Time-Sensitive Topic Derivation in Twitter

Robertus Nugroho[1](✉), Weiliang Zhao[1], Jian Yang[1], Cecile Paris[2],
Surya Nepal[2], and Yan Mei[1]

[1] Macquarie University, Sydney, Australia
{robertus.nugroho,yan.mei}@students.mq.edu.au,
{weiliang.zhao,jian.yang}@mq.edu.au
[2] CSIRO, Sydney, Australia
{cecile.paris,surya.nepal}@csiro.au

Abstract. Much research has been concerned with deriving topics from Twitter and applying the outcomes in a variety of real life applications such as emergency management, business advertisements and corporate/government communication. These activities have used mostly Twitter content to derive topics. More recently, tweet interactions have also been considered, leading to better topics. Given the dynamic aspect of Twitter, we hypothesize that temporal features could further improve topic derivation on a Twitter collection. In this paper, we first perform experiments to characterize the temporal features of the interactions in Twitter. We then propose a time-sensitive topic derivation method. The proposed method incorporates temporal features when it clusters the tweets and identifies the representative terms for each topic. Our experimental results show that the inclusion of temporal features into topic derivation results in a significant improvement for both topic clustering accuracy and topic coherence comparing to existing baseline methods.

Keywords: Temporal features in twitter · Topic derivation · Joint matrix factorization

1 Introduction

With about 288 million monthly active users and around 500 million tweets per day[1], Twitter is one of the most used social media platforms. Topic derivation from Twitter, to understand what people are talking about, is the foundation for a wide range of applications such as emergency, social awareness, health monitoring, and market analysis, and of interest to many organizations [1].

Topic derivation is the process of determining the main topic of every Twitter message (tweet) in a collection (to cluster the tweet based on topics) and choosing a set of terms to represent each topic [2]. Deriving topics from Twitter is a challenging task for several reasons: first, tweets are short (140 characters maximum) and often include informal language (e.g., emoticons, abbreviations)

[1] https://about.twitter.com/company, accessed 17 April 2015.

J. Wang et al. (Eds.): WISE 2015, Part I, LNCS 9418, pp. 138–152, 2015.
DOI: 10.1007/978-3-319-26190-4_10

and misspellings, leading to a sparsity problem when approaches only rely on term co-occurrences. Second, the Twitter environment is a highly dynamic one, with topics changing quickly over time.

Existing topic derivation methods based on term co-occurrences, such as LDA [3], PLSA [4] and NMF [5], suffer from the sparsity problem. Some have looked at addressing this problem, e.g., [6–8], by exploiting the relationship between correlated terms. However, they still only use the original tweet content, so that the problem remains. [9] proposed a method to incorporate static external resources to augment the tweet content. None of these approaches considered the information hidden in the interactions amongst posts in the Twitter environment. In their work, [10,11] went beyond terms and exploited content based social features such as hashtag, emoticons, and urls. In our previous work [2,12], we proposed topic derivation models that exploit both complex interaction features and content similarity. The intuition behind the use of interaction features such as *mention*, *reply*, and *retweet* to identify topics is that these features are typically employed to indicate that the posts are part of a conversation, and all posts pertaining to a conversation are likely to be on the same topic. Our experiments showed that, indeed, these models resulted in higher quality topics. To address the dynamic aspect of Twitter, some approaches have exploited temporal features, but only with respect to the tweet content or associated hashtags, e.g., [13–15]. To the best of our knowledge, the temporal features of the posts' *interactions* in Twitter have not been explored for topic derivation. This is what we propose to do in this paper.

