SOCI RANK: Identifying and Ranking Prevalent News Topics Using Social Media Factor

S. Suganya[1], R. Yazhini[2], V.Sharmila[3], Mrs. E. Elakiya[3], Mrs. R. Anusuya[4]
[1, 2] Dept of CSE
[3, 4] Assistant Professor, Dept of CSE
[1, 2, 3, 4] E.G.S. Pillay Engineering College, Nagapattinam,Tamil Nadu,India.

Abstract- *Mass media sources, specifically the news media, have traditionally informed us of daily events. In modern times, social media services such as Twitter provide an enormous amount of user-generated data, which have great potential 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 media, is considered valuable. However, even after noise is removed, information overload may still exist in the remaining data—hence , it is convenient to prioritize it for consumption. To achieve prioritization, information must be ranked in order of estimated importance considering three factors. First, the temporal prevalence of a particular topic in the news media is a factor of importance, and can be considered the media focus (MF) of a topic. Second, the temporal prevalence of the topic in social media indicates its user attention (UA). Last, the interaction between the social media users who mention this topic indicates the strength of the community discussing it, and can be regarded as the user interaction (UI) toward the topic. We propose an unsupervised framework—SociRank—which identifies news topics prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Our experiments show that SociRank improves the quality and variety of automatically identified news topic.*

Keywords- Media focus, User attention, User interaction.

I. INTRODUCTION

The mining of valuable information from online sources has become a prominent research area in information technology in recent years. Historically, knowledge that apprises the general public of daily events has been provided by mass media sources, specifically the news media. Many of these news media sources have either abandoned their hardcopy publications and moved to the World Wide Web, or now produce both hard-copy and Internet versions simultaneously. These 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. Micro blogs have become one of the most popular social media outlets. 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. The news media presents professionally verified occurrences or events, while social media presents the interests of the audience in these areas, and may thus provide insight into their popularity. Social media services like Twitter can also provide additional or supporting information to a particular news media topic.

In summary, truly valuable information may be thought of as the area in which these two media sources topically intersect. 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. 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. Additionally, news topics that perhaps were not perceived as popular by the mass media could be uncovered from social media and given more coverage and priority. For instance, a particular story that has been discontinued by news providers could be given resurgence and continued if it is still a popular topic among social networks. This information, in turn, can be filtered to discover how particular topics are discussed in different geographic locations, which serve as feedback for businesses and governments. A straightforward approach for identifying topics from different social and news media sources is the application of topic modelling. Many methods have been proposed in this area such as "Latent Dirichlet" allocation (LDA) [1] and probabilistic latent semantic analysis (PLSA) [2], [3]. Topic modelling is, in essence, the discovery of "topics" in text corpora by clustering together frequently co-occurring words. This approach, however, misses out in the temporal component of prevalent topic detection, that is, it does not take into account how topics change with time. Furthermore, topic modelling and other topic detection techniques do not rank topics according to their popularity by taking into account their prevalence in both news media and social media. 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. Moreover, SociRank undergoes an empirical framework, comprising and integrating several techniques, such as keyword extraction, measures of similarity, graph clustering, and social network analysis. The effectiveness of our system is validated by extensive controlled and uncontrolled experiments. To achieve its goal, SociRank uses keywords from news media sources (for a specified period of time) to identify the overlap with social media from that same period. We then build a graph whose nodes represent these keywords and whose edges depict their co-occurrences in social media. The graph is then clustered to clearly identify distinct topics. After obtaining well-separated topic clusters (TCs), the factors that signify their importance are calculated: MF, UA, and UI. Finally, the topics are ranked by an overall measure that combines these three factors.

II. LITERATURE SURVEY

T. Hofmann Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two[2]{mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class model.M.cateldi, L.Di Caro and c.shifanella illustrated Emerging topic Detection of Twitter based on Temporal and Social Term Evaluation[7] 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.Wartena and R. Brussee illustrated as, Topic Detection by Clustering Keywords 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. F. Archetti, P. Campanelli, E. Fersini, and E. Messina illustrated as**,** A hierarchical document clustering environment based on the induced bisecting k-means, 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.

