Dynamic Graph based Method for Mining Text Data

WAEL AHMAD ALZOUBI
Applied Science Department
Ajloun University College
Balqa Applied University
JORDAN

Abstract: An improved graph based association rules mining (ARM) approach to extract association rules from text databases is proposed in this paper. The text document in the proposed technique is read only once to look for the terms whose occurrences are greater than some threshold value, these terms are stored in a file with their frequencies, then they are represented as nodes in a weighted directed graph where edges represent relations between these terms, the edges will denote the associations between terms while the edges' weights denote the strength or confidence of these rules. The proposed method is called **D**ynamic **G**raph based **R**ule **M**ining from **T**ext (DGRMT) because the graph is built level by level according the length of a sentence (number of frequent terms). Weighted subgraph mining is used to ensure the efficiency and throughput of the proposed technique; only the most frequent subgraphs are extracted. The proposed technique is validated and evaluated using real world textual data sets and compared with one of the best graph based rule mining technique, which is algorithm for Generating Association Rules based on Weighting scheme (GARW). The results determine that the proposed approach is better than GARW on almost all textual datasets.

Keywords: Graph, Association rules, Text mining, Text database, Frequent terms, Term weight.

Received: April 24, 2020. Revised: September 1, 2020. Accepted: September 11, 2020. Published: September 23, 2020.

1. Introduction

Information from text documents can be easily stored, managed and retrieved by using digital methods, at the same time, there is no need to take care of hard-copy documents. The importance of automated text analysis increased in several computer applications, as information retrieval, document summarization, text classification, and text pattern mining. [1]

Several efforts were done to develop algorithms for text processing. One of the earliest methods for text representations is Bag of Words (BOW) [2]. This strategy is regarded as unsuccessful technique due to absence of connections among words in the text file, BOW model represents a text as a bag of its words regardless to the grammar or word order in the text documents but keeps the number of times each word appears in the bag. Another text representation model based on the BOW model is the vector space model (VSM). The disadvantages of VSM may be summarized by: a) No term's order in text documents. b) Terms are considered independent from each other. c) Documents with similar definitions but different terms may not be associated together. Whereas the advantages of VSM are: a) VSM is simple as it depends on linear algebra. b) No binary term weights. c) Partial matching of terms and sentences is permitted. Another model for retrieving information from text documents is the standard Boolean model (SBM)[3], SBM is considered as the first model for extracting information from text files since it mainly depends on Boolean logic and set theory where the documents are considered as set of terms, information extraction is based on whether or not the text documents contain the required terms. Some of the advantages of SBM are: a) Very simple to implement. b) Obvious concepts. While the main disadvantages of SBM are: a) All terms have the same weights. b) Difficulty in query conversion into Boolean expression. c) No documents' ranking.

The relationships between words are of great importance as these relationships help in the discovery of their meanings in the text, and so letting the analysis of texts to be done [3]. A graph-based representation of text was recommended as a solution to solve the shortcomings of BOW approaches to handle these problems.

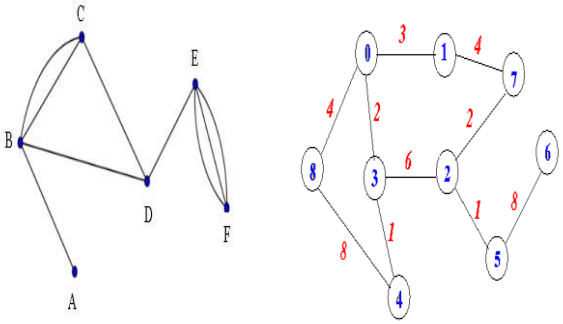
The main goal of this paper is to propose a dynamic method for building a directed association graph from single terms to n-long terms in order to mine strong association rules from text documents as displayed in section 5.

The rest of this paper is systematized as following: Section 2 talks about the graph theory. Some related works are briefly discussed in section 3. Section 4 talks about association rules mining. The proposed

graph based document rules mining technique (GDRM) is illustrated in section 5. The short experiments to prove the effectiveness of GDRM is displayed in section 6, conclusion and future works are displayed in section 7.

