Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 11, November-2018

Ranking News and Categorizing Based on User Interest and Various Factors

Pooja Walunj [1], Srushti Yewale[2], Anushka Sonawane [3], Nita Sahane [4], Prof. Gauri Bhange [5]

[1,2,3,4,5] Computer Engineering, Sinhgad Academy of Engineering, Kondhwa

Abstract — ***Now Days Important information from online sources has becomes a prominent research. The public of daily events has been provided by mass media sources, mainly the news media, have usually informed us of daily events. Today, online social media services such as Twitter contributing large amount of user generated data, which have great importance to contain informative news-related content. For these resources to be useful, we must find a way to filter noise and only capture the content that, based on its similarity to the news related data, however after noise is removed, the overload data still exists in the remaining data so, we need to prioritize it for consumption. For this, we can use three factors. we proposed an unsupervised method named as sociRank- which identifies news topic widespread in social media as well as the news media, and after that ranks them by using MF, UA, AND UI as relevance factors. First, the temporal prevalence of a topic (MF) of a topic. After that we are going to categorize all news location wise based of reviews or comments.***

The system will first use twitters dataset and filter out all the twits that are related to news. After filtration of news related twits, news topics are ranked based on factors like MF, UA, UI . In addition, the system will also focus on classifying news topics based on the Domain and location of user. It is expected that through the providing of filtered news, instead of reading unnecessary data user gets to read quality news depending on his interests and location.

Keywords: Information filtering, social computing, social network analysis, topic identification, topic ranking.

I. INTRODUCTION

Many of the news media sources have either abandoned their hard copy publications and moved to the World Wide Web, or now produce both hard-copy and Internet versions simultaneously. In social media, regular, no journalist users are able to publish unverified content and express their interest in certain events. One micro blogging service in particular, Twitter, is used by millions of people around the world, providing enormous amounts of user generated data. One may assume that this source potentially contains information with equal or greater value than the news media, but one must also assume that because of the unverified nature of the source, much of this content is useless. For social media data to be of any use for topic identification, we must find a way to filter uninformative information and capture only information which, based on its content similarity to the news media, may be considered useful or valuable. Unfortunately, even after the removal of unimportant content, there is still information overload in the remaining news-related data, which must be prioritized for consumption. To assist in the prioritization of news information, news must be ranked in order of estimated importance. The temporal prevalence of a particular topic in the news media indicates that it is widely covered by news media sources, making it an important factor when estimating topical relevance. This factor may be referred to as the MF of the topic. The temporal prevalence of the topic in social media, specifically in Twitter, indicates that users are interested in the topic and can provide a basis for the estimation of its popularity. This factor is regarded as the UA of the topic. Likewise, the number of users discussing a topic and the interaction between them also gives insight into topical importance, referred to as the UI. By combining these three factors, we gain insight into topical importance and are then able to rank the news topics accordingly.

We propose an unsupervised system SociRank which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Even though this paper focuses on news topics, it can be easily adapted to a wide variety of fields, from science and technology to culture and sports. To the best of our knowledge, no other work attempts to employ the use of either the social media interests of users or their social relationships to aid in the ranking of topics. More ever, socirank.

I. LITERATURE SURVEY

Title -: Topic Detection by Clustering Keywords

Authors: C. Wartena and R. Brussee

Description: We consider topic detection without any prior knowledge of category structure or possible categories. Keywords are extracted and clustered based on different similarity measures using the induced k-bisecting clustering algorithm. Evaluation on Wikipedia articles shows that clusters of keywords correlate strongly with the Wikipedia categories of the articles. In addition, we find that a distance measure based on the Jensen-Shannon divergence of probability distributions outperforms the cosine similarity. In particular, a newly proposed term distribution taking co-occurrence of terms into account gives best results.

Title -: A hierarchical document clustering environment based on the induced bisecting k-means

Authors: F. Archetti, P. Campanelli, E. Fersini, and E. Messina,

Description: The steady increase of information on WWW, digital library, portal, database and local intranet, gave rise to the development of several methods to help user in Information Retrieval, information organization and browsing. Clustering algorithms are of crucial importance when there are no labels associated to textual information or documents. The aim of clustering algorithms, in the text mining domain, is to group documents concerning with the same topic into the same cluster, producing a flat or hierarchical structure of clusters. In this paper we present a Knowledge Discovery System for document processing and clustering. The clustering algorithm implemented in this system, called Induced Bisecting k-Means, outperforms the Standard Bisecting k-Means and is particularly suitable for on line applications when computational efficiency is a crucial aspect.