III. ARCHITECTURE DIAGRAM

The architecture diagram concept based on Internet query top news and topics of the database of news articles and twitter of preprocessing interaction and team terms extraction used by preprocessing .The user use news topics of ranked news topics of content selection and ranking used by social graph construction .The node weighting used by (UA) User Attention estimation (UI) User Interaction estimation are used media focus estimation used by content selection and ranking. The key term graph clustering term frequency relevant term Identification of term similarity used . The graph clustering Between ness transitivity used.

3.

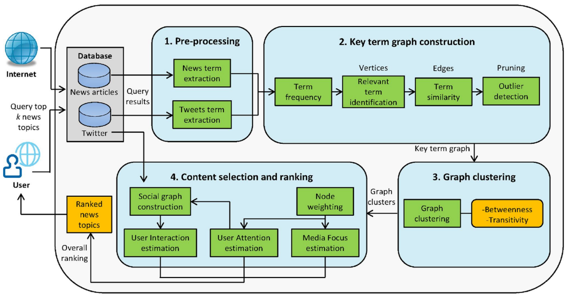
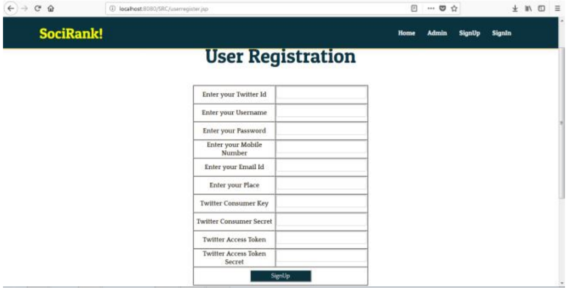


Figure 3.1 Architecture diagram

IV. MODULES

4.1 Registration module:

Registered user is a user of a website, program, or other system who has previously registered. Registered users normally provide some sort of credentials (such as a username or e-mail address, and a password) to the system in order to prove their identity: this is known as logging in. Systems intended for use by the general public often allow any user to register simply by selecting a register or sign up function and providing these credentials for the first time. Registered users may be granted privileges beyond those granted to unregistered users

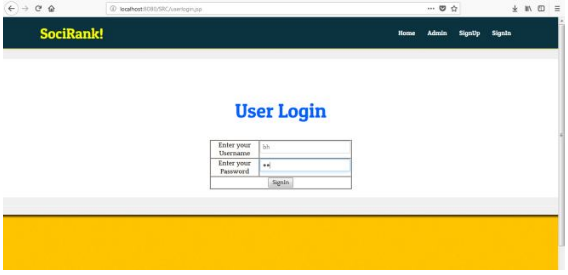


User Registration

Enter your Twitter Id	
Enter your Username	
Enter your Password	
Enter your Mobile Number	
Enter your Email Id	
Enter your Place	
Twitter Consumer Key	
Twitter Consumer Secret	
Twitter Access Token	
Twitter Access Token Secret	

4.2 Login module:

Logging in is usually used to enter a specific page, which trespassers cannot see. Once the user is logged in, the login token may be used to track what actions the user has taken while connected to the site.

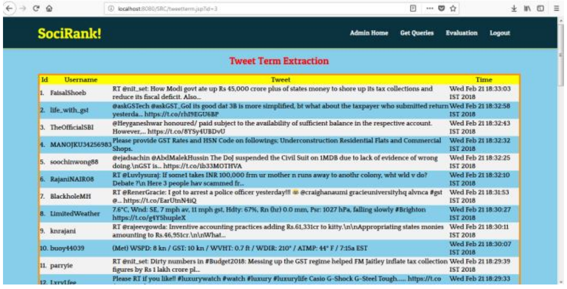


User Login

Enter your Username	bh
Enter your Password	••

4.3 Tweet post :

Used to post the tweet in twitter. Tweet post is a multiuser plug in which allows word press publishers to automatically tweet their new post to their tweet account.

