VOL 2 (2018) NO 3

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

BlogNewsRank: Finding and Ranking Frequent News Topics Using Social Media Factors

Harshitha H [#], Dr. Mohammed Rafi [#]

[#] #Department of Studies in CSE, University B.D.T College of Engineering (A Constituent college of VTU,Belagavi), India
E-mail: harshithadravid12@gmail.com

***Abstract*— In early days, mass media sources such as news media used to inform us about daily events. Now a days, social media services such as Twitter huge amount of user-generated data, which has a great potential to contain informative news-related content. For these resources to be useful, we have to find a way to filter noise and capture the content that, based on its similarity to the news media, is considered valuable. 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 mainly 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—BlogNewsRank—which identifies news topics prevalent in both social media and the news media, and then ranks them by relevance(frequency) using their degrees of MF, UA, and UI.**

***Keywords*— Topic identification, Topic ranking, Social network analysis, Keyword extraction, Co-occurrence similarity measures, Graph clustering.**

I. INTRODUCTION

Microblogs have become one of the most popular social media outlets. One microblogging 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.

To achieve its goal, BlogNewsRank 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. Background And Related Work

The main research areas applied in this paper include: topic identification, topic ranking, social network analysis, keyword extraction, co-occurrence similarity measures, and graph clustering.

A. *Topic Identification*

Many methods have been proposed for keyphrase extraction. Most of them are based on machine learning techniques.

Much research has been carried out in the field of topic identification--Two traditional methods for detecting topics are LDA [1] and PLSA [2], [3]. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. PLSA, similarly, is a statistical technique, which can also be applied to topic modeling.

Another trending area of related research is the detection of “bursty” topics (i.e., topics or events that occur in short, sudden episodes). Diao et al. [4] proposed a method that uses a state machine to detect bursty topics in microblogs. Their method also determines whether user posts are personal or refer to a particular trending topic. Yin et al. [5] also developed a model that detects topics from social media data, distinguishing between temporal and stable topics. These methods, however, only use data from microblogs and do not attempt to integrate them with real news. Additionally, the detected topics are not ranked by popularity or prevalence.

B. *Topic Ranking*

Another major concept that is incorporated into this paper is topic ranking. There are several means by which this task can be accomplished, traditionally being done by estimating how frequently and recently a topic has been reported by mass media.

Ranking is the central problem for many information retrieval applications, such as document retrieval and collaborative filtering. Recently a new research area is emerging in machine learning, which is called learning to rank. Learning to rank aims at automatically creating a model (function) that can perform ranking on instances, using training data and machine learning techniques. Many learning to rank methods have been developed and applied to information retrieval.

Shubhankar et al.[6] developed an algorithm that detects and ranks topics in a corpus of research papers. They used closed frequent keyword-sets to form topics and a modification of the PageRank[7] algorithm to rank them. Their work, however, does not integrate or collaborate with other data sources, as accomplished by BloggyNewsRank.

C. *Social Network Analysis*

In the case of UA, Wang et al. [8] estimated this factor by using anonymous website visitor data. Their method counts the amount of times a site was visited during a particular period of time, which represents the UA of the topic to which the site is related.

Additionally, we believe that the relationship between social media users who discuss the same topics also plays a key role in topic relevance. Kwan et al. [9] proposed a measure referred to as reciprocity, which attempts to detect the interaction between social media users and perceive their engagement in relation to a particular topic. Higher reciprocity means greater interaction between users, and thus topics with higher reciprocity should be considered more important because of their underlying community structure. We can inherently identify the power and influence of a well-structured community as opposed to a decentralized and unstructured one. Our method applies this logic to support the idea that higher reciprocity signifies greater importance.

D. *Keyword Extraction*

Concerning the field of keyword or informative term extraction, many unsupervised and supervised methods have been proposed. Unsupervised methods for keyword extraction rely solely on implicit information found in individual texts or in a text corpus. Supervised methods, on the other hand, make use of training datasets that have already been classified.

We use TextRank [10] to extract keywords from the news media sources. Furthermore, TextRank does not require training or any document corpus for its operation.

