ISBN: 978-1-63248-075-0 doi: 10.15224/ 978-1-63248-075-0-45

Political Trends Analysis in Social Networks Using Hidden Markov Model

Nazish Rafique, Zainab Nayyar, Khurram Mahmood

***Abstract*---- The main purpose of analyzing the social network data is to observe the behaviors and trends that are followed by people. How people interact with each other, what they usually share, what are their interests on social networks, so that analysts can focus new trends for the provision of those things which are of great interest for people so in this paper an easy approach of gathering and analyzing data through keyword based search in social networks is examined using NodeXL and data is gathered from twitter in which political trends have been analyzed and then its statistical calculation is done by applying hidden markov model on the data. As a result it will be analyzed that, what people are focusing most in politics.**

***Keywords*-- social networks, keyword searching, NodeXL, twitter, hidden markov model.**

I. Introduction

As the social networking media has introduced, it has evolved many new changes and trends to the society. It includes Facebook, twitter, LinkedIn, MySpace, IM, Skype, YouTube and many more. Now it has become an essential of life, people used to share their each and every activity with their friends through social networks. People connect with them through social networks. People not only used it for entertainment but it is also a source for public relationing, business activities [1] and outsourcing. It keeps us up to date from the most recent happenings. In short it reduces the distances and deepens the relationships by enhancing the communications.

With the passage of time the analysts started taking interest in analyzing this large web traffic [2] which has circulated due the emergence of social networks. By this analysis several things become unveiled like what the interests of people, their likings and disliking, which culture, trends and traditions are mostly followed by this social community, what are the basic and major problems related to health, education, defense etc. So for the analysis two approaches are followed user specific search and data specific search [3]. The former is used to gather data from users' perspective i.e. from the users' friend added on the social network and the latter is based on the general data that we get by querying through keywords.

There are several approaches for data tracking like self-involved approach, topic-based approach, actor based approach, random/exploratory approach and url-based approach. In self-involved approach when individuals wanted to know that how people think about them this can be analyzed by gathering tweets from twitter or to get comments, posts and pages from Facebook. Topic-based approach is followed when people wanted to track down opinions about them. In actor-based approach people analyze that how much an individual is influential in the society whereas in random/exploratory approach content is gathered and analyzed randomly. URL approach deals with the content hide behind the urls or hyperlinks [4]. Hidden markov model basically used to find out the probabilities on the data which gathered from different sources and to find out the hidden aspects of the data. In this paper political analysis is done on twitter and data is gathered which is then statistically analyzed by applying hidden markov model on the data.

II. Related Work

The literature review of this paper includes other concepts and methodologies that have been used for gathering data from social networks.

Vizster system [1] in 2005 gave a visual interface through which not only data could be analyzed from social networks but also it could be observed and monitors in the form of visuals. Vizster showed the graphical visuals of the data which was analyzed and browsed from the social network media.

Top-k query processing algorithm was applied in [3]. This algorithm gathers data in two dimensions first on the basis of tags and second on the basis of relationships among users. It is based on top-k threshold algorithm which set a threshold on the basis of which it made the keyword search. Furthermore in [15] a keyword based query was used over the SQL relational databases. Top-k best mapping approach was applied on tables, attributes and values. So that the answers generated from the relational databases was based upon keyword approach. This method was then called as a KEYRY approach. It also included the hidden markov model to find out the statistical analysis on the data found through keywords.

In [5], the author gathered the data by applying data mining techniques on instant messaging service on MSN, the search has been made in two ways, first the data which is gathered is based on user to user chats because the people who chat with each other mostly share their interests with each other, second the data is gathered on the basis of keywords which were entered on MSN search engine for searching different things. The author further applied Bayes' rule on the gathered data for calculating the probabilities.

In [6] and [11] referral web system was built to search reconstruction and analyze the social networks. As a result referrals were generated containing the results gathered from social networks. A social network is analyzed by a graph in which nodes represent individuals and edges represent the relationship between the individuals.

Social network query language was presented in [2] [7] for querying data from social networks. This language was very similar to SQL language as it gave a complete path of data storage. Its queries syntax also resembled to that of SQL and represented data in the form of tables just like in SQL. Similarly SNQL in [12] was discussed; these query languages were designed by following the mechanism used in SQL language.

Socio spatial graphs were used in [8] because the author combined both the data gathered from frequency and social networks. GSM and GPS were used for analyzing the user location and history over the mobile network. For storing the data socio spatial network algebra was designed for querying data. Data was also stored on graph databases (Neo 4j). It was observed that data storage on relational databases created a lot of redundancies but on graph databases no redundancies were observed.

