Identifying the News Topics Prevalent in Social Media and provide Ranking using SociRank Framework

Chandana Sankoju[1], N.V. Subba Reddy[2]
[1]PG Scholar Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad
[2]Associate Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad

Abstract

To predict interactions between social media and traditional news streams is becoming increasingly relevant for a variety of applications, including: understanding the underlying factors that drive the evolution of data sources, tracking the triggers behind events, and discovering emerging trends. Researchers have developed such interactions by examining volume changes or information diffusions, however, most of them ignore the semantical and topical relationships between news and social media data. Our work is the first attempt to study how news influences social media, or inversely, based on topical knowledge. We introduce a hierarchical Bayesian model that jointly models the news and social media topics and their interactions. We show that our proposed model can capture distinct topics for individual datasets as well as discover the topic influences among multiple datasets. By applying our model to large sets of news and tweets, we demonstrate its significant improvement over baseline methods and explore its power in the discovery of interesting patterns for real world cases.

Index Terms : social computing, Information filtering, social network analysis, topic ranking, topic identification

I. INTRODUCTION

Today, online social media, for example, Twitter have filled in as devices for sorting out and following social occasions. Understanding the triggers and moves in assessment driven mass social media information can give helpful bits of knowledge to different applications in the scholarly world, industry, and nonetheless, there remains a general absence of finding of what causes the problem areas in social media. Ordinarily, the purposes for the fast spread of information can be abridged as far as two classes: exogenous and endogenous components. Developing elements are the aftereffects of information dispersion inside the social network itself, in particular, clients get information principally from their online social network. Conversely, exogenous components imply that clients get information from outside sources initially, for instance, customary news media, and then bring it into their social network.

Albeit past works have investigated both the social media and outside news information datasets, couple of scientists have taken a gander at the endogenous and exogenous elements in light of semantical or topical learning. They have either looked to distinguish applicable tweets in light of news articles or basically corresponded the two information sources through comparable examples in the changing information volume. Still inside similar information source, there could be different elements that drive the development of information after some time. Exogenous factors over numerous datasets make breaking down the development and relationship among different information streams more troublesome. Watching social media and outside news information streams in an assembled casing can be a functional method for taking care of this issue. In this paper, we propose a novel topic model, News and Twitter Interaction Topic model (NTIT), that together learns social media topics and news topics and unobtrusively catch the impacts between topics. The instinct behind this approach is that before a client posts a message, he/she might be affected either by conclusions from his/her online companions or by articles from news organizations. In our new structure, a word in a tweet can be receptive to the topical impacts coming either from endogenous elements (tweets) or from exogenous components (news).

A direct approach for recognizing topics from various social and news media sources is the use of topic modeling. Numerous strategies have been proposed around there, for example, latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA). Topic modeling is, generally, the disclosure of topics in content corpora by grouping together every now and again co-happening words. This approach, in any case, passes up a great opportunity in the fleeting part of pervasive topic discovery, that is, it doesn't consider how topics change with time. Moreover, topic modeling and other topic recognition systems don't rank topics as per their ubiquity by considering their pervasiveness in both news media and social media.

We present an unsupervised framework SociRank which adequately recognizes news topics that are predominant in both social media and the news media, and then positions them by significance utilizing their degrees of MF, UA, and UI. Despite the fact that this paper centers around news topics, it can be effortlessly adjusted to a wide assortment of fields, from science and innovation to culture and games. To the best of our insight, no other work endeavors to utilize the utilization of either the social media interests of clients or their social connections to help in the ranking of topics. In addition, SociRank experiences an experimental system, including and incorporating a few methods, for example, catchphrase extraction, measures of comparability, chart grouping, and social network analysis. The viability of our framework is approved by broad controlled and uncontrolled experiments.

