Text Mining with Information Extraction of Predicting Diseases

P. Sumathi[1], R.Manicka chezian[2],

ABSTRACT

Text mining is a new and exciting research area that attempts to solve the information overload problem technique for automatically extracting association rules from collections of textual documents. Depending on keyword features for discover association rules amongst keywords labeling the documents. The main contributions of the technique are that it integrates XML technology with Information Retrieval scheme for keyword/feature selection that automatically selects the most discriminative keywords for use in association rules and use Data Mining technique for association rules discovery. It consists of three phases: Text Preprocessing phase (transformation, filtration, stemming and indexing of the documents), Association Rule Mining (ARM) phase (applying our designed algorithm for Generating Association Rules based on Weighting scheme GARW) and Visualization phase (visualization of results). Experiments applied on WebPages news documents related to the outbreak of the swine flu disease. The extracted association rules contain important features and describe the informative news included in the documents collection. The performance of the system compared with another system that uses the Apriori algorithm throughout the execution.

Keywords— Text mining, data mining, association rules mining, apriori, visualization.

I. INTRODUCTION

Access to a large amount of textual documents becomes more and more effective due to the growth of the Web, digital libraries, technical documentation, medical data,... These textual data constitute resources that it is worth exploiting. In this way knowledge discovery from textual databases or for short, text mining is an important and difficult challenge, because of the richness and ambiguity of natural language (used in most of the available documents). Therefore, the problem is the existing of huge amount of textual information available in textual form in databases and online sources. So the question is who is able to read and analyze it? In this context, manual analysis and effective extraction of useful information are not possible. We think the solution is that it is relevant to provide automatic tools for analyzing large textual collections by automatically find relevant information, analyze relevant information and structure relevant information. Text mining is an increasingly important research field because of the necessity of obtaining knowledge from the enormous number of text documents available, especially on the Web.

Text mining and data mining, both included in the field of information mining, are similar in some sense, and thus it may seem that data mining techniques may be adapted in a straightforward way to mine text. However,

1. P. Sumathi M.Sc.,M.Phil.,(Ph.D). Assistant Professor, School of IT and Science, Dr. G R D College of Science, Coimbatore, India. sumathi.p@grd.edu.in

2. R.Manicka chezian, Associate Professor, Department of Computer Science, NGM College Pollachi, India. chezian_r@yahoo.co.in

data mining deals with structured data, whereas text presents special characteristics and is unstructured In this context, the aims of this paper are to study particular features of text, to identify the patterns we may look for in text and to discuss the tools we may use for that purpose. In relation with the third point, we describe the text tool that we developed by adapting data mining technique[1].

Where, the analyzing and extracting useful information from documents written in natural language is very hard. We select some WebPages that are containing information news about the outbreak of the swine flu disease. The Motivations of choosing this domain are that

- Medical field is a general domain into which a great deal of effort in terms of knowledge management placed.
- Contain additional, valuable information which is comprehensive, up-to-date
- Our text mining system can more easily be adapted to this domain (because it contains many generic kinds of concepts or features)
- It does not require a domain expert to understand the features and concepts involved.