E. *Co-Occurrence Similarity*

Since establishing the importance of the word-pair cooccurrence distribution in the actual corpus of tweets is of more interest to us, we did not employ Bollegala’s or Chen’s semantic similarity methods. In this paper, we tested other similarity measures, and found that the Dice similarity measure provided the best results.

F. *Graph Clustering*

The main purpose of graph clustering in this paper is to identify and separate TCs. Newton proposed a new method to identify clusters based on modularity. Modularity is a measure designed to estimate the strength of division of a network into clusters. Networks that possess a high modularity value have dense connections between the nodes within each cluster, but sparse connections between nodes in different clusters. Newman’s new algorithm calculated modularity as it progressed, making it simple to find the optimal clustering structure. Given their effectiveness, the concepts of betweenness and transitivity are both applied into our graph clustering algorithm.

III. Blognewsrank Framework

The goal of our method—BlogNewsRank—is to identify, consolidate and rank the most frequent 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

vertices 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).

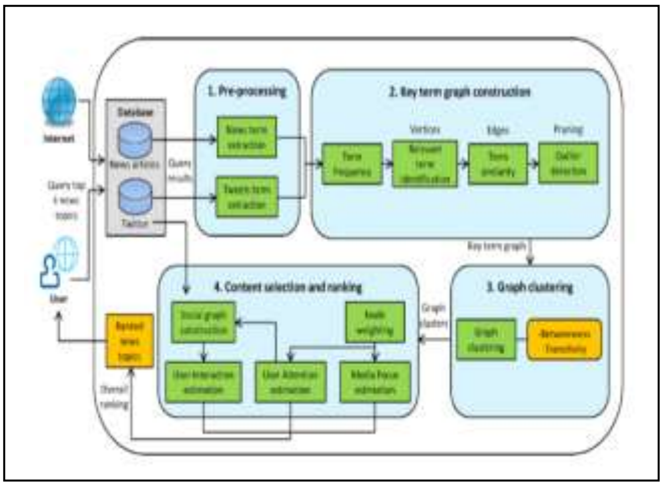

Fig. 1. Blognewsrank framework

IV. Experiments And Results

The testing dataset consists of tweets crawled from Twitter public timeline and news articles crawled from popular news websites.

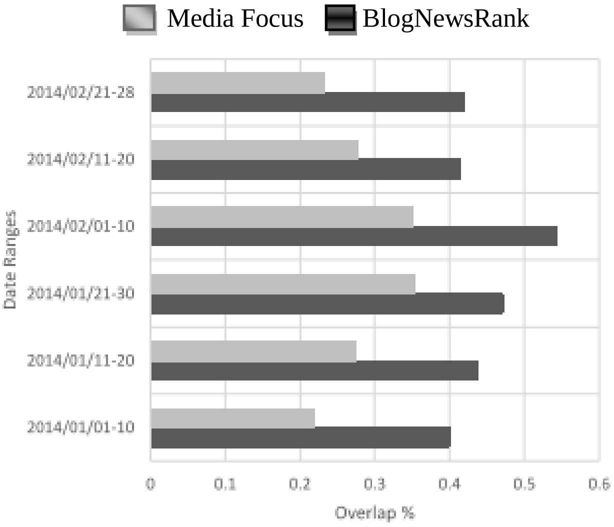

Fig. 2. Percentage of overlap between all voted topics and all topics selected by blognewsrank and mf.

A. *Method Evaluation*

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.

B. *Controlled Experiments*

In this section, we show experiments performed on some of the variables and components used in our method, namely, the co-occurrence measure, IQR coefficient, and node weighting. For the controlled experiments, the control dataset was divided into six partitions, with each partition representing ten days' worth of news and tweets.

Taking all results into consideration emphasizes the point that MF alone is a substandard estimator of what users find interesting or consider important, and should therefore not be used in this way. BlogNewsRank, on the other hand, proves to be more capable of performing this, and so we conclude that the information provided by BlogNewsRank can prove vital in commerce-based areas where the interest of users is paramount.

Conclusion

In this paper, we proposed an unsupervised method—BlogNewsRank—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.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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