II. LITERATURE REVIEW

Much research has been done in the field of topic identification—alluded to all the more formally as topic modeling. Two customary strategies for recognizing topics are LDA [1] and PLSA [2], [3]. LDA is a generative probabilistic model that can be connected to various undertakings, including topic identification. PLSA, comparatively, is a measurable procedure, which can likewise be connected to topic modeling. In these methodologies, be that as it may, worldly information is lost, which is foremost in distinguishing predominant topics and is a critical normal for social media information. Moreover, LDA and PLSA just find topics from content corpora; they don't rank in view of prevalence orprevalence. Wartena and Brussee [4] actualized a strategy to recognize topics by bunching catchphrases. Their strategy involves the grouping of watchwords—in view of various similitude measures—utilizing the inducedk-bisecting bunching calculation [5]. In spite of the fact that they don't utilize the utilization of diagrams, they do watch that a separation measure in view of the Jensen– Shannon dissimilarity (or information range [6]) of likelihood conveyances performs well. All the more as of late, examine has been directed in recognizing topics and occasions from social media information, considering transient information. Cataldiet al.[7] proposed a topic discovery system that recovers continuous developing topics from Twitter. Their strategy utilizes the arrangement of terms from tweets and model their life cycle as per a novel maturing hypothesis. Also, they consider social connections—all the more particularly, the specialist of the clients in the network—to decide the significance of the topics. Zhaoet al.[8] completed comparative work by building up a Twitter-LDA model intended to recognize topics in tweets. Their work, in any case, just thinks about the individual interests of clients, and not pervasive topics at a worldwide scale. Another slanting region of related research is the discovery of ―bursty‖ topics (i.e., topics or occasions that happen to put it plainly, sudden scenes). Diao et al. [9] proposed a technique that uses a state machine to identify bursty topics in microblogs. Their strategy additionally decides if client posts are close to home or allude to a specific slanting topic. Yin et al.[10] likewise built up a model that recognizes topics from social media information, recognizing fleeting and stable topics. These techniques, in any case, just utilize information from microblogs and don't endeavor to coordinate them with genuine news. Furthermore, the recognized topics are not positioned by fame or predominance.

Wanget al.[11] proposed a strategy that considers the clients' enthusiasm for a topic by assessing the measure of times they read stories identified with that specific topic. They allude to this factor as the UA. They likewise utilized a maturing hypothesis created by Chenet al.[12] to make, develop, and obliterate a topic. The existence cycles of the topics are followed by utilizing a vitality work. The vitality of a topic increments when it winds up mainstream and it lessens after some time except if it stays prevalent. We utilize variations of the ideas of MF and UA to address our issues, as these ideas are both legitimate and compelling. Different works have influenced utilization of Twitter to find news-related substance that may be viewed as imperative. Sankaranarayanan et al. [13] built up a framework called TwitterStand, which recognizes tweets that compare to breaking news. They achieve this by using a grouping approach for tweet mining. Phelan et al. [14] developed a suggestion framework that produces a positioned rundown of news stories. News are positioned in view of the co-event of mainstream terms inside the clients' RSS and Twitter channels. Both of these frameworks mean to recognize developing topics, however give no understanding into their prominence after some time. In addition, the work by Phelan et al. [14] just creates a customized ranking (i.e., news articles custom-made particularly to the substance of a solitary client), as opposed to giving a general ranking in light of an example everything being equal. In any case, these works furnish us with a reason for broadening the start of UA. Research has likewise been done in topic revelation and ranking from different areas. Shubhankar et al. [15] built up a calculation that distinguishes and positions topics in a corpusof explore papers. They utilized shut incessant catchphrase sets to frame topics and a change of the Page Rank [16] calculation to rank them. Their work, be that as it may, does not incorporate or team up with other information sources, as refined by SociRank.

III. SOCIRANK FRAMEWORK

The goal of our method—SociRank—is to identify, consolidate and rank the most prevalent topics discussed in both news media and social media during a specific period of time. The system framework can be visualized in Fig. 1. To achieve its goal, the system must undergo four main stages.

1) *Preprocessing:* Key terms are extracted and filtered from news and social data corresponding to a particular period of time.
2) *Key Term Graph Construction:* A graph is constructed from the previously extracted key term set, whose ver- tices represent the key terms and edges represent the co-occurrence similarity between them. The graph, after processing and pruning, contains slightly joint clusters of topics popular in both news media and social media.
3) *Graph Clustering:* The graph is clustered in order to obtain well-defined and disjoint TCs.
4) *Content Selection and Ranking:* The TCs from the graph are selected and ranked using the three relevance factors (MF, UA, and UI).

Initially, news and tweets data are crawled from the Internet and stored in a database. News articles are obtained from spe- cific news websites via their RSS feeds and tweets are crawled from the Twitter public timeline [41]. A user then requests an output of the top k ranked news topics for a specified period of time between date d_1 (start) and date d_2 (end).

