ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

Most Trending Topics with Pre-learned Knowledge in Twitter

Leena Patil[1], Kanchan Doke[2]

P.G. Student, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India[2]

ABSTRACT: Finding trending topic from the large amount of user-generated content (UGC) in social media helps to make easier lots of downstream applications of intelligent computing. Topic models, as one of the most powerful algorithms, have been mostly used to discover the similar patterns in text collections. The weakness of topic models is that they need documents with particular length to provide reliable statistics for generating coherent topics. Tweets from the users are mostly short and noisy, in twitter. Observations of word co-occurrences are inconceivable for topic model so obtain better results, previous work tried to incorporate prior knowledge to deal with this problem. However, this strategy is not practical for the fast evolving UGC in Twitter. We first cluster the users according to the retweet network, and the user's interests are mined as the prior knowledge, in this paper. Such data are then applied to improve the performance of topic learning. Users in the same community usually share similar interests, which will result in less noisy sub-data sets is the potential cause for the effectiveness. Our algorithm pre-learns two types of interest knowledge from the data set: the interest-word-sets and a tweet interest preference matrix. A dedicated background model is introduced further to judge whether a word is drawn from the background noise. Experiments on two real life twitter data sets show that our model achieves significant improvements over state-of-the-art baselines.

KEYWORDS: Topic model, social network, short texts.

I. INTRODUCTION

The tremendous amount of information generated by Online Social Networks (OSNs) has attracted enormous attention. Users in mobile social networks can share locations, textual content and videos with their friends, which raise great challenges for the existing data mining techniques. Topic modelling is one of the fundamental problems in the data mining applications. Statistical topic models, such as PLSA and LDA, provide powerful frameworks for analysing latent semantics underlying the news datasets. Naturally, researchers also apply them on social textual collections to discovering the fast evolving topics.

However, one important attribute of social texts is the extremely short length, which significantly deteriorates the performance of traditional topic models. In other words,the co-occurrence of words in tweets is not sufficient for topic models to discover latent patterns. Due to the ineffectiveness of traditional topic models on short texts, researchers tried to incorporate external knowledge to improve the topic modelling performance. Weng et al. propose to combine all the tweets of an individual into a single document. However, this approach does not reduce the noise inside. Conversely, it may make the word co-occurrences more puzzling. Some other studies also point out that combining or splitting documents contributes little to the final results of topic models. Zhao et al. propose the Twitter-LDA, which assumes that each tweet only has one topic. However, thisis not a reasonable hypothesis. For example, the short tweet, ‘‘Financing education is expensive for the government’’, is essentially related to two topics which are ‘‘Education’’ and ‘‘Economy’’. The strong constrain may deteriorate the modelperformance. Another train of thought is to incorporate prior knowledge, and several knowledge-based models have been proposed to optimize the basic LDAmodel. For example, MDK-LDA leverages synonym and antonym sets (called s-set) extracted from Word Net to

