A Text Mining Technique Using Association Rules Extraction

Nabil Ismail
Minoufiya University
121 PUBLICATIONS **253** CITATIONS

A Text Mining Technique Using Association Rules Extraction

Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey

Abstract**—**This paper describes text mining technique for automatically extracting association rules from collections of textual documents. The technique called, Extracting Association Rules from Text (EART). It depends on keyword features for discover association rules amongst keywords labeling the documents. In this work, the EART system ignores the order in which the words occur, but instead focusing on the words and their statistical distributions in documents. The main contributions of the technique are that it integrates XML technology with Information Retrieval scheme (TF-IDF) (for keyword/feature selection that automatically selects the most discriminative keywords for use in association rules generation) 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), A*ssociation 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 bird flu disease. The extracted association rules contain important features and describe the informative news included in the documents collection. The performance of the EART system compared with another system that uses the Apriori algorithm throughout the execution time and evaluating extracted association rules.

Keywords—Text mining, data mining, association rule mining

I. INTRODUCTION

THE 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 (TM), 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, data mining deals with structured data, whereas text presents special characteristics and is unstructured.

In this context, the aims of this paper are three: 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.

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 bird 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.

Since the volume of published online news about bird flu disease is expanding at an increasing rate because the virus of bird flu that is called H5N1 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 birds 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. The challenge is the multidimensionality of information sources of the disease like:

- Geographical spreading

Manuscript received, 6-12-2006.

H. Mahgoub, D. Roesner are with the Knowledge and Language Engineering Department, Faculty of Computer Science, Otto-von Guericke University Magdeburg, Germany (e-mail: mahgoub, roesner@iws.cs.uni-magdeburg.de)

N. Ismail, F. Torkey are with the Computer Science Department, Faculty of Computers and Information, Minufya University, Egypt (e-mail: nabil, fawzy@mufic.com)

- Spreading across species
- Countermeasures (treatments)

In this paper, we focus on the above two points. Following our previous work in [7], we present in this paper the use of association rule in TM. Association rules highlight correlations between features in the texts, e.g. keywords. A word is selected as a keyword if it dose 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 bird 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, bird 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. The scattering of features in text contribute to the complexity of define features to be extracted from text. These kinds of features relationships can be better described with our text mining system.

The reset of the paper is organized as follows. Section II presents our text mining system architecture. Experiments, interpretation and discussion are presented in section III. Section IV presents the related work. Section V provides conclusion and future work.

II. TEXT MINING SYSTEM ARCHITECTURE

The proposed text mining system, Extracting Association Rules from Text (EART) is shown in fig. 1. 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 (TF-IDF) (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 EART 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 EART system consists of three phases: *Text Preprocessing phase* (transformation, filtration, stemming and indexing of the documents), A*ssociation Rule Mining (ARM) phase* (applying GARW algorithm for generating association rules) and *Visualization phase* (visualization of results).

A. 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.

A1. 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.

A2. 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 (e.g., be)) 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.

A3. 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 [14, 15], 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 [13]. 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 [13]. 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.

