A Review of Graph Based Algorithms in Social Media Data Analytics

Dr. N. V. Muthu Lakshmi[1] ,T. Lakshmi Praveena[2]
Assistant Professor[1], Research Scholar[2],
Dept of Computer Science[1], Dept of Computer Science[2]
SPMVV, Tirupathi[1], SPMVV, Tirupathi[2]

Abstract--Globalization of data increased with evaluation of internet and social media sites. As the number of users increases in social media sites, the data from social media also increases in terms of billions or trillions. For example twitter and facebook has millions of connections so sharing information on these sites produce millions of data. The data generated by social media may be structured, semi structured and unstructured which can be commonly termed as variety of data. As social media data/information plays a vital role in analysing and predicting important conclusions for various users according to their interest of analysis. To predict these, social media data has to be analysed efficiently to find analytical results. The analysis of social media data can be done more efficiently by integrating graph theory algorithms with other analytical techniques like statistical, mining and probability techniques. So the graph theory plays a significant role in analysing the social media data. The social media data can be represented as graphs. Graph representation is easy to solve complicated analytical problems. Properties and Information of social media can also represent as a big graph where big graph is a sequence of graphs which capture dynamic properties of data. Particularly in social media data analytics, graph theory is being used successfully by adopting the relevant theorems in various algorithms for example extracting frequent substructures, pattern mining from big data. The most popular algorithms which adopted graph theory are shortest path analysis, optimal path analysis, path existence analysis and vertex centrality. The algorithms used to analyse big graphs are Page Rank algorithm which calculates relative importance of web pages, Random walk with Restart algorithm to measure proximity of vectors in graph, diameter estimation algorithm to estimate diameter and radius in large graphs. Connected Components algorithm is used to find connected nodes in graphs. There are various algorithms existing which adopted various graph theory concepts and these algorithms are used to extract different properties of nodes in graph and information about graph. In this paper, role of graph theory in computer science is discussed particularly in data analysis. The various algorithms which adopted graph theory concepts and theorems are presented in this paper. The role of graph theory in the analysis of social media data also termed as social media data analytics are discussed in this paper.

Keywords: Social Media Data, Graph Theory, Social Media Data Analytics, Big Graphs

1. INTRODUCTION

Globalization of data over the world is gradually increasing and it is proportionately related to the volume of data generated. The important and regularly accessed web sites are social media web sites which generate data in terms of millions to trillions. This data is referred as big data. Managing big data in regular systems with semi structured data is difficult. Different methods are used to represent social media data for analysis purpose. For example table structured datasets and graphs. The data on social media sites is in different formats like structured, unstructured and semi structured. The structured data is in the form of table structure and the unstructured data is multimedia data like audio, video and image data. Semi-structured data is combination of multimedia data and table data. Social media is generating large amount of data daily. For example facebook has 400 active users with an 120 friend ship connections for every month [1]. Graph theory and graph algorithms are best suited for social media data analysis [2].

|The right conclusions can be made by analysing large amount of social media data which may have different types of data in huge amounts. The various algorithms are applied over the data generated by social media to compute the analytical results quickly. In order to find the results easily and quickly in an efficient manner for the voluminous data the algorithms adopts various concepts which addresses complicated issues. One of the popular adopted technique incorporated in various algorithm is graph theory concepts. Graph theory concepts and theorems helps to solve many problems easily and efficiently.

In this paper, various important graph analysis methods are discussed such as shortest path analysis, optimal path analysis, path existence analysis and vertex centrality. The Page Rank algorithm which calculates relative importance of web pages and Random walk with Restart algorithm which measures proximity of vectors in graph are presented in this paper. The diameter estimation algorithm is used to estimate diameter and radius in large graphs. This paper also discusses how this diameter estimation algorithm helps in analysing social media data.

2. OVERVIEW OF GRAPH THEORY

A graph is a pair of vertices V, edges E. Graph can be defined as a set of objects connected with each other. As it is known that there are different graphs existing and some of them are simple graph, directed graph, directed edge graph, connected graph. A simple graph is a graph with no loops, directed graph is graph with directed edges and connected graph is graph with all edges connected with each other. The general properties of graphs are degree of graph which is the number of edges of a graph, In degree of graph is the number of edges entering into vertex, out degree of graph is the number of edges emitting from the vertex. Order property of graph is the number of vertices, rank property of graph is number of edges, vertices and components, diameter property of graph is the length of shortest path between two vertices [3].

Graphs can be represented in two different ways they are adjacency matrix and adjacency list. The adjacency matrix created with rows and columns where rows and columns are vertices of graph. The one row of adjacency matrix is with set of neighbours of a vertex[4]. The another way is adjacency list which is a collection of lists with its neighbours[4]. Adjacency list contains only neighbour vertices. It saves time and space but accessing and analysing graphs is difficult.

