© 2019 IJSRST | Volume 6 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : 10.32628/IJSRST196277

A Review on Identifying and Ranking Current News Topics

Neha Vijay Manwatkar [1], Prof. Jayant Adhikari[2], Prof. Rajesh Babu[3]
[1]M.Tech Scholar, Department of Computer Science and Engineering Tulsiramji Gaikwad-Patil College of Engineering and Technology Nagpur, Maharashtra, India
[2,3]Department of Computer Science and Engineering Tulsiramji Gaikwad-Patil College of Engineering and Technology Nagpur, Maharashtra, India

ABSTRACT

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, In present day times, internet based life administrations, for example, Twitter give a gigantic measure of client created information, which can possibly contain useful news-related substance. Twitter as a new form of social media can potentially contain much useful information, but content analysis on Twitter has not been well studied. Mass media sources such as news media used to inform us about daily events. For these assets to be helpful, we must find a way to filter noise and only capture the content that, in view of its closeness to the news media, is thought about significant. Be that as it may, even after noise is evacuated, data overload may even now exist in the rest of the information. Henceforth, it is advantageous to organize it for utilization. To accomplish prioritization, data must be positioned arranged by evaluated significance thinking about three variables. In the first place, the transient commonness of a specific point in the news media is a factor of significance, and can be viewed as the media center (MF) of a subject. Second, the fleeting predominance of the theme in social media demonstrates its client consideration (UA). Last, the communication between the internet based life clients who notice this theme demonstrates the quality of the network talking about it, and can be viewed as the client connection (UI) around the subject. We propose an unsupervised system SociRank which distinguishes news points common in both web-based social networking and the news media, and after that positions them by significance utilizing their degrees of MF, UA, and UI. Our analyses demonstrate that SociRank improves the quality and assortment of naturally recognized news points.

Keywords : Information Filtering, Social Computing, Social Network Analysis, Topic Identification, Topic Ranking.

I. INTRODUCTION

Now a day, extracting and mining valuable info from online sources has become an important zone in IT. Certainly, data that informs the general populace of step by step events has been given by expansive interchanges sources, explicitly the news media. Gigantic quantities of these news media sources have either surrendered their printed duplicate dispersions or moved to the World Wide Web, or now make both printed variant and Internet shapes at the same time. These news media sources are seen as trustworthy since they are conveyed by capable writers, who are viewed as in charge of their substance. On the other hand, the Internet, being a free and open social affair for information exchange,

has starting late watched an enamoring wonder known as web based systems administration. In web based systems administration, predictable, non-writer clients can disperse unsubstantiated substance and express their eagerness for explicit occasions. Microblogs have ended up being a champion among the most celebrated web based systems administration outlets. One microblogging organization explicitly, Twitter, is used by countless around the world, star viding colossal proportions of customer made data. One may acknowledge that this source conceivably contains information with equal or more essential impetus than the news media, anyway one ought to likewise expect that because of the unconfirmed thought of the source, a considerable amount of this substance is useless. For online long range informal communication data to be of any usage for point distinguishing proof, we should discover a way to deal with channel uninformative information and catch just information which, in light of its substance equivalence to the news media, might be viewed as helpful or important. The mining of important data from online sourceshas turned into a conspicuous research area in data.Innovation lately. Truly, learning that informs theoverall population of day by day occasions has beengiven by broad communications sources,specifically the news media. A significant numberof these news media sources have eithersurrendered their printed version productions ormoved to the World Wide Web, or now create bothprinted copy and Internet forms all the while. Thesenews media sources are viewed as reliable sincethey are distributed by proficient journalists, whoare considered responsible for their substance. Thenagain, the Internet, being a free and open discussionfor data trade, has as of late observed an interestingwonder known as online networking. In online networking, standard, non-journalist users candistribute unverified substance and express theirenthusiasm for specific occasions. The infiltrationof immense measure of data through the WorldWide Web (WWW) has made a developingrequirement for the advancement of procedures forfinding, getting to, and sharing learning. Thekeyphrases help perusers quickly comprehend, sortout, access, and offer data of an archive.Keyphrases are the expressions comprising of atleast one noteworthy words. Keyphrases can befused in the query items as subject metadatatoencourage data look on the web. A direct approach for recognizing Microblogs suchas Twitter reflect the common public's responses tomajor occasions. Bursty themes from microblogsuncover what occasions have pulled in the mostonline consideration. In spite of the fact that burstyoccasion discovery from content streams has beenconsidered some time recently, past work may notbe appropriate for microblogs since compared withother content streams such as news articles andlogical distributions, microblog posts are especiallyassorted and loud. To discover points that havebursty designs on microblogs, a theme demonstratethat at the same time captures two perceptions: posts distributed around the same time are morelikely to have the same theme, and postsdistributed by the same client are more likely tohave the same theme. The previous makes adifference discover event-driven posts while the lastmentioned makes a difference recognize andchannel out "individual" posts. Our tests on a hugeTwitter dataset appear that there are moresignificant and interesting bursty themes in the top-rank.