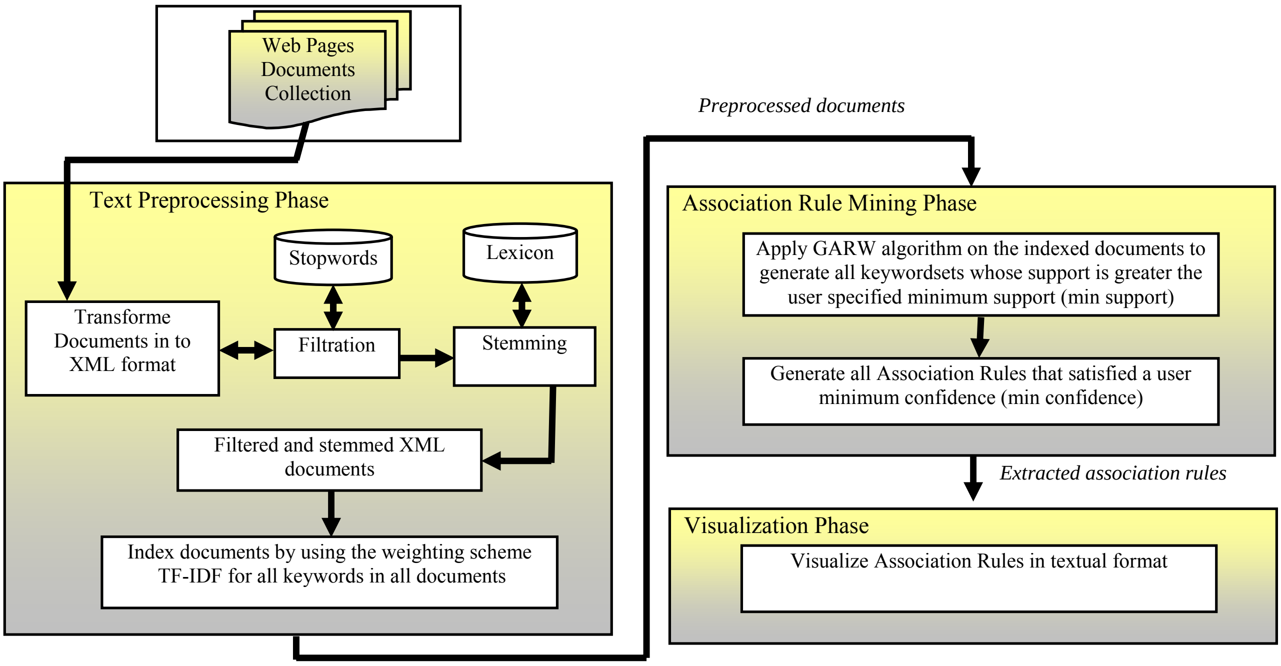

Web Pages Documents Collection
Preprocessed documents
Text Preprocessing Phase
Stopwords
Lexicon
Transforme Documents in to XML format
Filtration
Stemming
Filtered and stemmed XML documents
Index documents by using the weighting scheme TF-IDF for all keywords in all documents
Association Rule Mining Phase
Apply GARW algorithm on the indexed documents to generate all keywordsets whose support is greater the user specified minimum support (min support)
Generate all Association Rules that satisfied a user minimum confidence (min confidence)
Extracted association rules
Visualization Phase
Visualize Association Rules in textual format

Fig. 1 Text Mining System Architecture

The techniques for automated production of indexes associated with documents usually rely on frequency-based weighting schemes. The weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign higher weights to distinguished terms in a document, and it is the most widely used weighting scheme which is defined as (cf. [2] [9] [11]):

$$w(i,j) = tfidf(d_i,t_j) = \begin{cases} Nd_i,t_j * \log_2 \frac{|C|}{Nt_j} & if\ Nd_i,t_j \geq 1 \\ 0 & if\ Nd_i,t_j = 0 \end{cases} \qquad (1)$$

where $w(i,j) \geq 0$, Nd_i,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 least 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. The formula: $\log_2 \frac{|C|}{Nt_j}$ = log C-log Nt_j gives full weight to words that occur in one document (log C – log Nt_j=log C – log 1= log C). A word that occurred in all documents would get zero weight (log C- 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. The major advantage of an automated indexing procedure is that it reduces the cost of the indexing step.

1. Weight Constraints

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 TF-IDF (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 (for a user specified M). This is the criteria of using the weight constraints.

B. Association Rule Mining (ARM) 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 TM [7, 10, 11, 15, 16, 17, 18, 19]. Below we define and describe the association rules in the context of TM. Given a set of keywords $A = \{w_1, w_2, ..., w_n\}$ and a collection of indexed documents $D=\{d_1,d_2,...,d_m\}$, where each document d_i is a set of keywords such that $d_i \subseteq A$. Let W_i be a set of keywords. A document d_i is said to contain W_i if and only if $W_i \subseteq d_i$. An association rule is an implication of the form $w_i \Rightarrow w_j$ where $W_i \subset A, W_j \subset A$ and $W_i \cap W_j = \phi$. There are two important basic measures for association rules, support(s) and confidence(c). The rule $w_i \Rightarrow w_j$ has support *s* in the collection of documents *D* if *s*% of documents in *D* contain $w_i \cup w_j$. The support is calculated by the following formula:

$$Support\ (W_iW_j) = \frac{Support\ count\ of\ W_iW_j}{Total\ number\ of\ documents\ D} \qquad (2)$$

