Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (2) 221-226 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Mining Emerging News from Text Data

S. A. DAR, M. MUZAMMAL[++]*, H. ZAHEER, I. A. KOREJO**

Department of Software Engineering, Bahria University, Islamabad

Received 5th June 2015 and Revised 29th March 2016

Abstract: we live in the information age. There is so much information emerging over the internet that it is next to impossible to be able to go through all of it. This work is focused on extracting "interesting" information from the web. As a first step, we assume that newspapers report the most interesting information and thus propose a framework that is able to extract interesting information from the internet using the news feed from news websites. We collect RSS feed from a set of user-specified sources and thus obtain the title of the news from the RSS feed. Next, we remove the insignificant words from the news title and a tokenization procedure transforms the keywords into tokens. These tokens are combined to form sets of items. An itemset mining algorithm is implemented to extract most interesting patterns and a de-tokenization procedure is used to extract the most interesting news.

Keywords: Frequent Pattren Mining ; Emerging News; Sequential Pattern Mining

1. INTRODUCTION

With the increase in technological advances, more and more data is available in digital form. Nevertheless, most of this data is available in unstructured textual form thus making it essential developing better techniques that will enable extraction of interesting and useful information from the bulk textual data (Kanellis, 2006).

The process through which extraction of this information is executed is referred to as text mining. It is, however, imperative noting that text mining should not be confused with data mining since the two are distinct disciplines. The process involves various stages such as text pre-processing, text clean-up and post-processing. It is on this premise that text mining and text analytics has become an essential aspect of research (Srivastava, 2009).

In this paper, we will discuss a methodology on how to extract Emerging news from RSS online news feeds. Different existing text mining procedures and algorithms will be the primary areas of concern in the research.

From the definition, the process of text mining involves a system that analyses large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract useful or valuable information.

In this case, text mining process will be of great importance in retrieving emerging news from different online sources. Use of information retrieval will be of great significance in this process (Berry, 2010). This process will have an unlimited number of social and economic benefits. We did not sort out the information from different sources as there are already placed at a click. For instance, the mining of emerging news will have impacts on social and economic fields.

The society can learn emerging trends in the economy or the social world thus taking advantage for better lives. Through text mining and analytics, I will be able to extract new knowledge and hidden insights from the large online data set that may be of paramount importance to the society (Damerau, *et al* 2005).

2. BACKGROUND

The background of this work will be presented in two subsections. The first one will present the association rule clustering while the latter will be concerned with frequent pattern mining.

A. Association Rule Mining

Association rule in text mining has been in use for a long time (Hidber, 1999). This concept helps in uncovering the relationships among seemingly unrelated data in a relational database or another information repository. Identification of existence of certain relationships in a database can be very instrumental in decision making.

The rule is based on the assumption that there is a relationship between items in the database or repository (Tagarelli, 2011). There are instances whereby if an antecedent X happens, and then a consequent Y is also likely to happen.

[++]Corresponding author: e-mail: muzammal@bui.edu.pk
*Department of Computer Science, Bahria University, Islamabad,
**Department of Computer Science, Sindh University, Jamshoro , e-mail: imtiaz@usindh.edu.pk

In Association rule discovery, finding frequent item sets is computationally the most expensive step. Once the frequent itemsets have been found, implications of the form X → B are quite straight forward.

B. Frequent pattern mining

Frequent pattern mining problem was introduced by (Agrawal, 1994)**.** Many procedures have been proposed to solve the problem. Generally speaking, frequent pattern mining algorithms can be classified as apriori based candidate generation approaches (Agrawal, 1994) and pattern-growth based non-candidate generation approaches (Chen, *et al* 2007)**.**

Apriori based candidate generation approaches

The apriori algorithm was first proposed by (Agrawal, 1994) for association rule mining. These algorithm use a very useful property called the Apriori property, which is stated as follows:

For any item set of length k to be frequent, All of its (k-1)-sub item sets need to be frequent.