While taking conversations into account as in [2,12] can improve topic quality, conversations typically have a time element associated with them. So incorporating a temporal aspect when looking at the interactions might further help topic derivation. For example, two mentions of same users nearly at the same time are more likely to be about the same topic than two mentions of same users within a long time interval. In this paper, we investigate the temporal features of Twitter interactions and propose a topic derivation method that employs these features, building on our previous work [12]. This research is summarized as:

- We discuss the relationships between topics and interaction features (*mention*, *reply* and *retweet*) using a data set obtained by collecting tweets over a month. We found that the *mention* feature is time sensitive with respect to topic assignation.
- We model the time sensitivity of *mentions* as an exponential decay according to the time difference of two tweets with the same mention. The decay parameter is based on an analysis of tweets that include a mention. This time-sensitivity model is then incorporated in the tweet relationship model in order to affect the matrix inter-joint factorization for topic derivation.
- We conducted a comprehensive set of experiments to evaluate our new model with a Twitter dataset covering one-month tweets, using widely accepted evaluation metrics for topic derivation. The results show that the new time-sensitive method results in a significant improvement of the accuracy of tweet clustering and coherence between terms for topic representation comparing with well-known baseline methods and our previous work [12].

The rest of the paper is organized as follows. Section 2 describes an investigation on the temporal features of *mentions*, *replies*, and *retweets*. Section 3 proposes a topic derivation method that takes these features into account. Section 4 reports on our experiments, with first a discussion of the dataset, the baselines and the evaluation metrics. Related work is provided in Sect. 5, and we conclude in Sect. 6.

2 Temporal Features of Tweet Interactions

Twitter has evolved from a microblogging platform to a medium that also enables people to interact with each other in a conversation-like manner. A user can initiate a conversation by mentioning other users in his/her tweet, and a tweet can be "replied to" by other users, or retweeted to other users. These *mention*, *reply* and *retweet* features form interactions between users, often related to a particular topic. A *reply* is a clear turn in a discussion between users; a *retweet* resends the message. It is likely that both a message that contains a reply and one that contains a retweet are on the same topic as the original post. Two tweets which mention the same user are also likely to be on the same topic *if* they occur around the same time, but not necessarily otherwise. Time thus plays an important role when attempting to link tweets because they mention the same people. In this section we will analyze the impact of time on user interactions for the same topic based on *mention*, *reply* and *retweet*.

We investigate users' mention behavior by analyzing tweets in a Twitter dataset to see how time affects the connectivity between tweets. Using the Twitter's streaming API[2], we retrieved all tweets from the top 15 Twitter users in Australia[3] and all the tweets that mention those users during January 12, 2015 until February 12, 2015. Our data set consists of more than 6 million tweets and involves around 800 thousand users. The details of the dataset are shown in Table 1.

Our investigation starts with an analysis of individual user mentions at different level of granularity to see how the mentions are distributed over time. We then look at the topics in the dataset to see if there is a relationship between the mention distribution and the topics. We find that, for all users, when the number of mentions of a specific user rises at a particular time, most of the tweets at that time are on the same topic.

Figure 1 shows the distributions of the tweets that mention *@MrKRudd* in a 3 hr time interval. We can see that there are several fluctuations within different time intervals. We find that each peak in Fig. 1 (an indication of a sharp increase in the number of tweets mentioning *@MrKRudd*) is strongly related to a particular topic. For example, on January 22, 2015 at 7 am (22/7), most of the tweets mentioning *@MrKRudd* were talking about the *"plain packaging act"*. The tweets on January 31, 2015 at 1 pm were about *"Queensland votes"*,

[2] https://dev.twitter.com/streaming/overview.

[3] https://followerwonk.com/bio/?q_type=all&l=Australia, accessed January 11, 2015, ordered by number of followers.

Table 1. Top 15 Twitter users in Australia and all related tweets (i.e., tweets that involve these top 15 Twitter users, either by mentioning them, replying to them or retweeting their posts) between Jan 12, 2015 and Feb 12, 2015

Username	# related tweets	# of users involved	# of followers
@CodySimpson	388,970	69,246	7,384,541
@5SOS	2,068,129	258,292	6,619,112
@Calumn5SOS	2,330,628	340,686	5,154,177
@luke_brooks	583,999	56,908	2,242,597
@example	8,464	5,208	2,107,484
@KyrieIrving	46,896	33,311	2,064,137
@BrooksBeau	819,423	95,879	1,932,857
@jascurtissmith	3,318	1,368	1,831,271
@MrKRudd	2,249	1,553	1,524,455
@allisimpson	88,504	20,107	1,418,732
@claireholt	5,413	2,497	1,299,287
@MClarke23	2,442	1,525	1,293,651
@DarrynLyons	1,154	390	1,143,222
@hillsongunited	3,456	2,455	969,020
@imacelebrity	1,675	1,340	894,187
@JordanJansen	10,774	2,512	759,192

