Prioritizing News Topics Based on Social Media Factors

Malathi Kulkarni[1],Vishwanath R Hulipalled[2]

[1] Department of Computer Science and Engineering, REVA University, Bangalore, Karnataka, India
[2] Department of Computer Science and Engineering, REVA University ,Bangalore , Karnataka , India

e-mail: maltitanu777@gmail.com , vishwanath.rh@reva.edu.in

***Abstract*—** All media sources , particularly the news media have educated the every day 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 it 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 separating of information from assets has turned out to be driving examination region in Information Technology these days. From the past, day by day events has been given by media , that is news media. Lately numerous news media have left the printed version distributions and began distributing through World Wide Web or now there is both printed version just as web. The data from the news media are solid as they are confirmed and distributed by expert journalists, where as in online life , the data is unconfirmed and clients can distribute their very own advantage who are non writers.

In online networking these days Micro web journals are the mainstream outlets. For instance, Twitter which is utilized by gigantic number of individuals all through the world, which gives tremendous measure of information which is produced by client. One can say that this source emphatically contains the information which is comparable of more important than the news media and furthermore individuals expect that, it is unsubstantiated information , so content is futile or pointless. For theme recognizable proof in web based life information we should initially expel the superfluous information and catch just the information which is like news media then it tends to be said that it is increasingly profitable and helpful.

The news media gives us the checked information about the every day occasions by expert writers where as internet based life centers around client enthusiasm for specific zones. Twitter likewise gives us the extra data on explicit news media subject. Notwithstanding when the separating of the clamor information, there might be some substance over-burden in the remainder of the news associated data, which must be arranged in a systematic way so that it will be easy for usage.

To help in the ranking of news points,news should be ranked and arranged by determined significance. The fleeting predominance of a specific theme demonstrates that news media

Secured the point generally, which is a significant factor while computing the pertinence of the theme. This factor is called MF of the point. Twitter demonstrates to us the notoriety of the subject in which the clients express their advantage, this factor is alluded as UA of the theme. So also the points examined by clients and communication between them gives us a knowledge into topical significance. This factor is called User connection. Blend of these three variables, we get understanding into topical significance and after that assign rank for the news points.

A clear methodology for recognizing themes in different online networking sites and media news sources is the use of subject demonstrating. Numerous strategies have been proposed here, for instance some algorithms like Latent Dirichlet Allocation (*LDA)* [1] and Probabilistic Latent Semantic Investigation (*PLSA*) [2], [3]. Basically the Point demonstrating is, disclosure of topics from text corpora by grouping regularly frequently occurring words. This methodology, in any case, passes up a major opportunity in the fleeting segment of common point discovery, that is, it doesn't consider how subjects change with time. Moreover, theme demonstrating and other point recognition strategies don't rank subjects as per their position by considering their majority in both news media and web based life.

We implemented a framework which is unsupervised—NewsRank—that adequately distinguishes topics that are

pervasive in online networking as well as news media, and after that positions them by pertinence utilizing their three factors that is Media Focus, User Attention, and User Interaction. Despite the fact that this paper centers around news themes, it tends to be effectively adjusted to a various assortment of themes, from science and innovation to culture and sports. Supposedly, no other work endeavors to utilize the utilization of either the online networking interests of clients or their social connections to help in the positioning of themes. In addition, NewsRank experiences an experimental system, involving and incorporating a few procedures, for example, watchword extraction, proportions of closeness, Graph Clustering, and informal organization examination. The viability of our framework is approved by broad controlled and uncontrolled examinations.

To accomplish its objective, NewsRank utilizes watchwords from known news media sources(specified timeframe) to distinguish the cover with internet based life from that equivalent period. We at that point assemble a diagram whose hubs speak to these catchphrases and whose edges delineate their co-events in online life. The diagram is then bunched to plainly recognize particular themes. In the wake of getting admirably isolated point groups (TCs), the components that mean their significance are determined: MF, UA, and UI. At long last, the themes are positioned by a general measure that consolidates these three variables.

Rest of the paper is structured as follows, Section II contain the Literature Survey on Topic identification and some other research topics, 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 in section VI gives the conclusion.

II. LITERATURE SURVEY

In this paper the main research areas are :Topic identification, Topic Ranking ,Social network Analysis, Keyword extraction , co- occurrence similarity measures , and graph clustering.

A. Topic Identification

In Topic Identification there are many works some of them are LDA[1] and *PLSA[2][3]* called as Topic Modelling 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 System to recognize news topics by Grouping keywords. This sytem involves the grouping of keywords using k-bisecting clustering algorihm. One more method proposed by Cataldi *et al*. [5 on topic detection in which it gives the real-time trending topics from twitter using the novel aging theory. Zhao *et al*[6] proposed a method by implementing a Twitter LDA which is to identify the topics in tweets. This method focus only on personal user interests.