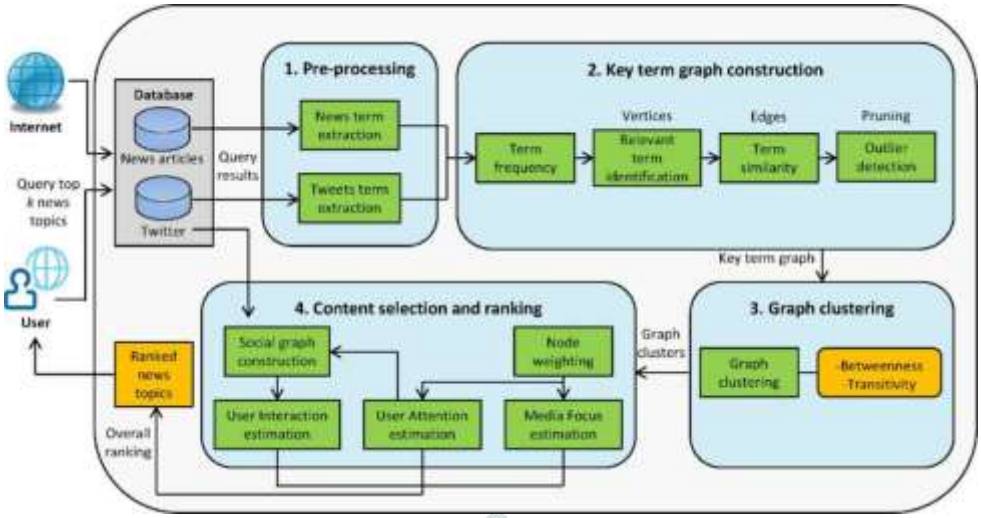

Fig. 1. SociRank framework.

IV. EMERGENCE OF TWITTER AS A NEWS MEDIA

Software engineering research network has dissected significance of online social media, specifically Twitter, as news spreading operator. Kwak et al. demonstrated the noticeable quality of Twitter as a news media, they demonstrated that 85% topics talked about on Twitter are identified with news. Their work featured the connection between client particular parameters v/s the tweeting movement designs, similar to analysis of the quantity of supporters and followees v/s the tweeting (re-tweeting) numbers. Zhao et al. in their work, utilized unsupervised topic modeling to analyze the news topic from Twitter versus New York Times (a conventional news spread medium). They demonstrated that Twitter clients are generally less keen on world news, still they are dynamic in spreading news of essential world occasions. Lu et al. indicated how tweets identified with news occasion on Twitter can be mapped utilizing vitality work. The strategies proposed act like novel occasion identification methods. The examination broke down 900 news occasions through 2010-2011. Castillo et al. performed subjective and quantitative analysis on online social media movement about news articles. They inferred that news articles depicting breaking news occasions have more monotonous social media responses, than top to bottom articles.

ANALYZING TWITTER DATA DURING REAL-WORLD EVENTS

The posts and action on Twitter, impacts and assumes an essential part in different certifiable occasions. Part of Twitter has been examined by PC researchers, analysts and sociologists for affect in reality. Twitter has advanced from being just a medium to impart clients' insights; to an information sharing and dispersal specialist; to engendering and coordination of alleviation and reaction endeavors. A portion of the prominent contextual analyses examined by PC researchers have been, Twitter exercises amid races, catastrophic events (like sea tempests, rapidly spreading fires, surges, and so on.), political and social uprisings (like Libya and Egypt emergency) and fear monger assaults (like Mumbai triple bomb impacts). Substance and client action examples of Twitter amid occasions have been investigated for both positive and negative viewpoints. A portion of the issues examined that outcome in terrible nature of information, nearness of spam and phishing posts, content spreading gossipy tidbits/counterfeit news, security rupture of clients by means of the substance shared by them and utilization of Twitter for proliferation and induction of detest among individuals. Analysts have utilized machine learning, information recovery, social network analysis and picture and video analysis to analyze and describing Twitter utilization amid true occasions.

