Study of Social Network Analysis using Big Data and Semantic Web Technologies

[a]Prof. Roshan M. Pote
Assistant Professor, Department of CSE
MGICOE & T, Buldhana, Maharashtra, India
roshan4892@gmail.com

Dr. Vilas M. Thakre
Head of Department
P.G.Department of Computer Science, SGB
Amravati University, Amravati, Maharashtra, India
vilthakare@yahoo.co.in

Prof. Mr. Shrikant P. Akarte
Assit. Professor, Department of CSE,
PRMIT&R, Badnera
Amravati.Maharashtra, India
s_akarte25@rediffmail.com

Abstract—Now days, important and popular topic among researchers is social network analysis. Because data from the social networks & another platforms are increasing tremendously and it is creating more interest in researches for social network analysis after introducing technologies like big data and semantic web technologies. For processing this large amount of data easily efficiently and effectively web technology plays an important role. Various different data processing frameworks are being used by the data researcher for analyzing, queering and integrating the data sets which are present on different places. But one problem was identified that the data which is available is not structured. It was semi structured or unstructured. Researchers are now finding the new methods to process the data in the cost effective ways. New technologies are required for SNA with integration of big data and semantic web technologies to improve the performance of existing methods and for developing new methods. This paper provides summarized information about recently developed SNA systems which will help researchers to find new ways. This paper presents the study of recently developed systems which will help the researchers for getting the directions.

***Keywords**-Big Data, Web technology.*

*****

I. INTRODUCTION

Computer Science and social science are brought together with the help of social network analysis. Nowadays this has become more popular amongst researchers who deal with big data and semantic web technologies. As social network make available large amount of the data, it opens new challenges. i.e. processing of the large data in sufficient time efficiently and combining all the dataset into single data set which will be compatible to the application. Social networks incorporates connected entities and relationships that exist between them, sometimes portrayed as a set of nodes and edges. A node is described as a illustration of a real world entity whereas an edge indicates a relation between those entities. In the same manner, a social network is portrayed and considered as a graph. According to Lee et al. Graph Mining Algorithms and Semantic web Technologies have common characteristics to analyze a given social network to extract hidden knowledge [1].SNA tries to come up with valuable and useful associations and knowledge that are hidden within the social knowledge.

II. LITERATURE REVIEW

There have been a lot of similar studies on SNA. Ostrowski developeda method to show that the structure of a network reflects the community dynamics for identification of influence and power among members [2]. Tsourakakis et al. analyzedhomophily and transitivity [3]. Mislove et al. studied computing the longest distance between the nodes [4]. Social data is growing quickly in size, selection and complexity whereas typically being unbroken in unstructured format .So as to beat the problem of analyzing large amounts of information that's in unstructured format, databases are being moved from relative to non-relational architecture. It is a very expensive and time consuming process to do analysis of a given social network. To resolve this performance drawback of large scale networks (datasets), researchers have started using parallel data processing platforms. Recently, another open source platform i.e. Spark has been developed which is more faster than Hadoop. This platform is based on the MapReduce algorithm & uses the utilization in-memory computing approach while processing data [5].

III. BIG DATA FRAMEWORK

All definitions regarding big data agree that it's large in size, unstructured and has totally different information formats. It’s a complex and difficult task to assemble, store, query, process, and analyze giant scale data. Big data is characterized by five Vs.: Volume (data size), speed (speed of information made and delivered), selection (data is comprised of heterogeneous sources), worth (extracting helpful and worthy information),and truthfulness (security, privacy, and trust) [6].

A. Database Related Technologies

To represent entities and the relationship between entities database models are used. There are two categories of No Sql related database technologies, i.e. Graph database and Tabular database. Tabular Databases: Some common examples of tabular databases are Hbase, Cassandra, Accumulo, and MongoDB. When data is needed to be accessed in real-time and randomly. Using Apache HBase is advantageous. Google, Facebook etc. use Hbase [7]. Barnawi et al. compared the performance of Giraph by using HDFS.The results show that as compare to HBase HDFS and Hive performed better. For managing very large amounts of structured data Cassandra is a distributed database system. Cassandra has been developed by Facebook initially [8]. CumulusRDF45, implemented on Cassandra, is competitive among other distributed RDF stores [9].