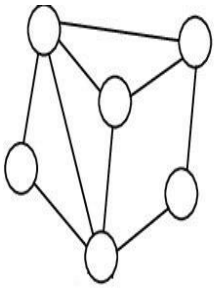
2. Graph Theory

Any graph G is defined as 2-tuple: $G = (V, E)$, where V is a finite non-empty set of vertices or nodes and E is a finite set of edges connecting each pair of vertices. A graph may be directed (digraph) which is a graph that has directed edges, or undirected, which is one in which edges have no direction. A graph in which many edges among the vertices are required is called *multi-graph,* where a graph in which all edges have a label that is positive integer is called *weighted graph*. Weighted graphs may be directed or undirected. Figure 1 demonstrates all these general types of graph.

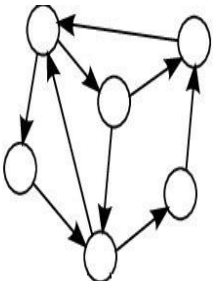


a) Multi-graph

b) Weighted-graph



c) Undirected graph



d) Digraph

Fig. 1: Main Types of Graphs

2.1 Text Documents As Graph Nodes

As mentioned earlier in the previous section, any graph consists of finite non-empty set of nodes or vertices and a finite set of edges to link these nodes together. In the graph representation of textual data, the nodes represent paragraphs, sentences, phrases, or words, where the edges of the graph capture various types of relationships between two or more nodes as semantic, syntactic relationships or co-occurrence network over the text. *The co-occurrence network*, is one of the most common techniques for text representation, in comparison to the BOW model, this model gives an important background to define relationships among words.

co-occurrence networks are the helpful connection of terms depending on their paired occurrence within a specified part of text document. Graph or networks are constructed by linking pairs of terms using some measures defining simultaneously existence of terms. For example, terms "Computer" and "Networks" may be said to "exist together" if they both found in a specific article. Another article may contain terms "Network" and "Security". Linking "Computer" to "Network" and "Network" to "Security" produces a co-occurrence system of these three terms.

Each text document $d \in D$ is represented as a graph $G_d = (V_d, E_d)$ where the nodes correspond to the terms t of the document and the edges represent co-occurrence relationships between terms. If G_d is directed then the actual flow of text is well-maintained, otherwise each edge represents co-occurrence of the connected terms regardless to their order. The weight of any edge reflects the number of co-occurrences of two terms in the document, term weights will be briefly discussed in the next sub section.

2.2 Term Weight

Assigning weights to the terms in a text document is a technique that occurs through the text indexing process to evaluate the value of each term. Term weighting is giving numerical values to terms that denote their rank in a document to improve retrieval efficiency.

Every term in the text document has a weight, the weight is the number of edges going inside the node in the graph of words. The text document is stored as a vector of weights in the direct and inverted index.

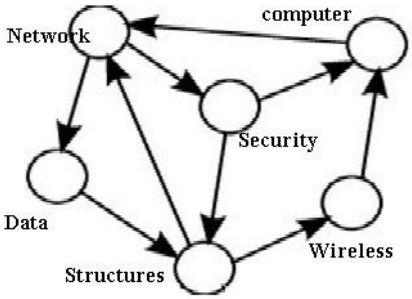


Fig. 2: An Example of Term Weights

The weight of each word in this example is computed as following, w (computer) = 2, w (Network) = 2, w (Security) = 1, w (Data) = 1, w (Structures) = 2, and w (Wireless) = 1.

Def1: Term frequency – inverse document frequency (TF-IDF): is one of the most popular term-weighting approaches used nowadays, it reflects the importance of a word in a document of textual database. It is normally used as a weighting factor in several fields as: information retrieval, text mining, and user modelling.

Def2: Let *t* denotes Term, *d* denotes document, textual database size *N*, term frequency *tf (t, d)*, document frequency *df(t)*, document length $|d|$, average document length *avg*, *s* is the slope parameter, then:

$$\text{TF-IDF (t, d)} = \left(\frac{1+\log(1+\log(tf(t,d)))}{1-b+b*\frac{|d|}{avg}}\right) * \log(\frac{N+1}{df(t)})$$

In the bag-of-word representation, term weight *(tw)* is usually defined as the term frequency or sometimes just the presence/absence of a term. In the graph-of-word representation, *tw* is the in-degree (number of edges going inside a node) of the vertex representing the term in the graph.

