Engineering World
Open Access Journal
ISSN: 2692-5079

# An Improved Graph based Rules Mining Technique from Text

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. To improve the accuracy of mining association rules from textual database, we represent the documents as graphs to allow for a much more expressive document representation than the traditional text representation approaches. Document sets are represented as graph sets to which the proposed graph mining algorithm is applied to extract frequent subgraphs, which are then further processed to produce feature vectors (one per document). 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. The results determine that the proposed approach is better than existing text mining algorithms on almost all textual datasets.

Keywords: Graph, Association rules, text mining.

## I 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 printed documents. The importance of automated text analysis increased in several computer applications, as information retrieval, document summarization, text classification, and text pattern mining. [5]

Several efforts were done to develop algorithms for text processing. One of the earliest methods for text representations is Bag of Words (BOW) [4]. This strategy is regarded as unsuccessful technique, as seen in other situations, and presents a variety of unplanned problems and weaknesses as a result to the absence of connections among words in the text file. Therefore BOW causes critical problems, from both semantic explanation and text processing viewpoints. The relationships between words are of great importance as these relationships help in the discovery of their meanings in the text, thus allowing the analysis of texts to be carried out [1]. A graph representation of text was recommended as a solution to solve the shortcomings of BOW approaches to handle these problems.

The rest of this paper is organized 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.

## II GRAPH THEORY

Any graph G is defined as 2-tuple: $G = (V, E)$, where $V$ is a finite 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.

ISSN: 2692-5079

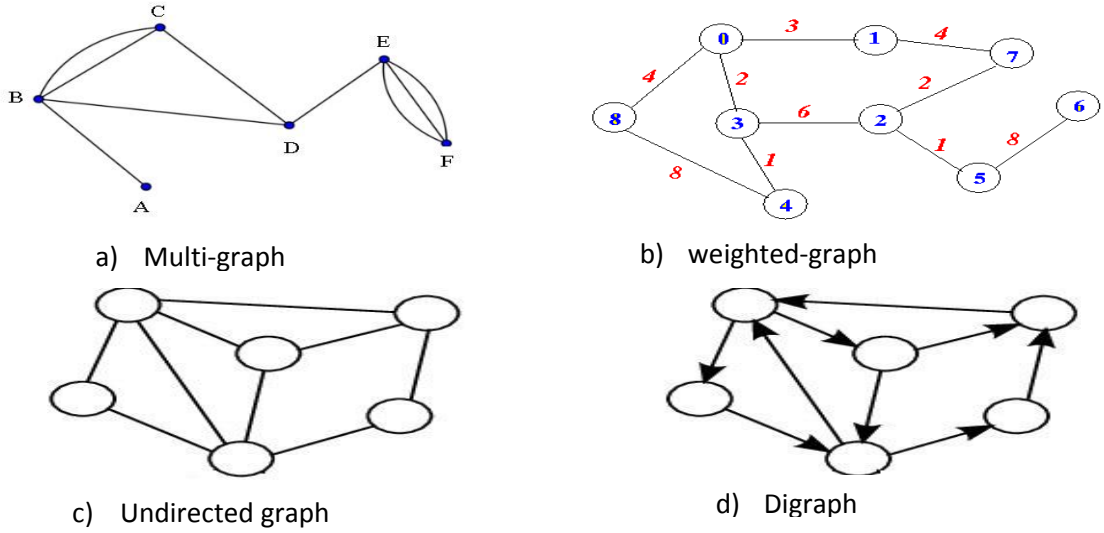



Figure 1: Main Types of Graphs

a. 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*, one of the most popular text representation forms has been implemented in several new systems. In comparison to the BOW model, this model provides an essential context to describe relationships among words. A text is basically represented as a graph.

*co-occurrence networks* are the helpful connection of terms based on their paired presence within a specified unit of text. Networks are generated by connecting pairs of terms using a set of criteria defining simultaneously existence of terms. For example, terms "Computer" and "Networks" may be said to "exist together" if they both appear in a particular 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.

b. Term Weight

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.