The rule $w_i \Rightarrow w_j$ holds in the collection of documents *D* with confidence *c* if among those documents that contain w_i, *c* % of them contain w_j also. The confidence is calculated by the following formula:

$$Confidence\ (W_i \setminus W_j) = \frac{Support\ (W_iW_j)}{Support\ (W_i)} \qquad (3)$$

An association rule-mining problem is broken into two steps: 1) generate all the keyword combinations (keywordsets) whose support is greater than the user specified minimum support (called minsup). Such sets are called the frequent keywordsets and 2) use the identified frequent keywordsets 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 keywordsets. This file contains all the keywords that satisfy the threshold weight value and their frequencies in each document. We summarize in Table 1 the notation used in the GARW algorithm.

TABLE I
NOTATION

k-keywordsets	An keyword set having *k*- keywordsets
L_k	Set of large *k*- keywordsets (that satisfy minimum support)
C_k	Set of candidate *k*- keywordsets (potentially large *k*- keywordsets)

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 TF-IDF 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-keywordset L_1 .

4. In $k \geq 2$, the candidate keywords C_k of size *k* are generated from large frequent (*k*-1)-keywordsets, L_{k-1} that is generated in the last step.

5. Scan the index file, and compute the frequency of candidate keywordsets C_k that generated in step 4.

6. Compare the frequencies of candidate keywordsets with minimum support.

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

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

C. 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.

III. Experiments, Interpretation and Discussion

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 and 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. This work correspond our previous work [7] in the following:

- Repesent documents in XML format
- Based on keyword features for extract association rules
- Automatic indexing process reduces the cost of the indexing step
- Using TF-IDF 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.

In addition, the differences are:

- 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.

A. Data Description

To investigate the use of EART 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 bird flu disease in the time from 3 July 2006 to 9 Oct. 2006. There are many sources of this news such as Reuters, BBC, Medical News Today, Yahoo news …etc. Some of this source news is geographical news that is about spreading of the virus in many countries and news about birds and humans 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. The experiments were performed on a Pentium 4, 2.2 GHz system running Windows XP professional with 512 MB of RAM.

B .The Description of the Extracted Association Rules

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 in section II. The EART 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. The text in figure 2 is a segment of an update about an outbreak of the bird flu disease in Egypt, from News-Medical Net.

News-Medical.Net
Bird flu back again in **Egypt**
Disease/Infection News
Published: Thursday, 28-Sep-2006
The Egyptian Ministry of Health and the World Health Organisation (WHO) have confirmed that another case of **avian flu** in **birds** has been found in the country. The latest case of H5N1 has been detected in Edfu a town near Aswan, in Upper Egypt. **Egypt** has suffered the worst **outbreak** of **avian 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 birds nationwide and of 14 **human** cases of bird flu found since mid-March, 6 have **died**. The last death was of a 75-year-old **woman** who **died** on the 18th May…

Fig. 2 An excerpt from a news document (abbreviated)

There are multiple features in this document and they are scattered widely in the text such as “bird flu”, “Egypt”, “outbreak”, “avian flu”, “poultry”, “birds”, “human”, “died”, and “woman”, etc. In this excerpt, there are many separate *mentions*-partial descriptions of the feature in text-describing victims infected with bird 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. Fig. 3 shows the graphical representation of bird flu’s outbreak relationships

