Identifying Dynamic Topics of Interest across Social Networks

Mohamed Salaheldin Aly, Abeer Al Korany
Department of Computer Science
Faculty of Computers and Information
Cairo, Egypt

***Abstract*—Information propagation plays a significant role in online social networks, mining the latent information produced became crucial to understand how information is disseminated. It can be used for market prediction, rumor controlling, and opinion monitoring among other things. Thus, in this paper, an information dissemination model based on dynamic individual interest is proposed. The basic idea of this model is to extract effective topic of interest of each user overtime and identify the most relevant topics with respect to seed users. A set of experiments on real twitter dataset showed that the proposed dynamic prediction model which applies machine learning techniques outperformed traditional models that only rely on words extracted from tweets.**

***Keywords*—Information propagation; topic modelling; dynamic user modelling; user behavior; machine learning; topic classification; social networks**

I. Introduction

Billions of users are now using different social networks (SN), SNs have proven to be effective in communication. Understating how information propagates across SNs became crucial to enhance the social networks and it attracted many businesses for the marketing value. Targeted advertisement along with many business applications in the past few years have proved to be very effective and to ensure the maximum efficiency researches have been studying information propagation in major social networks. This effort yields to develop different models that aim to predict how the information would propagate and its speed along with which users could be good candidates of becoming seeds for the information to propagate.

Information propagation depends on the users profiles which is represented by their interests, behaviour, and their position in the network which will affect their influence among other users. User's profiles contain a set of attributes that uniquely express each user like biography, age, gender, geographic location, hobbies, education history, and work information. While, other attributes that represent dynamic features with tagged time slots such as posts, comments and check-ins. Such information can be analysed in order to be used in different research areas such as: community detection, user recommendation(Abel et al., 2011; Blanco-Fernández et al., 2011). Studying user behaviour in SNs is quite complicated and the modelling for such behaviour has evolved drastically from how [Julia Stoyanovich 2008 et al.] [1] have simplified the user behaviour in their basic interests extracted from the tags they frequently use in the URLs they publish. Understanding that SNs users' behaviour is dynamic requires the consideration of the temporal factors [2], [3] when categorizing the user behaviour. This paper proposes a dynamic user modelling framework that aims to predict the candidate seeds (set of most influencing users) in the social network who will able to propagate information using topics of interest. The paper is organized as follows: Section II starts with discussing the related work. Section III introduces our dynamic user model, whereas Section IV explains the experimental setup used in building and validating that model. Section V discusses the results, and Section VI evaluates the model results. Section VII discusses the limitations while the paper is concluded in Section VIII.

II. Related Work

Various traditional approaches have been proposed for information propagation. Popular topic models such as Latent Dirichlet Allocation (LDA) [4] assumed that users could be classified according to the tags extracted from the topics they share and their similarities. Given that the behavior of the SNs users is not static and that it changes over time, many efforts went into understanding the effect of the temporal factor over the extracted interests. Qiaozhu Mei and ChengXiang Zhai in 2005 [5] explored the temporal text mining by utilizing the timestamps from the social posts extracted to identify different patterns in the topics extracted over time, proposing that adding the temporal factors with the understanding of the nature of the topics propagating may explain the themes that might follow and how they could influence other topics. This is more obvious when it comes to news as with an event happening thousands of articles are written and posted, however after this sudden burst for that particular event rests the summary of such events are the ones that are propagated afterwards [6], therefore understanding the lifecycle of a thread is important. Xuerui Wang, Andrew McCallum [7] later on proposed A Non-Markov continuous-time model of topical trends where the extracted topics from a document could be considered as a constant yet that only constitutes the meaning of that particular document and that time is a variable that affects the correlation between the keywords in documents with similar topics afterwards.

In the above related work the focus is on the topic modelling and understanding the impact of the temporal factor on in the information propagation, yet in social networks information propagation is not only associated with topics.