B. Topic Ranking

Wang *et al*.[7] developed a method that mainly focuses on interest of clients in a topic based on the number of times they go through the information on that specific topic. This is known as *UA* factor. An aging theory is developed by Chen *et al*.[8] which is based on the life cycle of the topics which is tracked 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 an algorithm that detects and ranks the topics of research paper and PageRank [11]Algorithm to rank them.

C. Social Network Analysis

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 idea of higher the reciprocity greater the importance.

D. Keyword Extraction

In Unsupervised methods there are some statistical measures of term informativeness such as 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 aim of this system is to identify and rank the news points discussed in both online media and news media. The system architecture shown in the Fig. 1. They are four main phases in this system.

1) Preprocessing: In first stage extraction of key terms in both news media and social media is carried out between given particular time period.

2) Key Term Graph Construction: A graph is construcyed using the previously extracted key terms where vertices are the key terms and edgers are co-occurrence similarity between them. After the processing graph contains topic clusters that are trending in both social and news media.

3) Graph Clustering: The Graph is then clustered to get very much defined disjoint *TCs.*

4) Content Selection and Ranking: In this stage the obtained *TCs* Selected and Ranked based on the three factors *MF,UA,UI*

IV. PROPOSED SYSTEM

An Unsupervised system is implemented—NewsRank which proficiently perceives news points that are available in both online life and the news media, and afterward positions them dependent on MF, UA, and UI. This System centers around news subjects, it is exceptionally simple to use in a different fields, from science and innovation to culture and sports. There is no other work has been actualized that attention on client interests or the social connections for positioning of themes. This work has the stages like pre-preparing , watchword extraction , Similarity of points between web-based social networking 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.

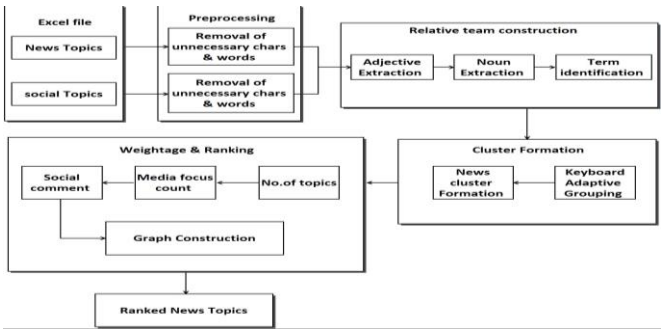
2) Process the tweets data and Media news

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. In English the main parts of speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection. 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 from various online sites from the period between 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 nirav 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.

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.

Fig 3. Detected Report

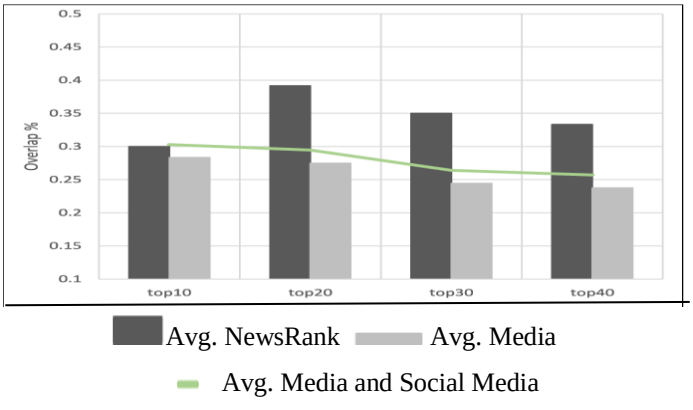


Fig.4 Average percentage of Overlap Between News Media and Social Media

VII. CONCLUSION

An unsupervised strategy is implemented, which recognizes news topic from both online networking and the news media, and ranking them by based on the three significance factors that is MF, UA, and UI . The transient commonness of a specific subject in the news media is viewed as the MF of a point, which gives us understanding into its broad communications prominence. The fleeting predominance of the theme in online networking, especially Twitter, shows client intrigue, and is called as its UA. At long last, the cooperation between the web based life clients who notice the subject demonstrates the quality of the network examining it, and is viewed as the UI. As far as we could possibly know, no other work has endeavored to utilize the utilization of either the interests of internet based life clients or the main goal is to build the relationship to help in ranking of news points.

Combined, removed, and positioned news subjects from verified news suppliers and people have a few benefits. The fundamental benefit is expanding the status and assortment of news recommender frameworks, just as finding covered up, prevalent subjects. Our framework can help news suppliers by giving criticism of points that have been suspended by the broad communications, however are as yet being talked about by the all inclusive community. NewsRank can likewise be stretched out and adjusted to different points other than news, for example, science, innovation, sports, and different patterns.

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.