Operations performed on graphs are categorised as elementary operations, high level operations and advanced operations. Elementary operations of graphs are adding and deleting a vertex or an edge. And finding adjacent and neighbours of a vertex. High level operations performed on graphs are finding degree of graph, finding connectivity between nodes using clustering coefficient algorithm, finding shortest path between nodes using Single source shortest path and all pairs shortest path algorithms and finding path between nodes using Depth first search and Breadth first search algorithms[4]. Advanced operations performed on graphs are required to analyse graphs. There are some algorithms which adopted graph theory concepts in analysing social media data are specified below.

- Finding relative importance of web pages using Page Rank algorithm
- finding proximity of vectors in graph using Random walk with Restart algorithm
- Finding connected nodes in graph using connected components algorithm.
- estimating diameter or radius between vertices of graph using diameter estimation algorithm.

3. APPLICATION OF GRAPHS IN SOCIAL MEDIA ANALYTICS

Graph theory is started with Euler who used the graphs to find the best path to cross over each of seven bridges for exactly once.

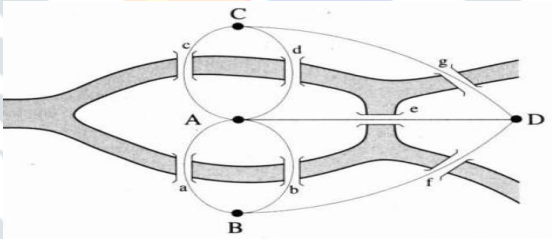

Fig 1 Euler path Graph

In 1805 Sir William Rowan Hamilton developed a toy based on finding optimal path by visiting all cities for only once.

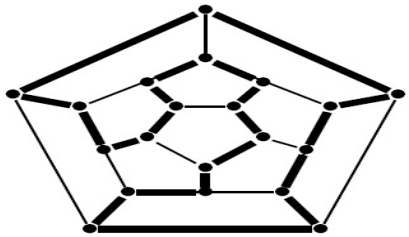

Fig 2 Hamilton Graph

Now the graphs are using for finding connections, communities, substructures. And for ranking websites. Graphs are useful in social networks like face book and twitter to find flow of information or opinions, finding most influenced person. Graph theory plays main role in finding spread of disease in medical field.

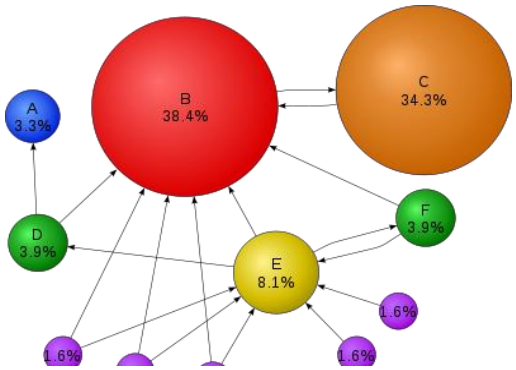

Fig 3: Simple PageRank graph

Navigation of online web sites is specified as a graph by specifying each page as a node and draw a line to another page if the pages has hyperlinks to navigate. These edges are directed edges and graph is a directed graph. These graphs are used in searching process and ranking the pages. The graph shows pages as nodes with percentage of page accessed. Using this graph, page rank is allotted for the page and this rank is used in searching the page.

Twitter data and users are specified as graph to find the flow of information. Users as nodes and data flow as edges. The direction of edge shows the direction of "following".

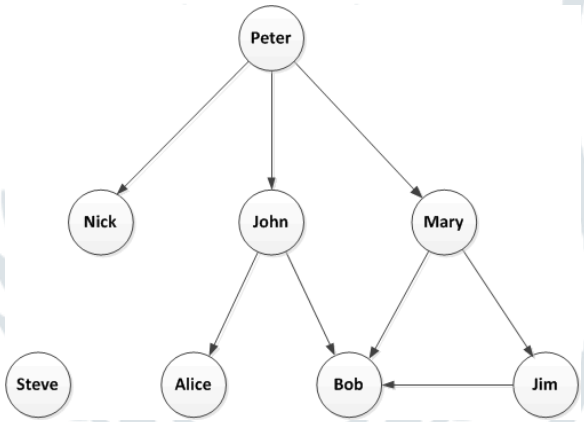

Fig 4: Simple twitter network

From the above graph Nick, John and Mary are followers of Peter. Bob is follower of John, Mary and Jim. Number of edges shows the number of followers or friends of a person. Using this most influenced person can be identified. The next section gives detailed explanation about these algorithms.

