ISSN: 2321-7782 (Online)
Volume 3, Issue 4, April 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Analysis of Twitter Data Credibility: A Survey

Rahul Bora[1]
Dept. of CSE
BMSCE
Bengaluru, India

Rahul Kumar[2]
Dept. of CSE
BMSCE
Bengaluru, India

Satyam Shankar Prasad[3]
Dept. of CSE
BMSCE
Bengaluru, India

Utkarsh Dev[4]
Dept. of CSE
BMSCE
Bengaluru, India

Pallavi G B[5]
Associate Professor, Dept. of CSE
BMSCE
Benagluru, India

Abstract: ***Conducting analytics over data generated by social networking sites such as Facebook and Twitter is challenging, due to the huge volume, variety and velocity of the data. There are various issues concerning data on social networking sites, one of them being examining the credibility of data constantly posted on them. In this paper we have done a detailed survey of mechanisms used in analysing the online data viz., data extraction, where streaming data is extracted and classified, data analytic, where analysis of data takes place and visualization where we put light on how curves and graph concerning data are plotted. Such analysis and survey helps in ensuring the integrity of the information.***

Keywords: OSNs, Analytics, Classifiers, Visualisation, Weka, Geo-tagging

I. INTRODUCTION

The rapid growth of the Internet, the widespread deployment of sensors and scientific advances such as the mapping of the human genome has resulted in a tsunami of data flooded across the internet. Today, the challenge has shifted from collecting a sufficient amount of data to understanding and gaining insight from the massive amount of data that are available. This paper deals with the collection of data from the social networking sites by extracting the efficient information and analyzing the data with the help of machine learning and various data visualization techniques.

Online Social Networks are the huge source of information. They have been a primary focus of the information retrieval and text mining, because it produces massive unstructured textual data and displays user relations in time. Twitter, for example, is a micro blogging OSN which allows a user to post his ideas in text, not more than 140 letters, called tweets.

There are over half a billion tweets posted every day. The tweets may be related to some events such as Assembly elections or even more sensitive such as a terrorist attack. Due to the sensitivity of the information, it is very important to check the credibility of the information, which otherwise may cause panic.

Research in data analytics and visualization is concerned with developing computational methods to extract knowledge from large, complex, interrelated data sets which spans machine learning, visualization and analysis of massive data sets.

In this paper, we have made a survey on search and information retrieval for documents and information on the OSNs, and on analyzing the available data. This process has three steps which are described in the next sections. The first step is to collect data from an OSN using APIs or through data crawlers (spiders) and storing it inside the databases. In the next step, we examine different tools and algorithms used to analyze the information and checking its authenticity. Further, we talk about various

visualization techniques that are used to showcase the result visually using histograms, pie-charts etc. Each of the three steps has been examined carefully and described in detail.

II. SURVEY

a) Data Collection

Twitter has released a set of API functions that support user information collection. In [1], Zi Chu et al., has diversified data sampling, by employing two methods to collect the dataset covering more than 500,000 users. The first method that they have used is Depth-First Search (DFS) based crawling. According to them, the reason they choose DFS is that it is a fast and uniformed algorithm for traversing a network and it implicitly includes the information about network locality and clustering. For each reached user, the authors record its follower list. Taking the following direction, the crawler continues with the depth constraint set as three. They customized their crawler with a core module of PHP cURL. Ten crawler processes work simultaneously for each seed. After a seed is finished, they move to the next.

In the second method, they used the public timeline API to collect the information of active users, increasing the diversity of the user pool. Twitter constantly posts the twenty most recent tweets in the global scope. The crawler calls the timeline API to collect the authors of the tweets included in the timeline.

We are familiar with the problem of big data constantly being generated on the social networking sites as millions of people post their views on it. In [2], Minas Gjoka et al., suggests a way to develop a practical framework for obtaining a uniform sample of user data in an online social network (OSN) by crawling its social graph. Such a sample allows estimating any user property and some topological properties as well. The four basic techniques they suggested were: Breadth-First Search (BFS), Random Walk (RW), Metropolitan-Hastings Random Walk (MHRW) and Re-weighting Random Walk (RWRW)

BFS sampling is known to introduce bias towards high degree nodes, which is highly non-trivial to characterize analytically. Random Walk (RW) sampling also leads to bias towards high degree nodes, but whose bias can be quantified by Markov Chain analysis and corrected via appropriate re-weighting (RWRW). Then, they consider the Metropolis-Hastings Random Walk (MHRW) that can directly yield a uniform stationary distribution of users. This technique has been used in the past for P2P sampling and recently used for few OSNs. Thus the two approaches that can produce approximately uniform samples are the Metropolis- Hasting random walk (MHRW) and a re-weighted random walk (RWRW).