Title -: Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation

Authors: M. Cataldi, L. Di Caro, and C. Schifanella,

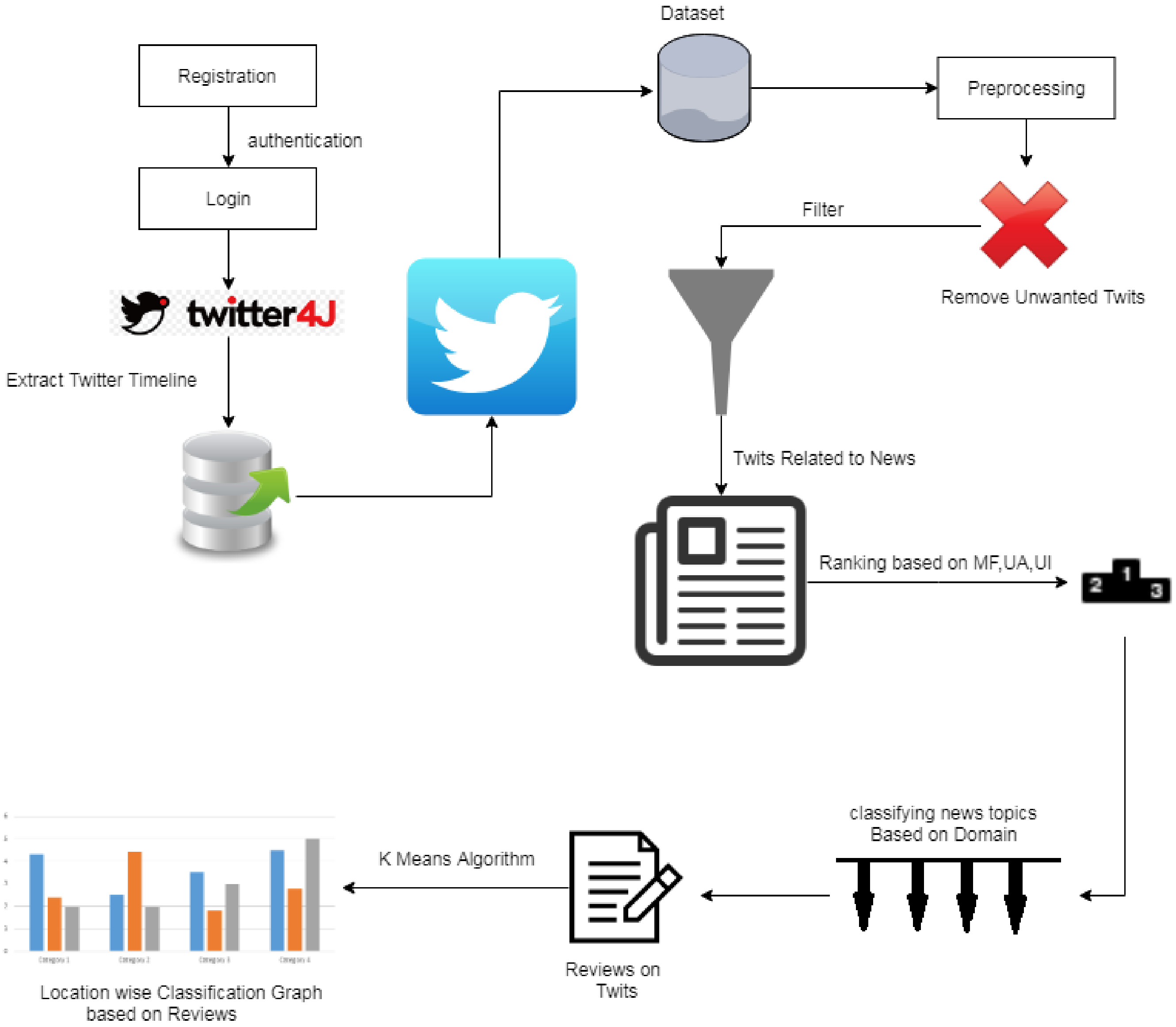
Description: Twitter is a user-generated content system that allows its users to share short text messages, called tweets, for a variety of purposes, including daily conversations, URLs sharing and information news. Considering its world-wide distributed network of users of any age and social condition, it represents a low level news flashes portal that, in its impressive short response time, has the principal advantage. In this paper we recognize this primary role of Twitter and we propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. First, we extract the contents (set of terms)of the tweets and model the term life cycle according to a novel aging theory intended to mine the emerging ones. A term can be defined as emerging if it frequently occurs in the specified time interval and it was relatively rare in the past. Moreover, considering that the importance of a content also depends on its source, we analyze the social relationships in the network with the well-known Page Rank algorithm in order to determine the authority of the users. Finally, we leverage a navigable topic graph which connects the emerging terms with other semantically related keywords, allowing the detection of the emerging topics, under user-specified time constraints. We provide different case studies which show the validity of the proposed approach.