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

generate more meaningful topics. It assumes that words are drawn based on not only topics but also s-sets. DF-LDA takes domain knowledge in the form of must-links (i.e., words should occur together) and cannot-links (i.e., words should occur together) to restrain the distributions of topics. All these models are based on one assumption that the knowledge introduced is correct, and it can also be easily obtained. However, the rapidly evolving data in social media makes it difficult to obtain proper prior knowledge. In this paper, we propose a novel topic model called SILDA (LDA with Social Interest). We develop the model from two ways to achieve better performance: one is to promote the distinctiveness of topics by incorporating the interest knowledge. The other is to reduce the background noise. In the promoting process, the interests are regarded as prior knowledge, which is very similar to the must-links. However, discovering proper prior knowledge for the rapidly evolving tweets is extraordinarily difficult. Here we propose to learn it from the dataset itself, and then apply it to guide the model inference. In order to make the learned knowledge reliable, we first divide the dataset into several less noisy sub-datasets. The main reason why a traditional topic model performs poorly is that the noisy and short tweets overwhelm the valid co-occurrence observations. With the partition, texts in the same sub-dataset would share similar topics, and thereby be more concentrated. Thus it is convincing that the learned knowledge is better. According to the users' relationships in twitter, we can conduct many kinds of partition methods. In this paper, we apply the re-tweeting behaviour of users. An individual retweet another only when he or she reads and approves the content. Users who are strongly connected by re-tweeting links are more probable to share similar interests. Our model mines two types of knowledge from the non-overlapping sub-datasets: the interest-word-sets and the tweet-interest preference matrix. In fact, most tweets are strongly related to the authors' interests. Thus in the generative process, we add a new latent variable s denoting the interests, and assume a new topic distribution of topics overs. In other words, tweets are assigned with a higher probability to the topics which are related to its author's interests. What is more, the interest-word-set will be promoted as a whole to make the final topics more coherent. In the noise removing process, a Bernoulli distribution is introduced to determine whether the word is from background noise. If the word has a higher probability drawn from the background model, it would be regarded as unnecessary information, thus it would remove from the learning phase. This step will remove many frequent but meaningless words such as the emoticons. Before going further, we discuss something about the partition methods. In our model, we treat the individuals as nodes, and assign an edge to two users if one person re-tweets any tweet of the other. The weight of an edge is defined as there-tweeting counts. The larger the weight is, the closer the two nodes are. Given the re-tweeting network, the simplest way for partition is to apply community detection algorithms. Many studies have been done in this field.

II. Related work

Traditional topic models, such as LDA and PLSA, provide powerful statistical frameworks to discover the latent topics in large text collections. Based on the observation of word co-occurrences, words with the same meanings are aggregated. Such unsupervised models are first proposed for news data which is rather longer than tweets. Previous studies noticed traditional topic models always perform poor on tweet datasets which are extremely short and noisy. However, witnessing the dramatic increase of online social media, many studies apply LDA as a basic method to explore the latent topical information in Twitter. Weng et al. combine all the tweets of an individual document to increase the documents length. However, it can hardly improve the model's performance. Some other work assumes that each tweet has exactly one topic. Actually, it is not a very reasonable hypothesis.

For example, a single tweet ``Financing education is expensive for the government'' is distinctly related to two topics ``Education'' and ``Economy''. Some other work which does not focus on social media datasets can give us some inspirations. The knowledge-based models are proposed to incorporate prior knowledges to optimize the topic modeling. Chen et al. leverage domain knowledge extracted from the WordNet to help analyze datasets. Ramage et al. restrain the documents only to choose the topics corresponding to the known labels to produce better topics in labeled datasets. However, finding proper prior knowledge for tweet datasets is an extremely difficult task. Moreover, incorrect knowledge always results in good looking results, which however may not _t the dataset itself. For example, topics with good word descriptions may be mainly affected by the prior knowledge but not by the dataset itself.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

In order to solve these problems, researchers try to take advantage of the knowledge underlying the data, such as social relations, citation links and temporal information. Typically, by adding the network information to the similar function of PLSA, Mei et al. propose a very general framework, called NetPLSA, to model a kind of problems in which the datasets have accompanied network structures. Wang et al. analyze the topic sentiment of tweets with hashtag (``#") which is a symbol manually de_ned as ``topic" in social media to express common interests. All these studies have strong relationships with our proposed model, but none of them focuses on improve the model performance on the short tweets. With increasing popularity of social media, making topic models produce more coherent topics is stimulating more and more interests. Social text analysis has been a hot research spot for quite a few years. Many techniques have been proposed to mine the implicit information hidden in the social networks.

Zhou et al. propose a probabilistic model to extract the e communities based on the content of communication documents. The retweeting which is a kind of subjective behavior of individuals is always applied to analyze the propagation of events. Meanwhile, it is also considered as a good representation of users' interest or the content preferences. Combining textual content with retweeting networks is a very interesting _eld to perform topic modeling. However, very little attention has been paid in this _eld. Another very important phase in our model is to cluster users in the retweeting network. Many algorithms have been proposed to discover communities.Wu et al. propose a very comprehensive demonstration for spectral clustering applied in community detection. Since detecting community is a complex but well studied _eld, we do not describe too much detail in this paper. To overcome the sparsity problem of social networks, we apply the smart local moving (SLM) algorithm whose ef_ciency has been demonstrated by many previous studies.

