Progressive Analysis on Twitter-Social Data Using Cloud

Soumya Terala,
Student of M.Tech, Specialization - Software Engineering
Padmasri Dr.B.V.Raju Institute of Technology
Hyderabad, India
terala.soumya@gmail.com

Niladri Sekhar Dey,
Asst. Professor, Dept of IT
Padmasri Dr.B.V.Raju Institute of Technology
Hyderabad, India
niladri.dey@bvrit.ac.in

Sujoy Bhattacharya,
Professor, Dept of IT
Padmasri Dr.B.V.Raju Institute of Technology
Hyderabad, India
sujoy.b@bvrit.ac.in

Abstract - As the rise of social networking, people started share information through different kinds of social media. Among all of them, twitter is an important resource. Twitter is a micro-blogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets". It was created in March 2006 by Jack Dorsey and launched in July. In this project we present a novel system which collects tweets from the users. With the information gathered from the tweets we shall be able to create a trend based on information. This research proposes a method of mining on tweets and allows the analyst to query for decision making data. we use cloud services such as Amazon Simple DB to store the data of the user as it is cost efficient, provides security and the data in cloud can never be prone to crashes rather than maintain the data in a local system .

Keywords- Twitter, Trend Analysis, Trend Analysis on Twitter, Simple DB Response time

I. INTRODUCTION

A **social network** is a social structure made up of a set of actors (such as individuals or organizations) and a complex set of the dyadic ties between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

II. MATERIALS AND METHODS

Twitter: Twitter is an online social networking service and micro blogging service that enables its users to send and read text-based messages of up to 140 characters, known as "**tweets**".It was created in March 2006 by Jack Dorsey and launched that July. The service rapidly gained worldwide popularity, with over 500 million registered users as of 2012, generating over 340 million tweets daily and handling over 1.6 billion search queries per day. Since its launch, Twitter has become one of the ten most visited websites on the Internet, and has been described as "the SMS of the Internet." Unregistered users can read tweets, while registered users can post tweets through the website interface, SMS, or a range of apps for mobile devices.

Trend Analysis: Trend analysis is a form of comparative analysis that is often employed to identify current and future movements of an investment or group of investments. The process may involve comparing past and current financial ratios as they related to various institutions in order to project how long the current trend will continue. This type of information is extremely helpful to investors who wish to make the most from their investments. The process of a trend analysis begins with identifying the category of the investments that are under consideration. For example, if the investor wishes to get an idea on the potential for making a profit with pork bellies, the focus will be on the performance of pork bellies in a commodities market. The trend analysis will include more than one supplier for the commodity, in order to get a more accurate picture of the current status of pork bellies on the market.

Trend Analysis on Twitter: Twitter is now the third most popular social network, behind Facebook and MySpace (Compete, 2009). A year ago, it has over a million users and 200,000 active monthly users sending over 3 million updates per day(TechCrunch, 2008). Those figures have almost certainly increased since then, with the

torrential streams of Twitter updates (or tweets), there's an emerging demand to sieve signals from noises and harvest useful information.

Here are five places to go for finding twitter trends.

1. Twitter Search – This is Twitter's own search feature. When you land on the page, you'll see a search box with the top trending topics underneath. Conduct a search and you'll get a list of tweets related to your topic. On the right side of the screen you'll see the top 10 trending topics on Twitter.
2. Tweetmeme – Tweetmeme aggregates all the popular links on Twitter and categorizes them for you.
3. What The Trend – What The Trend is unique. When you land on the home page, the first thing you'll see is a list of trending topics from around the world that are trending right now. You can easily click a link and see all trending topics for the day. You can also check the top Twitter trends by town and country. The links are on the right side of the page.
4. Trends Map – Trends Map shows you all the real-time trending topics geographically.
5. Twopular – Twopular shows you trending topics according to a specific time frame (now, 2 hours, 8 hours, day, week, etc.). Just click the appropriate tab.

III. SYSTEM FRAMEWORK

The Twitter framework allows your application to send Twitter requests on behalf of the user. The framework takes care of user authentication for you and provides a template for sending direct messages. Our framework is divided into four sections.

- **Data Collection:** We have set up the list of users manually, then go through the website of these users to collect their tweets. The data collected by the Atom reader to be stored in a database for further analysis.