Rather, user behavior as equally contributes to information propagation, especially that the behavior is not uniform. A Temporal Context-Aware Model for User Behavior was proposed by [8] which takes the two factors in consideration: 1) the users' interests 2) the temporal context in the topic selection. The model aimed for rating the nature of the user behavior (clicking, sharing, purchasing), it has an edge over previous studies as its able to differentiate between user oriented topics and temporal topics this enables the models to better understand the users' interests. The proposed model was tested on multiple social networks (Delicious, digg, movieLens and douban movies).

The Temporal Context-Aware Model was later on enhanced in [9] taking in consideration that users' interests across social networks are not stable yet the temporal factor has a huge impact on those interests. Given that a user's interests were capture at a point in time those interests will certainly change with changing his job, getting married or having a new born for example, hence users' interests are dynamic. The Dynamic Temporal context-Aware model considers the users' interests distribution across time to predict the likelihood of a user to interact with a social post at a certain point in time.

III. Dynamic Modeling of User

Information in social network is spread the interactions between different users or nodes. A node in a social network is an abstract representation of many features that identify it. Thus, users in a social network could be distinguished through several characteristics such as interests, behavior, activities, etc. Those characteristics are identified using either the content published by users or by analysis of their relationships through network links. Extracted content posted by the user is used to identify the user interest, while link-based features are used to identify the behavior and degree of influence between users. The proposed model utilize the content published by the users in order to predict the potential candidates to propagate specific content. As a case study, the proposed model was applied on Twitter dataset. The proposed model decomposes three main phases. The first phase aims to dynamically extract topic of interest of user. While, the second phase aims to classify users based on their topics of interest. The third phase identifies the topics to be spread by specific user within specific set of users by considering the effect of time. Each of those phases are described in the following subsections.

A. Extract User Dynamic Profile

The first phase in the proposed model is responsible for creating the dynamic user profile in terms of her/his topical interest. Thus, topics which represent interest of users within specific time interval are associated with their relevance (score). In micro blogging networks such as Twitter, the interests of a user could be extracted from two main sources: 1) the content that user publishes by her/himself, 2) the content that the user interacts with different neighboring in form of retweet and replies. Using both sources, interest of the user could be identified. It is significant to mention that the frequency of producing such content is also considered and used as a decaying factor to adjust the weights of the users' interests. As mentioned earlier, the proposed model differentiates between three types of topics of each users, actual topics, burst topics, and pattern topics. Actual topics corresponds to frequent topics published by user represent user interest. When breaking news or events occur, people can post tweets about breaking news and share with friends, which could not be considered to represent a user interest. Due to large number of people participating in such conversation and discussion, those tweets may become hot messages and the source of burst topics. While the third type could also be observed where the content is triggered by an event yet the behavior is repeated every specific period of time such as Halloween. In order to be able to differentiate between each topics per user, topics of each user is extracted and associated with time slices, then if a topic is only mentioned in a specific time slice and then disappear from user topic list, then this could be categorized as a temporal topic . While pattern topics are extracted if it appear in the same period of each year.

B. Topic Classification

After identifying set of actual content that represent the user interest, MALLET (MAchine Learning for LanguagE Toolkit) was applied to extract topics of interest with its associated relevance weight from each tweet. It uses a simple way to analyze unlabeled text, by defining a topic as a cluster if words with similar meanings and distinguish between uses of words with multiple meanings. A Java Wrapper was built to use MALLET to analyze the collected tweets using Naïve Bayes algorithm and divide them into a set of topics to be pushed to IBM Watson to label it. For each topic i we calculated its relevance score with respect to the target user during time interval t as shown in (1):

$$TopRelScr_{(i,t)} = \frac{\sum_{j=0}^{n} OCRelWeight_{(i,j)}}{\mathrm{N}} \quad (1)$$

Where n is the total number of occurrences of a topic i in tweets of the target user that are created within time interval t, N is the total number of topics contained in tweets of the target user within time interval t and OCRelWeight is the relative weight provided by MALLET of topic i for each one of its occurrence j in a tweet during time interval t. Finally, within each time interval t, each user's topical profile will be represented as a vector of topics associated with their relevance scores. It is significant to mention that, time plays an important role in calculating the topics relevance as well as the influence of a user. As time distribution of post behavior reflects massive users' behavior characteristic for burst and actual topics.

