News Rank: Ranking News Topics based on Social Media Factors

Malathi Kulkarni, Vishwanath.R.Hulipalled

Abstract—All media sources, particularly the news media have educated the everyday news. These days, internet based media like twitter gives us an immense amount of information that is created by client, which potentially contains news related information. These sources to be helpful we should remove undesirable information and concentrate just the information which is like the news media. Indeed, even though the unwanted information can still exist, so which is vital to give need to its usage. For this prioritization, information must be positioned utilizing three components. First Media Focus(MF) of the Topic which principally centers around both internet based life and news media, Next User Attention(UA) which depends on clients' interests and User Interaction(UI), which is on how client responds to that specific topic. This is an unsupervised framework NewsRank--- which find the news topics which is applicable in both news media and internet based life and after that ranking the news topics utilizing degree of three elements.

Keywords—Media focus, Prioritization,Unsupervised,,User Attention,User Interaction

I. INTRODUCTION

The extracting of data from resources has become leading research area in Information Technology nowadays. From the past, daily events has been provided by media that is news media. In recent days many news media have left the hard copy publications and started publishing through World or now there is paper copy as well as websites. The information from the news media are reliable as they are verified and published by professional journalists, where as in social media , the information is unverified and users are able to publish their own interest who are non journalists.

In social media nowadays Micro blogs are the popular outlets. For example, twitter which is used by huge number of people throughout the world, which gives huge amount of data which is generated by user. One can say that this source strongly contains the data which is similar of more valuable than the news media and also people assume that, it is unverified data , so content is useless or unnecessary. For topic identification in social media data we must first remove the unnecessary data and capture only the data which is similar to news media then it can be said that it is more valuable and useful.

The news media provides us the verified data about the daily events by professional journalists where as social media focuses on user interest in particular areas. Twitter also gives us the additional information on specific news media topic. Even when the filtering of the noise data, there may be some content overloaded in the rest of the news related data, which must be prioritized for utilization.

To help in the ranking of news points, it should be prioritized by determined significance. The temporal predominance of a specific theme shows that verified media news has covered the topic widely, which is an important factor when calculating the relevance of the topic. This factor is called MF of the topic. Twitter shows us the popularity of the topic in which the users express their interest, this factor is referred as UA of the topic. Similarly the topics discussed by users and interaction between them provides us an insight into topical importance. This factor is called User interaction. Combination of these three factors, we get insight into topical importance and then rank the news topics.

Merged, separated, and positioned news points from both expert news suppliers and people have a few advantages. The clearest benefit of the system is possibility to enhance the status and inclusion of appropriate model for suggesting the daily events that are trending, or online news encourages, including client fame criticism. Also, news subjects that maybe were not seen as prominent by the broad communications could be revealed from social media and given more inclusion and need. For example, a specific story that has been suspended by news suppliers could be given resurgence and proceeded in the event that it is as yet a well known theme among informal communities. This data, thusly, can be separated to find how specific points are talked about in various geographic areas, which fill in as input for organizations and governments

The remaining sections are structured as, Section II contain the related work of Topic identification and other research topics, and Section III contains the system architecture, framework of this model and its stages. Section IV contains the modules descriptions what are the methods used. Section V describes results and discussion and Conclusion in section VI.

II. RELATED WORK

In this paper the main research areas are :Topic recognition, Topic Grading ,Analysis of Social Network, Extraction of

Revised Manuscript Received on April 25, 2019.
Malathi Kulkarni, REVA University, India.
Vishwanath.R.Hulipalled, REVA University, India.

Keywords, Co- occurrence similarity measures , and graph clustering.

A. *Topic Recognition*

In Topic recognition there are many works some of them are LDA[1] and *PLSA[2][3]* called as Topic Modeling these are the two methods for topic detection. *LDA* and *PLSA* only identifies topics and do not rank the news topics based on popularity or prevalence.

Wartena and Brussee [4] proposed a strategy to identify news points by grouping watchwords. This system involves the grouping of watchwords using k-bisecting clustering algorithm. Cataldi *et al*. [5] implemented a method on news points detection in which it gives the trending topics in real time from online social media like twitter using the novel aging theory. Zhao *et al*[6] proposed a method by implementing a Twitter LDA which is to recognize the news points from tweets. This method focuses on personal interest of individual user.