We present a portion of the examination work done in applying client modeling strategies to break down conduct of clients on social networks. Yin et al. modeled client conduct utilizing two factors: the topics identified with clients' natural advantages and the topics identified with worldly setting . They made a latent class factual blend model, called Dynamic Temporal Context-Aware Mixture model (DTCAM). They assessed their framework on four expansive scale social media datasets. The creators exhibited how client modeling strategies can be adequately used to enhance the execution of recommender frameworks for social networks. Xu et al. presented a blended latent topic model to consolidate different variables to model clients' posting conduct on Twitter. The creators expected that a client's conduct is affected by three elements: breaking news, posts from social companions and client's advantage. They created and demonstrated that their model outflanks other client models in handling the perplexity of held-out substance and the nature of produced latent topics. Abel et al. built up a client modeling system for news proposals on Twitter utilizing in excess of 2 million tweets. The creators proposed diverse techniques for making hash tag-based, element based or topic-based client profiles utilizing semantic enhancement and worldly factors. Their outcomes demonstrated that thought of worldly profile examples can enhance proposal quality.

V. EXPERIMENTS AND RESULTS

The testing dataset consists of tweets crawled from Twitter public timeline and news articles crawled from popular news websites during the period between November 1, 2013 and February 28, 2014. The news websites crawled werecnn.com, bbc.com,

cbsnews.com, reuters.com, abcnews.com, and usatoday.com. Over the specified period of time, a total of 105 856 news articles and 175 044 074 bilingual tweets were
collected. After non-English tweets were discarded, 71 731 730 tweets remained. The dataset was divided into two partitions.

1) Data from January and February 2014 were used as the testing dataset, on which experiments were performed for the overall method evaluation.
2) Data from November and December 2013 were used as the control dataset, where experiments were performed to establish adequate thresholds and select measures that presented the best results.

Method Evaluation

The evaluation of topic ranking is quite challenging, as the interpretation of the results is generally subjective. However, in an attempt to show that the ranked topics are indeed those that users would prefer, a method for ranking popular news topics must be established.

TABLE I
SOME STATISTICS RELEVANT TO THE TESTING DATASET

Time period	# topics	Avg. tweets	Avg. news	Avg. users
2014/01/01–10				
2014/01/11–20				
2014/01/21–30				
2014/02/01–10				
2014/02/11–20				
2014/02/21–28				
Average	97	2567	16	894

Next, 20 ace's and doctoral understudies were solicited to see the titles from the main 10 news stories from every day and select the ones they considered important. Every member was required to choose at least two articles for every day and a greatest of each of the 10. The members' outcomes were then parceled into 12 date ranges: 1) November 1, 2013– November 10, 2013; 2) November 11, 2013– November 20, 2013; 3) November 21, 2013– November 30, 2013; 4) December 1, 2013– December 10, 2013; 5) December 10, 2013– December 20, 2013; 6) December 20, 2013– December 31, 2013; 7) January 1, 2014– January 10, 2014; 8) January 11, 2014– January 20, 2014; 9) January 21, 2014– January 31, 2014; 10)February 1, 2014– February 10, 2014; 11) February 11, 2014– February 20, 2014; and 12) February 21, 2014– February 28, 2014. As clarified before, the initial six information ranges were used for the controlled trials and the last six for the technique assessment.Fig. 3 shows the percentage of topics selected by SociRank and by MF that overlap with the voted topics. It can be seen that SociRank clearly outperforms MF in terms of overlap with the voted topics (i.e., those topics that users selected as the most important). This indicates that SociRank is better at discovering prevalent news topics that users find interesting when compared to a method that only utilizes data from the news media.

Topic Ranking Evaluation: Next, we assess the ranking of topics utilizing the SociRank and MF ranking equations, choosing just the best k topics from each approach. Once more, we compare the positioned topics for each time go with the voted topics. We assessed the best 10, 20, 30, and 40 topics for the two techniques and ascertained the level of topics that covered with the best 10, 20, 30, and 40 voted topics, separately. For example, on the off chance that we are comparing the best 10 topics utilizing SociRank with the voted topics, and five topics cover, at that point the cover rate would be 5/10. Fig. 6 demonstrates the level of cover (for every one of the time ranges) between the main 10 voted topics and the best 10 topics utilizing the SociRank and MF approaches. In the figure, the green line speaks to the MF cover rate in addition to 1 standard deviation. In the event that the SociRank bar outperforms this line, it demonstrates that there is a sufficiently huge distinction between the two cover rates for the change to be considered huge. Nonetheless, as can be found in the figure, this does not happen—there is no huge distinction between the two strategies in the best 10 positioned list.