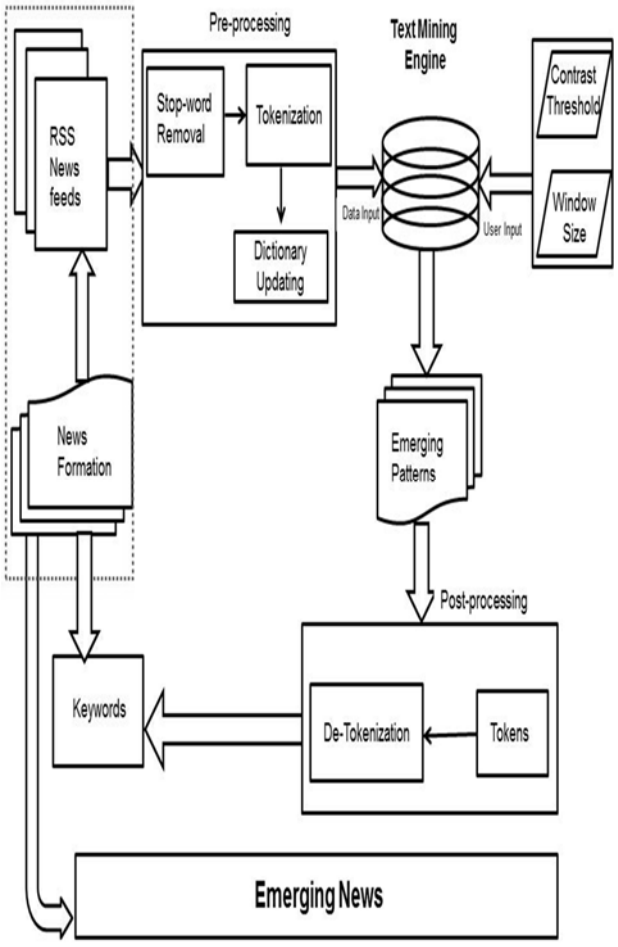

Fig. 1: Overview of the Framework

This means that if a *(k-1)*-subitemsets of a k-itemset is not frequent; the k-itemset can't be frequent. Thus, the use of the apriori property significantly reduces the search space and helps mining the frequent patterns efficiently.

C. Contrast Set Mining

The contrast set mining problem is defined as follows: Given a collection of *(item, value)* pairs where *item* ∈ *X* and *value* is the support count, the idea is to find such *(item, value)* pairs across two sets G and G' s.t. the item *i* in set G has a support count significantly different (higher) from the support count in set G.

Contrast set Mining is the process of extracting the difference between two sets of items and then reporting the items which are occurring more frequently in the second set. In this work, we define a time window *i* and mine the contrast sets across windows *(i-1, i)* to detect the emerging news.

3. OUR APPROACH

We design a web portal that reads news from user-defined news sources. This is done by way of implementing a set of web services (RSS feed) to extract latest news. A mining algorithm is implemented that processes the incoming text data from the news feed, and reports the Emerging news which is of "interest" to the user. For the purpose, techniques from data mining are used which include frequent pattern mining, emerging pattern mining, and contrast set mining.

RSS Feed

RSS (Rich Site Summary; originally RDF Site Summary; uses a family of standard web feed formats to publish frequently updated information, e.g., blog entries, news headlines, audio, video. An RSS document called "feed", "web feed" (Chen, 2008), or "channel" includes full or summarized text, and metadata, like publishing date and author's name.

Initially we obtain RSS News Feeds from a set of user-specified sources. News is extracted with following attributes: Title, Description, url, Category, Publishing Date, Image; and is stored in the news repository.

Pre-processing

Pre-processing is needed in order to prepare the input data for the Mining Engine. Initially Stop-words are removed from the text data in order to keep only meaning words for the mining purpose. And then Tokenization is performed, Tokenization is the process of converting keywords (obtained after stop-word removal) in to numeric format.After then A dictionary data structure is used in order to have a fast retrieval/updating.