B. *Topic Grading*

Wang *et al.*[7] developed a system that mainly focuses on interests of users on a particular news topic based on the number of times they visited that websites and read the news . This is called as UA factor. An aging theory is developed by Chen *et al.*[8] proposed an energy function which is mainly on the duration of the topics maintained by using an energy function. It is mainly based on creating and destroying the news.

Many other works on Twitter has been developed by Sankaranarayanan *et al.*[9] called TwitterStand which identifies breaking news on twitter. Shubhankar *et al*.[10] proposed a model that identifies and prioritize the main points of research paper and PageRank [11]Algorithm to rank them.

C. *Analysis of Social Network*

Kwan *et al*[12] proposed a method called reciprocity which detects the interaction between social media users on particular topic. This method is based on the methodology on which the reciprocity increases and also importance also increases.

D. *Extracting Keywords*

There are some analytical standards in unsupervised methods on term informativeness like term specificity, TFIDF ,Word frequency , n-grams , and word co-occurrence. Supervised methods like KEA and GenEx used extracting keywords. One more method TextRank [13] is used extract keywords from news media.

E. *Co-Occurrence Similarity*

Matsuo and Ishizuka[14] proposed a method of co-occurrence relationship between word pairs from a document.Chen *et al* [15] developed a method called novel co-occurrence similarity measures. This measure is known as co-occurrence double checking(CODC).One more method that uses page counts which is developed by Bollegala *et al*[16] to measeure the similarity between words.

F. *Graph Clustering*

In this paper Graph Clustering is used identify and separate TC's Topic clusters [4]. Matsuo *et al*[17] proposed a method for clustering of co-occurrence graphs. Newman Clustering [18] is used identify word clusters. In graph clustering Algorithm the concepts of betweenness and transitivity are used.

III. METHODOLOGY

The main objective of this proposed work is to find out and prioritize the news points describes in online social sites like twitter and news media. The system architecture shown in the Fig. 1. They are four main stages in this system.

1) *Preprocessing:* In first stage extraction of key terms in online social network as well as news media is carried out between given particular time period.
2) *Key Term Graph Construction:* A graph is constructed using key terms which is extracted in past where vertices and edges are are the key words and co-occurrence similarity between them respectively. After the processing graph contains topic clusters that are trending in both social and news media.
3) *Grouping of Graphs:* The Graph is then grouped to get very much determined separate TCs.
4) *Information Selection and Prioritizing:* In this stage the obtained TCs chosen and prioritized using MF,UA,UI components.

IV. PROPOSED SYSTEM

We implemented a strategy which is unsupervised—NewsRank efficiently recognizes news points that are present in online networking as well as news media, and after that positions them by degree of MF, UA, and UI. This System focuses on news topics, it is very easy to use in a various themes, like politics technology and science to culture and sports. There is no other work has been implemented that focus on user interests or the social relationships for ranking of topics. This work has the stages like pre-processing, keyword extraction , Similarity of topics between social media and news media.

V. MODULES DESCRIPTION

1) Uplaod Excel file

In the Upload Excel File Module, user has to select the file from the client machine which contains Tweets as well as media News and the file content will be sent to the server via URL in the form of

multipart, in the server side servlet receives the file content and write the file content in the folder of the application. From that folder it reads the file content and store the file content in to the database.

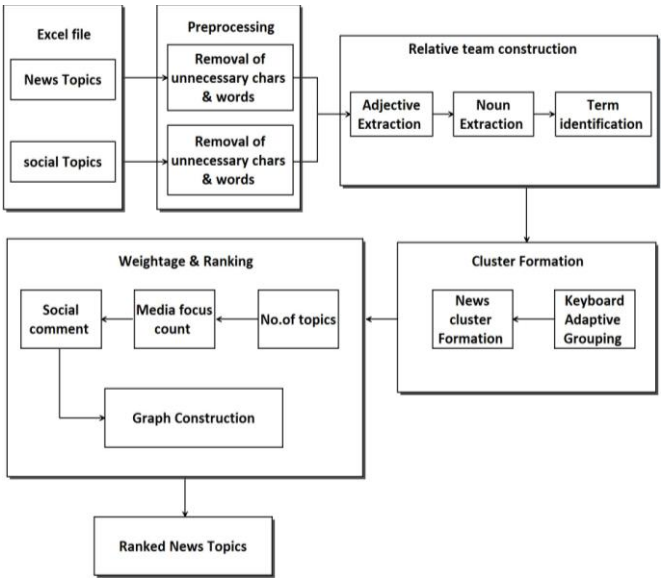
2) Process the tweets data and Media news

a) *Standford POS(Parts of speech) Tagger:* Using Stanford POS tagger, it is type in which every word is attached following with its well-formed activities. Noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection are main parts of speech in english. This method is used to find the nouns from sentences which is considered as Terms in our project.