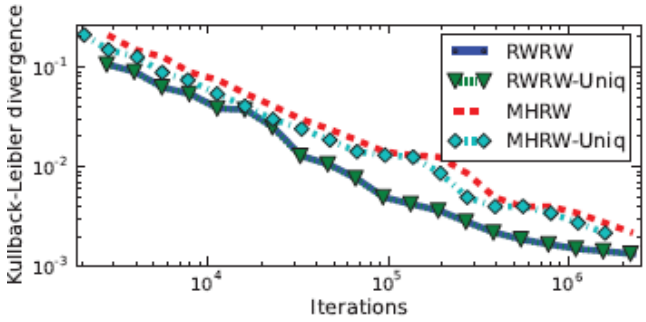

Fig. 1: The efficiency of RWRW and MHRW in estimating the degree distribution of Facebook(OSN), in terms of the Kullback-Leibler divergence.[2]

In [3], Changhyun Byun et al. introduce the design specifications and explain the implementation details of the Twitter Data Collecting Tool. The power of the cloud has to be used in order to crawl the data and PageRank algorithm is used. The process consisted of five steps:

1) First, queue and table are set up to maintain all user IDs that need to be crawled.

2) The users' and followers' IDs are saved in the SimpleDB, which is a service for storing structured data in the cloud.

3) Furthermore, different users' information is gathered for different instances simultaneously by using their own web service.

4) In the fourth step, the PageRank Algorithm is applied to rank users.

5) Finally, a web interface enables public users to access their data. As a result, they crawled 50 million users and 1.8 billion followers' information and analyzed Twitter users using the PageRank algorithm.

In [4], Xinyue Wang et al., mentions that by searching the information just by using the keywords results in a significant loss of relevant information. Thus they proposed an adaptive crawling model that detects emerging popular hashtags, and monitors them to retrieve greater amounts of highly associated data for events of interest. The proposed model analyzes the traffic patterns of the hashtags collected from the live stream to update subsequent collection queries. To evaluate this adaptive crawling model, we apply it to a dataset collected during the 2012 London Olympic Games.

The authors used adaptive crawling over baseline crawling technique, as in baseline method a set of keyword is set manually according to the event of interest. The system structure of the adaptive crawling model is similar to the baseline, except an additional Keyword Adaptation feature. In this model, the data collection process is started by the same set of predefined keywords as the baseline. The keyword adaptation feature enables the identifying of popular event-related hash tags by using the keyword Adaptation Algorithm. Then, those hash tags are added to the keywords retrieval set at the end of every time frame. Finally, a query that encodes all the words in the keywords set is sent to the Twitter API when the timer restart and another iteration of adaptation begin.

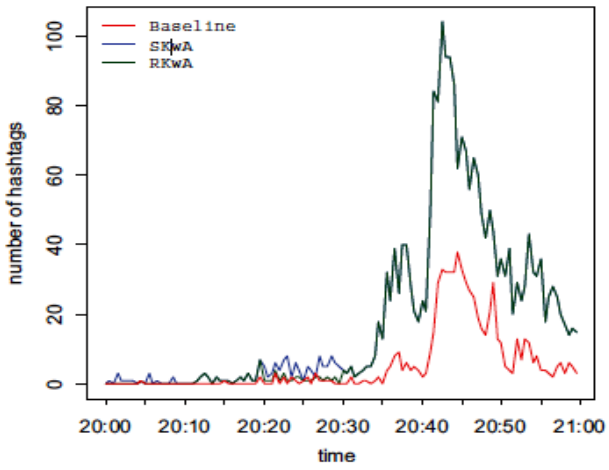

Fig. 2: Traffic pattern of #phelps in three datasets for Men's 4x100m Medley in London Olympics [4]

Once the data has been collected using the above described techniques, the next step involves classifying them using different classifiers in the required form for the further steps. The data classification techniques are discussed in details in the next section.

b) Data Classification

The development of data-mining applications such as clustering and classification has shown the need for machine learning algorithms to be applied to large scale data. In [5], Mohd Fauzi bin Othman et al., emphasize on Waikato Environment for Knowledge Analysis or in short, WEKA. WEKA is an open source software/tool which consists of a collection of machine learning algorithms for data mining and data classification tasks. It consists of various algorithm used for classification purpose such as Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. A fundamental review on the selected technique is presented for introduction purposes. The data breast cancer data with a total data of 6291 and a dimension of 699 rows and 9 columns will be used to test and justify the differences between the

classification methods or algorithms. Subsequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or other general applications.