In the main 20, 30, and 40 positioned records, in any case, the outcomes favored SociRank, as can be found in Figs. 4– 6, individually. In every one of these rundowns, the cover level of SociRank fundamentally outperforms that of the MF approach. Abridging our discoveries, Fig. 7 speaks to the normal cover rates of the main 10, 20, 30, and 40 positioned records for the two techniques. The figure obviously outlines that, except for the main 10 list, SociRank outflanked the MF technique by a critical edge.

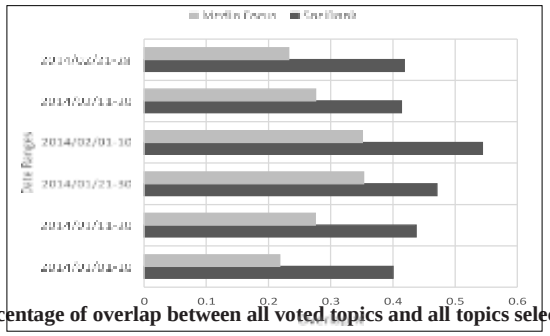

Fig. 2. Percentage of overlap between all voted topics and all topics selected by SociRank and MF.

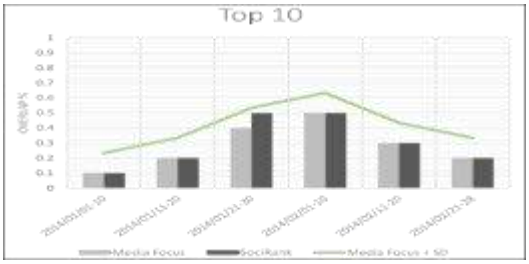

Fig. 3. Percentage of overlap between top 10 voted topics and top 10 topics selected by SociRank and MF.

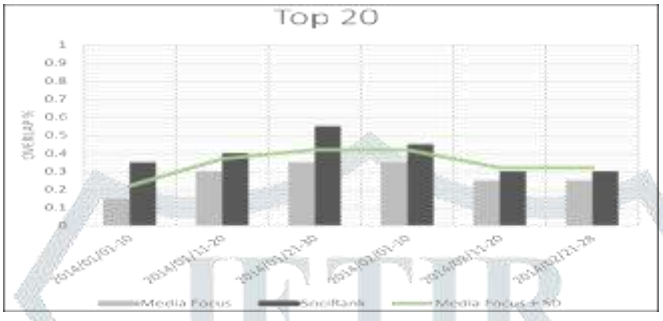

Fig. 4. Percentage of overlap between top 20 voted topics and top 20 topics selected by SociRank and MF.

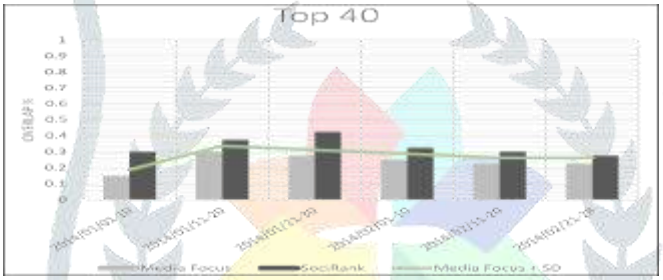

Fig. 5. Percentage of overlap between top 30 voted topics and top 30 topics selected by SociRank and MF.

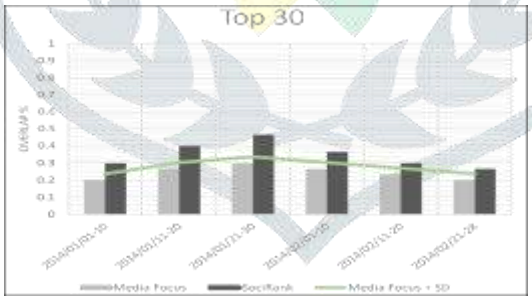

Fig. 6. Percentage of overlap between top 40 voted topics and top 40 topics selected by SociRank and MF.