b) *N-gram Technique:*This methodology is used to find the co-occurrence of the words in the sentences of tweets as well as media news and the Outlier detection.We are implementing two gram and three gram techniques.

c) *Cosine similarity:*This methodology is used to find the similarity between the sentences. If the cosine value of two sentences is 1 means, those are 100% similar, if it is 0.98 means 98% similar, this is useful to find that where the sentences related to the same terms.

d) *Group Clustering:*This methodology is used to create the Clusters with respect to the terms from the tweets as well as media news. By this methodology we will get the count of tweets and media news which laid in the cluster, by that we can achieve the Media Focus(MF) and User Interaction(UI).

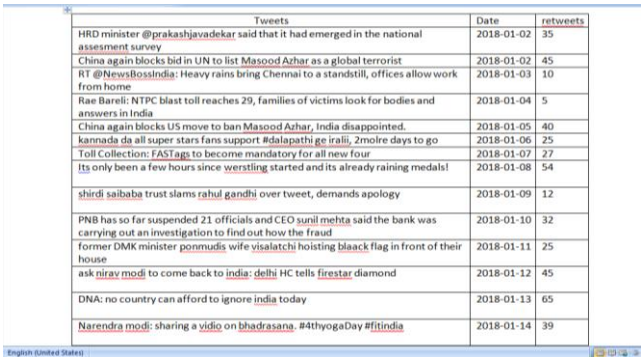


Fig,1 System Architecture

VI. RESULTS AND DISCUSSION

The dataset contains Tweets from twitter and news collected from various online sites like abcnews.com, timesofindia.com bbc.com from time period of January 1,2018 and April 9, 2018. The news websites are abcnews.com timesofindia.com bbc.com.

Uploading the excel file which contains tweets and another excel file which contains news articles. Then providing from and to dates to process (d_1 and d_2).



Tweets	Date	retweets
HRD minister @prakashjavadekar said that it had emerged in the national assesment survey	2018-01-02	35
China again blocks bid in UN to list Masood Azhar as a global terrorist	2018-01-02	45
RT @NewsBossIndia: Heavy rains bring Chennai to a standstill, offices allow work from home	2018-01-03	10
Rae Bareli: NTPC blast toll reaches 29, families of victims look for bodies and answers in India	2018-01-04	5
China again blocks US move to ban Masood Azhar, India disappointed.	2018-01-05	40
kannada da all super stars fans support #dalapathi ge iralii, 2molre days to go	2018-01-06	25
Toll Collection: FASTags to become mandatory for all new four	2018-01-07	27
Its only been a few hours since werstling started and its already raining medals!	2018-01-08	54
shirdi saibaba trust slams rahul gandhi over tweet, demands apology	2018-01-09	12
PNB has so far suspended 21 officials and CEO sunil mehta said the bank was carrying out an investigation to find out how the fraud	2018-01-10	32
former DMK minister ponmudis wife visalatchi hoisting blaack flag in front of their house	2018-01-11	25
ask niray modi to come back to india: delhi HC tells firestar diamond	2018-01-12	45
DNA: no country can afford to ignore india today	2018-01-13	65
Narendra modi: sharing a vidio on bhadrasana. #4thyogaDay #fitindia	2018-01-14	39

Figure 2 Tweets extracted from twitter

Figure 2. Shows the tweets that are extracted from Twitter and Dates and how many times the tweets are retweeted.

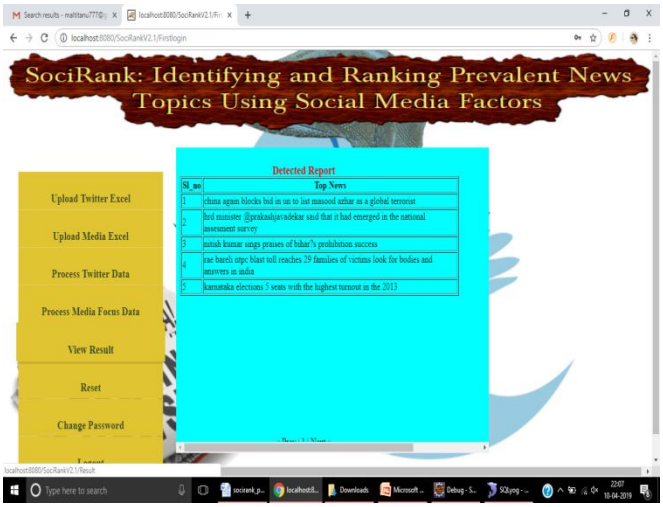


Fig 3. Detected Report

Figure 3 Shows the result that we got after processing the twitter data and media focus data. Top 5 Ranks From 1 January 2018 to 1 February 2018.