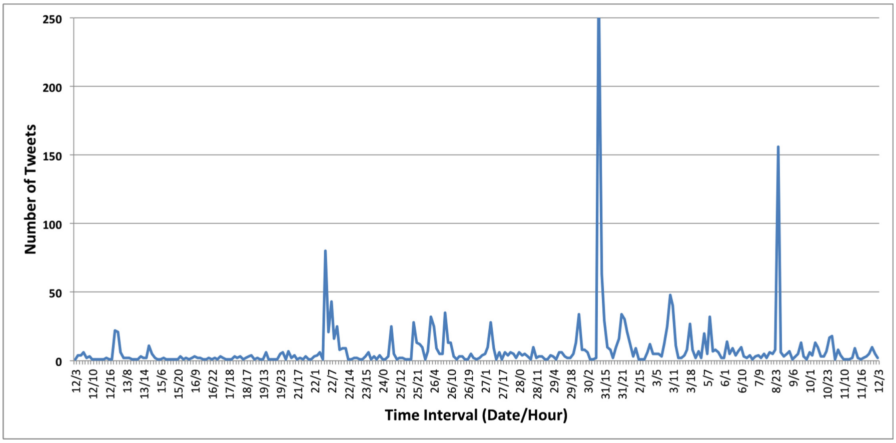
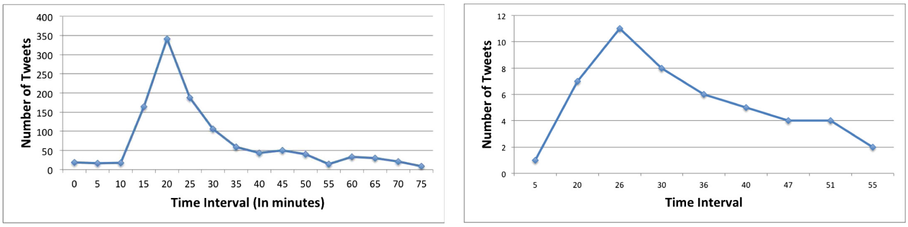


Fig. 1. Tweets mentioning user @MrKRudd with 3 h time interval

and the tweets on February 08, 2015 at 11 pm about "*the end of Kevin Rudd's leadership in February 2012*".



(a) @CodySimpson (b) @MClarke23

Fig. 2. Tweet distributions of tweets mentioning (a) *@CodySimpson* and (b) *@MClarke23* on 5 min time intervals within 1 h

The rises in the number of tweets with the same mention reaches their peak quickly and then slowly fade away (decay). Figure 2 shows the subset of the distributions of the tweets that mention (a) *@CodySimpson* and (b) *@MClarke23* on 5 min intervals. The specific distributions are different, reaching their peaks and decaying at different rates. What they have in common, however, is that each peak indicates a specific topic. The peak in Fig. 2a is related with the topic: "*Cody's birthday*"; and the peak in Fig. 2b is related with the topic: "*the absence of Michael Clarke on treatment issue*".