Flink system [9] was used to analyze the social networks online as it is a web based system. This system was helpful in finding the common interests, mutual friends and new friendships. Searching using Flink is made on the basis of health, education and relationships in the mentioned paper.

In [10] a twitter API was built to gather a data on the basis of keywords. There are several API's of twitter like streaming api, rest api and search api but rest api was used in that paper to get the data from twitter. Yioop search engine was also considered in that paper from which data gathered from social networks by just inserting appropriate keywords. As on Google data specifically from social networks was not readily available so the architecture of yioop is specially designed for getting data from the social networks.

In [13] and [14] the NodeXL software was used with Microsoft excel 2007, 2010 and 2013. It had an add-on feature which imports data from the social networks on the basis of given keywords.

The problem highlighted in [16] was that it was a big problem to apply individual web based search methods on collaborative web based search. So an automated technique named hidden markov model was used to apply collaborative searches on linear or temporal data. This approach was equally helpful in determining the search on individual web based search.

Coupled hidden markov model in [17] was used to cater with the problem of mentioning user profile activities along with their influence on their network structure. In hidden markov model each user was modeled as a hidden chain and the coupling between hidden chains showed the influence of users' network. The whole method was implemented on the twitter and the results shown the accuracy of the usage of coupled hidden markov model for getting data from social networks.

As the web moved to the decentralized approach the point of failure became very little and it became more scalable. This decentralization has improved the communication aspects among the people around the globe. Semantic search was the basic area of research which was rose due the initiation of decentralization. For this knowledge management solutions [18] were used which calculated the strengths of the users. However hidden markov model was applied to do the statistical calculations on the gathered data so that the hidden aspects from the gathered data could be observed.

Graph coupled hidden markov model [19] was used to model the spread of different infections via social networking. In that research mobile data of 84 people was gathered. The graphical coupled hidden markov model was the extended form of coupled hidden markov model, in which dependencies were created among the state and transitions of hidden markov models. The graphical coupled hidden markov model was applied when data changed with time. This sort of search was helpful in gathering the data about infectious diseases at individual level.

In [20] the broadcasted news stories were segmented the main reason behind this approach was the concept that the people which interact with each other are more likely to share their interests. Then the hidden markov models are used to map these segmented news stories over the users.

III. METHODOLOGY

A. Analyzing the political trends on Twitter:

To analyze the political trends in social networking websites, we have chosen ***twitter*** among them. For the purpose of data acquisition, we are using NodeXL Graph (Network Overview, Discovery, and Exploration for Excel).

In our framework, the first step is to acquire data from twitter. For this purpose, we have used the NodeXL data importer. There are numbers of sources available through which we can gather data on the network from popular network services available in the data import menu of NodeXL. As shown in figure 1.

Proc. of The Second Intl. Conf. On Advances in Applied Science and Environmental Technology - ASET 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-075-0 doi: 10.15224/ 978-1-63248-075-0-45

