Graphical Analysis for Big Data Analytics

Pathan Faisal Khan

***Abstract*—**

***Index Terms*— Big Data Analytics, Graph Analytics**

I. Introduction

WITH the introduction of the internet in the 1990s, there has been tremendous innovation in the tech industry. This changed the way organizations, businesses, governments function. It even changed the lifestyle of the people. Major contributions to the tech space were not until the early 2000s due to innovations in computational power and during this period, the volume of data generated with the introduction of social media and other services for the masses has risen a lot. Data is being created every second of the data. In 2013, Instagram users shared 3600 photos every minute, while in 2019, the number of photos shared every minute reached 46,740. The world internet population has increased from 2.5 billion to 3.7 billion [1]. It is estimated that by 2020, 40 trillion GB of data would be generated [2] which means internet user generates nearly 2.5 quintillion bytes of data every day [1]. Most of the data being generated is contributed by social media on which an average user spends 33% of his/her online time. This is why in 2019, there are 2.3 billion users active on Facebook [3].

With this vast amount of data, there was a need to develop more efficient and cost-effective data storage. This led to the introduction of the term Big Data in early 2005 [4]. Big data is the type of data that has a high variety, large volume, high velocity, greater veracity, and extreme value and is continuously growing on a large scale. These characteristics of the big data are referred to as the 5Vs. It will not be a surprise that the data is unstructured as it is being collected from multiple sources. Big data can be comprised of logs of the traffic coming in on a website, messages generated on a social media site, attributes of mouse clicks, details of products stored on an e-commerce website, medical data of a hospital, bank transactions, satellite data and many other sources which generate data.

Since generating data is an easier task than getting useful insights out of it, there was a need to emphasize on its analysis. But because of the sheer volume of high dimensional, unstructured and highly inconsistent data, running traditional methods for analysis might miss out on the hidden structures of the data. Thus, there was a need to devise powerful algorithms and provide high computational powers that can solve these problems. Due to the introduction of cloud computing and its scalable nature, researchers were able to develop algorithms to mine and make out meaningful insights from this data. With the right analysis methods, it can yield greater insights leading to stronger and strategic decisions. Using big data analysis, Netflix manages to easily save $1 billion every year [5]. Wikibon, an organization sharing tech-related knowledge, has estimated the market worth of big data analytics to a whopping $49 billion for the year 2019 [6].

Often data generated has relations among themselves. This data can be structured or unstructured or a mix of both. Since it is not feasible to understand these relations using the traditional big data analytics techniques, a better model had to be devised. A graph model was proposed to connect the data. Graphs are effective for analyzing, making recommendation systems and mining social networks. Due to the flexibility of this model it allows large quantities of information from many sources to be quickly absorbed and linked in ways that addressed the limitations in the source structures. A good way of representing the graph model is connections of a social media account; it represents a graphical structure with connections (edges) formed between different accounts (nodes/entities). This model enabled analyzing relationships and deducing interesting patterns between accounts (entities) in the structure. Graph analytics is the term used to define these methods of analysis. It is defined as an alternative to the conventional data warehouse model as a system for allowing analysts to check structured and unstructured data from different sources. Some business use cases of graph analytics include healthcare quality analysis, cybersecurity and correlation findings.

Under the supervision of Dr. Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, CO. Donegal

References

[1] "Resources - Whitepapers, Infographics & Webinars — Domo." [Online]. Available: https://www.domo.com/learn.
[2] C. Petrov, "Big Data Statistics 2019," Tech Jury. [Online]. Available: https://techjury.net/stats-about/big-data-statistics
[3] "Domo Resource - Data Never Sleeps 7.0." [Online]. Available: https://www.domo.com/learn/data-never-sleeps-7.
[4] "A Short History Of Big Data," Datafloq. [Online]. Available: https://datafloq.com/read/big-data-history/239.
[5] E. Team, "How Netflix Uses Big Data to Drive Success," inside BIGDATA, 20-Jan-2018.
[6] "Wikibon's 2018 Big Data Analytics Trends and Forecast," Wikibon Research, 28-Feb-2018. .
[7] "What is graph analytics?," IBM Big Data & Analytics Hub. [Online]. Available: https://www.ibmbigdatahub.com/blog/what-graph-analytics. [Accessed: 12-Dec-2019].
[8] A. Buluç and K. Madduri, "Parallel Breadth-First Search on Distributed Memory Systems," p. 12.

