International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

SOCIRANK IDENTIFYING AND RANKING PREVALENT NEWS TOPICS USING SOCIAL MEDIA FACTORS

SRIKANTH.M[1], RIHANA PARVEEN.SK[2], SIVA SAI.B[3], SRAVANI.M[4], VENU GOPAL REDDY.G[5]

[1]*B.Tech, M.Tech, Associate Professor Dept. of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P, India*
[2345] *B.Tech Students, Dept. of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P, 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 topics.*

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

1. 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, nonjournalist users are able to publish unverified content and express their interest in certain events.

1.1 Selection of Common News Topics

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.

1.2 Ranking them based on Social Media

A straightforward approach for identifying topics from different social and news media sources is the application of topic modeling. Many methods have been proposed in this area, such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA). Topic modeling 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 modeling 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.

2. Releated 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. Extensive work has been conducted in most of these areas.

2.1 Topic Identification

Much research has been carried out in the field of topic identification—referred to more formally as topic modeling. Two traditional methods for detecting topics are LDA and PLSA. 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. In these approaches, however, temporal information is lost, which is paramount in identifying prevalent topics and is an important characteristic of social media data. Furthermore, LDA and PLSA only discover topics from text corpora; they do not rank based on popularity or prevalence.

Wartena and Brussee implemented a method to detect topics by clustering keywords. Their method entails the clustering of keywords—based on different similarity measures—using the induced k-bisecting clustering algorithm. Although they do not employ the use of graphs, they do observe that a distance measure based on the Jensen–Shannon divergence (or information radius) of probability distributions performs well.

2.2 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.

2.3 Social Network Analysis

In the case of UA, Wang, 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. Our belief, on the other hand, is that, although website usage statistics provide initial proof of attention, additional data are needed to corroborate it. We employ the use of social media, specifically Twitter, as a means to estimate UA. When a user tweets about a particular topic, it signifies that the user is interested in the topic and it has captured her attention more so than visiting a website related to it.

2.4 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.

2.5 Co-Occurrence Similarity

Matsuo and Ishizuka suggested that the co-occurrence relationship of frequent word pairs from a single document may provide statistical information to aid in the identification of the document's keywords. They proposed that if the probability distribution of co-occurrence between a term x and all other terms in a document is biased to a particular subset of frequent terms, then term x is likely to be a keyword. Even though our intention is not to employ co-occurrence for keyword extraction, this hypothesis emphasizes the importance of co-occurrence relationships.

2.6 Graph Clustering

The main purpose of graph clustering in this paper is to identify and separate TCs, as done in Wartena and Brussee's work. Iwasaka and Tanaka-Ishii also proposed a method that clusters a co-occurrence graph based on a graph measure known as transitivity. The basic idea of transitivity is that in a relationship between three elements, if the relationship holds between the first and second elements and between the second and third elements, it also holds between the first and third elements. They suggested that each output cluster is expected to have no ambiguity, and that this is only achieved when the edges of a graph (representing co-occurrence relations) are transitive.

3. 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.

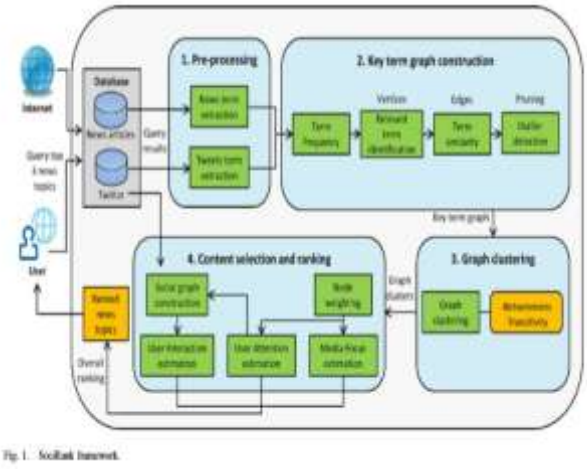
1) *Preprocessing*: Key terms are extracted and filtered from news and social data corresponding to a particular period of time.