4. GRAPH ALGORITHMS USED IN SOCIAL MEDIA ANALYTICS

The data generated from social media is very huge which consists of structured, semi structured and unstructured data. Social media data can be represented as graph for getting good analytical results easily and also this graph representation is simple and efficient to use to analyse social media data. Different algorithms are applied to determine various results in analysis process from the graph by predicting graph properties.. The most popularly used algorithms to analyse the social media graph are discussed in the following sub sections.

4.1. Page Rank Algorithm on Twitter Data Set

Page Rank algorithm is developed to find relative importance of web pages based on the links between the pages [6]. Page rank of a page is decided by finding the *in links* and *out links* of the page. *In link* is the receiving link from other pages of external sites. If a page has more *in links* then the page get more votes and that page is treated as rank 1 page or most prestigious page. *Out link* is the link specified by the page to pages of external sites. This algorithm is used in different applications like data mining, social network analysis and social media data analysis.

How page rank algorithm is applied on twitter data for analyses is specified here. In this algorithm, the proposed function is influence(X) which predicts the number of people who will read a tweet and also includes all retweets. If a person reads same message for two times then both are counted because of retweets. X is a member of followers of Y then there is 1/following(X) probability for X reads the tweet posted by Y. Following(X) is the set of people that X Follows [7].

Influence(X) = sum (1+p*influence(Y)/Following(Y))

where Y is the person in Followers of X.

Using this method influence of X is calculated depending on the influence of followers of X.

4.2 Random Walk with Restart in Search Engines

Random walk with restart (RWR) algorithm is used to find similarity between vertices of the graph[8]. Random Walk with Restart(RWR) is an algorithm to measure the proximity of nodes in graph. The proximity vector *rk* from node k satisfies the equation[8]:

$$rk = cMrk + (1 / c)ek$$

This algorithm is used to rank the pages in web graph and to re-index the pages. Random walk with restart algorithm is generally used with Page rank algorithm. RWR algorithm is successfully implemented and using by Google. RWR algorithm can be used in multimedia graphs to assign keywords to different types of media like images, videos and audios. RWR algorithm is applied on DBLP data to mine the research communities to search relevant conferences, authors and journals [9].

4.3 Diameter Estimation Algorithm

Diameter is the one of the important property of graph which is used in different graph algorithms. Diameter is the path length between two connected nodes of the graph. Maximum diameter is the outlier. The effective diameter is defined as minimum number of vertices required to connect each pair of vertices of graph. This diameter estimation is required to understand and design algorithms for web graphs and social media networks.

4.4 All-pair shortest paths Algorithm

The distance between two nodes in a social network is a useful feature for many applications. For example, it can be a feature to predict most influencer in the network and followers of a person. In this section a study on how to approximate efficiently the distance between any two nodes in the graph in the public-private networks is specified here. This algorithm is particularly interesting in finding optimal path and distance between two nodes. This path may change dramatically even if we add a single edge. This algorithm is very useful to perform shortest path analysis, optimal path analysis, path existence analysis in analysing social media networks & data.

4. CONCLUSION

Social media data and networks contain huge information in terms of millions and billions. So the graph representation of this huge data simplifies analysis process. This paper discussed about the graph based algorithms which are used in analysing social media data. At present page rank, RWR algorithms are using by google. This paper briefly explained the application of page rank and RWR algorithms for twitter data and for other community networks. So, for social media data analytics, graph based algorithms are very efficient in finding solutions in the analysis process. Further research is required to use graph algorithms effectively in social media data analytics.

REFERENCES

[1] Facebook, “User statistics,” February 2010. [Online]. http://www.facebook.com/press/info.php?statistics

[2] M. Newman, “The structure and function of complex networks,” *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] David F. Nettleton,Data mining of social networks represented as graphs, computersciencereview7(2013)1–34,Elseiver journal.

[4] T.h. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, second ed., MIT Press and McGraw–Hill ISBN: 0-262-53196-8, 2001.

[5] U Kang and Christos Faloutsos,Big Graph Mining: Algorithms and Discoveries Carnegie Mellon University fukang.

[6] U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos,PEGASUS: Mining Peta-Scale Graphs, publication in Knowledge and Information Systems

[7] http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

[8] Adams Wei Yu, Nikos Mamoulis , Hao Su, Reverse Top-k Search using Random Walk with Restart School of Computer Science, Carnegie Mellon University ,Department of Computer Science, The University of Hong Kong Computer Science Department, Stanford University.

[9] Osmar R. Zaïane,Jiyang Chen, Randy Goebel, Mining Research Communities in Bibliographical Data, Advances in Web Mining and Web Usage Analysis pp 59-76.

[10] Suchit Pongnumkul, Kazuyuki Motohashi, Random Walk-based Recommendation with Restart using Social Information and Bayesian Transition Matrices, International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 9, March 2015.