III. TOPIC MODELING ISSUE

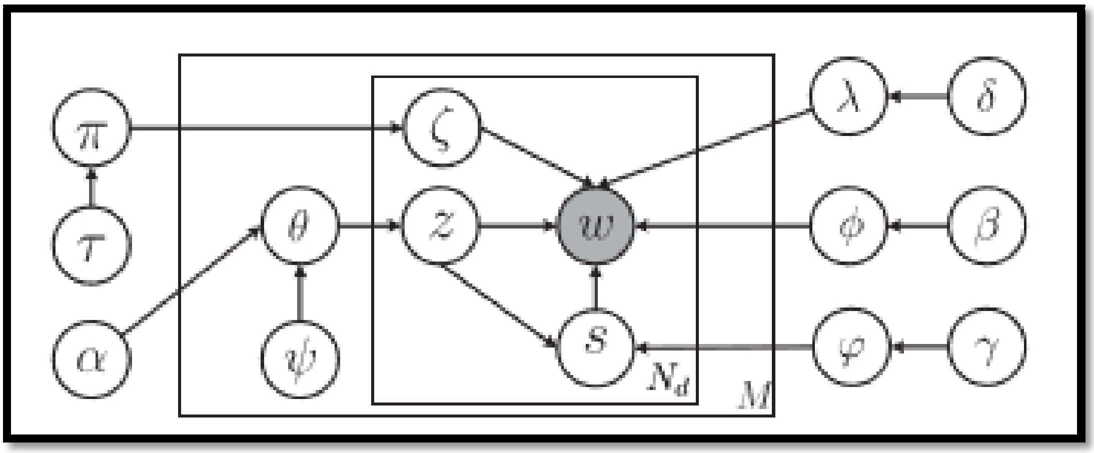

Figure 1: Topic modeling issue

In existing system topic modeling module has issue that it is unable to extract proper data from source. The structure of output is not matching with our required format. Problematic graphical model of topic modeling is as shown in previous figure.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

IV.LDA FRAMEWORK

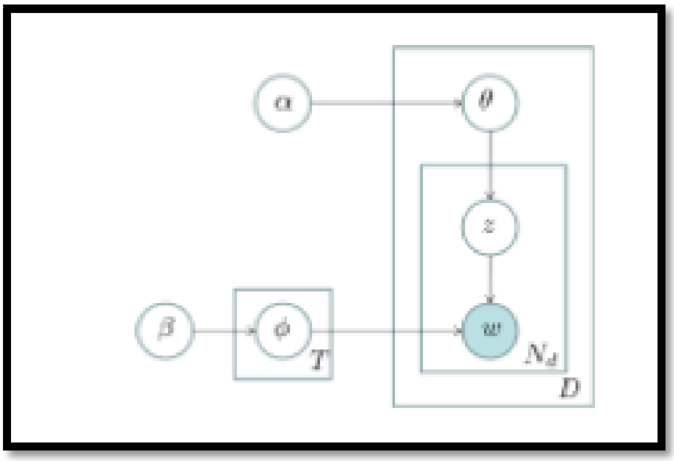

Figure 2: LDA framework

Graphical representation of LDA framework model is as shown in above figure. LDA framework is only used to analyse the twitter data. This framework is unable to extract the data in proper format.

Sr. No.	Paper	Methodology	Feature Set	Advantages	Limitations
1.	A smart local moving algorithm for large-scale modularity-based community detection	SLA Algorithm	Data extraction from source like twitter data sorce	Provides good results only in small scale systems	Large number of iterations due to huge data which cannot be handled.
2.	ETM: Entity Topic Models for Mining Documents Associated with Entities	Entity Topic Model (ETM)	To design topic model	This technique solve the problem of word co-occurrence in pairs of topic and entity model	Problem is occurred in entity information document of designing topic model.
3.	Predicting the content dissemination trends by repost behaviormodeling in mobile social networks	Social networks using PCA	Popularity of contents in social network	Tracks or captures the users behaviour	Works in iterative manner which results in loss of performance
4.	Large-Scale High-Precision Topic Modeling on Twitter	Integreative algorithm	Decision agregation	On the basis of majority voting, algorithm works	Unable to run in real-time close-loop iteration because of data-drift.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

