Survey Paper On User Analytics-based System using Social media Stream data

Ms. Sarika N. Raut

***Abstract*— Twitter, is one of the largest social media site that receives tweets in millions of data in every day in range of Zettabyte per year. This huge amount of raw data can be used for industrial or business. This paper provides a way of analyzing of big data such as twitter data using Apache Hadoop which will analyze the tweets on a Hadoop clusters. In this paper, we are going to talk how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. Twitter is an online web application which contains huge amount of data that can be a structured, semi-structured and un-structured data. Twitter is also difficult due to language that is used for comments. So here we are taking sentiment analysis, for this we are using Hive and its queries we have defined in the HQL (Hive Query Language). Here we have categorized this sentiment analysis into 3 groups like tweets that are having positive, moderate and negative comments.**

***Index Terms*— Hadoop, Big Data, Map Reduce, Twitter, HDFS, Tweets, Sentimental Analysis, Flume..**

I. INTRODUCTION

BIG data is the name used every where now a days in distributed paradigm on web. BIG data is the collection of sets of very huge amount of data in terabytes.

The upcoming of online social media and mobile communication technologies has triggered a rapid increase in the flow of user generated content of various forms.

People are express their reactions, fancies and predilections through social media by means of textual fragment of epigrammatic nature rather than writing long text. We call friends made through this traditional fashion as G-friends, which stands for geographical location-based friends.

In social media most of people adds their habits, daily activities, etc. on that data (habits, activities, comments) we found who is most matched to another person... we analyze that person to user by calculating polarity .

One challenge with existing social networking services is that large data is stored in hdfs means it is more scalable... Millions of tweets are generated every day on multifarious issues. We propose an unsupervised and domain-independent approach by using the polarity scores from three lexical resources-SentiWordNet 3.0, SenticNet 2 and SentislangNet. SentiWordNet contains polarity scores of uni-grams. for that we express positive or negative opinion.

II. RELATED WORK

In this paper, L. Page, S. Brin, R. Motwani, and T. Winograd [1] have taken on the audacious task of condensing every page on the World Wide Web into a single number, its PageRank. PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure. It found a number of applications for PageRank in addition to search which include traffic estimation, and user navigation. Also, It can generate personalized PageRanks which can create
a view of Web from a particular perspective.
Overall, our experiments with PageRank suggest that the structure of the Web graph is very useful for a variety of information retrieval tasks.

In this paper, P. Desikan, N. Pathak, J. Srivastava, and V. Kumar. [2] propose a method to incrementally compute PageRank for a large graph that is evolving. It have provided an approach to compute PageRank incrementally for evolving graphs. The key observation is that evolution of the Web graph is slow, with large parts of it remaining unchanged. By carefully delineating the changed and unchanged portions and the dependence across them, it is possible to develop efficient algorithms for computing the PageRank metric incrementally.

In this paper, W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger [3] addressed the problem of link recommendation in weblogs and similar social networks. It discussed the ground
features available in *LiveJournal*'s public user information pages and describe some graph algorithms for analysis of the social network.

In this paper E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell [4], presented a detailed description of the CenceMe architecture. Through our prototype implementation It have demonstrated successful integration with a number of popular off-the-shelf consumer computer communication devices and social networking applications.

In this paper, T. Huynh, M. Fritz, and B. Schiel [5] propose a novel method to recognize daily routines as a probabilistic combination of activity patterns. The use of topic models enables the automatic iscovery of such patterns in a user's daily routine.

In this paper, Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma [6] move towards understanding user mobility based on GPS data. A work aiming to infer transportation modes from GPS logs based on supervised learning is reported.

It consists of three parts: a change point-based segmentation
method, an inference model and a graph-based post-processing algorithm. First, It propose a change point-based segmentation method to partition each GPS trajectory into separate segments of different transportation modes. Second, from each segment, It identify a set of sophisticated features, which are not affected by differing traffic conditions. Later, these features are fed to a generative inference model to classify the segments of different modes. Third, It conduct graph-based postprocessing to further improve the inference performance.