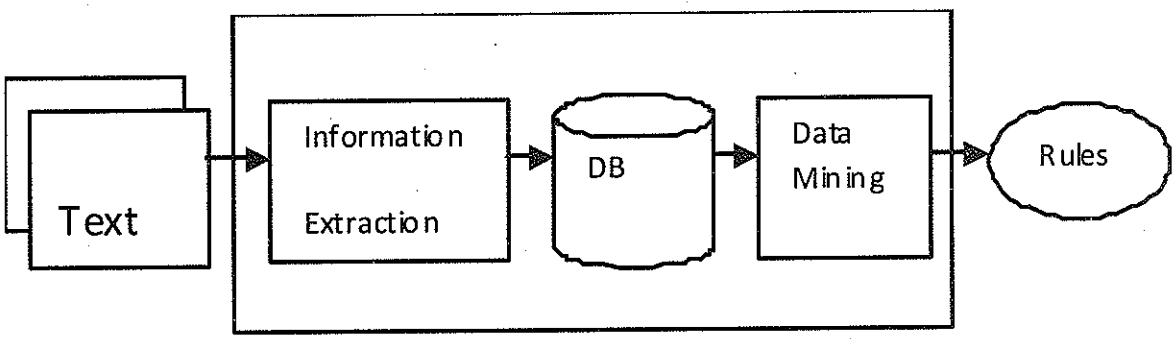

Figure 1

Since the volume of published online news about swine flu disease is expanding at an increasing rate because the virus of swine flu that is called H1N1 is speedily spreading in many countries in the world. With this explosive growth of news, it is extremely challenging to keep up-to-date with all of the new about cases of swine or humans that are infected or dead with the virus and the countries that the virus appears and spreading in it. There are many sources of this news such as newspapers, Reuters, BBC, CNN, Medical News Today, Yahoo news, World Health Organization web reports...etc. Some of this news are geographical news that are about spreading of the virus in many countries, news about humans infection , news about treatments that be used against the virus and also the new medicine discoveries research in this field. This news seems to be different and have a different kind of knowledge, so there are challenges to sharing of knowledge among these different topics[1]. The challenge is the multidimensionality of information sources of the disease like:

- Geographical spreading
- Spreading across species

Focusing on the points. Association rules highlight correlations between features in the texts, e.g. keywords. A word is selected as a keyword if it does not appear in a pre-defined stop-words list. Moreover, association rules are easy to understand and to interpret for an analyst or

may be for a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced. We have described in this paper a system for automatically extracting association rules from WebPages news documents that are about the outbreak of the swine flu disease. The system depends on word feature to extract association rules. For the infectious disease outbreak, the task is to track the spread of epidemics of infectious disease around the world. The system has to find many relationships between features such as the name of the disease, the location of the outbreak, the type of victim (e.g. human or animal) and the victim status (infected and dead). We ignore the order in which the words occur, but instead focusing on the words and their statistical distributions in text documents. In order to use the unordered words it is necessary to index the text. The index tends to be very large, so terms that are grammatically close to each other (like "disease" and "diseases") are mapped to one term via word stemming and terms that occur very often are removed by compiling stop word lists, so they do not interface with the data analysis. The extracted association rules identify the relations between features in the documents collection[2]. The scattering of features in text contribute to the complexity of define features to be extracted from text.

II. TEXT MINING SYSTEM ARCHITECTURE

The proposed text mining system It automatically discovers association rules from textual documents. The main contributions of the system are that, it integrates XML technology with an Information Retrieval scheme for feature selection that automatically selects the most discriminative features for use in association rules generation and with Data Mining techniques for association rules extraction. The system ignores the order in which the words occur, but instead focusing on the words and their statistical distributions. The system begins with selecting collections of documents from the web or internal file systems. The extracting association rules from text system consists of three phases: Text Preprocessing phase(transformation, filtration, stemming and indexing of the documents), Association Rule Mining phase and Visualization phase (visualization of results).

Text Preprocessing Phase

The goal of text preprocessing phase is to optimize the performance of the next phase: ARM. This phase begins with the transformation process of the original unstructured documents. This transformation aims to obtain the desired representation of documents in XML format. After that, the documents are filtered to eliminate the unimportant words (e.g. articles, determiners, prepositions and conjunctions, etc.) by using a list of stop words and after word stemming. The resulting documents are processed to provide basic information about the content of each document[2].

Transformation

The system accepts a different number of documents formats (doc, txt, rtf, etc.) and structures to convert them into the XML format amenable for further processing. In this work, we save the WebPages news as text documents and the text mining system transformed it into XML format.

Filtration

In this process, the documents are filtered by removing the unimportant words from documents content. Therefore, the unimportant words get discarded or ignored (e.g. articles, pronouns, determiners, prepositions

and conjunctions, common adverbs and non-informative verbs) and more important or highly relevant words are single out. We build a list of unimportant words called stop words, where the system checks the documents content and eliminate these unimportant words from it. In addition, the system replaces special characters, parentheses, commas, etc., with distance between words in the documents. After the filtration process the system does word stemming, a process that removes a word's prefixes and suffixes such as unifying both infection and infections to infection. We designed a stemming dictionary (lexicon) for the used medical domain.