C. Identify Imfluence Users

Identifying potential "Influencers" over time is not an easy task, it is required to understand the position of each user in the network at a given time slice. Certain nodes that are established around specific topics are the seed to create the burst in social media. For example for football pages they share hundreds of posts during the match day The user's position in the network is defined by his influence which could be captured in a microblogging network such as twitter using different attributes that are available in the public dataset. The number of users following a certain user could be a simple way to indicate how influential he could be, the number of times his posts are favored or retweeted or the number of times he's mentioned in different users posts. To even measure such influence in a certain time slice we factor in the frequency per

tweet per time slice. However and as by definition of a network, the position of a certain users cannot be only determined by his behavior yet also the neighboring nodes in the network, for example a user can have a lot follower yet they would be information seekers with passive behavior and would not contribute to propagating the created content, yet on the other hand a user with fewer followers yet very active on the network could have much more influence. Thus we propose that the user's influence could be measured by how much other users interact with the content he shares along with the position of his followers in the network.

For each user we determine the following, the number of times his tweets were favored in a time slice equals the summation of all tweets favorite count over number of tweets in a time slice.

IV. Experiment Setup

Twitter was selected as the social network to test the proposed model as it provides an easy to use API to extract data from public users. The API has its limitations yet enough data for testing purposes as it can provide all the tweets for a specific user during a specific time slice (with a limit of 3,000 tweets per user) along with the number of interactions on each tweet. The API also allows the retrieval of the list of friends for each user (with the limit of 5,000 per user) along with the total number of friends and followers.

A. Extracting Seed Users

For the purpose of this study a random sample of 1,000 public users was extracted using a java application to collect the data using twitter API and save them in an SQL relational database to facilitate the analysis. The sample was collected only form one location (Liverpool – UK) for two main reasons: 1) allow a better understanding of the context of the researched sample to facilitate the understanding of the contextual trends. 2) Understanding the influence of each node in the surrounding neighbors in the geographical network. The friends were also extracted for each user with the limitation of 5,000 users per user, accordingly 1,401,801 user where collected out of 3,884,033 in the 1000 users friends' lists. We selected the ones who were active during a specific time period which started from 1/1/2015 till 1/1/2016 regardless of their rate of tweet as our sample.

B. Extracting Users' Tweets

The 1,000 users collected had in total 12,793,079 Tweets. Given that the Twitter API has a limitation of around 3,000 tweets per user, only 2,248,181 tweets were collected. The tweet could be a retweet and accordingly the retweeted flag allows the differentiation between the content that is actually generated by the user and the content that the user shares from his network. Table I represent the summary of frequency of extracted tweets.

TABLE I. SUMMARY OF FREQUENCY OF EXTRACTED TWEETS

Tweet Type	Percentage
User Replies	15%
Tweet Replies	13%
Retweets	30%
Original Tweets	42%

C. Topic Modeling

One of the important challenges with the collected data is to be able to extract topics from the text for each tweet and differentiate them into corresponding types for each user. In a network like Twitter the issue becomes particularly complicated as the character limitation restricts users' accordingly they use abbreviations or slang that is difficult to classify. Understanding the content is not straight forward as for example in twitter the tweets are very short (with a maximum of 140 characters), accordingly even using different topic extractors such as Open Calais or IBM Watson the accuracy of the topics extracted is not reliable.

TABLE II. EXTRACTED TWEETS CATEGORIZATION

Category	Number of Tweets
Sports	541,516
art and entertainment	461,434
business and industrial	168,769
food and drink	158,400
law, govt and politics	122,273
Travel	101,655
Uncategorized	88,829
technology and computing	74,203
family and parenting	64,997
Society	62,130
Education	57,764
Shopping	55,835
Science	51,743
health and fitness	50,760
News	43,606
hobbies and interests	39,657
style and fashion	23,513
religion and spirituality	16,989
home and garden	15,985
Pets	15,450
Finance	15,245
automotive and vehicles	12,223
real estate	5,205
Grand Total	**2,248,181**