Small scale websites have turned out to be a standout among the most prevalent online networking outlets. One small scale blogging administration specifically, Twitter, is utilized by a great many individuals around the globe, giving tremendous measures of client created information. One may accept that this source conceivably contains data with equivalent or

more noteworthy incentive than the news media, however one should likewise expect that in view of the unverified idea of the source, quite a bit of this substance is futile. For online networking information to be of any utilization for point identification, we should find an approach to filter uninformative data and catch just data which, in light of its substance similitude to the news media, might be viewed as helpful or profitable.

Online social systems have ended up greatly prevalent; various locales permit clients to associate and share substance utilizing social joins. Clients of these systems frequently set up hundreds to indeed thousands of social joins with other clients. As of late, analysts have proposed looking at the movement organize-a organize that is based on the real interaction between clients, Or maybe than simple fellowship-to recognize between solid and frail joins. While introductory ponders have driven to bits of knowledge on how an action organize is basically distinctive from the social organize itself, a common and vital perspective of the movement arrange has been neglected: the reality that over time social joins can develop more grounded or weaker.

A clear approach for recognizing themes fromdiverse social and news media sources is theapplication of subject modeling. Numerousstrategies have been proposed in this region, such asidle Dirichlet allotment (LDA) and probabilisticidlesemantic investigation (PLSA). Subjectmodeling is, in pith, the disclosure of “topics” incontent corpora by clustering together regularly co-occurring words. This approach, in any case, missesout in the worldly component of predominant themelocation, that is, it does not take into account howsubjects alter with time. Besides, point modelingand other theme location procedures do not ranksubjects agreeing to their ubiquity by taking intoaccount their predominance in both news media andsocialmedia

Firstly, the data is taken from various databases i.e News articles and social networking websites and sorted for the process to start. Now the query results are preprocessed. The preprocessing is followed by key term graph construction. The key term graph is sent for further process called graph clustering. The graph clusters are proceeded for content selection and ranking and now based on the relevance factors the rank of topics is determined. Programmed key phrase extraction strategies have by and large taken either supervised or unsupervised approaches. Supervised strategies extricate key phrases by utilizing at raining report set, in this way obtaining information from a worldwide collection of texts.

To help in the prioritization of news data, news must be positioned in arrange of evaluated significance. The worldly predominance of a specific point in the news media shows that it is broadly secured by news media sources, making it an imperative figure when assessing topical pertinence. This figure may be alluded to as the MF of the theme. The worldly predominance of the point in social media, specifically in Twitter, demonstrates that clients are interested in the point and can give a premise for the estimation of its ubiquity.

We introduce an unsupervised systemSociRankwhich 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 oursystem is validated by extensive controlled and uncontrolled experiments.

The section I explains the Introduction of SociRank. Section II presents the literature review of existing systems and Section III present proposed system implementation details Section IV presents experimental analysis, results and discussion of proposed system. Section V concludes our proposed system. While at the end list of references paper are presented.