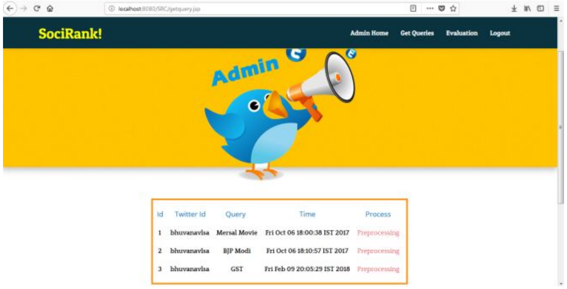


Tweet Term Extraction

Id	Username	Tweet	Time
1.	FaisalShoeb	RT @nit_set: How Modi govt ate up Rs 45,000 crore plus of states money to shore up its tax collections and reduce its fiscal deficit. Also...	Wed Feb 21 18:33:03 IST 2018
2.	life_with_gst	@askGSTech @askGST_GoI its good dat 3B is more simplified, bt what about the taxpayer who submitted return yesterda... https://t.co/rhI9EGU6BP	Wed Feb 21 18:32:58 IST 2018
3.	TheOfficialSBI	@Heyganeshwar honoured/ paid subject to the availability of sufficient balance in the respective account. However,... https://t.co/8YSy4UBDvU	Wed Feb 21 18:32:43 IST 2018
4.	MANOJKU34256983	Please provide GST Rates and HSN Code on followings: Underconstruction Residential Flats and Commercial Shops.	Wed Feb 21 18:32:32 IST 2018
5.	soochinwong88	@ejadsachin @AbdMalekHussin The DoJ suspended the Civil Suit on 1MDB due to lack of evidence of wrong doing.\nGST is... https://t.co/ib33MOTHVA	Wed Feb 21 18:32:25 IST 2018
6.	RajaniNAIR08	RT @Luvlysuraj: If some1 takes INR 100,000 frm ur mother n runs away to anothr colony, wht wld v do? Debate ?\n Here 3 people hav scammed fr...	Wed Feb 21 18:32:10 IST 2018
7.	BlackholeMH	RT @RenerGracie: I got to arrest a police officer yesterday!!! 😂 @craighanaumi gracieuniversityhq alvnca #gst @... https://t.co/EarUtnN4iQ	Wed Feb 21 18:31:53 IST 2018
8.	LimitedWeather	7.6°C, Wnd: SE, 7 mph av, 11 mph gst, Hdty: 67%, Rn (hr) 0.0 mm, Psr: 1027 hPa, falling slowly #Brighton https://t.co/g4Y5hupleX	Wed Feb 21 18:30:27 IST 2018
9.	knrajani	RT @rajeevgowda: Inventive accounting practices adding Rs.61,331cr to kitty.\n\nAppropriating states monies amounting to Rs.46,951cr.\n\nWhat...	Wed Feb 21 18:30:11 IST 2018
10.	buoy44039	(Met) WSPD: 8 kn / GST: 10 kn / WVHT: 0.7 ft / WDIR: 210° / ATMP: 44° F / 7:15a EST	Wed Feb 21 18:30:07 IST 2018
11.	parryie	RT @nit_set: Dirty numbers in #Budget2018: Messing up the GST regime helped FM Jaitley inflate tax collection figures by Rs 1 lakh crore pl...	Wed Feb 21 18:29:39 IST 2018
12.	Lxrylfee	Please RT if you like!! #luxurywatch #watch #luxury #luxurylife Casio G-Shock G-Steel Tough..... https://t.co	Wed Feb 21 18:29:33

4.4Query Tweet:

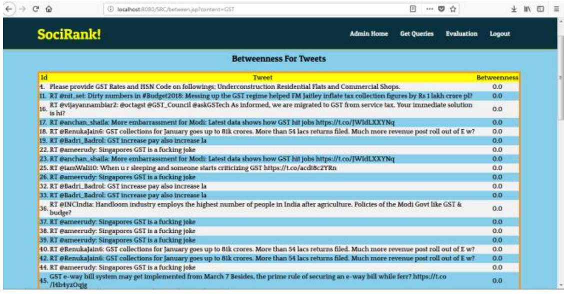
Query post() is a way to alter the main query word press uses to display posts. It does this by putting the main query to one side, and replacing it with.