Using topic extractors is crucial to also allow the classification of topics extracted and understanding the areas of interest of each users on different levels. Thus, MALLET was used to analysis unlabeled text, and the Java Wrapper analyses the collected tweets and divide them into 500 topics, each topic having the top 50 significant keywords. The number of

iterations was set to 2000 to refine the results as much as possible, 19,272,829 tokens were found in all the tweets collected and used to train the model and create the 500 different topics. After training the model it was then used to go through all the tweets and assign each tweet to the most relevant topic with a relevancy score. Since the 500 topics were in the form of a cluster of related keywords yet not labelled, each cluster was then pushed to IBM Watson to label it. Watson API offers different configuration settings to get the desired output, for the purpose of our experiment the categories, entities and concepts were selected each with a set limit of three. Watson was only able to Identify 468 topics out of the 500 giving an output of 154 category for all the tweets collected. The summary of the categorized tweets is shown in Table II forming 20 categories.

V. Topic Classification

A. Extracting Bursty Topics

We take an example from the collected dataset to better understand the differentiation between the temporal topics and the topics that are based on user interest. Since the data collected is only in Liverpool, we take the hashtag "Cunard175" where Cunard liner a Britannia ship, left British waters bound for America marked its 175th anniversary in Liverpool, the event was on May 2015 by looking at the normal distribution of the topic over 2015 in Fig. 1 we find the following:

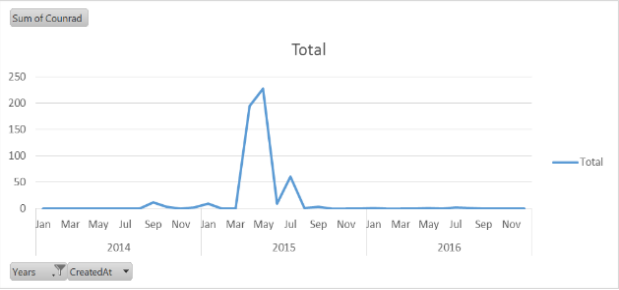

Fig. 1. Normal Distribution of the Topic Over 2015.

The topic was purely generated by temporal trigger in April and May 2015. Thus, it could not be considered as one of the interests of any of the users who has shared it. Thus, for each year, burst topics were detected and eliminated from all users' topics of interest using the following algorithm.

If number of tweets for topic is greater than twice the calculated median mark as burst topic Another example would be the sudden burst in the "/law, govt and politics/law enforcement/police" interest, in Table III we can see the significant score in April 2016.

TABLE III. Significant Score of Different Topics in April 2016

Topic	Median	Score	YYYY-MM
/law, govt and politics/law enforcement/police	57	2.56	2016-01
/law, govt and politics/law enforcement/police	57	2.28	2016-02
/law, govt and politics/law enforcement/police	57	2.46	2016-03
/law, govt and politics/law enforcement/police	57	9.05	2016-04

To understand the reason for the burst we start looking into major events or news that are relevant to the topic identified and explore different possibilities. For example, Hillsborough disaster which was a human crush at Hillsborough football stadium in Sheffield, England on 15 April 1989, during the 1988–89 FA Cup semi-final game between Liverpool and Nottingham Forest. The resulting 96 fatalities and 766 injuries makes this the worst disaster in British sporting history which came shortly after the 27th anniversary of the lethal crush at the FA Cup semi-final between Liverpool and Nottingham Forest, vindicated the bereaved families. The number of tweets increased started increasing from January until it reached 9 times its median in April 2016. Similarly thirty nine topics were identified to have busts throughout 2016.

B. Extracting Pattern Topics

The second type could also be observed where the content is triggered by an event yet the behavior is repeated every specific period of time. We take Halloween as example where and check the normal distribution over four years of data, we notice that every year around October there is a spike in the number of mentions for this topic as shown in Fig. 2.