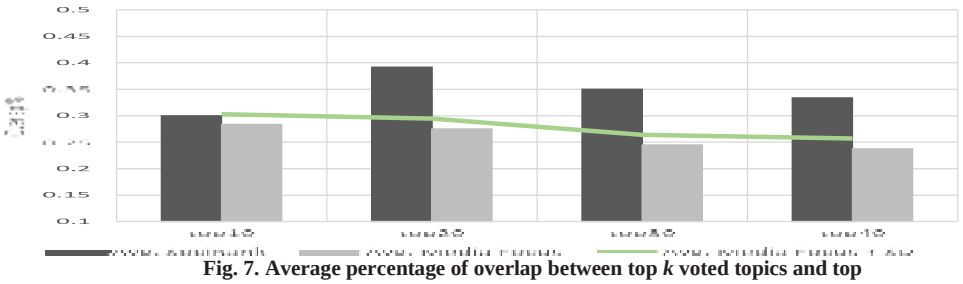

Fig. 7. Average percentage of overlap between top *k* voted topics and top *k* topics selected by SociRank and MF.

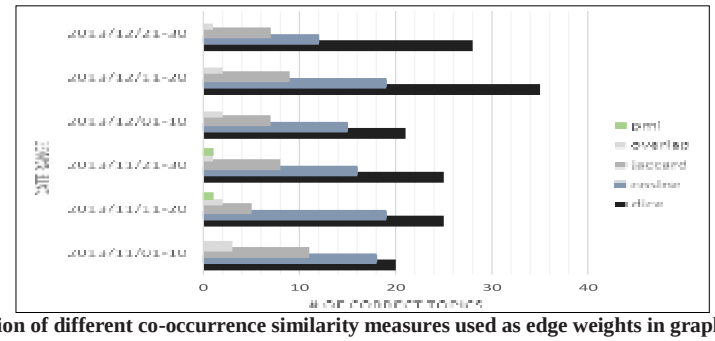

Fig. 8. Evaluation of different co-occurrence similarity measures used as edge weights in graph *G* (Section III-B3).

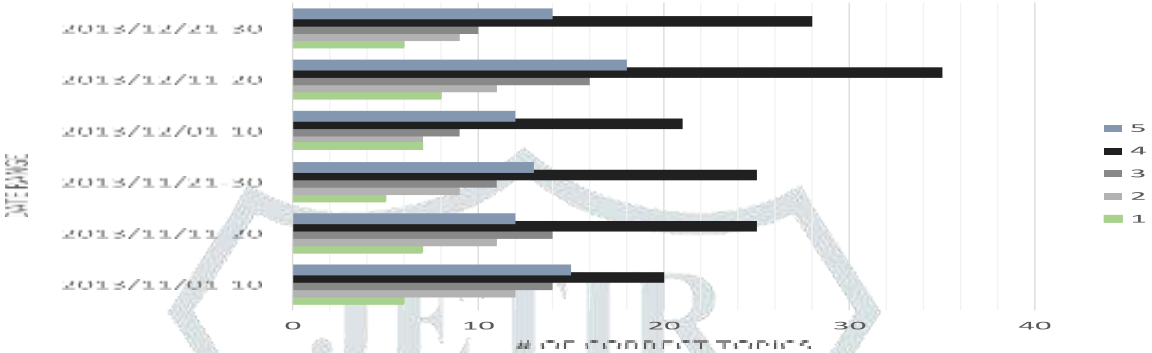

Fig. 9. Evaluation of different values for IQR coefficient *c* in the outlier detection formula (9).

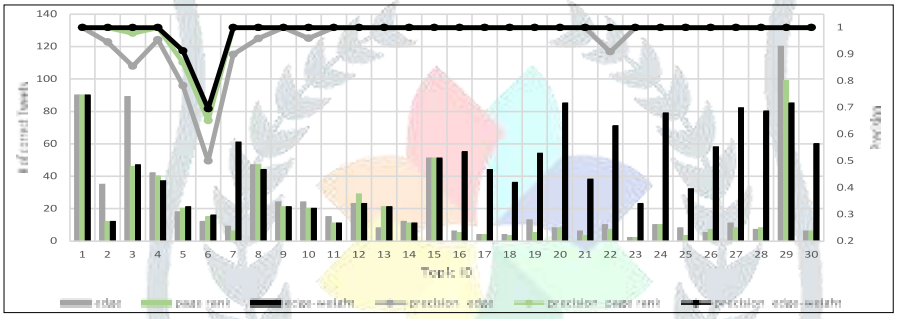

Fig. 10. Evaluation of different node-weighting approaches (Section III-D1).