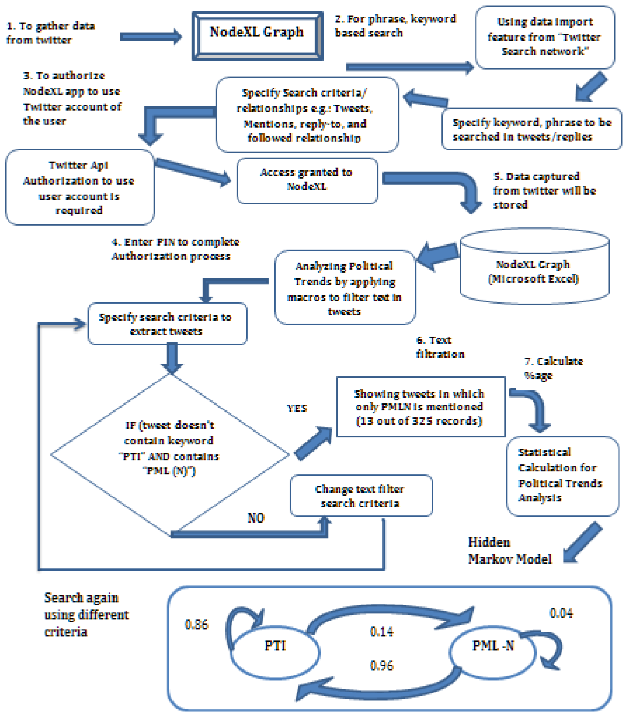

1. To gather data from twitter
NodeXL Graph
2. For phrase, keyword based search
Using data import feature from "Twitter Search network"
3. To authorize NodeXL app to use Twitter account of the user
Specify Search criteria/ relationships e.g.: Tweets, Mentions, reply-to, and followed relationship
Specify keyword, phrase to be searched in tweets/replies
Twitter Api Authorization to use user account is required
Access granted to NodeXL
5. Data captured from twitter will be stored
4. Enter PIN to complete Authorization process
Analyzing Political Trends by applying macros to filter text in tweets
NodeXL Graph (Microsoft Excel)
Specify search criteria to extract tweets
6. Text filtration
7. Calculate %age
IF (tweet doesn't contain keyword "PTI" AND contains "PML (N)")
YES
Showing tweets in which only PMLN is mentioned (13 out of 325 records)
Statistical Calculation for Political Trends Analysis
Change text filter search criteria
NO
Hidden Markov Model
Search again using different criteria
0.86
PTI
0.14
0.96
PML -N
0.04

Fig. 1 work flow of extracting tweets.

We can search the twitter data on the basis of two networks:

i) **From twitter user's network:** Which imports twitter data on the basis of user's followings.

ii) **From twitter search network:** Which imports twitter data on the basis of keywords, search terms which are involved in "tweets", "reply-to" and "mention" relationships.

We have selected the second option because the focus of our research is **keywords based search**. So, the second step is to specify keyword of search term which should be included in tweets or replies. This will add a vertex for each person who tweeted this search term or who was replies to or mentioned in those tweets. Search can be done for any string of characters which may also include operators like "OR". The limit of search tweets is up to 18,000 which can rarely be achieved because of age limit of twitter messages. We have specified the relationship/edge which is used to describe the connection between two twitter users formed by following, replying or mentioning one another. Moreover, another column is added to show the tweets of users which include the specified keyword(s).

After specifying the search criteria the next step is the authorization from twitter API to use user account to get the data by NodeXL, for this, the user will be directed to twitter authorization webpage. A PIN code will be shown which requires to be copied in NodeXL graph, after that access will be granted to this app to get the data. The NodeXL Search network starts its data collection by performing a query against twitter search service at http://search.twitter.com

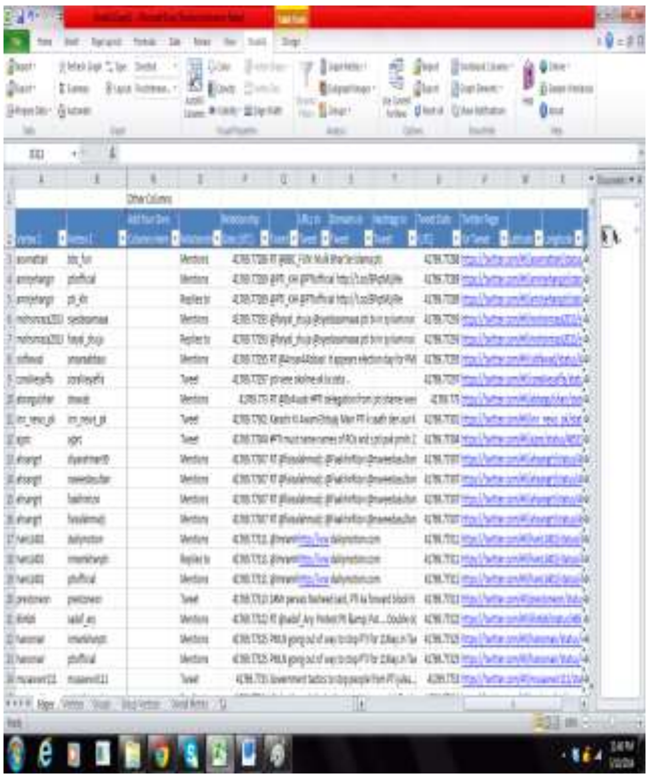
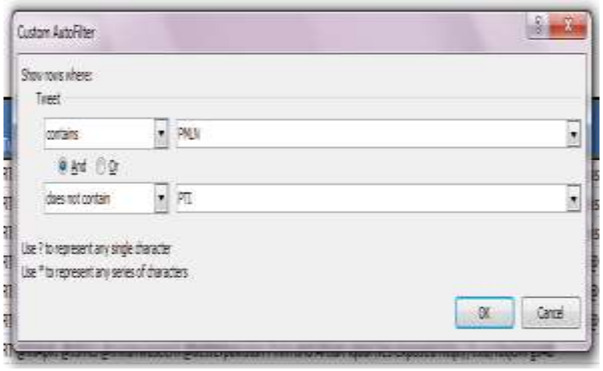