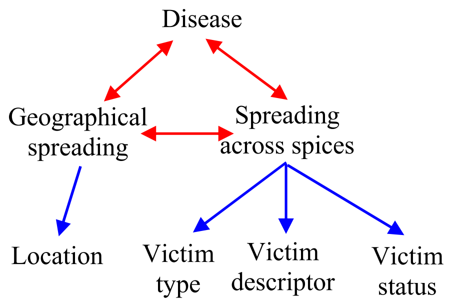
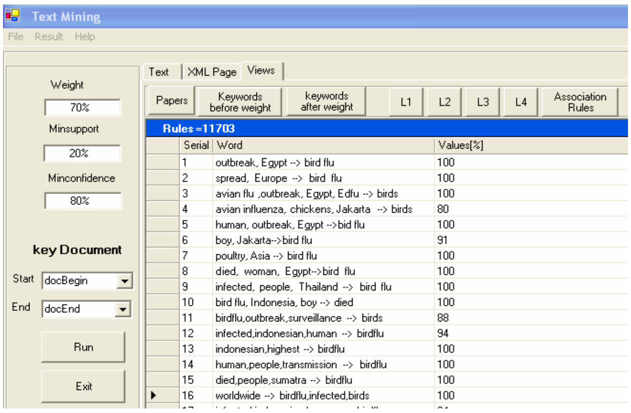


Fig. 3 Disease outbreak relationships

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”, “bird” 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. Fig. 4 shows a snapshot of the resultant association rules extracted by the EART system (using a weight 70%, support 20%, and confidence threshold 80%), where the number presented at the end of each rule is the rule’s confidence.



Weight 70%
Minsupport 20%
Minconfidence 80%
Start docBegin
End docEnd

Rules =11703

Serial	Word	Values[%]
1	outbreak, Egypt --> bird flu	100
2	spread, Europe --> bird flu	100
3	avian flu ,outbreak, Egypt, Edfu --> birds	100
4	avian influenza, chickens, Jakarta --> birds	80
5	human, outbreak, Egypt -->bid flu	100
6	boy, Jakarta-->bird flu	91
7	poultry, Asia --> bird flu	100
8	died, woman, Egypt-->bird flu	100
9	infected, people, Thailand --> bird flu	100
10	bird flu, Indonesia, boy --> died	100
11	birdflu,outbreak,surveillance --> birds	88
12	infected,indonesian,human --> birdflu	94
13	indonesian,highest --> birdflu	100
14	human,people,transmission --> birdflu	100
15	died,people,sumatra --> birdflu	100
16	worldwide --> birdflu,infected,birds	100

Fig. 4 Resultant association rules

1. The Argumentation of the Thresholds Chosen

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 keywordsets and selecting the final rules. 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 avian flu) that cannot appear if we chose high support value and chose high threshold confidence value 80% to extract the more interesting rules.

2. Interpretation of the Extracted Association 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 bird 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. Sample of abstractions and their corresponding extracted association rules are shown below. The following rules represent the relation between the disease and its spreading location:

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

bird flu, outbreak --> Egypt	100%
outbreak, Egypt --> bird flu	100%
spread, Thailand --> bird flu	100%

In addition, the EART extracts more informative association rules than the above rules where it extracts more relationships between features such as the disease, its spreading location and victim:

- **<disease> <location> --> < victim >**

avian flu ,outbreak, Egypt, Edfu --> birds	100 %
bird 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, Egypt -->bid flu	100 %
woman, Egypt--> bird flu	100%
boy, Jakarta-->bird flu	91%
China, farmer --> bird flu	100%
Indonesia chickens --> bird flu	86%
poultry, Asia --> bird 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 >**

bird flu, Indonesia, boy --> died	100 %
bird flu, Indonesia, birds --> died	100 %

- **<victim status> <victim> <location> --> < disease >**

died, woman, Egypt-->bird flu	100%
died, boy, Jakarta --> bird flu	91%
died, China, farmer --> bird flu	100 %
died, poultry, Asia --> bird flu	91 %
infected, people, Thailand --> bird flu	100%
infected, poultry, Surveillance --> bird 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.

C. Evaluation of the EART system

We design another system for extracting association rules from text by using the Apriori algorithm. This system corresponds to the EART system in the following processes:

- Transformation of documents into XML format
- Filtration and stemming of the transformed documents

After the above processes, this system uses the Apriori algorithm [12] for generating association rules. It called Apriori-based system.