II. LITERATURE REVIEW

Author D. M. Blei, A. Y. Ng, and M. I. Jordan, describe latent Dirichlet allocation (LDA)[1], a generative probabilistic model for collections of discrete data such as text corpora. They says LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. They also present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model. LDA is a simple model, and although we view it as a competitor to methods such as LSI and pLSI in the setting of dimensionality reduction for document collections and other discrete corpora, it is also intended to be illustrative of the way in which probabilistic models can be scaled up to provide useful inferential machinery in domains involving multiple levels of structure. Indeed, the principal advantages of generative models such as LDA include their modularity and their extensibility. As a probabilistic module, LDA can be readily embedded in a more complex model a property that is not possessed by LSI. In recent work we have used pairs of LDA modules to model relationships between images and their corresponding descriptive captions.

T. Hofmann proposed a novel method for unsupervised learning, called Probabilistic Latent Semantic Analysis (PLSA) [2], which is based on a statistical latent class model. Author argued that this approach is more principled than standard Latent Semantic Analysis, since it possesses a sound statistical foundation. Tempered Expectation Maximization has been presented as a powerful fitting procedure. Also they experimentally verified the claimed advantages achieving substantial performance gains. Probabilistic Latent Semantic Analysis has thus to be considered as a promising novel unsupervised learning method with a wide range of applications in text learning and information retrieval.

Author T. Hofmann presented a novel method for automated indexing [3] based on a statistical latent class model. This approach has important theoretical advantages over standard LSI, since it is based on the likelihood principle, defines a generative data model, and directly minimizes word perplexity. It can also take advantage of statistical standard methods for model fitting, over fitting control, and model combination. The empirical evaluation has clearly confirmed the benefits of Probabilistic Latent Semantic Indexing which achieves significant gains in precision over both, standard term matching and LSI. Further investigation is needed to take full advantage of the prior information provided by term

weighting schemes. Recent work has also shown that the benefits of PLSA extend beyond document indexing and that a similar approach can be utilized, e.g., for language modeling and collaborative filtering.

Mario Cataldi, Luigi Di Caro proposed a novel topic detection technique [4] that permits to retrieve in real-time the most emergent topics expressed by the community. First, they 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 emergingif 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.

In this paper author present a Knowledge Discovery System (KDS) [5] for document processing and clustering. The clustering algorithm implemented in this system, called Induced Bisecting k-Means, outperforms the Standard Bisecting k-Means and is particularly suitable for on line applications when computational efficiency is a crucial aspect. Because of 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. So used some methods like Clustering algorithms are of crucial importance when there are no labels associated to textual information or documents. The aim of clustering algorithms, in the text mining domain, is to group documents concerning with the same topic into the same cluster, producing a flat or hierarchical structure of clusters.

The aim of a linguistic science is to be able to characterize and explain the multitude of linguistic observations [6] circling around us, in conversations, writing, and other media. Part of that has to do with the cognitive side of how humans acquire, produce, and understand language, part of it has to do with understanding the relationship between linguistic utterances and the world, and part of it has to do with understanding the linguistic structures by which language communicates. In order to approach the last problem, author proposed that there are rules which are used to structure linguistic expressions.

Author empirically compare the content of Twitter with a traditional news medium, New York Times, using unsupervised topic modeling. They use a Twitter-LDA model [8] to discover topics from a representative sample of the entire Twitter. They then utilized text mining techniques to compare these Twitter topics with topics from New York Times, taking into consideration topic categories and types. They also study the relation between the proportions of opinionated tweets and retweets and topic categories and types. Our comparisons show interesting and useful findings for downstream IR or DM applications.

Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma [11] proposed a novel automatic online algorithm for news topic ranking based on an aging theory, using both media focus and user attention. Both media focus and user attention varies as time goes on, so the effect of time on topic ranking has already been included. Inconsistency exists between media focus and user attention, which is analyzed and quantitatively measured in this paper. Topics are ranked by the combination of their media focus and user attention values online automatically. Related news stories of topics are provided for users' quick access. Empirical evaluation on the topic ranking

result indicates that the proposed topic ranking algorithm reflects the influence of time, the media and users.

Wanget al.[11] proposed a method that takes into account the users' interest in a topic by estimating the amount of times they read stories related to that particular topic. They refer to this factor as the UA. They also used an aging theory developed by Chenet al.[12] to create, grow, and destroy a topic. The life cycles of the topics are tracked by using an energy function. The energy of a topic increases when it becomes popular and it diminishes over time unless it remains popular. We employ variants of the concepts of MF and UA to meet our needs, as these concepts are both logical and effective.