Id	Twitter Id	Query	Time	Process
1	bhuvanavlsa	Mersal Movie	Fri Oct 06 18:00:38 IST 2017	Preprocessing
2	bhuvanavlsa	BJP Modi	Fri Oct 06 18:10:57 IST 2017	Preprocessing
3	bhuvanavlsa	GST	Fri Feb 09 20:05:29 IST 2018	Preprocessing

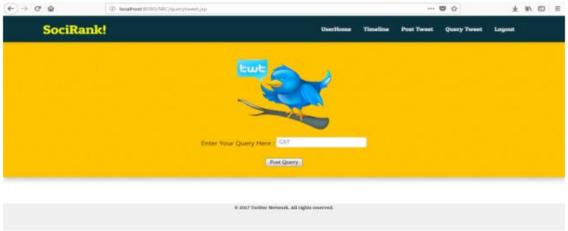
4.5 Evaluation:

Used To present the output. Thus the evaluation report typically Documents the failure and success of a project providing a detailed record of the estimated and actual schedule and budget.



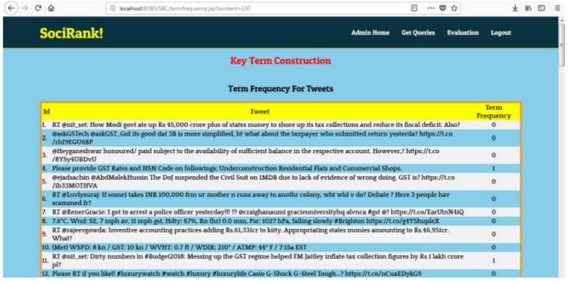
Betweenness For Tweets

Id	Tweet	Betweenness
4.	Please provide GST Rates and HSN Code on followings; Underconstruction Residential Flats and Commercial Shops.	0.0
11.	RT @nit_set: Dirty numbers in #Budget2018: Messing up the GST regime helped FM Jaitley inflate tax collection figures by Rs 1 lakh crore pl?	0.0
16.	RT @vijayannambiar2: @octagst @GST_Council @askGSTech As informed, we are migrated to GST from service tax. Your immediate solution is hi?	0.0
17.	RT @anchan_shaila: More embarrassment for Modi: Latest data shows how GST hit jobs https://t.co/JWIdLXXYNq	0.0
18.	RT @RenukaJain6: GST collections for January goes up to 81k crores. More than 54 lacs returns filed. Much more revenue post roll out of E w?	0.0
19.	RT @Badri_Badrol: GST increase pay also increase la	0.0
22.	RT @ameerudy: Singapores GST is a fucking joke	0.0
23.	RT @anchan_shaila: More embarrassment for Modi: Latest data shows how GST hit jobs https://t.co/JWIdLXXYNq	0.0
25.	RT @iamWali10: When u r sleeping and someone starts criticizing GST https://t.co/acdt8c2YRn	0.0
26.	RT @ameerudy: Singapores GST is a fucking joke	0.0
32.	RT @Badri_Badrol: GST increase pay also increase la	0.0
33.	RT @Badri_Badrol: GST increase pay also increase la	0.0
36.	RT @INCIndia: Handloom industry employs the highest number of people in India after agriculture. Policies of the Modi Govt like GST & budge?	0.0
37.	RT @ameerudy: Singapores GST is a fucking joke	0.0
38.	RT @ameerudy: Singapores GST is a fucking joke	0.0
39.	RT @ameerudy: Singapores GST is a fucking joke	0.0
40.	RT @RenukaJain6: GST collections for January goes up to 81k crores. More than 54 lacs returns filed. Much more revenue post roll out of E w?	0.0
42.	RT @RenukaJain6: GST collections for January goes up to 81k crores. More than 54 lacs returns filed. Much more revenue post roll out of E w?	0.0
44.	RT @ameerudy: Singapores GST is a fucking joke	0.0
45.	GST e-way bill system may get implemented from March 7 Besides, the prime rule of securing an e-way bill while ferr? https://t.co /l4b4yzOqjg	0.0