II. PROPOSED SYSTEM

We propose an unsupervised system SociRank which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Even though this paper focuses on news topics. News media sources are considered reliable because they are published by professional journalists, who are held accountable for their content. On the other hand, the Internet, being a free and open forum for information exchange, has recently seen a fascinating phenomenon known as social media. In social media, regular, non-journalist users are able to publish unverified content and express their interest in certain events. Consolidated, filtered, and ranked news topics from both professional news providers and individuals have several benefits. The most evident use is the potential to improve the quality and coverage of news recommender systems or Web feeds, adding user popularity feedback.

In this project we propose k means model in which we gather news information from Twitter and on the basis of news popularity we categories news in different domain and then rank it using k-means algorithm put important news always top of the dashboard of social media site. Also displays news with respect to user domain so they get only those news which they want. As well as we are classifying tweets based on few pre-existing approaches. The work extends basic classification models by incorporating a new feature of location. The expected results will show how these features can achieve decent performance improvement, as now a better result analysis voicing the opinion of people in every region of the country has been made. The system also encounters the problem of population concentration. Like for example a state like utter Pradesh which is so populated will have a more dominant feelings or sentiments over particular News topic in the sentiment analysis and would silence down the voice of the state with less population like Arunachal Pradesh. The point to determine the overall sentiment region wise.

III. SYSTEM ARCHITECTURE



Modules:

1. Registration – Here we are Creating Registration from Creating Account for New User.

2. Login – Existing User can Login into Our System.

3. Extract Twitter Dataset – We are Extracting Twitter Dataset using Twitter 4J.

4.Preprocessing – Here We will Process on that Dataset and we will remove unwanted Twits/Twits which are not related to News.

5. Ranking – After that we will Rank that Twits based on MF, UA and UI.

6. Classification – Using reference of hash tag Domain wise Classification will be Done here. For ex. Sports Related News, Politics related News etc …

7. Segmentation – Using the Reviews we will Classify the Twits as per positive and negative Reviews.

And after that We will Generate the Location wise Result Graph.

Advantages:

- Detection and removal Unwanted/Not Related to News Twits.
- User will get News Related Twits as per Domain interest.
- Popular news always displays on top.
- User can easily choose different domain and get trending news.
- We will get Positive Negative Feedback from Peoples which are belongs to different Locations.

VIII. Hardware Requirement

System	: Intel I3.
Hard Disk	: 40 GB.
Monitor	: 15 VGA Colour.
Mouse	: Logitech.
Ram	: 4 GB.

IX. Software Requirement

Operating system	: Windows XP Professional/7LINUX.
Coding language	: JAVA/J2EE.
IDE	: Eclipse Kepler.
Database	: MYSQL

X. CONCLUSION

In this paper, we proposed a method named as SociRank which is used for identification of news related topics, and after that give ranking to them using MF, UA, and UI as relevance factors. The temporal prevalence of a topic in the news media is considered the MF of a topic, which provide us acuity into its mass media popularity. The temporal prevalence of the topic in social media, specifically Twitter, indicates user interest, and is considered its UA. One of its main uses is increasing the quality and variety of news recommender systems, as well as discovering hidden, popular topics. Our system can help to news providers by providing feedback of topics that have been discontinued by the mass media, and we are classifying and clustering Location wise positive and negative feedbacks, So We can get Proper Result or location wise opinion of peoples for Twits.

REFERENCES

[1] O. Phelan, K. McCarthy, and B. Smyth, “Using Twitter to recommend real-time topical news,” in *Proc. 3rd Conf. Recommender Syst.*, New York, NY, USA, 2009, pp. 385–388.

[2] 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.

[3] K. Sarkar, M. Nasipuri, and S. Ghose, “A new approach to keyphrase extraction using neural networks,” *Int. J. Comput. Sci. Issues*, vol. 7, no. 3, pp. 16–25, Mar. 2010.

[4] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, ―A unified model for stable and temporal topic detection from social media data,‖ in Proc. IEEE 29th Int. Conf. Data Eng. (ICDE), Brisbane, QLD, Australia, 2013,pp. 661–672. [11] C. Wang, M. Zhang, L. Ru, and S. Ma, ―Automatic online news topic ranking using media focus and user attention based on aging theory,‖ in Proc. 17th Conf. Inf. Knowl. Manag., Napa County, CA, USA, 2008,pp. 1033–1042.

[5] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, ―Life cycle modeling of news events using aging theory,” inMachine Learning: ECML 2003. Heidelberg, Germany: Springer Berlin Heidelberg, 2003,pp. 47–59.

[6] 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.

[7] O. Phelan, K. McCarthy, and B. Smyth, ―Using Twitter to recommend real-time topical news,‖ inProc. 3rd Conf. Recommender Syst., New York, NY, USA, 2009, pp. 385–388.

[8] 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.

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