Author J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling investigate the use of Twitter to build a news processing system, called TwitterStand [13], from Twitter tweets. The idea is to capture tweets that correspond to late breaking news. The result is analogous to a distributed news wire service. The difference is that the identities of the contributors/reporters are not known in advance and there may be many of them. Furthermore, tweets are not sent according to a schedule: they occur as news is happening, and tend to be noisy while usually arriving at a high throughput rate. Some of the issues addressed include removing the noise, determining tweet clusters of interest bearing in mind that the methods must be online, and determining the relevant locations associated with the tweets.

Author E. Kwan, P.-L. Hsu, J.-H. Liang, and Y.-S. Chen, have introduced and evaluated Keyword-based Evolving Graph Sequences [17] for event identification purposes, and demonstrated how social structure in social streams data can be utilized for event identification. Furthermore, also proposed the use of a hidden link (hidden relationship) for event identification. The experimental results show the usefulness of our approach in identifying real-world events in social streams.

In this paper author propose a new method of computing term specificity, based on modeling the rate of learning of word meaning in Latent Semantic Analysis (LSA) [18]. We analyze the performance of this method both qualitatively and quantitatively and demonstrate that it shows excellent performance compared to existing methods on a broad range of tests. They also demonstrate how it can be used to improve existing applications in information retrieval and summarization.

System Architecture

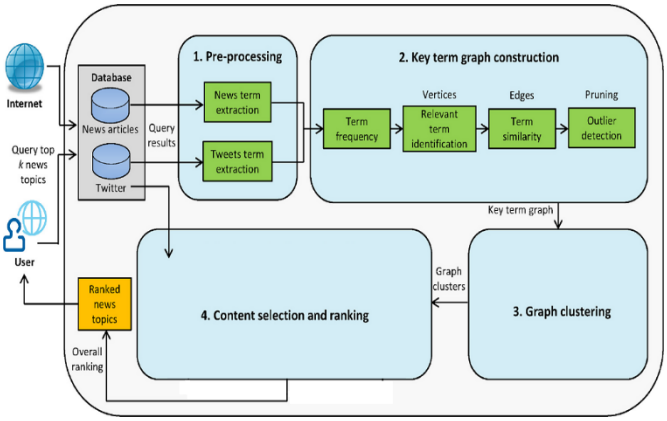

Fig 1. System Architecture

III. RESULT AND DISCUSSIONS

A. Experimental Setup

All the experimental cases are implemented in Java in congestion with Netbeans tools and MySql as backend, algorithms and strategies, and the competing classification approach along with various feature extraction technique, and run in environment with System having configuration of Intel Core i5-

6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM

IV. CONCLUSION

We proposed an unsupervised system SociRank which distinguishes news subjects unavoidable in both web based systems administration and the news media, and after that positions them by thinking about their MF, UA, and UI as essentialness factors. The transient transcendence of a particular topic in the news media is seen as the MF of a point, which gives us understanding into its wide correspondences reputation. The worldly predominance of the subject in online networking, specifically Twitter, demonstrates client between est, and is viewed as its UA. At long last, the connection between the online networking clients who say the theme shows the quality of the group talking about it, and is viewed as the UI. To the best of our insight, no other work has endeavored to utilize the utilization of either the interests of online networking clients or their social connections to help in the positioning of points. Solidified, filtered, and positioned news themes from both expert news suppliers and people have a few benefits.

V. 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]. Mario Cataldi, Luigi Di Caro" Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation" MDMKDD'10, July 25th, Washington, DC, USACopyright 2010 ACM 978-1-4503-0220-3

[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 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. LongPapers, 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 ACMSIGSPATIAL 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. ChapterAssoc. Comput. Linguist., 2009, pp. 530-538.

Cite this article as :

Neha Vijay Manwatkar, Prof. Jayant Adhikari, Prof. Rajesh Babu, "A Review on Identifying and Ranking Current News Topics", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 2, pp. 412-419, March-April 2019.

Journal URL : http://ijsrst.com/IJSRST196277