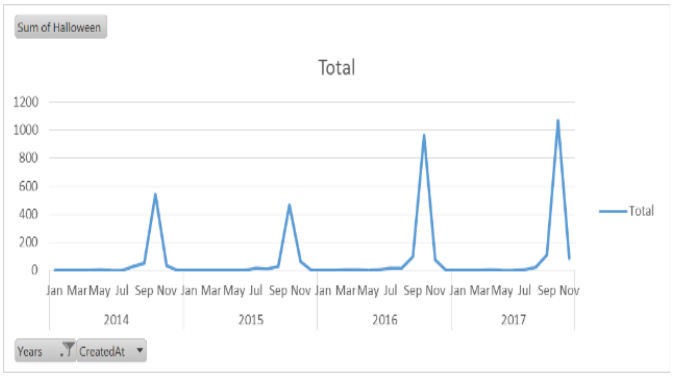

Fig. 2. Example of Pattern Topics (Halloween).

C. Extracting Actual Topics of Interest

Finally, the remaining topics of each user are considered her/his topic of interest. For example, Liverpool FC, this topic is constantly mentioned by different users over time and is not a temporal topic although bursts could be observed in some time slices, however those bursts could be attributed to certain contests in the context of Liverpool FC as shown in Fig. 3.

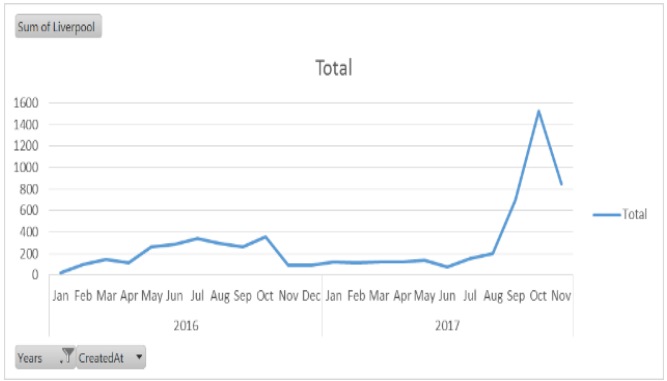

Fig. 3. Example of Actual Topics of Interest.

D. Refining the Sample

To be able to test the proposed model the data collected had to be refined to ensure that for each user, only actual topic of interest would be used in prediction. Thus we identify active users that had tweets during 2015, 2016 and 2017. The proposed model would be applied to extract user interests on 2015 and 2016 and use the results to feed in the overall model and run it on 2017 for evaluation. Accordingly we choose 631 users having tweets in all three years and we start detecting their interests by first categorizing the different tweets to see where they fall in the three categories mentioned above.

E. Identifying Dynamic User Interests

Fig. 4 shows the change of percentage of each interest from the overall interests of one user across time. For example, "Shopping and gifts" had 73% of the overall interests and further explore as the profile is for a famous footballer where the algorithm was responsive for the event accordingly gave it a high percentage among the interests while the month after the curve a had a dive equivalent to the hike it had in February 2014. As a result the topic would not have a score increase yet and thus, decay factor should be considered over time. For example in Fig. 4, the hike in February 2014 affects the percentage the topic has in the user's interests despite the fact that the interaction with the topic for slices after was minimum.

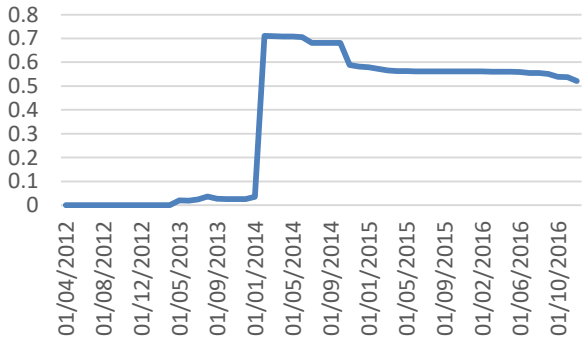

Fig. 4. Change of One Topic of Interest of One User Over Four Year Without Considering the Decay Factor.