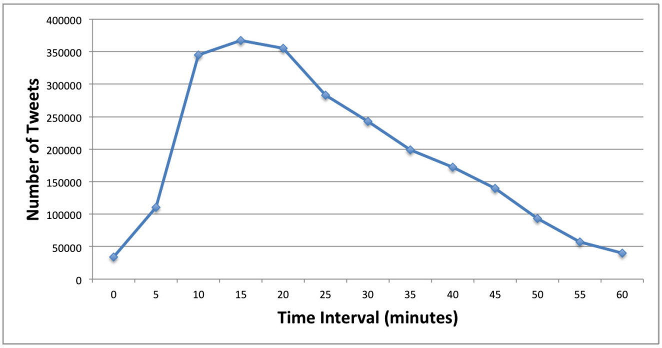
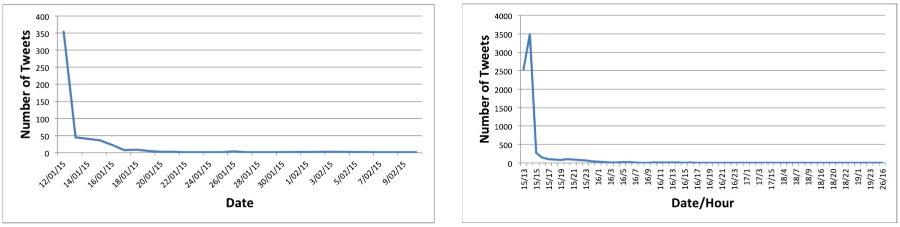


Fig. 3. The sum of all fluctuations in all tweet mention distributions with 5 min time interval

We performed a statistical analysis on all the variations of the tweet distributions, using a 5 min interval. We sum up the number of tweets from all users by choosing the subset of the tweet distributions starting from the closest lowest point before a peak and ending at the lowest point after the peak. Figure 3 shows this sum. We can see from the figure that most of mentions related to a particular topic reach a peak in about 15 min and then gradually fade away. An exponential function is adopted to model the process of fading away. We calculate the half-life of the exponential decay, which is how long the mention frequency decays from its peak to the peak's half value, using the following formula:

$$a = i_{t_{max}/2} - i_{t_{max}} \tag{1}$$

where $i_{t_{max}}$ is the time when the tweet mention distribution reaches its peak, and $i_{t_{max}/2}$ is the time when the tweet mention distribution reaches half of the peak value after the peak. In Fig. 3, the number of tweets in the highest point (t_{max}) is 367,368, and it is reached after 15 min ($i_{t_{max}}$). Then, $i_{t_{max}/2}$ can be calculated as the time to reach 183,684 after the peak, which is 37 min. So, a for Fig. 3 will be 22 min (1,320 s). This a will be used in the exponential function that models time in the mention behavior in Twitter in the next section.



(a) RT of @CodySimpson

(b) Reply to @5SOS

Fig. 4. Tweet distributions of a. retweet to a tweet by (a) *@CodySimpson* and (b) reply to a tweet by *@5SOS* within 1 month period

In contrast to the mention behavior, the topic relationship of a *reply* or a *retweet* with respect to the original tweet is not affected by time. As expected, the analysis of the dataset shows that a *retweet* or a *reply* could occur much after the original tweet and still be on the same topic.

Figure 4a shows the tweet distributions of a retweet to a tweet by *@CodySimpson*: ("*It's the 11th back home in Aus. I m officially 18.*"). The tweet was retweeted for 494 times in total, with 354 retweets on the first day, 22 on the third day, and the remaining scattered over time. Irrespective of the time elapsed, the retweets are still on the same topic. Figure 4b shows the tweet distribution of the replies to a tweet by @5SOS ("*Getting lots and lots of ideas for songs! Ready to write a new record!!*"). The total number of replies was 7414 tweets, with a peak on the first day but continuing the following day (with still 291 replies on the next day).

3 Topic Derivation

Our aim is to improve the quality of topic derivation in Twitter. Following our previous work [12], we classify the social interactions present in Twitter messages as interactions based on people and actions. In this paper, we first improve our interaction model to incorporate a time aspect. We then incorporate the new model into a matrix inter-joint factorization process to simultaneously achieve the clustering of the tweets based on topics and the identification of representative terms for each topic.

