Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.394 – 400

RESEARCH ARTICLE

EXTRACT ASSOCIATION RULES TO MINIMIZE THE EFFECTS OF DENGUE BY USING A TEXT MINING TECHNIQUE

Atif Amin*[1], Ramzan Talib[2], Salman Raza[3], Saima Javed[4]

[*1,2,3] College of Computer Science & Information Studies,
Government College University Faisalabad, Pakistan

[4] National College of Business Administration and Economics, Lahore, Pakistan

[1] atisalar@gmail.com

ABSTRACT:

Nowadays, dengue (vector-borne tropical viral diseases) has been become the greatest scourge of humankind and there consequence has more impact than any other pathogen in shaping the human genome. Generally in Pakistan specifically in Punjab (Pakistan) Dengue is emerging as one of the major public-health problem. Federal & Provincial Health Governments are taking all possible steps on "War-Footing" to recover such type of diseases. They are determines in making special possibilities to face such type of problems in advance. From WWW, digital libraries, World Health Organization (WHO) and other news sources it is estimated that about 2.5 billion people, or 40 percent of the world's population, live in areas where there is a risk of dengue transmission because Dengue flourishes in urban poor areas, suburbs and the countryside but also affects more affluent neighborhoods in tropical and subtropical countries. As of November 2011, it has killed over 300 people in the last several months and over 14,000 are infected by this mosquito-borne disease. Majority of the people infected are from the Lahore area in Punjab, Pakistan. As a matter of fact, if a virus attacks on somewhere, what will be its next target in geographically aspects, because dengue virus will be spread from one place to other from contaminated water and mosquito. Therefore by using above said data sources, performing some preprocessing techniques such as transformation, filtration, stemming and indexing of the documents and then applying data mining techniques our system will not only helps to identify geographical spreading patterns of the viruses but it also helps to suggest proactively next geographical location where virus has most probability to attach so that government can take remedy measures.

Keywords—Text mining, data mining, association rule mining, Dengue

I. INTRODUCTION

Over the past decades, several infectious diseases have increased in incidence and expanded into new geographic areas. There are multiple factors that contribute to the spread of disease, including increasing urban population density, more international travel, and widespread international import/export of goods. We will consider a disease which is the most spreading disease over the world as well as in Pakistan that is Dengue. Aedes aegypti (Dengue) is very comfortable laying its eggs in standing water inside or outside people's homes, allowing it to thrive in urban environments. Poor urban areas, where people store water inside their houses or have no screens on windows or doors, are perfect breeding grounds for aeg0ypti.

Requirement is that there should be some rules and suggestions by following them we can save our environment from such type of viruses and viral diseases. ("In today's technology-driven environment, it is critical for hospitals to have a path to the high quality products and services they need at a lower cost," said Greg Knapp, vice president, Support Services and Client Relations, Novation). There is a computer technique (i.e. Text Mining) that will be helpful to find out the proper solutions to cover areas from such type of viruses before their attacks.

Different types of sources are there from where concern data can be fetched like newspapers, BBC, CNN, Medical News, Yahoo News, and World Health Organization web reports etc. In these sources some news are geographical news that consist the content about spreading of the virus in many countries. Some news consist the content of treatment that should be used against the virus and some about medicine discoveries research of related field. This shows that different type of data contain the different knowledge. So, the challenge is that how to share the knowledge among these different topics and find the related things that will be beneficial.

By Using Text Mining technique we will describe a system that extracts the concern information automatically from WebPages news documents that contain the outbreak news of Dengue. The System will make association rules from extracted data that fully dependent on word features. For this system have to evaluate the data in different aspects to find many type of relationships between features such as name disease, the location where the news outbreak, the type of disease, condition of disease and the status of disease. System will ignore the order in which data occur. It will make an index in which related words will be added, those words grammatically close to each other like "disease" and "diseases". Then make some association rules to define the relationship between features in the document collection.