TABLE II
COMPARISON OF RANKED TOPIC LISTS PRODUCED BY THE MF AND SOCIRANK METHODS

Topics selected for time period 2014/02/01–10

#	media focus	SociRank
1	sochi ceremony olympics winter opening olympic	sochi ceremony olympics winter opening olympic
2	seymour hoffman philip actor apartment phillip manhatta	seymour hoffman philip actor apartment phillip manhattan
3	homs aleppo beirut convoy syria civilian humanitarian ev	broncos seahawks denver seattle ↑ (+1)
4	broncos seahawks denver seattle	bieber arrest justin race ↑ (+6)
5	ukrainian ukraine kiev opposition bulatov protester	bruno mars halftime show
6	giraffe zoo copenhagen danish inbreeding marius	woody farrow allen dylan mia abuse ↑ (+1)
7	woody farrow allen dylan mia abuse	united manchester fulham
8	knox extradition amanda lengthy	zimmerman boxing george celebrity match
9	israeli palestinian gaza	homs aleppo beirut convoy syria civilian humanitarian evacua
10	bieber arrest justin race	giraffe zoo copenhagen danish inbreeding marius ↓ (-4)

Table II demonstrates the best 10 topics acquired by the MF and SociRank approaches for the February 1, 2014– February 10, 2014 day and age. In the table, the cells with the new topics are shaded in green and the cells with the topics that were expelled from the MF list are shaded in red. As can be seen, utilizing every one of the three components delivered a rundown with three new topics that did not show up in the MF list. Moreover, a significant number of the topics in the SociRank list either climbed or down in position as compared with the MF list.

Mulling over all outcomes underscores the point that MF alone is a substandard estimator of what users find fascinating or consider critical, and ought to along these lines not be utilized as a part of thusly. SociRank, then again, turns out to be more equipped for playing out this,

and so we conclude that the information gave by SociRank can demonstrate imperative in commerce-based zones where the enthusiasm of users is central.

VI. CONCLUSION

Our model includes jointly topic modeling on multiple data sources in an asymmetrical frame, which benefits the modeling performance for both long and short texts. 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.

We present the results of applying model to two large-scale datasets and show its effectiveness over non-trivial baselines. Based on the outputs of model, further efforts are made to understand the complex interaction between news and social media data. Through extensive experiments, we find following factors: 1) even for the same events, focuses of news and Twitter topics could be greatly different; 2) topic usually occurs first in its dominant data source, but occasionally topic first appearing in one data source could be a dominant topic in another dataset; 3) generally, news topics are much more influential than Twitter topics.

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, "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 hier- archical 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 detec- tion 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.

[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 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.

[12] 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.

[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 recom- mend 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 hypertex- tual 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 search- ing of literary information," IBM J. Res. Develop., vol. 1, no. 4, pp. 309–317, 1957.

[21] J. D. Cohen, "Highlights: Language- and domain-independent automatic indexing terms for abstracting," J. Amer. Soc. Inf. Sci., vol. 46, no. 3, pp. 162–174, 1995.

[22] 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.

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

[24] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extrac- tion," in Proc. 4th ACM Conf. Digit. Libr., Berkeley, CA, USA, 1999, pp. 254–255.

[25] P. D. Turney, "Learning algorithms for keyphrase extraction," Inf. Retrievel, vol. 2, no. 4, pp. 303–336, 2000.

[26] J. Wang, H. Peng, and J.-S. Hu, "Automatic keyphrases extraction from document using neural network," in Advances in Machine Learning and Cybernetics. Heidelberg, Germany: Springer, 2006, pp. 633–641.

[27] T. Jo, M. Lee, and T. M. Gatton, "Keyword extraction from docu- ments using a neural network model," in Proc. Int. Conf. Hybrid Inf. Technol. (ICHIT), vol. 2. 2006, pp. 194–197.

[28] 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.

[29] C. Zhang, "Automatic keyword extraction from documents using condi- tional random fields," J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.

[30] G. Figueroa and Y.-S. Chen, "Collaborative ranking between supervised and unsupervised approaches for keyphrase extraction," in Proc. Conf. Comput.