3.1 Relationship Between Tweets

A tweet is defined as a tuple of $t = \langle P_t, RTP_t, C_t, i_t \rangle$, where $P_t \subset P$ is the union of the author and people mentioned in the tweets, RTP_t the reply and retweet information, $C_t \subset C$ the set of the terms contained by the tweet, and i_t the timestamp of the tweet. We denote a relationship between two tweets t_i and t_j as $R(t_i, t_j)$. A zero value (0) of R means that there is no relation between them, and a higher value indicates the relationship is stronger. The relationship R includes three components: interactions based on people $(po(P_{t_i}, P_{t_j}))$, common actions $(act(RTP_{t_i}, RTP_{t_j}))$, and content similarity $(sim(C_{t_i}, C_{t_j}))$. It is defined as follows:

$$R(t_i, t_j) = po(P_{t_i}, P_{t_j}) + act(RTP_{t_i}, RTP_{t_j}) + sim(C_{t_i}, C_{t_j}) \ . \tag{2}$$

Interaction based on people $po(P_{t_i}, P_{t_j})$ is defined as the number of common *mentioned* people in the tweets t_i and t_j divided by the total number of people *mentioned* in both tweets. As discussed in Sect. 2, time affect the topic behavior in tweet mentions distributions. Tweets that mention similar users within a particular period are more likely to share the same topic. So, for the interactions based on people, we add a temporal factor $f(i_{t_i} - it_j)$. The people-based interaction is calculated as follows:

$$po(P_{t_i}, P_{t_j}) = \frac{|P_{t_i} \cap P_{t_j}|}{|P_{t_i} \cup P_{t_j}|} f(i_{t_i} - it_j) \tag{3}$$
$$where \quad f(i_{t_i} - it_j) = e^{-\frac{1}{a}|i_{t_i} - i_{t_j}|},$$

$f(i_{t_i} - it_j)$ is the exponential function that models time in the mention behavior in Twitter. Its parameter, a, was defined in the previous section. $f(i_{t_i} - it_j)$ controls the decay rate of the temporal effect.

The interaction based on user actions, denoted as $act(RTP_{t_i}, RTP_{t_j})$, is based on the *retweet* and *reply* relationship between two tweets. As already mentioned, time does not have an effect on these relationships. If tweet A is a *retweet* or *reply* of tweet B (or vice versa), or if both tweets are *replying* to or *retweeting* the same tweet, $act(RTP_{t_i}, RTP_{t_j})$ will be 1 (indicating a strong relationship), otherwise it is 0. We denote a *retweet* or *reply* of tweet t as RTP_t.

$$act(RTP_{t_i}, RTP_{t_j}) = \begin{cases} 1, (RTP_{t_i} = t_j) \ or \ (t_i = RTP_{t_j}) \\ \quad\quad or \ (RTP_{t_i} = RTP_{t_j}) \\ 0, \ otherwise \end{cases} \tag{4}$$

As there are a large number of self-contained tweets (i.e., tweets with no relation to any other tweet), our model for topic derivation also takes content similarity between tweets into account. Before calculating the content similarity, we perform some preprocessing steps to remove all irrelevant terms/characters and stop words. As tweets are short, two tweets sharing at least one (non-stop) word are likely to be on the same topic. For this purpose, $sim(C_{t_i}, C_{t_j})$ denotes the similarity between tweet t_i and t_j, which is measured by *cosine similarity* [16].

$$sim(C_{t_i}, C_{t_j}) = \frac{C_{t_i}.C_{t_j}}{\|C_{t_i}\|\|C_{t_j}\|} \ . \tag{5}$$

The values of all the relationships among the tweets form a tweet-to-tweet relationship matrix $A \in \mathbb{R}^{m \times m}$, where $a_{ij} = f(R(t_i, t_j))$. $f(R(t_i, t_j))$ is a *sigmoid function* [17] to normalize the value of $R(t_i, t_j)$ for a better relationship distribution.

$$f(R(t_i, t_j)) = \begin{cases} \frac{1}{1+e^{-R(t_i,t_j)}}, R(t_i, t_j) > 0 \\ 0, \ otherwise \end{cases} \tag{6}$$

By incorporating a time factor in the people-based interactions, we obtain a more accurate tweet-to-tweet relationship matrix. This matrix will be used to improve the topic derivation by jointly factorizing it with tweet-to-term matrix, as discussed in the next section.