II. ARCHITECTURAL MODEL OF TEXT MINING SYSTEM

The purposed system extracts the association rules from text that is shown in fig.1. The association rules automatically discovers from textual data. The system uses the XML technology with Information Retrieval scheme that is Data Mining Techniques to select the discriminative features that will be helpful for extracting association rules. The system focuses on the words and their statistical distribution but fully ignore the arrangement of words.

The system divided into three distinct phases:

1. Processing Phase of Text
2. Mine the Association Rule
3. Visualization

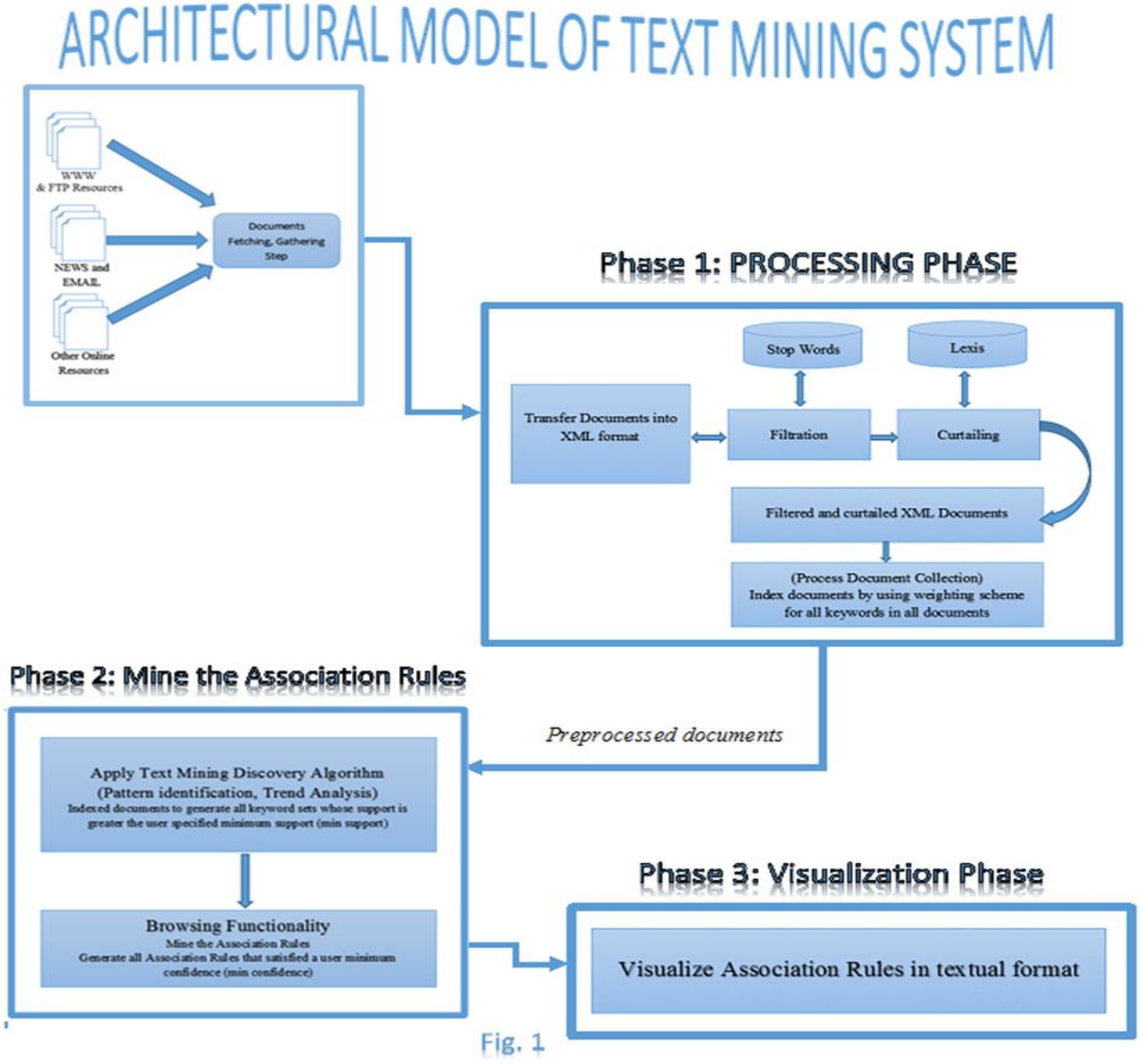

Fig. 1

1. PROCESSING PHASE OF TEXT

The main purpose of this phase is to provide the concern data to make the better performance of following phase in which association rules will be mined. In this phase firstly text gathers from different sources for transformation. In transformation obtain the data in XML format. After that, eliminate the irrelevant and unimportant words (filtration) like grammatical words (e.g. prepositions, conjunctions, articles and determiners, etc.). The resulting documents are processed to provide basic information about the content of each document. Processing phase also has some sub steps (transformation, filtration, stemming and indexing of the documents).

A1. Renovation (Transformation)

Our system will accepts a different number of documents formats (rtf, txt, doc, etc.) and convert these files into XML format acquiescent for further processing. System save the WebPages news and text documents and our text mining system transformed the documents into XML format.

A2. Categorization

In this step, the documents are filtered by removing the unimportant words from documents content. Therefore, all the unimportant words discarded or ignored by the system. (e.g. articles, pronouns, common adverbs, determiners, prepositions and conjunctions and all non-informative verbs). More important and highly relevant words are pointed out as single word. We create a list of Unimportant words that called stop words, system checks the documents content and eliminate all the unimportant words that are listed in stop words and in addition, the system replace the special characters, parentheses, commas, etc., with the spaces among words in the converted document. After completion of categorization process the system does the word stemming, in this process system removes the word's prefixes and suffixes (like a word infection and infections, system unifying them into infection)

A3. Indexing Process