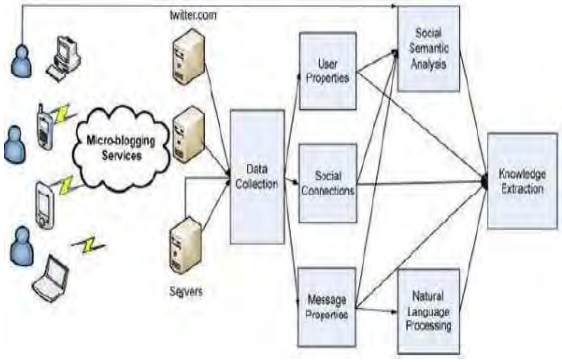


Fig1:Overview of Data Collection

- **Data Categorization:** The collected data should be categorized based on these factors Age, Relationship status, Status update, Interests, Photos, Shares, Travel, Friends, and Events etc.
- **Data Analysis:** Analyze the data based on the recent activities or by calculating the average. By analyzing the data we can add more interesting features to the site by doing this many people can attracted to the site and even the popularity of the site can also be increased.
- **DataPrediction:** We can predict the data based on data analysis. For example based on the places he/she travelled we can predict which kind of places he/she likes most and we suggest them about those places.

IV. PERFORMANCE EVALUATION APPLICATION

Below figure shows that system architecture for data analysis. Here Twitter contains twitter 4j API. Here twitter data analysis is developed to understand the relation between multiple participants called twitter users. Twitter users generates some data during some operations this data is called tweets.

To test the performance of SimpleDB, a simple application was built during this research work, which allows multiple operations on social networking data. The application maintains a mirrored copy of the data in SimpleDB and performs all the operations on this DB The complete application has many parts which are shown in Fig2 and all of them are explained below.

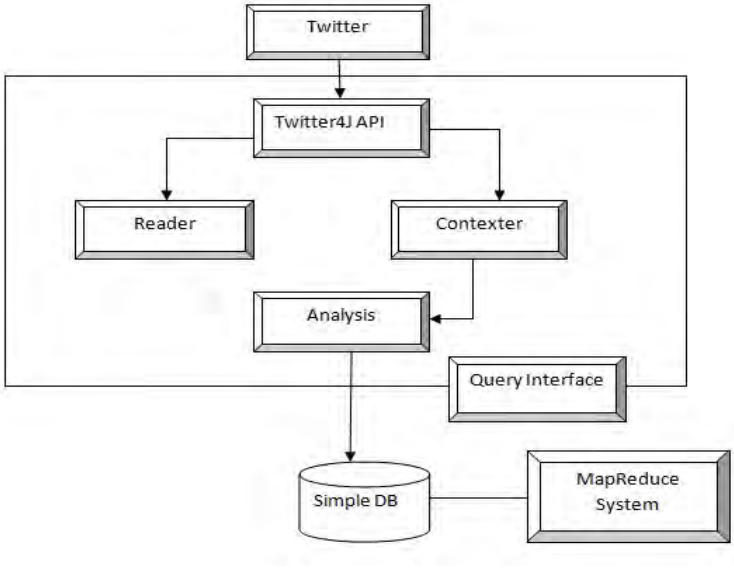


Fig2 : Components of an application

Twitter 4J API: Twitter 4J is an API for java. But it is unofficial java library for Twitter API. But in this application it is more helpful for working with Twitter Services, because it contains number of classes relating to twitter services.

Reader: Reader is a component of twitter 4j, in our application reader component is used for reading purpose. i.e it reads all the tweets from your timeline and stores all the tweets in your database.

Contexter: Contexter is a component of twitter 4J, in this application contexter component is used for searching purpose. i.e. if your database contains a large amount of data then this contexter component searches for a particular keyword among large amount of data from your database.

Analysis: Here we can analyze the data based on the recent updates of twitter data. By analyzing the twitter data we can add more interesting features to the website. By doing this lot of people can attracted to the site. By doing this the popularity of the site also be increased.

Query Interface: Query Interface means sending queries to the database and this interface is interfaces between the query GUI and the database. It also takes care of the predefined queries, result sets etc..

Simple DB: Amazon Simple DB is a highly available and flexible non-relational data store that offloads the work of database administration. Developers simply store and query data items via web services requests and Amazon Simple DB does the rest.

MapReduce System: MapReduce program consists of a Map() procedure, which performs filtering and sorting and Reduce() procedure that performs summarization operation. We used MapReduce library for simple DB called Amazon Elastic Cloud MapReduce. In our application we are not using this MapReduce System but it is similar to contexter component so we are discussing this MapReduce System here.

V. ROLE OF CLOUD COMPUTING – SIMPLEDB RESPONSE TIME

To test the performance of Simple DB, a simple application was built during this research work, which allows multiple operations on social networking data. The application maintains a mirrored copy of the data in Simple DB [9] and performs all the operations on this DB. The complete application has many parts which are shown in Fig2 and all of them are explained above.