Bayesian networks are a powerful probabilistic representation. Their use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability.

Radial basis function (RBF) networks have a static Gaussian function as the nonlinearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centered. The key to a successful implementation of these networks is to find suitable centers for the Gaussian functions.

A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited.

Single conjunctive rule learner is one of the machine learning algorithms and is normally known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, and at the same time being as small as possible.

To make a prediction about an unknown point, the nearest neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point accordingly to some distance metric.

In [3], Changhyun Byun et al. also provide an analysis of Twitter data gathered by crawler about Super Bowl in a case study. The case study aims to address the question of how people use Twitter and to assess the power of Twitter in creating consumer interest in brands and commercials. The main objective of this study is to find the relationship between Twitter and Super Bowl ads by analyzing data on Twitter. They also suggested embedding a rule based engine in the Twitter Data Collecting Tool to filter the data.

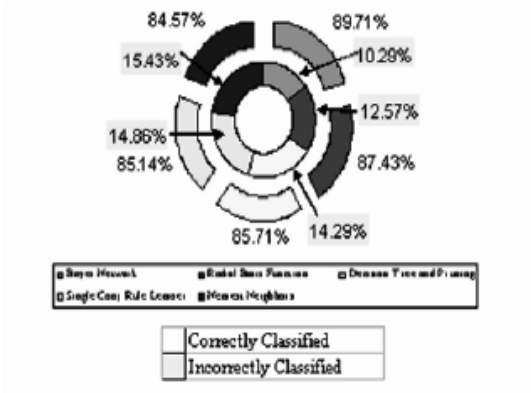

Fig. 3: Classification Results of using Breast Cancer WEKA [5]

Algorithm (Total Instances, 175)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Time Taken (seconds)	Kappa Statistic
Bayes Net.	89.7143 (157)	10.2857 (18)	0.19	0.7858
Radial Basis Function	87.4286 (153)	12.5710 (22)	0.53	0.7404
Decision Tree and Pruning	85.7143 (150)	14.2857 (25)	0.23	0.7019
Single Conj. Rule Learner	85.1429 (149)	14.8571 (26)	0.15	0.6893
Nearest Neighbors	84.5714 (148)	15.4286 (27)	0.81	0.6860

Table 1: Simulation result of each algorithm using breast cancer data [5]

In the first phase, the authors analyzed the relevance between tweets about the car-related companies and Super Bowl commercials. First, they filtered tweets that are related to Super Bowl car-related commercials using keywords and rules. The keywords are based on the name of companies that aired commercials during the Super Bowl and actors' and car names from the commercials. This is because users mostly mentioned the companies, the cars, the actors, or characters in the commercials when they composed tweets about the commercials. For this reason, they made a rule that if a tweet includes at least one of the keywords, it is a tweet about the company.

The twitter data analytics was used for analysing data in the Korean presidential election in 2012. It can be used for real-time events. In [6], Min Song et al., describes crawling and analysing the real-time data. They used three important mining techniques: multinomial topic modelling, network analysis, and term co-occurrence retrieval:

Topic modeling is one such document-modeling technique; LDA, one of the earliest topic modelling techniques, is based on a graph model and assumes a Dirichlet prior-based topic distribution. That is, LDA represents documents as mixtures of topics that spit out words with certain probabilities.

In network Analysis: Users in Twitter, however, can form relationships with others by following them without the other party's consent. Hence, focusing only on the follow/following connection cannot properly reflect the traits of Twitter user networks. In addition, dynamic changes of user networks accompanied by changes in real-world social issues must be identified.

Term co-occurrence retrieval. Given a query, the system retrieves the list of terms that co-occur with the query term. Once the list is obtained, it sorts co-occurring terms by co-occurrence frequency and can display the result.