3.2 Matrix *inter-joint* Factorization for Topic Derivation

We incorporate time into the Non-Negative Matrix inter-joint Factorization (*NMijF*) process described in [12]. We denote the resulting new method as *tNMijF*. Like the method on which it is based, *tNMijF* is an *inter-joint* factorization of a non-negative symmetric matrix $A \in \mathbb{R}^{m \times m}$ and another non-negative matrix $V \in \mathbb{R}^{m \times n}$ within a unified process. In our implementation, matrix A is the new tweet-to-tweet relationship matrix discussed in previous section (which includes a temporal aspect), and V is the tweet-to-term matrix which contains the relationship between tweets and the unique terms appearing in all tweets in the dataset. Each element in V is calculated using the *tf-idf* function described in [18]. We briefly describe the process here. More details can be found in [12].

The tweet-to-tweet matrix A is factorized to the tweet-topic matrix W as a base and W^T as the coefficient matrix. *Within the same process*, the tweet-to-term matrix V is factorized to the shared tweet-topic matrix W and topic-term matrix Y as the coefficient. In this method, matrix A and V share the tweet-topic matrix W. Hence, by implementing *tNMijF*, we can directly retrieve the main topic of a tweet from the tweet-topic matrix W *and* the top-n representative terms for each topic from the topic-term matrix Y within a unified process.

Tweet-to-tweet matrix A is much more dense than the tweet-to-term matrix V. At the best case (all terms are connected), the density of A will be equal to V. Sparsity of V could heavily penalized the quality of topic derivation. So, to handle this problem, the effect of matrix V in the factorization process to retrieve matrix W needs to be reduced. We implement the scale parameter α to control the effect in every iteration to achieve the objective function.

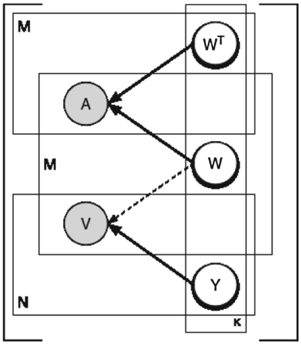


Fig. 5. Graphical Model of *tNMijF*

The inter-joint factorization process in *tNMijF* aims at finding the minimum divergence ($\mathscr{D}$) of $A \approx WW^T$ and $V \approx WY$. The graphical model for *tNMijF* is shown on Fig. 5, with the following objective function ($\mathscr{T}_{tNMijF}$):

$$\begin{aligned}\mathscr{T}_{tNMijF} &= \mathscr{D}(A\|WW^T) \ + \ \alpha\mathscr{D}(V\|WY) \\ &= \sum_{im} d(a_{im}|(ww^T)_{im}) + \alpha \sum_{mn} d(v_{mn}|(wy)_{mn})\end{aligned} \tag{7}$$

where there exists at least one element w and y in matrices W and Y such that $w \geq 0$ and $y \geq 0$, and the scaling parameter α satisfies $0 \leq \alpha \leq 1$.

For each element wise divergence, we employs *Kullback-Leibler divergence*:

$$\begin{aligned}d(a_{im}|(ww^T)_{im}) &= a_{im} \log \frac{a_{im}}{(ww^T)_{im}} - a_{im} + (ww^T)_{im}, and \\ d(v_{mn}|(wy)_{mn}) &= v_{mn} \log \frac{v_{mn}}{(wy)_{mn}} - v_{mn} + (wy)_{mn}\end{aligned} \tag{8}$$

In each iteration, we apply the following multiplicative update rules to every element in latent matrices W and Y to minimize $\mathscr{T}_{tNMijF}$:

$$\begin{aligned}\hat{w}_{i,k} &= w_{i,k} \frac{(\sum_{m=1}^{M} \frac{a_{i,m}}{(ww^T)_{i,m}} w^T_{k,m} + \alpha \sum_{n=1}^{N} \frac{v_{i,n}}{(wy)_{i,n}} y_{k,n})}{\sum_{m=1}^{M} w^T_{k,m} + \alpha \sum_{n=1}^{N} y_{k,n}}, \\ and \quad \hat{y}_{k,n} &= y_{k,n} \frac{(\sum_{m=1}^{M} \frac{w_{k,m}}{(wy)_{k,m}} w_{k,m})}{\sum_{m=1}^{M} w_{k,m}}\end{aligned} \tag{9}$$

4 Experiments

We now describe our experiments with the new model that is time sensitive. We first present our dataset, followed by the baseline methods and the evaluation metrics we employed. Then, we provide the results with a discussion.