5.	Twitterrank: Finding topic-sensitive influential Twitterers	LDA framework	Analysis of data	LDA framework is used to analyse the large amount of data	LDA framework only analyse the data, unable to extract the data from sources.
6.	Emerging Topic Detection for Organizations from Microblogs	Real Time Framework	Analysis od data	Real Time Framework is used to crawling the data, discover topics and to identify the topics.	Real Time Framework only analyse the data.
7.	Personalization and Context-awareness in Social Local Search: State-of-the-art and Future Research Challenges	Local Search	Location-based services	Based on current user location, data emerged quickly.	Give best result locally only.
8.	Influence Propagation Model for Clique-Based Community Detection in Social Networks	Biased density metric	Active users	Detect interaction with neighbourhood	Communications between active users only
9.	INDEPENDENT COMPONENT ANALYSIS IN MULTIMEDIA MODELING	Blind source separation method	Independent component analysis ICA	Multimedia data intelligently processed	Only analyse the data
10.	Community detection : topological vs topical	Community detection approach	Topology based and topic based	Analyse community data	Only analysis of community data is done

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

V. REIMPLEMENTING THE PROPOSED SYSTEM

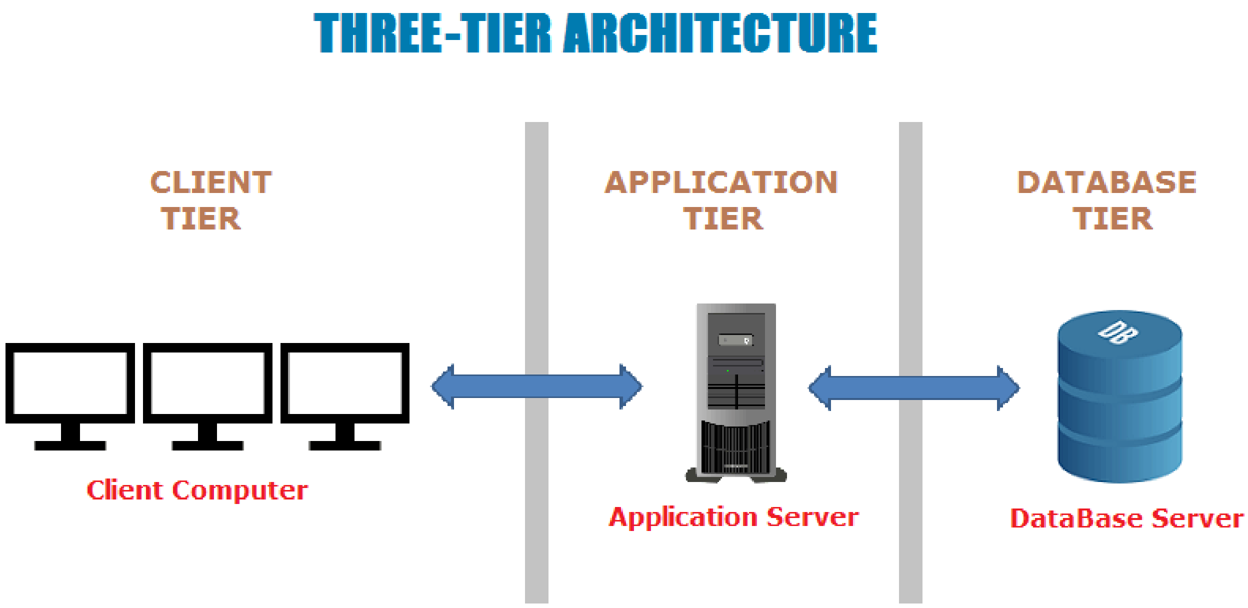

Figure 3: Three-tier architecture

In this system we follow three tier system architecture as shown in previous figure. The twitter data / content added to the database using twitter API or form. The heading, content, hashtags and cities are added to database using form.Using User Interface to add data we send request to application server to add data to database.Application server forward the request which come from user interface i.e. client to the database server.After storing data to database, database server sends success response to the application server.If data is not stored in database, then database server send error response to the application server.Then application server sends success or error response to the user interface i.e. client.When we fetch the data from database, we send request to the application server to fetch data.Application server forward the data fetch request to the database server.Database server process the coming request and sends success response to the application server.If database server not process the request successfully then database server sends error response to the application server.Then application server sends the success / error response to the user interface that is client.