To assay the performance of the EART system we compare it with the Apriori-based system throughout the execution time and evaluating extracted association rules. The experiments are applied on both systems at the same corpus and the same threshold values support and confidence. Relative to the EART system, at threshold weight 70%. Fig. 5 shows the execution time comparison between the Apriori-based system and EART system for the 100 documents collection. This graph is obtained for a minimum confidence of 80 %. It can be seen that our EART system based on the GARW algorithm always outperforms the Apriori-based system for all values of minimum support. Where the GARW algorithm reduces the execution time in comparable to the Apriori algorithm because it does not make multiple scanning on the original documents like Apriori but it scan only the generated XML file which contains all the keywords that satisfy the threshold weight value and their frequencies in each document.

It can be observed that the performance of both algorithms largely depends on the number of frequent keywordsets. For lower values of minimum support, it is expected to have many frequent keywordsets 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 keywordsets created in the Apriori-based system caused the large gap between this system and the EART system at lower values of minimum support. The reason of this is that the Apriori-based system generates all frequent keywordsets from all keywords in the documents that are important and unimportant. This leads to extract interesting and uninteresting rules.

In contrast, the EART system based on the GARW algorithm generates all frequent keywordsets from only important keywords in the documents based on the weighting scheme. Where, the weighting scheme plays an important role for selecting important keywords for using in association rules generation. This leads to extract the more interesting rules at short time.

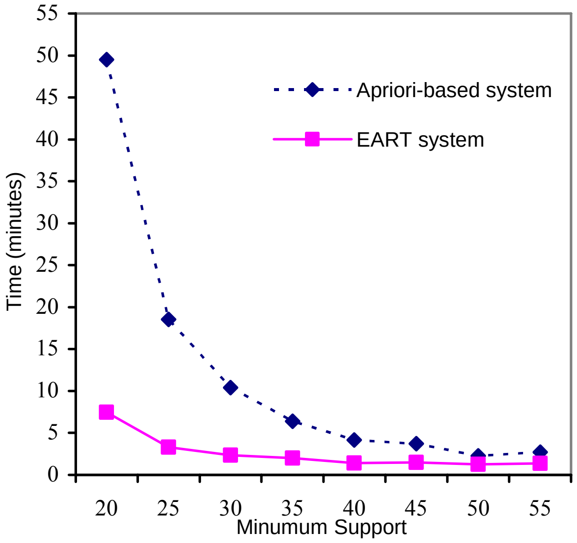


Fig. 5 EART system Vs Apriori-based system

Finally, we can notice that the difference of the execution time between the two systems is in minutes because we applied them on a small corpus. However, this difference will increase to become in hours especially in the Apriori-based system when we apply them on large corpus of documents.

IV. RELATED WORK

Earlier works on mining association rules from text have explored the use of manually assigned keywords. Where, they used keywords as features for generation of association rules [14, 15]. The drawbacks of approaches that use manually assigned keywords are that: (1) it is time consuming to manually assign the keywords (2) the keywords are fixed (i.e., they do not change over time or vary based on a particular user) (3) if the keywords are manually assigned , they are subject to discrepancy (4) the textual resources are constrained to only those that have keywords. Therefore, in our work we considered automated indexing (to overcome the above drawbacks) of the textual document base in order to allow the use of association extraction techniques on a large scale.

In [11] the authors presented two examples of text mining tasks, association extraction and prototypical document extraction, along with several related NLP techniques. In the case of association extraction task, they had extracted association rules from a collection of indexed documents. A way of finding information in a collection of indexed documents by automatically retrieving relevant associations between keywords was presented. In addition, there are several researchers [4, 5, 8] applied existing data mining techniques to discover episode rules from text. Where Episode rule mining is used for language analysis because it preserve the sequential structure of terms in a text document. However, in our work we focus on extraction of association rules that get the relations of the existing of the keywords in text ignoring the order in which these keywords occur. Mining association rules in temporal document collection and performing the various steps in the temporal text mining process are described in [10].

The authors in [18] presented a text mining technique that discovers association rules from documents for a particular user. It derives a user's background knowledge from his background documents, and exploits such knowledge in the form of association rules. In addition, TF-IDF is applied to select significant noun phrases from each target document. In [20] the authors evaluate the effectiveness of the weighting schemes for keyword extraction for gene clustering. The result produced from TF-IDF weighted keywords outperformed those produced from normalized z-score weighted keywords. This result is corresponding of our point of view of the importance and the effectiveness of using TF-IDF for weighing keywords in documents collection.