Fig. 2 NodeXL displaying connections between Twitter Users who posted a tweet containing the search term

As shown in the figure above, Data will be acquired showing the vertices which indicates the usernames, description about them, tweets, followers, followings, number of total tweets, tweets URLs, top domains in tweets, top hash tags in tweets, top words in tweets, top words in tweets and many other properties fields are available. Each "edge" between vertices represents different kind of relationships created through twitter. NodeXL creates four different types of Twitter edges from which it collects data: follows replies, mentions and tweet.

Next step is to analyze the gathered data on the basis of keyword based approach by applying text filter on tweets. For this, we have created a Macro to record/control the text filtering. To analyze the political trends in Pakistan we have considered two political parties e.g.: PTI and PML (N), To find out that which political party is more focused in conversations and interest of people. Different search criteria have been applied to filter tweets on the basis of given keywords "PTI" and "PML (N)". As shown in the figure below:



Custom AutoFilter

Show rows where:

Tweet

contains PMLN

● And ○ Or

does not contain PTI

Use ? to represent any single character

Use * to represent any series of characters

Fig. 2 Tweets filtration criteria using macros

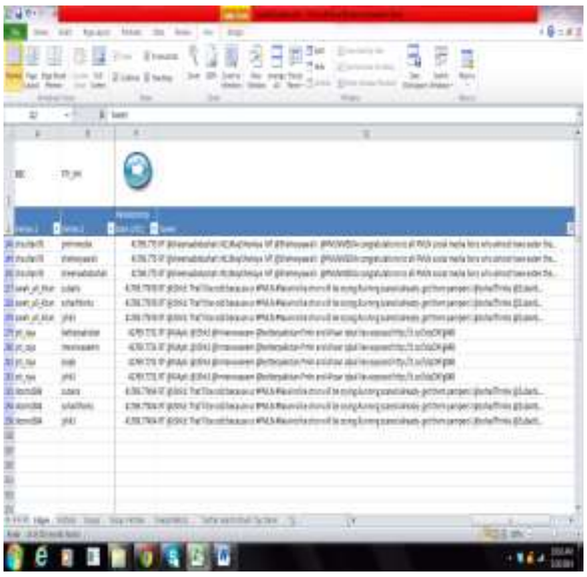


Fig.3 extracted tweets containing the keyword "PML (N)" only.

On the basis of records found, it has shown that out of 325 records (which were stored before in NodeXL); there are only 10 tweets in which people have shared their views about PML (N). As shown in fig.3.

These results are helpful to analyze the top conversations on political parties on twitter and to explore the political trends. From these results, we can predict the future probability of tweets about both parties. As it is known that transition matrix:

$$\begin{bmatrix} X & 1\text{-}x \\ 1\text{-}x & x \end{bmatrix}$$

Expression -------- 1

According to the analysis:

PTI= 85.5%

PML-N= 4%

By dividing these values by 100, the values are converted into decimal point values. Now, the values are placed in the transition matrix the missing values which are showing the transitions in the trend can get by subtracting them from 1 as shown in expression 1;

Transition Matrix =

$$\begin{array}{c c} & \begin{array}{cc} \textbf{PTI} & \textbf{PML-N} \end{array} \\ \begin{array}{c} \textbf{PTI} \\ \textbf{PML-N} \end{array} & \begin{bmatrix} 0.86 & 0.14 \\ 0.96 & 0.04 \end{bmatrix} \end{array}$$

On the basis of above mentioned matrix, the transitions which were hidden are known and a hidden markov model is designed on the basis of the values extracted from the matrix

In figure 4, it is shown that 86% of tweets are about PTI and there are 14 % chances that the people can involve in tweets related to PMLN. 4% tweets are about PMLN and there are 96% chances of moving to PTI.

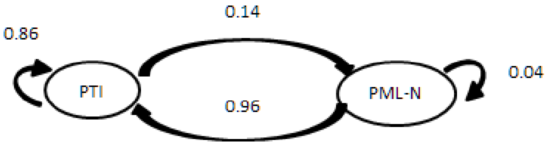


Fig. 4 hidden markov model

IV. Results

For the extraction of tweets to know the political trends on twitter, 325 tweets were gathered on the basis of keywords e.g.: PTI, PMLN. We have analyzed that 4% of people have conversation involving PML (N) only and 85.5% of people have conversation involving PTI only.