In [7], Shamanth Kumar et al., have not only discussed the crawling methods but also how to store them and fetch them from the data base. They also discussed how to use the machine learning to analyze the huge sets of data extracted from the website and then finally visualizing them graphically. The authors also provide details about extracting multi-fields such as username and location from the Twitter along with the tweets, so that they can be used for further analysis. Location information on Twitter is available from two different sources:

- Geo-tagging information: Users can optionally choose to provide location information for the Tweets they publish. This information can be highly accurate if the Tweet was published using a smart phone with GPS capabilities.
- Profile of the user: User location can be extracted from the location field in the user's profile.

Approximately 1% of all Tweets published on Twitter are geo-located. This is a very small portion of the Tweets, and it is often necessary to use the profile information to determine the Tweet's location. The location string obtained from the user's profile must first be translated into geographic coordinates. Typically, a gazetteer is used to perform this task. A gazetteer takes a location string as input, and returns the coordinates of the location that best correspond to the string.

Once the data is classified and relevant data is extracted in required format, different tools and algorithms are used for their analysis. This is the most important step as it's here we test the integrity of the information. The analytics is followed by visualization where we discuss about different tools and methods to picture the result. These methods are describe in details in the next section.

c) Data Analysis and Visualization

Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individual actors within the network) and ties. These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines. The ties represent relationships between the individuals, such as Facebook friendships, email correspondence, hyperlinks, or Twitter responses. These are the bulk sources of information for social network analysis. These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines.

Mark Smith and others [9] has introduced Nodexl. It is a free and open source network analysis and visualization software package for Microsoft excel 2007/2010. It is a revolutionary graphics program that synthesizes and cluster social network data. NodeXL has the ability to synthesize the collected data, for example, Twitter feeds, and produce a relevant graphic and report. It creates maps that make sense of social media, and can be used to conclude powerful results from data that may otherwise seem

a huge collection of meaningless data. Michael Lieberman [11] has explained the clusters found in a twitter social graph that can be identified using Nodexl. They are of the following 6 types:

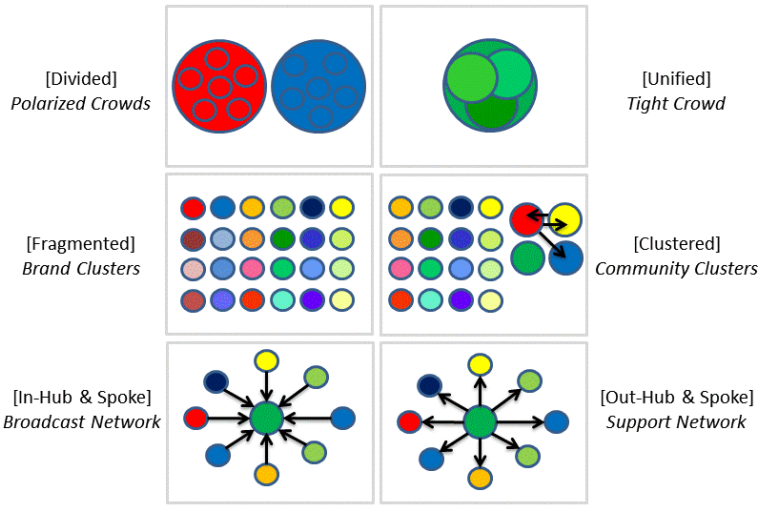

Fig 4: clusters in a twitter social graph [11]

- Polarized crowds emerge when two groups are divided on their opinion about an issue. It is most often seen in political topics.
- In group Network is seen in a tightly knit group of people like those in a conference or a bunch of classmates. This type of network rarely ventures outside its membership.
- Brand Clusters are created when people using the same brand share information about the product. It can be used by the manufacturer to reach out to his customers.
- Community Clusters are medium sized discussions about a topic of common interest. For example people discussing their opinions on a political figure running for elections.
- In a Broadcast network an individual account dominates the map. For example the official Twitter account of Prime Minister Narendra Modi.
- Support Networks are good at customer service. They discuss the solution to the problems faced by customers.

This is important as understanding the type of crowd we are dealing with helps us understand the value of the data. For e.g., If data comes through the Broadcast network, it has utmost importance and is authenticate.