2.3 Graph-Based Text Representation

The vector space model (VSM) which depends on the bag-of-words approach is widely used model to represent text files, but VSM doesn't deal with the order of the terms in the document or about the borders between sentences or paragraphs. And so, it is highly required to develop a strong and scalable method to represent the information extracted from text documents and allow visualization and query of such information. A graph based text representation model is proposed to take care of the order, co-occurrence and frequency of the terms in a document. The proposed model is applied to discover implicit associations between two or more words (terms) in a large database of texts.

In the models of graph-based text representation, a text is represented as a graph where the set of vertices (nodes) denote the terms and the set of edges denote relationships between these frequent terms. One of the most former fields that use graph to represent text is Natural Language Processing (NLP) [4]. One of the main goals of graph-based text representation methods is to simplify the extraction of association rules from these documents.

2.4 Frequent Subgraph Mining (FSM)

Frequent Subgraph Mining (FSM) deals with databases of graphs. Because of the ease with which data can be represented as graph formats, there has been much interest in the mining of graph data. The objective of FSM is to extract all the frequent subgraphs in a given dataset, whose occurrence counts are above a specific threshold. The problem of FSM can be defined as following:

Given a graph dataset D = $\{G_0; G_1; \ldots; G_n\}$, support(g) denotes the number of graphs (in D) in which g is a subgraph. The problem of frequent subgraph mining is to find any subgraph g such that support(g) is greater than minSup where minSup is a minimum support threshold predefined by the user [5].

3. Related Works

[6] illustrates a comprehensive study and comparison of graph based text mining and the application domains that use these tools

Various approaches have been applied to deal with this problem. An Apriori-based algorithm used to discover all frequent (both connected and disconnected) substructures was proposed by [7, 8] developed FSG, a method using adjacent representation of graph and an edge-growing strategy to find all frequent connected subgraphs. In another work, [9] proposed gSpan which is the first algorithm that explores depth first search in frequent subgraph mining.

Many factors determine the efficiency of any text classification algorithm, the main common factors that must be taken into consideration are the time required to accomplish this task and the order of terms, some of these studies are [10, 11, 12, 13].

4. Association Rules Mining (ARM)

Association rules mining (ARM) approach was first introduced in [14], ARM is defined as the automatic discovery of pairs of element sets that tend to appear together in a general framework [15].

Def3: Let X be a set of keywords, such that X = $\{w_1, w_2, \ldots, w_n\}$ and a collection of indexed documents D = $\{d_1, d_2, \ldots, d_k\}$, where each document d_i is a finite set of keywords ($d_i \subseteq X$). A text document d*i* contains W_i if and only if $W_i \subseteq d_i$.

An association rule is an inference of the form $W_i \rightarrow W_j$ where W_i and $W_j \subset X$ and they are disjoint. There are two important basic measures for association rules, support(s) and confidence(c). the support of the rule $W_i \rightarrow W_j$ in documents' database is the percentage of documents that have W_i or W_j or both of them to the total number of documents in the database, formally, the support formula is given as:

$$\text{Support } (W_i \rightarrow W_j) = \frac{Support\ count\ of\ w_i\ and\ w_j}{Total\ number\ of\ documents}$$

Whereas the confidence of the rule $W_i \rightarrow W_j$ is computed by this formula:

$$\text{Confidence } (W_i \rightarrow W_j) = \frac{\text{Support } (Wi \rightarrow Wj)}{\text{Support } (Wi)}$$

The association rule-mining process consists of two steps:
1) looking for all keyword combinations (term sets) whose support is greater than the user specified minimum support. Such sets are called the frequent term sets.
2) Using the frequent term sets to extract the association rules that satisfy a user specified minimum confidence. This step is straightforward.

5. The Proposed Dynamic Graph based Rule Mining from Text (DGRMT)Technique

In this paper, an improved graph based method for Generating Association Rules from database of documents has been proposed, the proposed method is **D**ynamic **G**raph based **R**ule **M**ining from **T**ext (DGRMT). The DGRMT method scans the text file containing the generated frequent term sets only once. This file holds all the terms whose weight is greater than the threshold weight value.
Figure 3 displays the steps of the proposed DGRMT algorithm, where *N* denote the number of terms that satisfy the predefined threshold weight value, these terms are stored in a file with their frequencies in all documents. The data in the file will be in table form that contains *N* rows and 4 columns, these columns contain document id, frequent terms, the frequency of each term, and their value of TF-IDF.