V. Conclusion

In this paper, we have discussed keyword based approach to analyze the political trends on twitter. For this, NodeXL was used among other available tools and API's because it provides an opportunity to both novice and experts to analyze the trends on social networking websites more easily. As it requires less effort for the configuration and preparation to collect data from twitter. Moreover with the use of simple Macros, we can control the text filtering to extract tweets which involves the required keywords.

VI. References:

[1] Jeffrey Heer and danahboyd (2005). "Vizster: Visualizing Online Social Networks." *IEEE Symposium on Information Visualization (InfoVis 2005)*. Minneapolis, Minnesota, October 23-25.

[2] Sara Cohen, LiorEbel and Benny Kimelfeld. A Social Network Database that Learns How to Answer Queries.

[3] Ralf Schenkel, Tom Crecelius, MounaKacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Gerhard Weikum. Efficient Top-k Querying over Social-Tagging Networks. ACM. July 20–24, 2008.

[4] Stefan Stieglitz and Linh Dang-Xuan. Social media and political communication: a social media analytics framework. 25-August-2012. Springer.

[5] Parag, Singla.Matthew, Richardson. Yes, there is a Correlation - From Social Networks to Personal Behavior on the Web. WWW-08 (pp. 1 - 7).

[6] Henry Kautz, Bart Selman, and Mehul Shah (1997).
ReferralWeb: Combining Social Networks & Collaborative Filtering. *Communications of the ACM*, 30 (3) 1997, volume 40, pp. 27-36.

ISBN: 978-1-63248-075-0 doi: 10.15224/ 978-1-63248-075-0-45

[7] Royi Ronen, OdedShmueli: SoQL: A Language for Querying and Creating Data in Social Networks. ICDE 2009: 1595-1602.

[8] YerachDoytsher, Ben Galon, YaronKanza. Querying Geo-social Data by Bridging Spatial Networks and Social Networks. Pages 39-46, **GIS** Geographic Information System ACM New York. 2010-11-02.

[9] Peter Mika. Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. Elsevier Science journal. 14 May 2005.

[10] VijethPatil. Keyword Search in Social Networks. 4-1-2012.

[11] Bin Yu and Munindar P. Singh. Searching Social Networks. Pages 65-72. AAMAS 2002 Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA) ACM. July 14–18, 2003.

[12] Mauro San Martin, Claudio Gutierrez and Peter T. Wood. SNQL: A Social Network Query and Transformation Language.

[13] Hansen, D., Shneiderman, B., & Smith, M. (2010). Analyzing Social Media Networks with NodeXL: Insights from a connected world.

[14] Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A. &Gleave, E. (2009),"Analyzing (Social Media) Networks with NodeXL", In C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies. Springer.

[15] Sonia Bergamaschi, Francesco Guerra, Silvia Rota and Yannis Velegrakis. A Hidden Markov Model Approach to Keyword-based Search over Relational Databases. Pages 411-420. ER'11 Proceedings of the 30th international conference on Conceptual modeling. 2011-10-31.

[16] Zhen Yue, Shuguang Han and Daqing He. Modeling Search Processes using Hidden States in Collaborative Exploratory Web Search. Pages 820-830. 17th ACM conference on Computer supported cooperative work & social computing. 2014-02-15.

[17] Vasa than Raghavan, Greg ver Steeg, Aram Galstyan and Alexander G. Tartakovsky. Coupled Hidden Markov Models For User Activity In Social Networks. Pages 1-6. IEEE International Conference on Multimedia and Expo Workshops (ICMEW). 2013.

[18] Shimaa M. El-Sherif, Armin Eberlein and Behrouz Far. Calculating the Strength of Ties of a Social Network in a Semantic Search System Using Hidden Markov Models. Pages 2755 – 2760. IEEE International Conference on Systems, Man, and Cybernetics (SMC). 9-12 Oct 2011.

[19] Wen Dong, Alex Sandy Pentland and Katherine A. Heller. Graph-Coupled HMMs for Modeling the Spread of Infection. Twenty-Eighth Conferences on Uncertainty in Artificial Intelligence. 16 Oct 2012.

[20] Alessandro Vinciarelli and Sarah Favre. Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models. Pages 261-264. The 15th international conference on mealtime