Dr. Sanjay Silakari and others in their paper [12] has identified the various methods of machine learning to analyze the collected data. The main method that is used is predictive analysis. Using this we can predict the future outcome based on analyzing data collected previously. It includes two phases:

1. Training phase: Learn a model from training data.
2. Predicting phase: Use the model to predict the unknown or future outcome

The various models that can be used for predictive analysis are

- Linear regression has the longest, most well-understood history in statistics, and is the most popular machine learning model. It is based on the assumption that a linear relationship exists between the input and output variables
- Logistic Reasoning: In this classification problem, the output is binary rather than numeric. We can imagine doing a linear regression and then compressing the numeric output into a 0..1 range using the logit function 1/(1+e-t).
- Bayesian Network and Naïve Bayes From a probabilistic viewpoint, the predictive problem can be viewed as a conditional probability estimation; trying to find Y where P(Y | X) is maximized. From the Bayesian rule, P(Y | X) == P(X | Y) * P(Y) / P(X). It is equivalent to finding Y where P(X | Y) * P(Y) is maximized. Let's say the input X

contains 3 categorical features— X1, X2, X3. In the general case, we assume each variable can potentially influence any other variable. Therefore the joint distribution becomes: P(X | Y) = P(X1 | Y) * P(X2 | X1, Y) * P(X3 | X1, X2, Y).

» K nearest neighbors: A contrast to model-based learning is K-Nearest neighbor. This is also called instance-based learning because it doesn't even learn a single model. The training process involves memorizing all the training data. To predict a new data point, we found the closest K (a tunable parameter) neighbors from the training set and let them vote for the final prediction.

III. CONCLUSION

We have made a thorough study on various techniques and steps used in analysing the credibility of data present on various OSNs. The analysis of specialized literature helps us to examine the credibility and truthfulness of data available on an OSN. We would like to continue our study by applying these techniques for twitter, one of the most popularly used social networking site where thousands of people post their tweets every day and hence ensuring its credibility becomes an essential.

ACKNOWLEDGEMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

References

1. Sushil Jajodia, Zi Chu, Steven Gianvecchio and Haining Wang "Who is Tweeting on Twitter: Human, Bot, or Cyborg?" ,In Proc. 26th Annual Computer Security Applications Conf. (ASAC), 21–30
2. Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou "Practical Recommendations on Crawling Online Social". IEEE journal on selected areas in communications, Vol. 29, No. 9, October
3. Changhyun Byun, Hyeoncheol Lee, Yanggon Kim, Kwangmi Ko Kim "Automated Twitter Data Collecting Tool and Case Study with Rule-based Analysis" In Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services (IIWAS '12). pp. 196-204.
4. Xinyue Wang, Laurissa Tokarchuk, Félix Cuadrado and Stefan Poslad " Exploiting Hashtags for Adaptive Microblog Crawling" ,In ASONAM.
5. Mohd Fauzi bin Othman,Thomas Moh Shan Yau "Comparison of Different Classification Techniques Using WEKA for Breast Cancer" ,IFMBE Proceedings Vol.15, 2007, pp.520-523.
6. Min Song, Meen Chul Kim, Yoo Kyung Jeong "Analyzing the Political Landscape of 2012 Korean Presidential Election in Twitter" Intelligent Systems, IEEE Volume 29 , Issue 2 , Mar.-Apr. 2014
7. Shamanth Kumar, Fred Morstatter, Huan Liu "Twitter data analytics" Springer Publications, New York, NY, USA
8. Caroline Haythornthwaite, Maarten de Laat "Social Networks and Learning Networks: Using social network perspectives to understand social learning", Paper presented at the 7th International Conference on Networked Learning, Aalborg, Denmark.
9. Oshini Goonetilleke, Timos Sellis, Xiuzhen Zhang, Saket Sathe "Twitter Analytics: A Big Data Management Perspective", ACM SIGKDD Explorations Newsletter - Special issue on big data archive, Volume 16, Issue 1, June 2014, Pages 11-20, DOI: 10.1145/2674026. 2674029
10. Hansen, Derek, Ben Shneiderman, and Marc A Smith "Analyzing Social Media Networks WithNodeXL: Insights from a connected world" (2001).
11. Michael Lieberman "Visualizing Big Data: Social Network Analysis" , In Digital Research Conference (2014)
12. Dr.SanjaySilakari, N Mishra "Predictive Analytics: A Survey, Trends, Applications, Oppurtunities& Challenges", International Journal of Computer Science and Information Technologies, vol. 3, no. 3, pp. 4434-4438, 2012.