VII. CONCLUSION

For the above described proposed work we implemented an unsupervised strategy for identifying points in news from both online site like Twitter and the news media and ranking them based on their MF,UA,UI factors. As of now no work has been done which focus on the users interest and their social relationship. This system also gives report of news pints which are terminated by all the medias , while somel people are discussing about that topic. This method can also be used in main topics like science, sports, politics, technology and many more.

REFERENCES

[1] D.M. Blei, A. Y. Ng, and M. I. Jordan , "Latent Dirichlet allocation", j. Mach. Learn. Res., vol. 3. Pp. 993-1022, jan 2003

[2] T. Hofmann , "Probablistic laten semantic Analysis" in proc. 15th conf. uncertainity Artif.intell., 1999,pp. 289-296.

[3] T. Hofmann , "Probablistic laten semantic Analysis" in proc.22nd Annu. Int. ACM SIGIR conf. res. Develop. Inf. Retrieval , Berkeley,CA. USA,1999,pp.50-57.

[4] C. Wartena and R. Brussee, "Topic Detection by Clustering Keywords", in proc. 19th Int. Workshop Database Expert Syst. Appl.(DEXA),Turin, Italy. 2008, pp. 54-58.

[5] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proc. 10th Int. Workshop Multimedia Data Min. (MDMKDD), Washington, DC, USA, 2010, Art. no. 4. [Online]. Available: http://doi.acm.org/10.1145/1814245.1814249.

[6] W.X. Zhao et al., "Comparing Twitter and traditional media using topic models," in Advances in information retrieval . Heidelberg, Germany: Springer Berlin Heidelberg, 2011 , pp.338-349.

[7] C. Wang M Zhang. L.Ru and S.Ma, "Automatic online news topic ranking using mediafocus and user attention based on aging theory." In proc 17th conf. Inf .Knowl.Manag., Napa County, CA,USA,2008, pp.1033-1042.

[8] C.C Chen, Y-T. Chen, Y.Sun, and M.C. Chen, " Life cycle Modeling of news events using aging Theory," in Machine Learning. ECML 2003.Heidelberg, Germany: Springer Berlin Heidelberg, 2003, pp. 47-59

[9] J. Sankaranarayanan , H. Samet, B. E. Teitler , M. D. Lieberman, and J. Sperling, "Twitterstand: News in Tweets," in proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst., Seattle, WA, USA, 2009, pp 42-51.

[10] K. Shubhankar , A.P.Singh , and V.Pudi, " An efficient algorithm for topic ranking and modeling topic evolution ," in Database Expert Syst. Appl., Toulouse,France,2011, pp, 320-330

[11] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," Comput. Netw., vol. 56, no. 18, pp. 3825–3833, 2012.

[12] E. Kwan, P.-L. Hsu, J.-H. Liang, and Y.-S. Chen, "Event identification for social streams using keyword-based evolving graph sequences," in Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min., Niagara Falls, ON, Canada, 2013, pp. 450–457.

[13] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proc. EMNLP, vol. 4. Barcelona, Spain, 2004.

[14] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," Int. J. Artif. Intell. Tools, vol. 13, no. 1, pp. 157–169, 2004.

[15] H.-H. Chen, M.-S. Lin, and Y.-C. Wei, "Novel association measures using Web search with double checking," in Proc. 21st Int. Conf. Comput. Linguist. 44th Annu. Meeting Assoc. Comput. Linguist., 2006, pp. 1009–1016.

[16] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using Web search engines," in Proc. WWW, Banff, AB, Canada, 2007, pp. 757–766.

[17] Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka, "Graph-based word clustering using a Web search engine," in Proc. Conf. Empir. Methods Nat. Lang. Process., 2006, pp. 542–550.

[18] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Nat. Acad. Sci., vol. 99, no. 12, pp. 7821–7826, 2002.