III. INDEXING

The filtered and stemmed XML documents are then index by using the weighting scheme. If the textual data is indexed, either manually or automatically, the indexing structures can be used as a basis for the actual knowledge discovery process. As a manual indexing is a time-consuming task it is not realistic to assume that such a processing could systematically be performed in the general case. Automated indexing of the textual document base has to be considered in order to allow the use of association extraction techniques on a large scale. Techniques for automated production of indexes associated with documents can be borrowed from the Information Retrieval field. Each document is described by a set of representative keywords called index terms. An index term is simply a word whose semantics helps in remembering the document's main themes[3]. It is obvious that different index terms have varying relevance when used to describe document contents in a particular document collection. This effect is captured through the assignment of numerical weights to each index term of a document.

$$W(i,j) = \text{tfidf}(di, tj) = \begin{cases} N di, tj * log2\,(Ntj) & \text{if } N di, tj \geq 0 \\ 0 & N di, tj = 0 \end{cases}$$

where w(i,j) e" 0, Ndi , t j denotes the number the term t j occurs in the document d i (term frequency factor),Nt j denotes the number of documents in collection C in which t j occurs at leleast once (document frequency of the term t j) and %C% denotes the number of the documents in collection C. The first clause applies for words occurring in the document, whereas for words that do not appear(Nd i , t j =0), we set w(i,j)=0. Document frequency is also scaled logarithmically. A word that occurred in all documents would get zero weight (logC-log Nt j =log C –log C =0). This weighting scheme includes the intuitive presumption that: the more often a term occurs in a document, the more it is representative of the content of the document (term frequency) and the more documents the term occurs in, the less discriminating it is(inverse document frequency). Once a weighting scheme has been selected, automated indexing can be performed by simply selecting for each document the keywords that satisfy the given weight constraints[4]. The major advantage of an automated indexing procedure is that it reduces the cost of the indexing step.

IV. WEIGHT CONSTRAINS

The notation of term relevance with respect to a document collection is a central issue in Information Retrieval. We assign for each keyword its score (weight value) based on maximal with respect to all the documents in the collection. Our aim is to identify and filter the keywords that may not be of interest in the context of the whole document collection either because they do not occur frequently enough or they occur in a constant distribution among the different documents. Our system uses a statistical relevance-scoring function that assigns a score to each keyword based on their occurrence patterns in

the collection of documents, and the top N taken as the final set of keywords to be used in the ARM phase. The system sort the keywords based on their scores and select only the top N frequent keywords up to M % of the number of running words[5]. This is the criteria of using the weight constraints.

Association Rule Mining Phase

This phase presents a way for finding information from a collection of indexed documents by automatically extracting association rules from them. Association rules have already been used in text mining .Given a set of keywords A= { w1,w2,w3.....wn} and collection of indexed documents D= {d1,d2,d3.....dn} where each document d I is a set of keyword. A document di is said to contain Wi if and only if Wi & di . An association rule is an implication of the form WiÒ! W j. There are two important basic measures for association rules, support(s) and confidence(c). The rule WiÒ! W j has support s in the collection of documents D if s% of documents in D contain Wi*"W j . The support is calculated by the following formula

$$\text{Support } (Wi\ W\,j) = \frac{\text{Support count of } WiW\,j}{\text{Total number of documents } D}$$

The rule Wi W j holds in the collection of documents D with confidence c if among those documents that contain Wi, c % of them contain W j also. The confidence is calculated by the following formula

$$\text{Confidence } (Wi\ /W\,j) = \frac{\text{Support } (W\,iW\,j)}{\text{Support } (W\,i)}$$

An association rule-mining problem is broken into two steps:

1) Generate all the keyword combinations(keyword sets) whose support is greater than the user specified minimum support (called minsup). Such sets are called the frequent keyword sets

2) Use the identified frequent keyword sets to generate the rules that satisfy a user specified minimum confidence (called minconf). The frequent keywords generation requires more effort and the rule generation is straightforward. We design an algorithm for Generating Association Rules based on Weighting scheme (GARW). The GARW algorithm does not make multiple scanning on the original documents but it scans only the generated XML file during the generation of the large frequent keyword sets[5]. This file contains all the keywords that satisfy the threshold weight value and their frequencies in each document.

