and more methods for the text mining applies search algorithms on the labels assigned to each document. These labels might be keywords extracted from the document or just a list of words in the document. In text miner Method search algorithms is applied on terms (meaningful sequence of words such as department of computation) combined with the events (meaningful set from terms, such as in a financial domain, purchasing between company A and B) extracted from the documents. Authors of this article believe that the most important Factors of characteristic that describe a document are terms and events expressed in the document. This information is kept in a table named the e. Information extraction is an important technology that has a pre-processing step. In this method once extracts information, and then the information can be stored in the database and search for query and is summarized in natural language. First Necessary step is linguistic preprocessing (linguistics). The Step consists a number of linguistic techniques, such as, tokenization, part of speech tagging, and....

This method is made up from two components, text analysis and data mining. The first component converts semi-structured data into more structured data stored in databases. And the second component applies Data mining techniques on the output of the first component. Goal of this method is, managing the information (classification of documents in the appropriate category) and data mining to discover useful knowledge. So In this method terms and events is extracted and stored in the database. Then a suitable clustering algorithm (using the algorithm Rock and the concept link) is applied on the obtained database and documents are grouped so that similar documents are in one group. Then an appropriate classification algorithm (decision tree) is applied for the more validate the results of clustering and better utilization from discovered knowledge. Comparison of the criterion combination methods of precision and recall are shown for the combination different methods in Table 2. Method matching filler in the table have came, a method that always extracts sub-fields that are famous filler for a specific gap

Text mining applications

There are variety definitions from Text-mining, so it's not surprising that are exist various opinions about the applications from text mining. Hence we try examining that the number of accepted applications from this process and do not have tried in accordance these applications with the previous expressed definitions for the text mining.

While text-mining is a new field but textual data analysis software such as SAS and SPSS from late 1990 by suppliers are available. Among the most common applications of text mining can be named search engines where users type a phrase or word (which may have misspelled) and the search engines by a large repository of documents that have, find most relevant documents. Among other applications of text mining is as follows.

Identification spam: anglicizing the title and the content of e-mail to determine if it is spam or not.

Supervision: means supervising the behavior of a person or a group of people-are hidden.

A project called E supervises telephones, the Internet and other communication devices to identify terrorism (wikipedia)

Identification of aliases: aliases in medical care is analyzed to identify frauds. For example an account may is presented in the name of J. Smith, John Smith and Smith, John. In this way or by other methods is possible claimant abuse and provide a lot of demands of premium under the various aliases.

Summarizing: order of summarizing, is the process of creating a set of text basic concepts in a few lines. In this type of text mining, it seems that the new information from text does not obtain because the author probably knew what wanted to say and summarizing of his writings, do not add new information. Although this work may more simplify for users to evaluate the contents of documents and speeds them in path arriving to what they need.

The relationships between concepts

Among facts that can be received from a texts collection is correlation and dependency of some concepts with the others concepts. These facts for example can say that the occurrence of some words may be dependent to the appearance of some other words. Purpose is that whenever you see first words set, can expect to see the second set of words we too. This concept is borrowed from the Data Mining in database.

Finding and analysis of trends: To describe this application you assume that you are the manager of a commercial company. Obviously you should always supervise your competitors' activity. This could be any kind of information you have taken from news, bourse transactions or documents generated by same competitor companies. Currently the information increasingly is increasing; the management of these resources is not possible just by the help of eyes. Text mining allows it that automatically finds new trends and changes. In fact, what essentially should expect from Text Mining is that among a range of news tell you what the news is relevant to what you want. In the meantime, what news is new, what progress do in your work and how is current trends are and interests and with what process changes. Using this information, director can benefits only from discovered information for evaluating the competitor status.

That from website of university in the address gate.ac.uk is available. Another platform is supported by JBM, is called UIMA4 in the address research.ibm.com/UIMA is available. In addition to these tools, a number of vendors, data mining, text mining capabilities within its software packages offer. Because the area is still under research and development capabilities of the software will change rapidly. A list of tools and popular vendors data mining is in this form as well as following commercial software are software for text mining software:

1) Aero Text and suitable application for content analysis

2) Intensity: independent text mining software, integrated a host of using for processing natural language

3) Basic Technology: provides an analysis model of appropriate text to identify the language, the ability of search in more than 20 languages, existence extraction, and efficient search for translated entities

4) Autonomy: classification software, clustering text mining

5) Expert System SpA: products and suitable technologies for developers and managers of knowledge

Conclusion

This was an overview on main applications and the methods are used in text mining. Although a wide range of applications for this technology is conceivable. However this is a young field growing that will help us take advantage of the knowledge contained in unstructured text, later work about the methods will be used from NLP.

In this field, also is an idea that in it the Human Plausible Reasoning will be used. It's quite natural that we use from such logical framework in text mining when we use understanding of text.

References:

1. Karami, Mahtab, the use of data mining and text mining tools in agility of health and medical care organizations, scientific research quarterly, period 10, No. 30.
2. J.Froelich,S.Ananyan, and D.L Olson, "Business Inteligence Through Text Mini."Business Intelligence Journal, Vol.10,No. 1, Winter 2005. p.43-50; and Gain Full Valuefrom Text Response, spss.com/ textanalysis_surveys/ (accessed April 2006).
3. M. Rajman. “Text Mining, Knowledge extraction from unstructured texual data”. Proc. of EUROSTAT Conference, Francfort (Deutchland), may, 1997.
4. R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.
5. H. Karanikas and B. Theodoulidis, ‘Knowledge discovery in text and text mining software’, Technical report, UMIST - CRIM, Manchester, 2002.
6. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In 7th Int. Conf. on Information and Knowledge Managment, 1998.