Sr. No.	Paper	Methodology	Feature Set	Advantages	Limitations
1.	Most Trending Topics with Pre-learned Knowledge in Twitter	Automated Data Extraction Using Twitter API	Huge amount of Data in the form of topics, contents, hashtags, cities, etc.	Using Twitter API, we can easily fetch data with secure format.	We have to buy Twitter API.
2.	Most Trending Topics with Pre-learned Knowledge in Twitter	Fetching Twitter API data from database	We can fetch large amount of secure data from database.	Data is secure. No need to add/provide extra security.	Limit the exposure of data to avoid data hacking.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

VI. PSEUDO CODE

I. Step 1: Insert twitter data / tweets into database using twitter API.
II. Step 2: Fetch heading, content, hashtags and cities of tweets from use of twitter API.
III. Step 3: Store heading, content, hashtags and cities into database fetch from twitter API.
IV. Step 4: Enter word which want to search in search box.
V. Step 5: Enter search word as name of city, heading, content or hashtags.
VI. Step 6: Word which want to search is checking into database with heading,content, hashtag or city.
VII. Step 7: If search word match found with heading, content, hashtag or city thenrelated data will display.
VIII. If($word == $row['data'])
IX. Then display the related data.
X. Else
XI. Data will not display.

VII. SIMULATION RESULTS

The aid of this system are as follows:

I. It proposes to mine knowledge from the dataset itself, and then leverage it to promote the performance of topic modelling. We call such process learning twice from the data.
II. By introducing the latent interests, it can handle the cross-community interests and multiple senses problems. With a background model, it can reduce the noise, and produce more coherent topics.
III. Proposed model and state-of the- art baselines datasets are compared by the complete evaluation.

Screenshot 1: The input page for entering the details such as city name, heading of tweet, content of tweet and the hashtags, as shown in following screenshot:

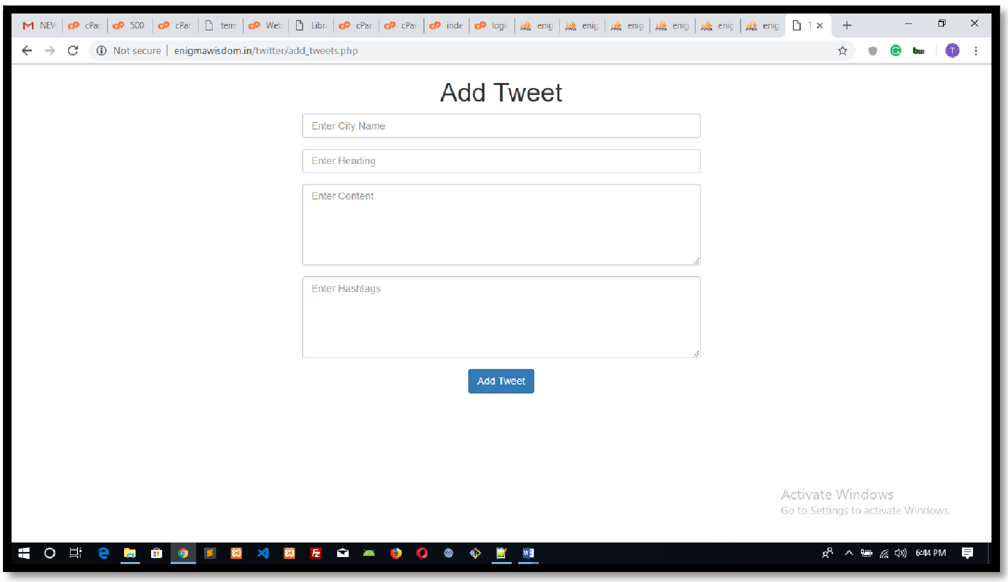

Figure 4: Form to fill information

Screenshot 2: The screenshots show how it looks after entering the details of tweet such as city name, heading of tweet, tweet information and hashtags:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