The GARW algorithm is as follows:

1. Let N denote the number of top keywords that satisfy the threshold weight value.

2. Store the top N keywords in index XML file along with their frequencies in all documents, their weight values and documents ID. Four XML tags for all keywords (<doc-id>, <keyword>, <keyword-frequency>, <TF-IDF>) index the file.

3. Scan the indexed XML file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequent 1-keyword set L.

4. In 2 e" k , the candidate keywords k C of size k are generated from large frequent (k-1)-keyword sets, that is generated in the last step.

5. Scan the index file, and compute the frequency of candidate Keyword sets k C that generated in step 4.

6. Compare the frequencies of candidate keyword sets with minimum support.

7. Large frequent k-keyword sets k L, which satisfy the minimum support, is found from step 6.

8. For each frequent keyword set, find all the association rules that satisfy the threshold minimum confidence.

Visualization Phase

The extracted association rules can be reviewed in textual format or tables, or in graphical format. In this phase, the system is designed to visualize the extracted association rules in textual format or tables.

V. EXPERIMENTS

The EART system is a user-friendly application developed in order to simplify experimenting with rule mining in textual documents collection. The EART system is essentially a process consisting of three operations:

1) Loading the document collection.

2) Let the user enters the three-threshold values weight, support and confidence.

3) Let the system perform the operations and presents the result, i.e., the association rules generated based on the documents collection, operations, and measures or parameters. In this section, we describe the argumentation for the thresholds chosen, the interpretation of the extracted association rules and the evaluation of the EART system[6].

Repesent documents in XML format, Based on keyword features for extract association rules, Automatic indexing process reduces the cost of the indexing step, Using scheme is very important to filter the unimportant keywords in the context of the whole documents collection. The number of the top N of keywords is always greater than the M% of the running keywords

Previous work applied on the Medline abstracts that are more scientific so they require a domain expert to understand the features and concepts involved. However, this extended work applied on webpages news documents that are understandable for any reader. Previous work depends on the analysis of the keywords in the extracted association rules through the co-occurrence and without co-occurrence of the keywords in one sentence in text. Our work here ignores the order in which the words occur, but instead focusing on the words and their statistical distributions in documents collection. Where the extracted association rules contain important features and they describe the important news included in the documents.

Data Description

To investigate the use of the system to extract association rules from text, we applied it on a selected sample of 100 recent WebPages news that are related to the outbreak of swine flu disease in the time from 3 April 2009 to 9 Nov. 2009. There are many sources of this news such as Reuters, BBC, Medical News, Yahoo news ...etc. Some

of this source news is geographical news that is about spreading of the virus in many countries and news about human's infection. The collection of the 100 documents (corpus) is 440 KB in size and contained 30000 single words. Each document contained on average 300 single words. After the filtration process, the collection of documents contained 9500 single word. The system implemented using C# language.

Finding association rules in text document can be useful in a number of contexts. For example, investigations, and in general understanding affect of events in the real world. Extraction of association rules was achieved by using the GARW algorithm. The system extracted association rules depending on the analysis of relations between the keywords in the text documents collection. This analysis has been done through the scattering of the features in the text[7]. The text in is a segment of an update about an outbreak of the swine flu disease in USA, from News-Medical Net.

Sample news:

Swine flu back again in USA Disease/Infection News Published: Thursday, 27-Sep-2009The USA Ministry of Health and the World Health Organisation (WHO) have confirmed that another case of flu in pigs has been found in the country. The latest case of H1N1 has been detected in fredonia a town near Aswan ,in USA. USA has suffered the worst outbreak of flu so far this year apart from Asia, and although the disease was largely brought under control, fears remain of a renewed outbreak. An outbreak in mid-February among poultry led to the culling of at least 20 million pigs nationwide and of 16 human cases of swine flu found since mid-March, 6 have died. The last death was of a 75-year-old woman who died on the 15th March...