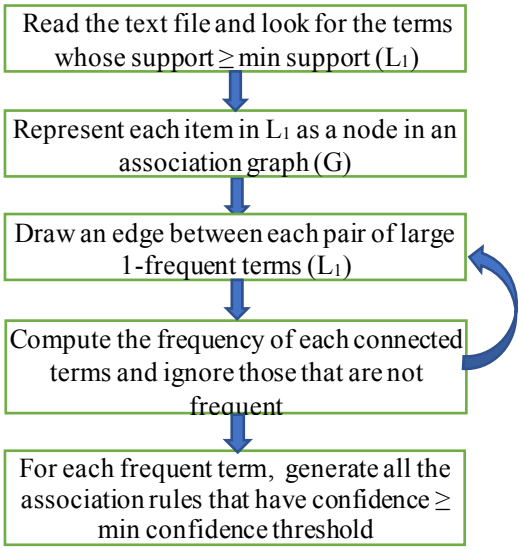


Fig. 3: the steps of the proposed DGRMT algorithm

Step 3 is repeated until no more edges can be added to the graph, i.e. all frequent terms and the relationships among them are found. DGRMT is dynamic with respect to the ability of building the document graph progressively starting from single frequent terms, i.e. those that have occurrences greater than some predefined threshold. This graph must be directed to reflect the order of the frequent terms. After that, those single terms are merged to make a sentence of length equals two. The predefined minimum support is reassigned and then any 2-term sentence with frequency (weight) less than the new value of minimum support is ignored while the others are used to add new edges to the constructed graph and this process continues until no more frequent n-term sentences are found, where n ia an integer whose value is greater than two.

6. Experiments

The main purposes of the experiments presented in this section is to test the proficiency of DGRMT in extracting strong association rules, and to assess its efficiency on several text analysis and text mining tasks.
The proposed DGRMT method is compared with one of the best graph based text mining algorithms, that is, the Association Rules based on Weighting

algorithm (GARW) [16]. Both GARW and DGRMT scans the documents only once but GARW concentrates only on the keyword sets that are stored in XML file, while DGRMT takes in consideration all words but abbreviations, the file for the proposed technique consists of the terms only together with their frequencies in each document.
the input to the proposed system is the minimum support threshold to extract the frequent terms and then the system requires the minimum confidence to extract only strong rules from the file containing the frequent terms, the output is the time required to get the desired rules.
The experiments have been carried out using a database of documents that contains 250 documents is 1120 KB in size and the total number of single words is about 55000. Each text document consists of 220 single words. After the purification process, the number of single words is minimized to 17417. The proposed DGRMT algorithm use the same platform as GARW to assure that the comparisons are reasonable. The experiments were performed on a Core i5, 3.8 GHz system running Windows 7 with 8 GB of RAM. The running time is reduced and the strongness of the extracted rules is increased using the proposed DGRMT in comparable to the GARW algorithm.

Execution Time (Minute)		Min Support (%)
GARW	DGRMT	
19	7	15
15	4	20
14	3	25
12	3	30
9	2	35
7	2	40
6	2	45
5	1	50

Table 1: Comparison between DGRMT and GARW

As shown in table 1, the time required to mine association rules from text documents is decreased dramatically using the proposed technique.