V. CONCLUSION AND FUTURE WORK

This paper has presented a new 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 (GARW). We compared the performance of our system that based on the GARW algorithm with a system that use Apriori algorithm. We noticed that the GARW algorithm reduces the execution time in comparable to the Apriori algorithm. Therefore, our system performed well against the one that we compared. In addition, the EART 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 (pharmaceuticals), their effectiveness and side effects. Moreover, we intend to visualize the extracted association rules in graphical representation in two or three-dimension association networks.

REFERENCES

[1] B. Lent, R. Agrawal, and R. Srikant, "Discovering trends in text Databases," *KDD'97,* 1997, pp.227-230.

[2] C. Manning and H Schütze, Foundations of statistical natural language processing (MIT Press, Cambridge, MA, 1999).

[3] G. W. Paynter, I. H. Witten, S. J. Cunningham, and G. Buchanan, "Scalable browsing for large collections: a case study," *5th Conf. digital Libraries,* Texas, 2000, 215-218.

[4] H. Ahonen, O. Heinonen, M. klemettinen, and A. Inkeri Verkamo, "Mining in the phrasal frontier," in *Proc. PKDD'97.1st European Symposium on Principle of data Mining and Knowledge Discovery, Norway, June,* Trondheim, 1997.

[5] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Inkeri Verkamo, "Applying data mining technique for descriptive phrase extraction in digital document

collections," in *Proc. of IEEE Forum on Research and technology Advances in Digital Libraries,* Santa Barbra CA*,* 1998.

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

[7] H. Mahgoub,"Mining association rules from unstructured documents" in *Proc. 3rd Int. Conf. on Knowledge Mining, ICKM,* Prague, Czech Republic, Aug. 25-27, 2006, pp. 167-172.

[8] H. Mannila, H. Toivonen and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data Mining and Knowledge Discovery,* 1(3), 1997b, pp. 259-289.

[9] J. Paralic and P. Bednar, "*Text mining for documents annotation and ontology support* (*A book chapter in: "intelligent systems at service of Mankind," ISBN 3-935798-25-3, Ubooks,* Germany, 2003*).*

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

[11] M. Rajman and R. Besancon, "Text mining: natural language techniques and text mining applications", in *Proc. 7th working conf. on database semantics (DS-7), Chapan &Hall IFIP Proc. Series.* Leysin, Switzerland Oct. 1997, 7-10.

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

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

[14] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (KDT)", in *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, 1995.*

[15] R. Feldman and H. Hirsh, "Mining associations in text in the presence of background knowledge," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Portland, USA*,* 1996.

[16] R. Feldman and M. Fresko, Y. Kinar, Y Lindell, O. Liphstat, M. Rajman, Y. Schler, O. Zamir, "Text mining at the term level," in *Proc. 2nd European symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Vol. 1510,* Nantes *pp 65-73.*

[17] R. Feldman and M. Fresko, Y. Kinar, Y Lindell, O. Liphstat, M. Rajman, Y. Schler, O. Zamir, "Knowledge management: a text mining approach," in *Proc. of th 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98),* Basel, Switzerland, *29-30 Oct. 1998.*

[18] X. Chen and Y. Wu, "Personalized knowledge discovery: mining novel association rules from text" Available: www.siam.org/meetings/sdm06/proceedings/067chenx.pdf

[19] Y. Kodratoff, "Knowledge discovery in texts: a definition, and applications," in *Proc. of th 2nd Int., symposium, ISMS'99, Vol. 1609 of LNAI,* Warsaw, Pol*. Springer,* Berlin Heidelberg New York*, pp 16-29.*

[20] Y. Liu, S. Navathe, A. Pivoshenko, A. Dasigi, R. Dingledine, B. Ciliax, "Text analysis of Medline for discovering functional relationships among genes: evaluation of keyword extraction weighting schemes," *Int. J. Data Mining and Bioinformatics, Vol. 1*, No 1, 2006.