III. SYSTEM OVERVIEW

As it can have seen existing system drawbacks, here we are going to overcome them by using Big Data problem statement. So here we are going to use Hadoop, for getting raw data from the Twitter we are using Hadoop online streaming tool using Apache Flume.

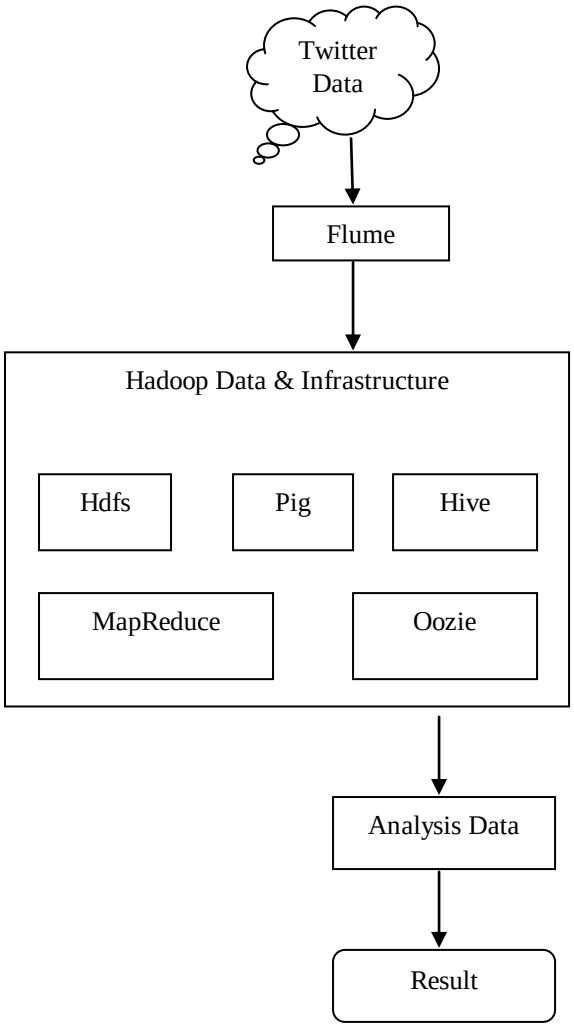

fig. 1 Architecture of Twitter Analytics

First the raw data of Twitter accessed by flume. The flume is contained source, channel and sink. The source accesses the data of Social media such as Twitter. Then the channel transmits that data to sink for extraction. After that the sink stores the data to centralized storage like, hdfs.

Secondly, Hdfs can give the data that is stored in it to Hadoop Framework. Hadoop Framework are nothing but map reduce, hive, pig.

Hadoop Framework are analyze twitter data using specific algorithms. The spacific algorithms are Market Basket Analysis Algorithm, Apriori Algorithm and naive bayes algorithm for sentiment analysis. After that resulted analysis data stores in hdfs, and show result to user..

IV. CONCLUSION

In this paper, only the polarity scores from SenticNet to analyze twitter sentiments. To evaluate our system on large-scale field experiments and done some analysis on the tweets and the most number of tweet ids. There are several ways to define and analyze the social media data such as Twitter.

REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
[2] P. Desikan, N. Pathak, J. Srivastava, and V. Kumar. Incremental page rank computation on evolving graphs. Proc. of WWW, pages 1094-1095, 2005.
[3] W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis.
[4] E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell. Cenceme-Injecting Sensing Presence into Social Networking Applications. In Proc. of EuroSSC, pages 1–28, October 2007.
[5] T. Huynh, M. Fritz, and B. Schiel. Discovery of Activity Patterns using Topic Models. In Proc. of UbiComp, 2008.
[6] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding Transportation Modes Based on GPS Data for Web Applications. ACM Transactions on the Web (TWEB), 4(1):1–36, 2010.

Sarika N. Raut is a Student in the Department of Computer Engineering, JSPM ICOER,Wagholi. Savitribai Phule, Pune University, Pune.
Mobile No.::9405348361