Enter Your Query Here GST

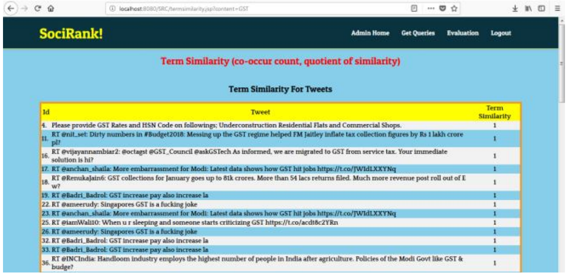
© 2017 Twitter Network. All rights reserved.



Key Term Construction

Term Frequency For Tweets

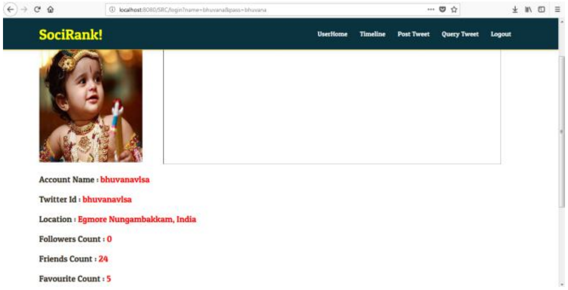
Id	Tweet	Term Frequency
1.	RT @nit_set: How Modi govt ate up Rs 45,000 crore plus of states money to shore up its tax collections and reduce its fiscal deficit. Also?	0
2.	@askGSTech @askGST_GoI its good dat 3B is more simplified, bt what about the taxpayer who submitted return yesterda? https://t.co /rhI9EGU6BP	0
3.	@Heyganeshwar honoured/ paid subject to the availability of sufficient balance in the respective account. However,? https://t.co /8YSy4UBDvU	0
4.	Please provide GST Rates and HSN Code on followings; Underconstruction Residential Flats and Commercial Shops.	1
5.	@ejadsachin @AbdMalekHussin The DoJ suspended the Civil Suit on 1MDB due to lack of evidence of wrong doing. GST is? https://t.co /ib33MOTHVA	0
6.	RT @Luvlysuraj: If some1 takes INR 100,000 frm ur mother n runs away to anothr colony, wht wld v do? Debate ? Here 3 people hav scammed fr?	0
7.	RT @RenerGracie: I got to arrest a police officer yesterday!!! ?? @craighanaumi gracieuniversityhq alvnca #gst @? https://t.co/EarUtnN4iQ	0
8.	7.6°C, Wnd: SE, 7 mph av, 11 mph gst, Hdty: 67%, Rn (hr) 0.0 mm, Psr: 1027 hPa, falling slowly #Brighton https://t.co/g4Y5hupleX	0
9.	RT @rajeevgowda: Inventive accounting practices adding Rs.61,331cr to kitty. Appropriating states monies amounting to Rs.46,951cr. What?	0
10.	(Met) WSPD: 8 kn / GST: 10 kn / WVHT: 0.7 ft / WDIR: 210° / ATMP: 44° F / 7:15a EST	0
11.	RT @nit_set: Dirty numbers in #Budget2018: Messing up the GST regime helped FM Jaitley inflate tax collection figures by Rs 1 lakh crore pl?	1
12.	Please RT if you like!! #luxurywatch #watch #luxury #luxurylife Casio G-Shock G-Steel Tough...? https://t.co/nCuaEDykG9	0



Term Similarity (co-occur count, quotient of similarity)