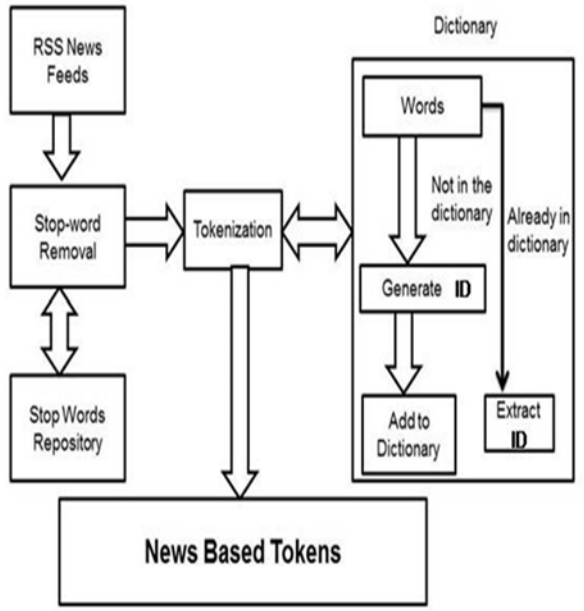

Fig. 2: Pre-Processing

A. Stop-word Removal

A stop-word is a word that is of no significant importance and should be filtered during the text pre-processing. For example, words like "is", "we", "you", "are", "me" etc. appear frequently in any text document. These words carry little to no semantic weight and thus are unlikely to help in obtaining some useful information in the text mining task. Eliminating such words saves sizeable space for document processing.

B. Tokenization

In lexical analysis, tokenization is the process of breaking stream of text into words, phrases, symbols, or other meaningful elements called tokens. The aim of tokenization is to process the words such that the words are replaced by Numeric Tokens. We maintain a dictionary of words and assign an ID to every word added to the dictionary. Thus a string of words is transformed to a string of numeric tokens. After stop-word removal, the actual words of the news feed are tokenized.

C. Dictionary updating

The actual words are checked in the dictionary and if the word is already present in the dictionary, then we extract the ID of that word, and if not then generate a new ID for the word and add that word to the dictionary. For example, a segment of the dictionary looks like as shown in **(Fig-3)**.

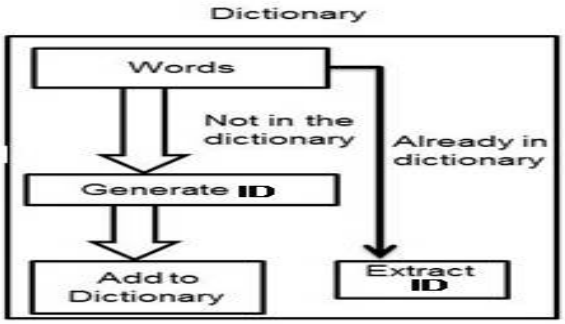

Fig. 3: Dictionary Updating

Text Mining Engine

Text Mining Engine performs some tasks like Frequent patterns, Contrast set mining and Frequent-Term-Based clustering.

Frequent patterns are itemsets that appear in a dataset with frequency no less than a user-specified minimum support threshold.

Contrast set mining is closely related to association rule mining, and utilizes some of the terminology and notation of association rule mining.

Contrast set mining process in done by Frequent Pattern set of window *i* (current) and frequent pattern set of window *i*-1 (previous). Contrast set mining computes that whether the patterns which are frequent in the current window were frequent or otherwise in the previous window; and thus finds Emerging patterns using the frequency count in current and previous window.

For frequent patterns that is more frequent in the current window w_i and were not present in the previous window, the count in previous window w_{i-1} is taken as zero. And a Time window is a user-specified time duration after which a Frequent Pattern Mining algorithm is triggered. For example, if the user specifies time window w=3, then every 3 hours Frequent Patterns Mining algorithm is triggered for the data obtained in the previous time window.

In Frequent-Term-Based clustering, frequent patterns are clustered based on the similarities is Frequent patterns. Thus the longest frequent pattern with highest support is considered as a candidate for Emerging news.

The output of the engine is clusters of tokens which are passed on to the post-processing phase for news formation.