4.1 Dataset

To evaluate our new mothod for topic derivation in Twitter, we employed a data set collected between 03 March 2014 and 07 March 2014 using the Twitter Streaming API. We call this dataset the *TweetMarch*. It includes 729,334 tweets involving 509,713 users all over the world. It contains 12,221 reply tweets, 101,272 retweets, and the rest are self-contained tweets.

We only used tweets in English in the experiments. A pre-processing is employed to remove irrelevant terms or characters (emoticons, punctuations, and terms that less than 3 characters), and stop-words. Then, all terms are lemmatized and all tweets are tokenized. Hashtags are kept unchanged.

Four people manually labeled around 120,000 tweets from the first subset of *TweetMarch* dataset as an evaluation set. From the labeled data, we observe that the *TweetMarch* dataset covers a wide range of topics, from politics and traveling to life entertainment and school activities.

4.2 Evaluation Metrics

For the evaluation purposes, we used several baseline methods:

- *NMijF*. This is our previous model. It takes into account tweet' interactions and employs a non negative inter-joint factorization, but it is not time-sensitive. We use this method as a baseline to see the impact of the temporal features. While we have already shown that *NMijF* improves on the next three baselines, *TNMF*, *LDA* and *NMF*, we still include them for completeness sake.
- *TNMF* [6]. This topic derivation method incorporates a term correlation matrix to improve the quality of the result using matrix factorization techniques.
- *LDA* [3]. The most popular method in topic derivation. It has a“bag of words” assumption and works solely on the content of the document.
- *NMF* [5]. This is the basic method of matrix factorization. It directly factorizes the tweet-to-term matrix into topic-tweet and topic-term matrix.

We conducted the evaluations on both the quality of the clusters and the topics produced by all the methods. The quality of the clusters is measured through their accuracy with respect to our manually labeled classes. We compared the clustering result K from N tweets with an evaluation set of classes C. In particular, three metrics were used in the evaluation on cluster quality [18]: Pairwise *F Measure*, *Purity* and *Normalized Mutual Information (NMI).*

We used the Pairwise *F-Measure* to measure the accuracy of the clustering result by analyzing the harmonic mean of both precision and recall. In this metrics, precision p is defined as the fraction of pairs of tweets correctly put in the same cluster, and recall r is the fraction of actual pairs of tweets that were identified. The formula of pairwise F-Measure is shown in Eq. 10 below:

$$F = 2 \times \frac{p \times r}{p + r} . \tag{10}$$

Purity is calculated by assigning each cluster in K to class in C, and then counting the number of correctly assigned elements divided by the total of elements in all clusters. A *Purity* value of 1 indicates a perfect clustering, whereas a *Purity* value of 0 means low quality clustering.

$$purity(K, C) = \frac{1}{N} \sum_{i} \max_{j} |k_i \cap c_j| \ . \tag{11}$$

NMI measures the mutual information shared between clusters and classes $I(K;C)$, normalized by the entropy of clusters $H(K)$ and classes $H(C)$. Similar to *Purity*, the value of *NMI* will be ranged between 0 and 1.

$$NMI(K, C) = \frac{I(K;C)}{[H(K) + H(C)]/2} \ . \tag{12}$$

To evaluate the quality of the representative terms for each topic, we used the *topic coherence*, $Co(k, W)$, for a topic described by its topic-term [19]. It measures the readability of all terms that represent the topic by evaluating the frequency of pair of terms in the same tweet over the original dataset. It is described by the following equation:

$$Co(k, W) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{T(w_m, w_l) + 1}{T(w_l)}, \tag{13}$$

where $w_m, w_l \in W$; $T(*)$ and $T(*, *)$ are document frequency and co-document frequency functions, representing the number of tweets which contain a given term or a pair of two terms respectively; M is the size of the set W of topic-term.