Cloud computing is the use of computing resources that are delivered as a service over a network The name comes from the use of a cloud-shaped symbol as an abstraction for the a complex infrastructure. cloud computing applies remote services to a user's data, software and computation. Here Cloud DB is connected to this application, where this appliction sends all the queries to the database. This application equipped with multi threading of queries. Hence we also included an incremental type of multi-threading code, where initially the code will batch only 5 queries per second and then for each second 2 new threads will be generated. Hence after running the application for 1 hour, total number of queries will be approximately 7000 with a internet connection of 10 MBPS. With this setup, we recorded the response time for Simple DB(AWS).

Below formula is used for finding Simple DB Query Response Time. Result of Simple DB Response Time is given in below table.

$$Q_R = DB_S * Q_N * 4 + N_S, \quad N_S = 10$$

Where, Q_R is Query Response Time,

DB_S is size of the database domain to scan,

Q_N is number of Queries to be fired in a second,

N_S is the network speed to be considered at 0 for local

VI. RESULTS

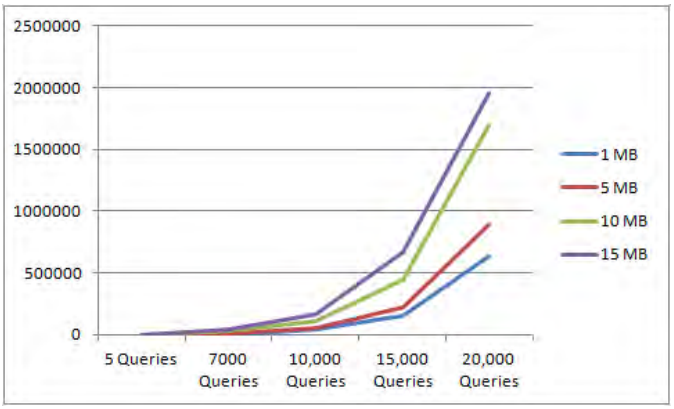
Above formulation was used to calculate the simple DB Response Time for number of Query instances. Here we calculate the response time in milliseconds.

Table: Simple DB Response Time

Simple DB	Query Instances				
	5 Queries	*7000 Queries*	*10,000 Queries*	*15,000 Queries*	*20,000 Queries*
1 MB	20	2.8 K	40 K	160 K	640 K
5 MB	20	14 K	56 K	224 K	896 K
10 MB	20	28 K	112 K	448 K	1702 K
15 MB	20	42 K	168 K	672 K	1952 K

All the time values are in ms, where k denotes thousand

Result of this table in graphical format is shown in Figure 3.



Xaxis- Query Instances
Yaxis- Query Response Time

Fig 3: Query Vs Dataset size

VII. CONCLUSION

In this paper, we present a trend analysis system for monitoring trends on technology domain from Twitter. This system not only collects data, but also provides functions for further information exploration by data mining and friendly user interface display. There are four main functions presented in this system: top news, trending topics, active users, and top sources. Top news and trending topics demonstrate two different ways to display our mining results, users can also use filters, search function, or top keyword/hashtag to interact with the mining result. The function of presenting active users and top sources not only provides the sight of future trend prediction, but also provides valuable information for detecting opinion leaders among celebrities and source media. By the ability of analyzing celebrities' discussion, and display mining result in an user friendly visualized website, our system provides a new way for users to discover the trend of hot topics in the future.

The performance of SimpleDB is analyzed on a social network dataset, which is large in volume and the effect of the number of queries on the same dataset is studied. It is proven that there is a large difference in terms of query response time is highly dependent on the network speed or the network bandwidth, throw which AWS is accessed.

VIII. FUTURE WORK

In this paper, we demonstrate a trend analysis tool on users from Twitter. This Structure can also be applied to different domain and different social network site. According to mining results, the contributions of each user and source media are thereby revealed.

ACKNOWDEMENT

The research work is being carried in the Cloud Computing Center in Padmasri Dr. B. V. Raju Institute of Technology, Hyderabad and we would like to express our greatful thank to Dr.Sujoy Bhattacharya for giving his valuable suggestions.

REFERENCES

[1] e. Articles. (2012, July 13). Top 15 Most Popular Social Networking Sites | July 2012. Available: http://www.ebizmba.com/articles/social-networking-websites

[2] S. Wasserman and K. Faust, Social network analysis : methods and applications. Cambridge; New York: Cambridge University Press, 1994.

[3] O. Phelan, K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news," presented at the Proceedings of the third ACM conference on Recommender systems, New York, New York, USA,2009.

[4] Technology Trend Analysis Tool using Twitter As a Source by vi-Chun Lin, Department of Information Technology, Ping-che Yang,Institute for Information Industry, Wen-Tai Hsieh, Department of Information Management, Seng-cho T. Chou, Department of Information Management,National Taiwan University , Taipei, Taiwan