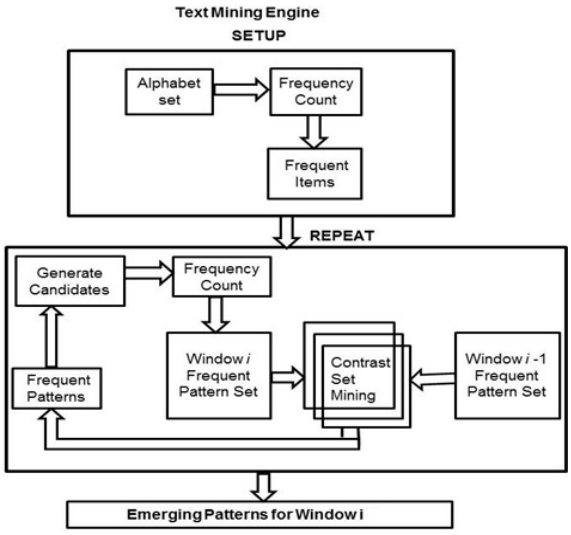

Fig. 4: Text Mining Engine

Post-Processing

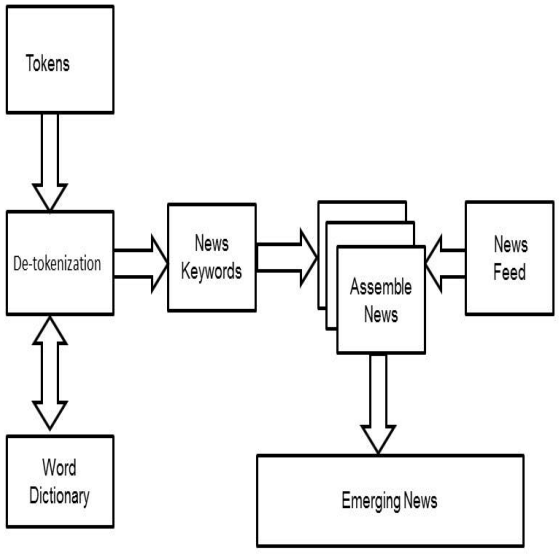

Fig. 5: Post Processing

In News formation keywords are used to establish news which is similar in the news feed (Kim, 2008). This is done by Frequent Pattern based clustering. Assemble news is a matching scheme is used to search similar news based on the extracted keywords.

Emerging News

Emerging News is the most frequent news for the current time window and is the output of the proposed framework.

4. ALGORITHMS

Now we discuss the algorithms for performing different tasks. First of all, gathering the RSS News feeds from different sources, then the Pre-processing on the data for tokenization, after that the Text mining engine processes data and gets emerging news. In Post-processing, De-tokenized data to get the actual news feed items. And in the end News formation gathers and assembles all the news feeds and show the Emerging news in that time**.**

5. EVALUATION

The process of data input will also be of paramount importance in our approach. To subscribe to an online RSS news feed, we will need to have a news aggregator or a feed reader. By the help of a feed reader, it will be possible to subscribe to and view as many news feeds as possible thus making our experiment a worthy course. The news reader will enable automatic retrieval of news updates thus making timely delivery on any news update as soon as they are published. To make more effective, we will use a web-based feed readers that will be compatible with our browser to enhance effectiveness and efficiency in the extraction of emerging news (Xu, *et al* 2013)**.** It will also be imperative to note that this RSS feeder will give us an opportunity to get only the required news and only in a formatted code.

It is prudent noting that all news received by news reader is stored in a semi-structured set of the database. The RSS will help sort out in different sets for the purpose of Pre-processing. This is an experimental process that will involve the use of keywords to derive the ID of the news item. The sorting process of the news item will involve the use of the hottest news headline received by our browser. In this case, the process involves extracting news regarding the current state in Syria. To get the essential news, it is imperative to drop keywords such as ‘’the, on, has.’’ From the news, ‘’ the on-going Syrian conflict has displaced millions’’. After the removal, the remaining words are’’ Syrian, conflict, displaced, millions. This helps easier extraction of the news item from the text mining engine that determines the frequency of news item.

The text mining engine will also be a significant tool in the analysis and explains the parameter setting of our approach. After sentence splitting, the next step in the experiment will involve the tokenization process. This is the stage that will involve generation of hot news item intended in the mining process. It is after the generation of hot news that we set a frequency that the news emerges for example in a period of three hours. With a specific time of three hours, one can determine the algorithms that happen to a certain frequent pattern. It is on this premise that it will be possible for us to sort out the frequent item generated by the text mining engine. In essence, the approach used in the data mining process is extracting emerging news from RSS news feeder of XML database through a text mining engine.

Algorithm 1 : Preprocessing

1. **Input : w_i News feed**
2. **$w_{i-1,T}$ News Frequency**
3. **θ_E Emerging Threshold**
4. **Output: Emerging News n_E**
5. **Repeat**
6. **Load Window w_i**
7. **For each news *n* in w_i**
8. **Generate Tokens n_T**
9. **for all n_T Generate $w_{i,T}$**
10. **Find top-k Longest items in $w_{i,T}$**
11. **Compute θ_E using $w_{i,T}$ and $w_{i-1,T}$**
12. **Report n_E**

Algorithm 2 : Stop-word Removal

1. **Input: News.Title**
2. **Output: News.keywords**
3. **for all term *t* in News.Tilte**
4. **If t ∈ s-list (stop-word List) then**
5. **Continue else**
6. ***n*-words ← News.Title**
7. **end if**
8. **Return *n*-words**