In this process, the system index the filtered and stemmed XML documents by using weighting scheme. A textual data can be indexed manually or automatically on the basis of knowledge discovery process. As a manual process for indexing is a time-consuming process [14, 15], it is not considered realistic as could systematically be performed in general case. Automated indexing by a system of a textual document considered in order to allow the use of association extraction technique on 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].

- Weight Constraints

In information retrieval process the document should assign notation for relevance with respect to document. We assign for each keyword its score as weight value based on maximal TF- IDF (maximal with respect to all the documents in the collection). Our main purpose is to identify and filter the keyword that will be helpful to make

association rules. 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 [16].

2. MINE THE ASSOCIATION RULE

In this step a way is identify to find the information from collection of indexed documents that is already processed by automatically extracting association rules from them. To identify the results Association rules have already been used in Text Mining [7, 10, 11, and 15]. Here we describe the association rules in the context of Text Mining. In this phase an algorithm is used to find out the related words that are frequently used and to generate the confidence on these words that will be helpful to make association rules. We design an algorithm for making association rules based on weighting scheme. Algorithm processed the XML files that already filtered in processing phase and make association rules. [16]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	A keyword set having k- keywordsets
kL	Set of large k- keywordsets (that satisfy minimum support)
kC	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 1 L.

4. In 2 ≥k, the candidate keywords k C of size k are generated from large frequent (k-1)-keywordsets, 1 −kL that is generated in the last step.

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

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

7. Large frequent k-keyword sets kL , 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.

3. VISUALIZATION

In this phase visualize the association rules to make sure that they will be helpful to make the decisions. The extracted association rules can be reviewed in textual format or tables or in graphical format. A system is also designed to visualize the extracted association rules in textual format or tables. 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 [16].

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] H. Mahgoub, D. Rösner, and Torkey "A Text Mining Technique Using Association Rules Extraction" in International Journal of Computational Intelligence www.waset.org Winter 2008