Term Similarity For Tweets

Id	Tweet	Term Similarity
4.	Please provide GST Rates and HSN Code on followings; Underconstruction Residential Flats and Commercial Shops.	1
11.	RT @nit_set: Dirty numbers in #Budget2018: Messing up the GST regime helped FM Jaitley inflate tax collection figures by Rs 1 lakh crore pl?	1
16.	RT @vijayannambiar2: @octagst @GST_Council @askGSTech As informed, we are migrated to GST from service tax. Your immediate solution is hi?	1
17.	RT @anchan_shaila: More embarrassment for Modi: Latest data shows how GST hit jobs https://t.co/JWIdLXXYNq	1
18.	RT @RenukaJain6: GST collections for January goes up to 81k crores. More than 54 lacs returns filed. Much more revenue post roll out of E w?	1
19.	RT @Badri_Badrol: GST increase pay also increase la	1
22.	RT @ameerudy: Singapores GST is a fucking joke	1
23.	RT @anchan_shaila: More embarrassment for Modi: Latest data shows how GST hit jobs https://t.co/JWIdLXXYNq	1
25.	RT @iamWali10: When u r sleeping and someone starts criticizing GST https://t.co/acdt8c2YRn	1
26.	RT @ameerudy: Singapores GST is a fucking joke	1
32.	RT @Badri_Badrol: GST increase pay also increase la	1
33.	RT @Badri_Badrol: GST increase pay also increase la	1
36.	RT @INCIndia: Handloom industry employs the highest number of people in India after agriculture. Policies of the Modi Govt like GST & budge?	1



Account Name : bhuvanavisa

Twitter Id : bhuvanavisa

Location : Egmore Nungambakkam, India

Followers Count : 0

Friends Count : 24

Favourite Count : 5

V. CONCLUSION

In this paper, we proposed an unsupervised method SociRank which identifies news topics prevalent in both social media and the news media, and then ranks them by taking into account their MF, UA, and UI as relevance factors. The temporal prevalence of a particular topic in the news media is considered the MF of a topic, which gives us insight into its mass media popularity. The temporal prevalence of the topic in social media, specifically Twitter, indicates user interest, and is considered its UA. Finally, the interaction between the social media users who mention the topic indicates the strength of the community discussing it, and is considered the UI. To the best of our knowledge, no other work has attempted to employ the use of either the interests of social media users or their social relationships to aid in the ranking of topics. Consolidated, filtered, and ranked news topics from both professional news providers and individuals have several benefits. 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 aid news providers by providing feedback of topics that have been discontinued by the mass media, but are still being discussed by the general population. SociRank can also be extended and adapted to other topics besides news, such as science, technology, sports, and other trends.

REFERENCE

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “LatentDirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[2] T.Hofmann,“Probabilistic latent semantic analysis,” in Proc. 15th Conf. Uncertainty Artif. Intell., 1999, pp. 289–296.

[3] T. Hofmann, “Probabilistic latent semantic indexing,” 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] F. Archetti, P. Campanelli, E. Fersini, and E. Messina, “A hierarchical document clustering environment based on the induced bisecting k-means,” in Proc. 7th Int. Conf. Flexible Query Answering Syst., Milan, Italy, 2006, pp. 257–269. [Online]. Available:http://dx.doi.org/10.1007/11766254_22.

[6] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[7] 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 MultimediaDataMin.(MDMKDD),Washington, DC, USA, 2010, Art.no.4.[Online]. Available:http://doi.acm.org/10.1145/1814245.1814249.

[8] 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.

[9] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, “Finding bursty topics from microblogs,” in Proc. 50th Annu. Meeting Assoc. Comput. Linguist. Long Papers, vol. 1. 2012, pp. 536–544.

[10] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, “A unified model forstable 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.

[12] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, “Life cyclemodeling of news events using aging theory,” in Machine Learning:ECML 2003. Heidelberg, Germany: Springer Berlin Heidelberg, 2003,pp. 47–59.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] K. Kireyev, “Semantic-based estimation of term informativeness,” in Proc. Human Language Technol. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist., 2009, pp. 530–538.

[19] G. Salton, C.-S. Yang, and C. T. Yu, “A theory of term importance in automatic text analysis,” J. Amer. Soc. Inf. Sci., vol. 26, no. 1, pp. 33–44,1975.

[20] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information,” IBM J. Res. Develop., vol. 1, no. 4, pp. 309–317, 1957.