There are multiple features in this document and they are scattered widely in the text such as "swine flu", "USA", "outbreak", " flu", "poultry", "pigs", "human", "died", and "woman", etc[8]. In this excerpt, there are many separate mentions-partial descriptions of the feature in text- describing victims infected with swine flu. The aim of this work is to find relations between the features and represent them in association rules form to give the end user or the reader of news the useful information about the outbreak of disease. In the case of extracting these relations, we do not take into account the order in which the keywords occur. Our attention is finally paid to extract useful news information from documents based on abstractions that describe the relationships between features in text.

In our scenario of association rules extraction, we observe the following features and our system get the relationships between them

disease: disease name
location : continent, country, city
victim type: e.g. "human", "pig" and "animal"
victim descriptor: e.g., "people", 'boy", "poultry" "pig"...etc.
victim status: dead, infected, sick

We have many of relations between features per document; this means we have many of association rules to be extracted

In text mining in general, a very number of association rules will be found. So the measures like support and confidence are important when creating keyword sets and selecting the final rules[9]s. However, the problem is that we may find the important keywords which have frequently appeared recently but not discovered because the height of support and confidence threshold values. So, one of our purposes is to find these informative

keywords to extract more informative rules. In our experiments, we choose low threshold support value 20% to extract important keywords (such as swine flu) that cannot appear if we chose high support value and chose high threshold confidence value 80% to extract the more interesting rules.

We present some of our association rules abstractions that describe the relations between features in text. In addition, they give information about geographical spreading and spreading across species of the swine flu disease. The extracted association rules get the relations of the existing of the keywords in text documents collection ignoring the order in which these keywords occur. The system concentrates on the distribution of features in text to get the rules that are more useful and give information[9]. The following rules represent the relation between the disease and its spreading location:

- **<location> —> < disease > or**
- **< disease > —> <location>**

swine flu, outbreak —> USA	100%
outbreak, USA —> swine flu	100%
spread, Thailand —> swine flu	100%

<disease> <location> —> < victim >

swine flu ,outbreak, USA, fredonia —> pigs	100 %
swine flu, Jakarta —>human	85 %

The following rules represent the relation between the disease, its spreading location and victim:

[(<victim><location>)or(<location><victim>)]—><disease>

human, outbreak, USA —>swine flu	100 %
woman, USA—> swine flu	100%
boy, Jakarta—>swine flu	91%
China, farmer—> swine flu	100%
Indonesia pigs —> swine flu	86%
poultry, Asia —> swine flu	100%

More informative association rules than the above rules represent the relationships between features such as the disease, its spreading location, victim, and victim status as follows:

<disease> <location> < victim > —> < victim status >

Swine flu, Indonesia, boy —> died	100 %
Swine flu, Indonesia, swines—> died	100 %

<victim status> <victim> <location> —> < disease >

died, woman, Egypt—>swine flu	100%
died, boy, Jakarta —> swine flu	91%
died, China, farmer —> swine flu	100%
died, woman, Egypt-->swine flu	100%
died, boy, Jakarta —> swine flu	91%
died, China, farmer—> swine flu	100%

It can be noticed that the extracted association rules include the most important features and informative news of the domain in the documents collection. We design another system for extracting association rules from text by using the Apriori algorithm

VI. ALGORITHM

Input: D is the set of records.
Output: Lk is the frequent k-item sets.

```
F unction Apriori (D)
L1 := Find Frequent Item sets(D).
k := 2.
while (Lk 16 = ;) do
begin
Ck := GenerateCandidates(Lk-1).
forall records r 2 D do
forall c 2 Ck do
if c †"sof t r
c:count := c:count + 1.
Lk := All candidates in Ck
with minimum softsups.
k := k + 1.
end
Return k Lk..
```

While counting the occurrences of all items, we measure the similarity of every pair of items and construct an m × m matrix similar(i, j), where m is the total number of items in the database. To determine frequent 1 itemsets, the soft-supports of all items are computed. Intuitively, we construct a cluster of items containing the items similar to each given item, and sum the support of all items in the cluster. The similarity table is used to eûciently retrieve similar items.