ISSN: 2692-5079

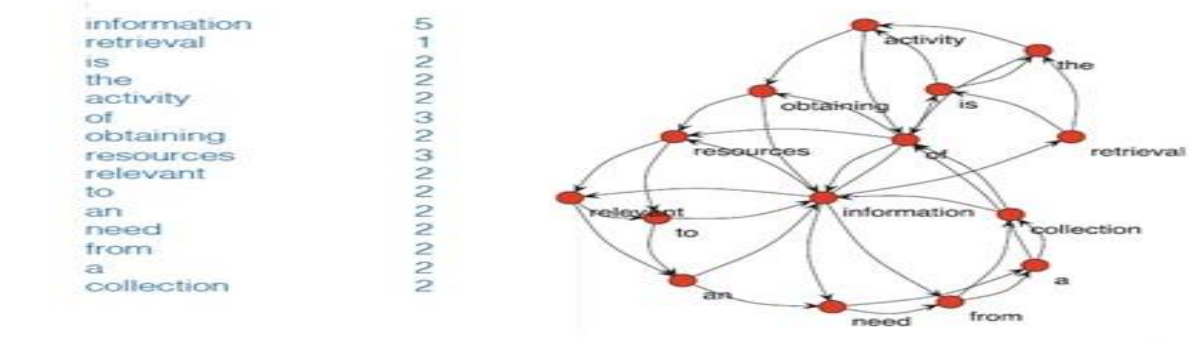



Figure 2: An Example of Term Weights

The weight of each word in this example is displayed in the left part of the figure, where w (activity) = 2, w (information) = 4, and so on.

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.

c. 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 graph text representation models, a text is represented as a graph containing a set of vertices (nodes) and a set of edges representing relationships between nodes. One of the most former fields that use graph to represent text is Natural Language Processing (NLP), it has focused on language understanding techniques such as part of speech tagging, rather than text mining tasks like text classification [6]. One of the main goals of graph-based text representation methods is to simplify the extraction of association rules from these documents.

There are six main types of graph based text representation based on their functionality, which include standard representation, simple representation, N-distance-representation, N-simple distance representation, absolute frequency representation, and relative frequency representation. A simple representation is adopted in this paper.

d. 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 = {$G_0$; $G_1$; ….;$G_n$}, 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

ISSN: 2692-5079

support(g) is greater than minSup where minSup is a minimum support threshold predefined by the user [3].

## III. RELATED WORKS

[14] 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].

## IV. ASSOCIATION RULES MINING (ARM)

Association rules mining (ARM) approach was first introduced in [3], ARM is defined as the automatic discovery of pairs of element sets that tend to appear together in a general framework [12].

Def3: Let X be a set of keywords, such that X = $\{w_1, w_2, ..., w_n\}$ 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:

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

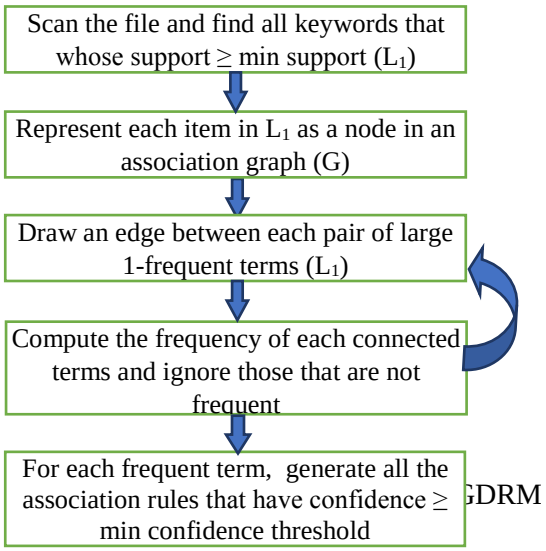
Whereas the confidence of the rule $W_i \rightarrow W_j$ is computed by this formula:

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

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.

## V. THE PROPOSED GRAPH BASED DOCUMENT RULE MINING (GDRM)TECHNIQUE

In this paper, an improved graph based method for Generating Association Rules from database of documents, the proposed method is **G**raph based **D**ocument **R**ule **M**ining (GDRM). The GDRM method scans the file containing the generated frequent term sets only once. This file holds all the terms that satisfy the threshold weight value and their frequencies in each document.
Figure 3 displays the steps of the proposed GDRM 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.




GDRM

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.

## VI. EXPERIMENTS AND CONCLUSIONS

ISSN: 2692-5079

The main purposes of the experiments presented in this section is to test the proficiency of GDRM in extracting strong association rules, and to assess its efficiency on several text analysis and text mining tasks. The proposed GDRM method is compared with one of the best graph based text mining algorithms, that is, the Association Rules based on Weighting algorithm (GARW) **[2].** Both GARW and GDRM scans the documents only once but GARW concentrates only on the keyword sets that are stored in XML file, while GDRM 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 document contained on average 220 single words. After the filtration process, the number of single words is minimized to 17417. The proposed GDRM 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 execution time is reduced and the strongness of the extracted rules is increased using the proposed GDRM in comparable to the GARW algorithm.

| Execution Time (min) | | Min Support % |
|---|---|---|
| GARW | GDRM | |
| 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 GDRM and GARW
As shown in figure 1, the time required to mine association rules from text documents is decreased dramatically using the proposed technique.

References

[1] Himani Raina, Omais Shafi. *Analysis Of Supervised Classification Algorithms*. International Journal Of Scientific & Technology Research Volume 4, Issue 09, September 2015.

[2] 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.

[3] 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.

[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] 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.

[6] 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

[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 Conferenceon 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 Hasan3. *Text Classification Using Data Mining*. ICTM 2005.

[14] 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.