7. Conclusion and Future Work

A graph representation of text is an efficient paradigm that can successfully denote the relationships among the terms and the structure of the text file. Text reported in a graph representation is significant because it can be used in most text operations from lexical analysis of documents until query matching. This paper has presented an improved method for extracting strong association rules from text documents based on the weights of the terms, i.e. number of occurrences for each term. The proposed technique is called **D**ynamic **G**raph based **R**ule **M**ining from **T**ext (DGRMT). DGRMT is dynamic with respect to the ability of building the document graph progressively starting from single frequent terms, i.e. those that have occurrences greater than some predefined threshold. This graph must be directed to reflect the order of the frequent terms. After that, those single terms are merged to make a sentence of length equals two. The predefined minimum support is reassigned and then any 2-term sentence with frequency (weight) less than the new value of minimum support is ignored while the others are used to add new edges to the constructed graph and this process continues until no more frequent n-term sentences are found, where n ia an integer whose value is greater than two. We observed that the proposed DGRMT algorithm reduces the execution time in comparable to the **G**enerating **A**ssociation **R**ules based on **W**eighting scheme (GARW) algorithm.
For future work, we want to apply the proposed algorithm to find the most meaningful rules by taking in consideration the features of any frequent terms and to excerpt the more suitable association rules that have better meaning. Moreover, we plan to visualize the mined association rules in graphical representation in two-dimension association networks.

References

[1] S. S. Sonawane, Dr. P. A. Kulkarni. *Graph based Representation and Analysis of Text Document: A Survey of Techniques*. International Journal of Computer Applications (0975 8887) Volume 96 - No. 19, June 2014.
[2] Surabhi Lingwal, Bhumika Gupta. *A Text Mining Approach for Automatic Classification Of Web Pages*. Proc. of the Second Intl. Conf. on Advances in Electronics, Electrical and Computer Engineering - EEC 2013. ISBN: 978-981-07-6935-2 doi:10.3850/ 978-981-07-6935-2_52.
[3] Lashkari, A.H.; Mahdavi, F.; Ghomi, V. (2009). *A Boolean Model in Information Retrieval for Search Engines*. 2009 International Conference on Information Management and Engineering. 3 – 5 April 2009. Kuala Lumpur Malaysia.
[4] G. Salton, A. Wong, and C. S. Yang. *A vector space model for automatic indexing.*

Communications of the ACM, 18(11):613{620, 1975.
[5] Himani Raina, Omais Shafi. *Analysis Of Supervised Classification Algorithms*. International Journal Of Scientific & Technology Research Volume 4, Issue 09, September 2015.
[6] Ilkay Yelmen, Metin Zontul, Oguz Kaynar, Ferdi Sonmez, *A Novel Hybrid Approach for Sentiment Classification of Turkish Tweets for GSM Operators*, International Journal of Circuits, Systems and Signal Processing, pp. 637-645, Volume 12, 2018.
[7] Inokuchi, A., Washio, T., & Motoda, H. 2000. *An Apriori-based algorithm for mining frequent substructures from graph data*. In Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'00), pp 1 – 11.
[8] Kuramochi, M. & Karypis, G. 2001. *Frequent subgraph discovery*. In Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 313–320.
[9] Yan, X. & Han, J. 2002. *gSpan: graph-based substructure pattern mining*. Technical Report UIUCDCS-R-2002-2296, Department of Computer Science, University of Illinois at Urbana-Champaign.
[10] Srihari, S. 2011. *Principles of data mining*. University at Buffalo. The State University of New York.http://www.cedar.buffalo.edu/~srihari/CSE626/Lecture-Slides/Ch1-Part1-Introduction.pdf, 2011, pp 1 – 41.
[11] Wang, J. 2009. *Data warehousing and mining: concepts, methodologies, tools, and applications*. USA: Information Science Reference, pp. 303-335.
[12] Majeed, S. K. & Abbas, H. K. 2010. *An improved distributed association rule algorithm*. Eng.& Tech. Journal, Vol.28, No.18, 2010, pp. 5695 – 5710.
[13] S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan. *Text Classification Using Data Mining*. ICTM 2005.
[14] Agrawal, R., Imielinski T. & Swami A. 1993. *Mining Association Rules between Sets of Items in Large Databases*. Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, pp. 1 – 10.
[15] Ahmed Hamza Osman & Omar Mohammed Barukub. *Graph-Based Text representation and Matching: A Review of the State of the Art and Future Challenges*. IEEE Access. Volume 8, 2020. Pp 87562 -87583.
[16] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey. *A Text Mining Technique Using Association Rules Extraction*. World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:6, 2008.
[17] Guo-cheng, Niu and Zhen Hu. *Evaluation and Health Status Prediction Method of Beer Filling Production Line Based on Data Mining Technology*. International Journal of Circuits, Systems and Signal Processing, pp. 306-311, Volume 13, 2019.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)