2) *Key Term Graph Construction*: A graph is constructed from the previously extracted key term set, whose 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).

Initially, news and tweets data are crawled from the Internet and stored in a database. News articles are obtained from specific news websites via their RSS feeds and tweets are crawled from the Twitter public timeline . A user then requests an output of the top k ranked news topics for a specified period of time between date d1 (start) and date d2 (end).



3.1 Preprocessing

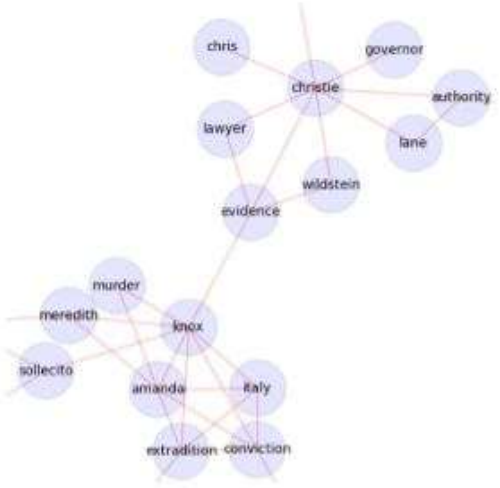
In the preprocessing stage, the system first queries all news articles and tweets from the database that fall within date d1 and date d2. Additionally, two sets of terms are created: one for the news articles and one for the tweets, as explained below.

1) *News Term Extraction*: The set of terms from the news data source consists of keywords extracted from all the queried articles. Due to its simple implementation and effectiveness, we implement a variant of the popular TextRank algorithm to extract the top k keywords from each news article.2 The selected keywords are then lemmatized using the WordNet lemmatizer in order to consider different inflected forms of a word as a single item. After lemmatization, all unique terms are added to set N. It is worth pointing out that, since N is a set, it does not contain duplicate terms.

2) *Tweets Term Extraction*: For the tweets data source, the set of terms are not the tweets' keywords, but all unique and relevant terms. First, the language of each queried tweet is identified, disregarding any tweet that is not in English. From the remaining tweets, all terms that appear in a stop word list or that are less than three characters in length are eliminated. The part of speech (POS) of each term in the tweets is then identified using a POS tagger. This POS tagger is especially useful because it can identify Twitter-specific POSs, such as hashtags, mentions, and emoticon symbols.

3.2 Key Term Graph Construction

In this component, a graph G is constructed, whose clustered nodes represent the most prevalent news topics in both news and social media. The vertices in G are unique terms selected from N and T, and the edges are represented by a relationship between these terms. In the following sections, we define a method for selecting the terms and establish a relationship between them. After the terms and relationships are identified, the graph is pruned by filtering out unimportant vertices and edges.



3.3 Graph Clustering

Once graph G has been constructed and its most significant terms (vertices) and term-pair co-occurrence values (edges) have been selected, the next goal is to identify and separate well-defined TCs (subgraphs) in the graph. Before explaining the graph clustering algorithm, the concepts of betweenness and transitivity must first be understood.

$$\text{betweenness}(e) = \sum_{i,j \in V} \frac{\sigma(i,j|e)}{\sigma(i,j)}$$

$$\text{transitivity}(G) = \frac{\#\text{triangles}}{\#\text{triads}}$$

3.4 Content Selection and Ranking

Now that the prevalent news-TCs that fall within dates d1 and d2 have been identified, relevant content from the two media sources that is related to these topics must be selected and finally ranked. Related items from the news media will represent the MF of the topic. Similarly, related items from social media (Twitter) will represent the UA—more specifically, the number of unique Twitter users related to the selected tweets. Selecting the appropriate items (i.e., tweets and news articles) related to the key terms of a topic is not an easy task, as many other items unrelated to the desired topic also contain similar key terms.

4. 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.

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

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

[8] 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.

[9] 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.