Algorithm 3 : Tokenization

1. **Input : *n*-words**
2. **Output: *t*-words**
3. ***t*-words ← ""**
4. **for all terms *t* in *n*-words**
5. **id ← Lookup(t , *i*Dictionary)**
6. ***t*-words ← t-word +id**
7. **Return *t*-words**

Algorithm 4 : Emerging Patterns

1. **Input : L_i , L_{i-1}**
2. **Output: L_E**
3. **L_i ← SortDecending (L_i)**
4. **for all I ∈ L_i**
5. **if support (I) in L_i > support (I) in L_{i-1} then**
6. **L_I ← DiffSupport (I , L_i , L_{i-1})**
7. **Sort- L_I {Length , Support , Time}**
8. **L_E ← Topk(L_I)**
9. **end if**

Algorithm 5 : Frequent Pattern Mining

1. **Input : A database *D* and a support threshold θ**
2. **Output: All frequent patterns *X* with support at least θ**
3. ***j* ← 1**
4. **L_1 ← ComputeFrequent – I- Patterns (D)**
5. **while $L_j \neq \theta$ do**
6. **C_{j+1} ← Join L_j with itself**
7. **Perform apriori pruning on C_{j+1}**
8. **for all X ∈ C_{j+1} do**
9. **Compute Support (X , D)**
10. **end for**
11. **L_{j+1} ← all frequent patterns X ∈ C_{j+1} {s.t. Support (X, D) ≥ θ }**
12. ***j*← *j*+1**
13. **end while**
14. **Stop and output L_1 U …U L_j**

Algorithm 6 : News Formation

1. **Input : L_E**
2. **Output: N_E**
3. **N ← Extract (L_E , News.Titles)**
4. **N ← sort + (N, Time, Length)**
5. **n_E ← Top-*k*(N)**
6. **Return n_E**

6. CONCLUSIONS

The objective of this work is to give an in-depth analysis of text mining process. As mentioned already, more and more data is available in digital form. Increased globalization has necessitated the urge for collecting emerging information from all over the world. With most of this information being in unstructured textual form, it is imperative that we design better techniques that will enable extraction of emerging and interesting information from the textual data. This call For extensive data pre-processing and post-processing that will enable extracting the emerging information for the interest of the user. It is nevertheless prudent noting that the unstructured text mining process is not an easy process and has a significant number of challenges. One of the issues is to come up with a unified framework for the mining task that is able to process unstructured data. We want to focus on this aspect in future.

REFERENCES:

Kanellis, P., (2006) Digital crime and forensic science in cyberspace.: IGI Global, NY, USA.

Srivastava, A. N., and M. Sahami, (2009) Text mining: Classification, clustering, and applications.: CRC Press, Netherland.

Berry, M. W., and J. Kogan, (2010) Text mining: applications and theory. NewYork: John Wiley \& Sons, NY, USA.

Damerau, S., N. I. Weiss, Z. Tong, and J. Fred, (2005) Text Mining, 2nd ed., Kormentan J., Wiley & Sons, NY, USA. .

Hidber, C., (2011) "Online Association Rule Mining," SIGMOD Rec., vol. 28, no. 8, 145—156, USA.

A. Tagarelli, XML Data Mining: Models, Methods, and Applications: IGI Global, NY, USA.

Agrawal, R., and R. Srikant, (1994) "Fast Algorithms for Mining Association Rules in Large Databases," in Proceedings of the 20th International Conference on Very Large Data Bases, 487-499, 1994, Chicago, USA.

Chen, H., J. Peng, S. Kuo, and H. Cheng, (2007) "An efficient incremental mining algorithm-QSD," Intelligent Data Analysis, vol. 11, no. 4, 62-78, Washington DC, USA.

Chen, J., (2008) Guardian UK-webfeeds. [Online]. http://www.theguardian.com/help/insideguardian/2008/oct/22/full-fat-rss-feed-upgrade, accessed: 15 Sep, 2015.

Kim, S. L., and J. Han, (2008) "News Keyword Extraction for Topic Tracking," in Fourth International Conference on Networked Computing and Advanced Information Management, 554-559, Beijing, China.

Xu, J., M. Yasinzai, and B. Lev, (2013) Emerging Trends in Large-Scale Computing, Electrical and Information Technology Springer Books Series, Springer, 2013, NY, USA.