[9] V. N. Rao and V. Kumar, “Parallel depth first search. Part I. Implementation,” Int J Parallel Prog, vol. 16, no. 6, pp. 479–499, Dec. 1987.
[10] M. Naumov, A. Vrielink, and M. Garland, “Parallel Depth-First Search for Directed Acyclic Graphs,” in Proceedings of the Seventh Workshop on Irregular Applications: Architectures and Algorithms - IA3’17, Denver, CO, USA, 2017, pp. 1–8.
[11] M. Kranjčević, D. Palossi, and S. Pintarelli, “Parallel Delta-Stepping Algorithm for Shared Memory Architectures,” arXiv:1604.02113 [cs], Apr. 2016.
[12] J. W. Kim, H. Choi, and S.-H. Bae, “Efficient Parallel All-Pairs Shortest Paths Algorithm for Complex Graph Analysis,” in Proceedings of the 47th International Conference on Parallel Processing Companion, New York, NY, USA, 2018, pp. 5:1–5:10.
[13] L. Fitina, J. Imbal, V. Uiari, N. Murki, and E. Goodyear, “An Application of Minimum Spanning Trees to Travel Planning,” vol. 12, p. 11, 2010.
[14] L. L. Sz, “Random Walks on Graphs: A Survey,” p. 46.
[15] M. Sharir, “A strong-connectivity algorithm and its applications in data flow analysis,” Computers & Mathematics with Applications, vol. 7, no. 1, pp. 67–72, Jan. 1981.
[16] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Phys. Rev. E, vol. 76, no. 3, p. 036106, Sep. 2007.
[17] R. Tarjan, “Depth-First Search and Linear Graph Algorithms,” SIAM J. Comput., vol. 1, no. 2, pp. 146–160, Jun. 1972.
[18] A. Monge and C. Elkan, An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. 1997.
[19] Y. An, J. Janssen, and E. E. Milios, “Characterizing and Mining the Citation Graph of the Computer Science Literature,” Know. Inf. Sys., vol. 6, no. 6, pp. 664–678, Nov. 2004.
[20] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech., vol. 2008, no. 10, p. P10008, Oct. 2008.
[21] H. Lu, M. Halappanavar, and A. Kalyanaraman, “Parallel Heuristics for Scalable Community Detection,” arXiv:1410.1237 [physics], Oct. 2014.
[22] T. Schank and D. Wagner, “Finding, Counting and Listing All Triangles in Large Graphs, an Experimental Study,” in Experimental and Efficient Algorithms, vol. 3503, S. E. Nikoletseas, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 606–609.
[23] H. C. Johnston, “Cliques of a graph-variations on the Bron-Kerbosch algorithm,” International Journal of Computer and Information Sciences, vol. 5, no. 3, pp. 209–238, Sep. 1976.
[24] S. C. Antoro, K. A. Sugeng, and B. D. Handari, “Application of Bron-Kerbosch algorithm in graph clustering using complement matrix,” presented at the International Symposium On Current Progress In Mathematics And Sciences 2016 (ISCPMS 2016): Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016, Depok, Jawa Barat, Indonesia, 2017, p. 030141.
[25] L. C. Freeman, “Centrality in social networks conceptual clarification,” Social Networks, vol. 1, no. 3, pp. 215–239, Jan. 1978.
[26] Phillip Bonacich Reviewed, “Power and Centrality: A Family of Measures,” American Journal of Sociology, vol. 92, no. 5, pp. 1170–1182, 1987.
[27] C. F. A. Negre et al., “Eigenvector centrality for characterization of protein allosteric pathways,” Proc Natl Acad Sci U S A, vol. 115, no. 52, pp. E12201–E12208, Dec. 2018.
[28] L. Katz, “A new status index derived from sociometric analysis,” Psychometrika, vol. 18, no. 1, pp. 39–43, Mar. 1953.
[29] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web.,” 11-Nov-1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/. [Accessed: 12-Dec-2019].
[30] G. Sabidussi, “The centrality index of a graph,” Psychometrika, vol. 31, no. 4, pp. 581–603, Dec. 1966.
[31] S. P. Borgatti, “Centrality and network flow,” Social Networks, vol. 27, no. 1, pp. 55–71, Jan. 2005.
[32] L. C. Freeman, “A Set of Measures of Centrality Based on Betweenness,” Sociometry, vol. 40, no. 1, pp. 35–41, 1977.
[33] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A Comprehensive Survey on Graph Neural Networks,” arXiv:1901.00596 [cs, stat], Dec. 2019.
[34] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini, “The Graph Neural Network Model,” IEEE Trans. Neural Netw., vol. 20, no. 1, pp. 61–80, Jan. 2009.
[35] “Special Issue on Graph-based Methods for Large Scale Financial and Business Data Analysis - Call for Papers - Elsevier.” [Online]. Available: https://www.journals.elsevier.com/pattern-recognition/call-for-papers/graph-based-methods. [Accessed: 09-Dec-2019].