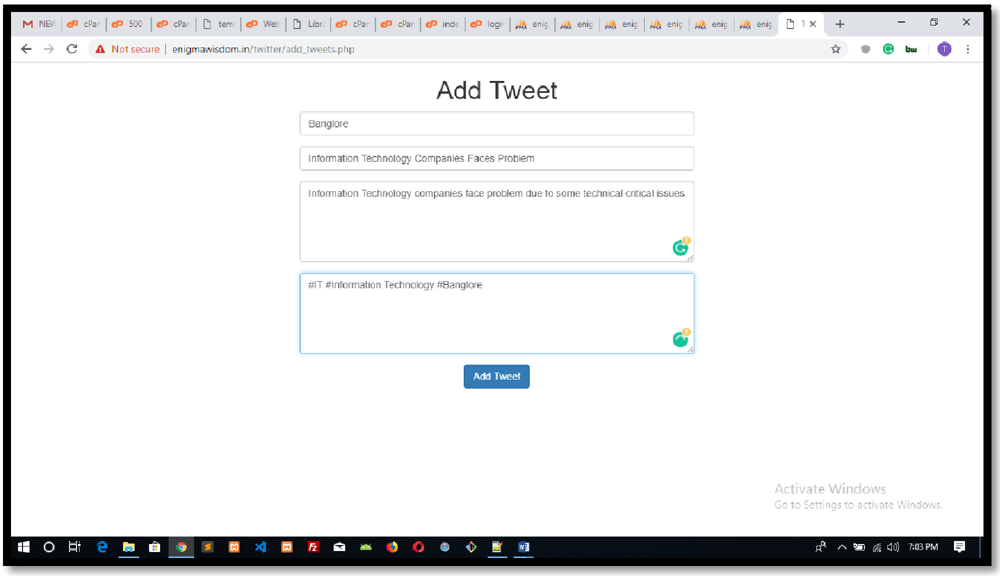

Figure 5: Form after entering information

Screenshot 3: The screenshot shows search criteria based on city name or hashtags. When we enter a city name or hashtag we get the latest tweets based on city name or hashtag:

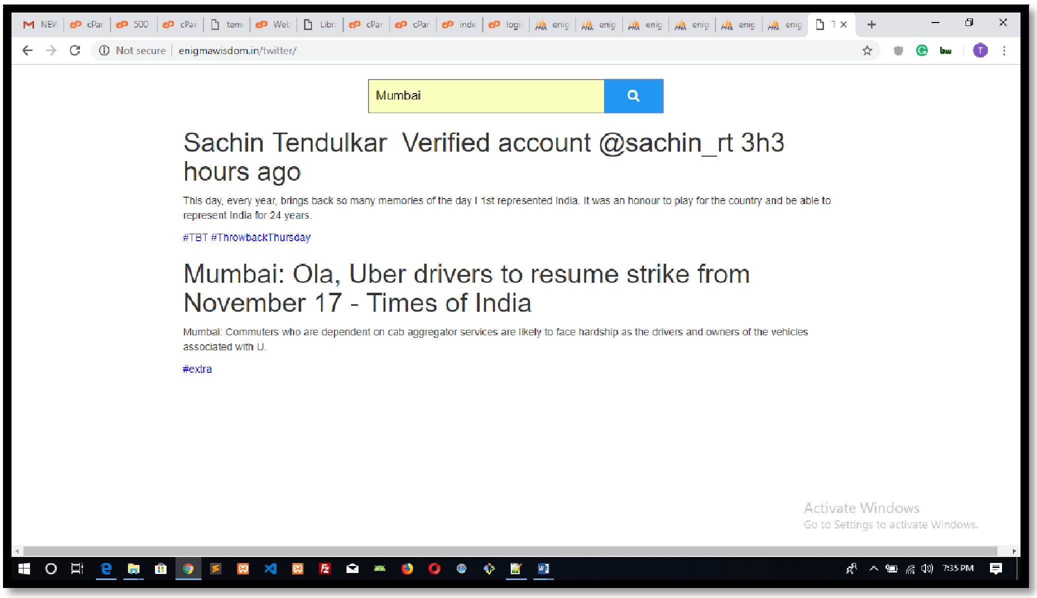

Figure 6: Output after search

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

VII. CONCLUSION AND FUTURE WORK

- **CONCLUSION**

This paper presents the simplest method to fetch the data from Twitter. An API will be used to get the data from Twitter and to store into the database. The paper presents the simplest method to extract the data from the database dependent on the location, topic of interest and # tags. This facilitates the removal of unnecessary data so users can view their interested tweets real quickly. Also the overhead on machine to run complicated algorithms is reduced by implementing simplest method. It overcomes the issue of existing system like topic modeling which is unable to extract the data in desired manner. On the other hand proposed system provides expected results as required.

- **FUTURE WORK**

I. Automatic Location Based Search: The user location can be fetched directly using Google Location Tracking API. The Search will be carried out on the basis of current location of the user. This will eliminate the need for user to each time select the location and can improve the experience.

II. Behavioural Search: The search can be added up with the ability to scan for user's behavior. This can be achieved by checking for user's login session or ip address and MAC address combination. The topics searched by the user are divided into the different categories like sports, entertainment, news or into different channels etc. and are stored into the database. The database maintains a separate table for login session and ip-MAC combination. Later whenever user goes for searching, system will show suggestions on the basis of last search. By implementing behavioural search, user's experience can be further improved.

REFERENCES

[1] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, ``Discovering coherent topics using general knowledge," Proc. CIKM, New York, NY, USA, 2013, pp. 209218. [Online].
Available: http://doi.acm.org/10.1145/2505515.2505519
[2] D. M. Blei and J. D. Lafferty, ``Dynamic topic models," in Proc. ICML, New York, NY,USA, 2006, pp. 113120. [Online].
Available: http://doi.acm.org/10.1145/1143844.1143859
[3] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, ``Emerging topic detection for organizations frommicroblogs," in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2013, pp.4352. [Online].
Available: http://doi.acm.org/10.1145/2484028.2484057
H. Kim, Y. Sun, J. Hockenmaier, and J. Han, ``Etm: Entity topic models for mining documents associated with entities," in Proc. ICDM, 2012, pp. 349358. [Online]. Available: http://dblp.unitrier.de/db/conf/icdm/icdm2012.html#KimSHH12
[5] D. Andrzejewski, X. Zhu, and M. Craven, ``Incorporating domain knowledge into topic modeling via dirichlet forest priors," in Proc. ICML, New York, NY, USA, 2009, pp. 2532. [Online].
Available: http://doi.acm.org/10.1145/1553374.1553378
[6] G. Miao et al., ``Latent association analysis of document pairs," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 14151423.
[7] X. Lu, Z. Yu, B. Guo, and X. Zhou, ``Predicting the content dissemination trendsby repost behavior modeling in mobile social networks," J. Netw. Comput. Appl., vol. 42, pp. 197207, Jun. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804514000599
[8] G. A. Miller, ``WordNet: A lexical database for English," Commun. ACM, vol.38, no. 11, pp. 3941, 1995. [Online].
Available: http://doi.acm.org/10.1145/219717.219748
[9] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, ``Optimizing semantic coherence in topic models," Proc. Conf. Empirical Methods Natural Lang. Process. Stroudsburg, PA, USA, 2011, pp. 262272. [Online].
Available: http://dl.acm.org/citation.cfm?id=2145432.2145462
[10] D. Ramage, D. Hal, R. Nallapati, and C. D. Manning, ``Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in Proc. Conf. Empirical Methods Natural Lang. Process., Stroudsburg, PA, USA, 2009,pp. 248256.[Online].
Available: http://dl.acm.org/citation.cfm?id=1699510.1699543