B. Parallel Processing In Big Data

Effective Big Data Mining requires scalable and efficient solutions that are also easily accessible to users at all levels of expertise. Distributed systems provide an infrastructure that can enable efficient and scalable Big Data Mining. Such systems, made up of organized clusters of commodity hardware, process large volumes of data in a distributed fashion. Hadoop , an open source implementation of MapReduce , is the most widely-used platform for large-scale distributed data processing. Hadoop processes data from disk which makes it inefficient for data mining applications that often require iteration. Spark is a more recent distributed framework that works with Hadoop and provides in-memory computation called Resilient Distributed Datasets (RDDs) that allows iterative jobs to be processed much faster, hence making it a more appropriate base for data mining. Graph process frameworks provide a surroundings to analyze social networks by permitting developers to develop scalable and fault-tolerant applications in a distributed manner. The challenges on graph process are constructing networks by exploitation unstructured data, analyzing the structure of the network, and inferring hidden data. Graph analytics have evidenced to be valuable tools in determination these challenges [10]. Some of the known technologies are MapReduce, programming paradigm, Hadoop framework, Spark, and Flink [11].

IV. Semantic Web Technology

Basically there are two technologies are present in semantic web. “Resource Description Framework” (RDF) which is a model definition for representing and defining associations between resources and another is “RDF Schema” (RDFS) and “Web Ontology Language” (OWL) for vocabularies that are used to represent the semantics of resources. The main good thing about the semantic net is to confirm the ability and to create certain that information is in machine readable format thus it enhances the potential of huge data analytics.

The graph-based representation of social data is done by semantic web by pushing the RDF standard. Linked information has seemed to overcome the issues like integrating different datasets from completely different domains. Ontologies, the key technologies during this context, give a additional versatile and distributed approach for representing data.

A. A Tripple Store

A triple store may be seen as a database for RDF triples. It is capable of storing and querying RDF triples in a repository by employing a query language. RDF triples may be represented as graphs, thus RDF is outlined as a graph data model, SPARQL is defined as a graph query language, and triple stores are outlined as a graph information system, respectively. The performance of triple stores has high importance particularly for real-time applications, Saleem et al. developed a customizable framework to benchmark triple stores, namely possible [12].Traditional triplestores run on single machine, and a few of the renowned examples are jena TDB, RDF4J(formerly Sesame), RDFSuite, AllegroGraph, RDF-3X, Hexastore [13], and RDFox [14].

B. Prototype Systems & Frameworks

Flink system has been developed by Mika et al. to extract, mixture and visualize social networks data. It uses semantic web technologies for reasoning on personal information [11]. Wang et al. developed an application that can question data that's extracted from Facebook user profiles [15]. Link Probe could be a paradigm designed by Chen et al. to predict the existence of links among members in social networks [16].

C. Vocabularies of RDF

A set of core ontologies describing certain domain or application areas in terms of semantic Web protocols (RDF, RDFS, OWL)is known as RDF vocabulary. FOAF, SKOS, SIOC, SCOT, MOAT,SemSNI , ActOnto , are the Some of the frequently usedRDF vocabularies in Social Network Analysis applicationsAclOnto, and InterestOnto and OntoSNA .FOAF (Friend of a Friend) ontology is used for therepresentation of their friendships and user profiles in social networks. FOAF can be used to merge personal information that is extracted from different sources [17]. Pankong et al.proposed a framework for semantic social network analysisby using FOAF [18].

V. Conclusion

Analysis of large scale social data isa challenging task .Big data not only deals with the data size but also it is about heterogeneity and diversity in the data formats and sources. Optimizing relations between individuals in social network has created interested due to repaid increase in the social networks & its data. The information about the active and effective technologies which are currently running and are being tested for different results are studied .By using existing ontologies Social data can be represented as RDF triples. a collection of RDF triples is called as RDF graph. Queries should be translated to jobs that can run in parallel so that to achieve benefit from Big Data Technologies. Decentralized RDF triple stores are the key technologies to handle large scale RDF data efficiently.

References

[1] S. Lee, S. R. Sukumar, and S.-H. Lim, “Graph mining meets the semantic web,” in *Data Engineering Workshops(ICDEW), 2015 31st IEEE International Conference on*. IEEE, 2015, pp. 53–58.

[2] D. A. Ostrowski, “Semantic social network analysis for trend identification,” in *Semantic Computing (ICSC), 2012*

IEEE Sixth International Conference on. IEEE, 2012, pp. 178–185.

[3] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos, “Doulion: counting triangles in massive graphs with a coin,” in *Proceedings of the 15th ACM SIGKDD internationalconference*

on Knowledge discovery and data mining. ACM, 2009, pp. 837–846.