Linguist. Speech Process. (ROCLING), 2014, pp. 110–124.
[31] 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.
[32] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic simi- larity between words using Web search engines," in Proc. WWW, Banff, AB, Canada, 2007, pp. 757–766.
[33] D. Szymkiewicz, "Etude comparative de la distribution florale," Rev. Forest, vol. 1, 1926

[34] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.
[35] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," Comput. Linguist., vol. 16, no. 1, pp. 22–29, Mar. 1990. [Online]. Available: http://dl.acm.org/ citation.cfm?id=89086.89095.
[36] P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Lausanne, Switzerland: Impr. Corbaz, 1901.
[37] H. Iwasaka and K. Tanaka-Ishii, "Clustering co-occurrence graph based on transitivity," presented at the 5th Workshop Very Large Corpora, 1997, pp. 91–100.
[38] 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.
[39] 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.
[40] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Phys. Rev. E, vol. 69, no. 6, 2004, Art. no. 066133.
[41] Twitter. [Online]. Available: http://www.twitter.com, accessed Feb. 2014.
[42] K. Gimpel et al., "Part-of-speech tagging for Twitter: Annotation, fea- tures, and experiments," in Proc. 49th Annu. Meeting Assoc. Comput. Linguist. Human Lang. Technol. Short Papers, vol. 2. Portland, OR, USA, 2011, pp. 42–47.
[43] A. J. Viterbi, "Error bounds for convolutional codes and an asymptoti- cally optimum decoding algorithm," IEEE Trans. Inf. Theory, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[44] M. Hubert and S. Van der Veeken, "Outlier detection for skewed data," J. Chemometr., vol. 22, nos. 3–4, pp. 235–246, 2008.
[45] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," Soc. Netw., vol. 30, no. 2, pp. 136–145, 2008.
[46] D. B. Johnson, "Efficient algorithms for shortest paths in sparse net- works," J. ACM, vol. 24, no. 1, pp. 1–13, Jan. 1977. [Online]. Available: http://doi.acm.org/10.1145/321992.321993.
[47] R. W. Floyd, "Algorithm 97: Shortest path," Commun. ACM, vol. 5, no. 6, p. 345, Jun. 1962. [Online]. Available: http://doi.acm.org/10.1145/367766.368168.
[48] Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao, "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications" Vol V, Issue II,Febrauary 2014 [ISSN : 2321-3469]
[49] Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Analysis on Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management", Vol VII, Issue X, November 2014 [ISSN : 2249-4510]
[50] Shoban Babu Sriramoju, "Mining Big Sources Using Efficient Data Mining Algorithms" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2, Issue 1,January 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[51] Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Processing" in "International Journal of Information Technology and Management" Vol V, Issue I, August 2013 [ISSN : 2249-4510]
[52] Dr. Shoban Babu Sriramoju, "A Review on Processing Big Data" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol-2, Issue-1, January 2014 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[53] Shoban Babu Sriramoju, Dr. Atul Kumar, "An Analysis around the study of Distributed Data Mining Method in the Grid Environment : Technique, Algorithms and Services" in "Journal of Advances in Science and Technology" Vol-IV, Issue No-VII, November 2012 [ISSN : 2230-9659]
[54] Guguloth Vijaya, A. Devaki, Dr. Shoban Babu Sriramoju, "A Framework for Solving Identity Disclosure Problem in Collaborative Data Publishing" in "International Journal of Research and Applications", Volume 2, Issue 6, 292-295, Apr-Jun 2016 [ISSN : 2349-0020]
[55] Monelli Ayyavaraiah, Shoban Babu Sriramoju, "A Survey on the Approaches in Targeting Frequent Sub Graphs Mining" in "Indian Journal of Computer Science and Engineering (IJCSE)", Volume 9, Issue 2, Apr-May 2018 [e-ISSN : 0976-5166 p-ISSN : 2231-3850], DOI : 10.21817/indjcse/2018/v9i2/180902024
[56] Ramesh Gadde, Namavaram Vijay, "A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP" in "International Journal of Research in Science and Engineering", Vol-3, Issue-6, Nov-Dec 2017, 92-99 [ISSN : 2394-8299].
[57] Monelli Ayyavaraiah, "Review of Machine Learning based Sentiment Analysis on Social Web Data" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 4,Issue 6,March 2016 [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]
[58] Siripuri Kiran, 'Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification', International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X,Volume 4 Issue 2, pp.829-831, January-February 2018. URL : http://ijsrst.com/IJSRST1841198
[59] Ajmera Rajesh, Siripuri Kiran, " Anomaly Detection Using Data Mining Techniques in Social Networking" in "International Journal for Research in Applied Science and Engineering Technology", Volume-6, Issue-II, February 2018, 1268-1272 [ISSN : 2321-9653], www.ijraset.com