It can be observed that the performance of algorithms largely depends on the number of frequent keyword sets. For lower values of minimum support, it is expected to have many frequent keyword sets and this number will decrease as the minimum support increases.

Therefore, the execution time decreases as the minimum support increases in both systems. The large number of candidate keyword sets created in the Apriori-based system caused the large gap between this system and the system at lower values of minimum support. The reason of this is that the Apriori-based system generates all frequent keyword sets from all keywords in the documents that are important and unimportant[10]. This leads to extract interesting and uninteresting rules.

VII. CONCLUSION

This paper has presented a text mining technique for automatically extract association rules from collection of documents based on the keyword features. The system has been designed to accept documents with different structures and formats to transform them into the structured form and it is domain-independent so it is flexible to apply on different domains. The system can be applied on all or specific parts of documents. In addition, it is designed to automatically index documents by labeling each document by a set of keywords that satisfy the given weight constraints based on the weighting scheme. We designed an algorithm for association rules generation based on the weighting scheme. We compared the performance of our system that based on the algorithm with a system that use Apriori algorithm reduces the execution time in comparable to the Apriori algorithm. Therefore, our system performed well against the one that we compared. In addition extracted more interesting rules than the other compared system. We plan to extend our text mining system to use the concept features to represent text and to extract the more useful association rules that have more meaning. In addition, we intend to conduct experiments on the medical domain where in this case, we will focus on the disease treatments their effectiveness and side effects.

REFERENCES

[1] R. Feldman and I. Dagan, *"Knowledge discovery in textual databases(KDT)"*, in Proc. 1St Int. Conf. on Knowledge Discovery and Data Mining, 1999

[2] B. Lent, R. Agrawal, and R. Srikant, *"Discovering trends in text Databases,"* KDD'97, 1997.

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval (Addison-Wesley, Longman publishing company, 1999).

[4] Y. Kodratoff, *"Knowledge discovery in texts: a definition, and applications,"* in Proc. of th 2ndInt., symposium, ISMS'99, Vol. 1609 of LNAI, Warsaw, Pol. Springer, Berlin Heidelberg New York.

[5] K. Norvag, T. Eriksen, and K. Skgstad, *"Mining association rules in temporal document collections,"* Available: http://www.idi.ntnu.no/~noervaag/papers/ISMIS2006.pdf

[6] H. Mannila, H. Toivonen and A. I. Verkamo, *"Discovery of frequent episodes in event sequences,"* Data Mining and Knowledge Discovery.

[7] H. Karanikas and B. Theodoulidis, *"Knowledge discovery in text and text mining software,"* Technical Report, UMIST Departement of Computation, January 2002..

[8] R. Agrawal and R. Srikant, *"Fast algorithms for mining association rules,"* In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo editors, Proc. 20thInt. conf. of very Large Data Bases, VLDB, Santigo, Chile, 1994, 487-499.

[9] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Inkeri Verkamo, *"Applying data mining technique for descriptive phrase extraction in digital document International Journal of Information and Mathematical Sciences 4:1 2008 27 collections,"* in Proc. of IEEE Forum on Research and technology Advances in Digital Libraries, Santa Barbra CA, 1998.

[10] C. Cardie. Empirical methods in information extraction. AI Magazine, 18(4):65–79, 1997.

Author's Bibliography

Mrs.P.Sumathi M.Sc.,M.Phil., working as Assistant Professor with eight years of experience in teaching, She has presented five papers in national level conference and two papers in journals. She has participated in seminars/ conference/workshops/refresher courses at national Level. Now she is pursuing Ph.D Computer Science in Dr. Mahalingam center for research and development at NGM College Pollachi.

Dr. R.Manicka chezian M.Sc., M.S. Ph D in Computer Science. He served as a Faculty Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989.Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published fifty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.