[4] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMMconference on Internet measurement*. ACM, 2007, pp. 29–42.

[5] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social bigdata: Recent achievements and new challenges," *InformationFusion*, vol. 28, pp. 45–59, 2016.

[6] T. Das and P. M. Kumar, "Big data analytics: A frameworkfor unstructured data analysis," *International Journal ofEngineering Science & Technology*, vol. 5, no. 1, p. 153,2013..

[7] A. Bhardwaj, A. Kumar, Y. Narayan, P. Kumar *et al.*, "Bigdata emerging technologies: A casestudy with analyzingtwitter data using apache hive," in *2015 2nd InternationalConference on Recent Advances in Engineering &Computational Sciences (RAECS)*. IEEE, 2015, pp. 1–6.

[8] A. Lakshman and P. Malik, "Cassandra: a decentralizedstructured storage system," *ACM SIGOPS OperatingSystems Review*, vol. 44, no. 2, pp. 35–40, 2010.

[9] G. Ladwig and A. Harth, "Cumulusrdf: linked datamanagement on nested key-value stores," in *The 7thInternational Workshop on Scalable Semantic WebKnowledge Base Systems (SSWS 2011)*, 2011, p. 30.

[10] W. M. Campbell, C. K. Dagli, and C. J. Weinstein,"Social network analysis with content and graphs," *LincolnLaboratory Journal*, vol. 20, no. 1, pp. 61–81, 2013.

[11] P. Mika, "Flink: Semantic web technology for the extractionand analysis of social networks," *Web Semantics: Science,Services and Agents on the World Wide Web*, vol. 3, no. 2,pp. 211–223, 2005.

[12] M. Saleem, Q. Mehmood, and A.-C. N. Ngomo, "Feasible:A feature-based sparql benchmark generation framework,"in *International Semantic Web Conference*. Springer, 2015,pp. 52–69.

[13] C. Weiss, P. Karras, and A. Bernstein, "Hexastore: sextupleindexing for semantic web data management," *Proceedingsof the VLDB Endowment*, vol. 1, no. 1, pp. 1008–1019, 2008.

[14] Y. Nenov, R. Piro, B. Motik, I. Horrocks, Z. Wu, andJ. Banerjee, "Rdfox: A highly-scalable rdf store," *International Semantic Web Conference*. Springer, 2015,pp. 3–20.

[15] R.-C. Wang, T.-H. Su, C.-P. Ma, S.-H. Chen, and H.-H.Huang, "Social network data retrieving using semantictechnology," in *Computer Software and ApplicationsConference Workshops (COMPSACW), 2013 IEEE 37thAnnual*. IEEE, 2013, pp. 322–327.

[16] H. Chen, W.-S. Ku, H. Wang, L. Tang, and M.-T. Sun,"Linkprobe: Probabilistic inference on large-scale socialnetworks," in *Data Engineering (ICDE), 2013 IEEE 29thInternational Conference on*. IEEE, 2013, pp. 290–301.

[17] P. Mika, "Social networks and the semantic web," in*Proceedings of the 2004 IEEE/WIC/ACM InternationalConference on Web Intelligence*. IEEE Computer Society,2004, pp. 285–291.

[18] N. Pankong, S. Prakancharoen, and M. Buranarach, "Acombined semantic social network analysis framework tointegrate social media data," in *Knowledge and SmartTechnology (KST), 2012 4th International Conference on*.IEEE, 2012, pp. 37–42.

.

Authors and Affiliations

Prof. Roshan M. Pote

Presently working as Assistant Professor in ,Department of CSE, at Mauli Group of Institutions College of Engineering & Technology.

Dr. V. M. Thakare

Dr. Vilas M. Thakare is Professor and Head in Post Graduate department of Computer Science and engg, Faculty of Engineering & Technology, SGB Amravati university, Amravati. He is also working as a co-ordinator on UGC sponsored scheme of e-learning and m-learning specially designed for teaching and research. He is Ph.D. in Computer Science/Engg and completed M.E. in year 1989 and graduated in 1984-85.

Prof. Mr. Shrikant P. Akarte

Presently working as Assistant Professor in ,Department of CSE, Prof. Ram Meghe Institute OfTechnology and Research, Badnera, Amravati. Saint Gadgebaba Amravati University, Amarvati, Maharashtra,
India – 444701.