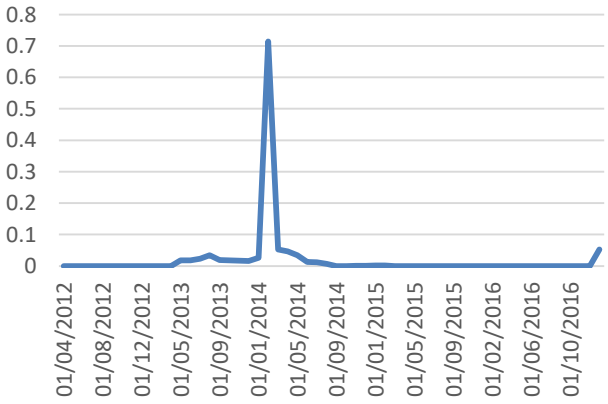

Fig. 5. Change of One Topic of Interest of Same User Over Four Year Using a Decay Factor.

While when considering the decay factor, the topic appears in slice yet with a lower score as shown in Fig. 5.

New score =

$$currentScore - \frac{tweetsInPreviousSlice}{TweetsInSlice} X \frac{(tweetsInPreviousSlice+1)}{2}$$

The above method ensures the reversal of any burst effect for any topic by calculating the ratio of tweets in the current slice to the tweets in the one preceding it for the same topic and using it as a multiplier to half the number of tweets in the previous slice.

VI. Evaluation

We then evaluate the results of the model by calculating the accuracy of the predicted number of tweets per topic for each user by applying the following for each tweets in 2016 and January 2017:

- For each user use the model score once with the decay factor and once without, to predict the number of tweets per topic for January 2017.
- Collect the actual tweets over January 2017 for the same users and use MALLET and IBM Watson to categorize them per topic
- Compare the actual number of tweets per category published for each user with both predicted scores by calculating the accuracy score for each.

The results show that the without the decay factor the model is 46% accurate while after applying the decay factor the model becomes more accurate as expected with 60% accuracy.

VII. Limitations

The public data that could be extracted from twitter for modeling is limited not only when it comes to quantity but also the behavioral data that could be crucial for this research such as the tweets favored or retweeted by each user. Such data could significantly enhance the model by factoring in those attributes in the weighting process.

VIII. Conclusion

The dynamic behavior of users across social networks makes it extremely challenging to predict user interests and perfectly understand how information propagates across social networks, However it is possible to reduce the factors that might decrease the accuracy of the predictions which we tried to do in this paper by understanding the nature of the interests of each users and eliminating all behavior that is not considered steady enough to predict future tweets. With this understanding and with a flexible design for the predictive model it is possible to use machine learning to profile users using their different activities and enhance their experience on social networks by displaying the most relevant content along with the utilization of the marketing value.

References

[1] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. A Study of the Benefit of Leveraging Tagging Behavior to Model Users´Interests in

del.icio.us. In AAAI Spring Symposium on Social Information Processing, 2008.

[2] Abel, F., Gao, Q., Houben, G.-j. and Tao, K. (2011) 'Analyzing User Modeling on Twitter for Personalized News Recommendations', in *User Modeling, Adaption and Personalization*, pp.1-12.

[3] Blanco-Fernández, Y., López-Nores, M., Pazos-Arias, J.J. and García-Duque, J. (2011) 'An improvement for semantics-based recommender systems grounded on attaching temporal information to ontologies and user profiles', *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 1385-1397.

[4] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Machine Learn. Res. 3, 993–1022

[5] Qiaozhu Mei and ChengXiang Zhai Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. ACM 1-59593-135-X/05/0008

[6] Qiming Diao, Jing Jiang, Feida Zhu, Ee-Peng Lim. Finding Bursty Topics from Microblogs. ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1 Pages 536-544

[7] Xuerui Wang, Andrew McCallum 2006 Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. ACM SIGKDD-2006 August 20-23, 2005, Philadelphia, Pennsylvania, USA

[8] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, Zi Huang. Temporal Context-Aware Model for User Behavior. Proceeding SIGMOD '14 Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data Pages 1543-1554

[9] Jia Dong Zhang and Chi Yin Chow. Ticrec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations. IEEE Transactions on Services Computing, (1):1–1, 2015.