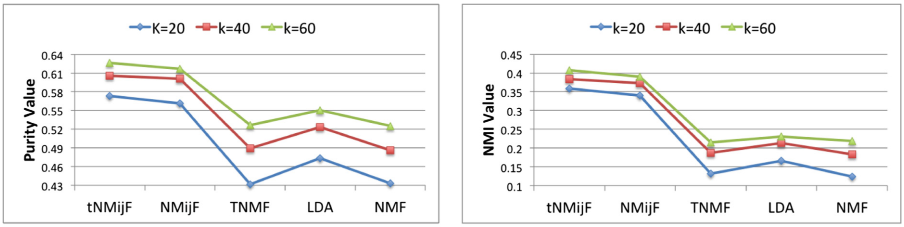
4.3 Results and Discussion

To see the performance of our method on a different number of topics, we used k = 20, 40, and 60 as input for each experiment with every method. We run all methods for 20 times over the dataset and tuned all parameters for the best performance. The average density (non zero element) of the tweet-to-term matrix V is only 0.08 %, which is far below our tweet-to-tweet relationship matrix with 32.64 % density. The scaling parameter $\alpha = 0.1$ was found to be the best for all of the matrix inter-joint factorization processes as the matrix V is very sparse. This α value ensures that the sparsity of V does not heavily penalize the topic-tweet matrix W and still gives good results when factorizing the topic-term matrix Y.

Table 2 shows the results of the pairwise *F-Measure* metrics. It can be seen that the inclusion of time improves both precision and recall in comparison to the baseline methods for all values of k (the number of topics). *tNMijF* consistently provides the best results for both precision and recall, with a positive trend over increasing values of k. Our previous work, *NMijF*, which does not take time into account, also outperforms the other baseline methods. However, as k increases, the improvement in precision and recall lessens. In contrast, the new method proposed in this paper gives a consistent improvement of the precision and recall for all k values.

Table 2. *Precision, Recall and F-Measure* for topics $k = 20, 40, 60$

Method	k=20			k=40			k=60		
	p	r	F-m	p	r	F-m	p	r	F-m
tNMijF	**0.407**	**0.236**	**0.298**	**0.444**	**0.264**	**0.330**	**0.481**	**0.292**	**0.361**
NMijF	0.396	0.218	0.280	0.417	0.227	0.293	0.418	0.227	0.293
TNMF	0.276	0.079	0.123	0.335	0.051	0.088	0.381	0.043	0.078
LDA	0.310	0.084	0.132	0.369	0.057	0.099	0.404	0.047	0.084
NMF	0.271	0.072	0.114	0.336	0.047	0.083	0.405	0.039	0.072



(a) (b)

Fig. 6. (a) Purity evaluation results and (b) NMI evaluation results

This cluster evaluation is confirmed by other two metrics: *Purity* and *NMI*. Figure 6a shows the evaluation results using the *purity* metrics, and Fig. 6b shows the results of the NMI evaluation. In the purity evaluation, our proposed method *tNMijF* gives about 5 % improvement over our previous work, and 15–30 % over the other baseline methods. For the NMI evaluation, *tNMijF* results in roughly a 5 % improvement compared to *NMijF*, and 90–200 % improvement over the other methods, *TNMF*, *LDA* and *NMF*. We conclude that the introduction of a temporal aspect leads to an obvious improvement over other methods for the accuracy of the topic derivation process.

For the topic coherence evaluation, we use the metric defined in Eq. 7 and take the top-10 terms to represent each topic from the topic-term matrix Y. Figure 7 shows the result of the topic coherence evaluation. We can see that, for a small number of topics (k=20), all methods have quite a good performance. When the number of topics becomes bigger, however, the topic coherence with our new method *tNMijF* reduces only slightly in comparison to the baseline methods which have significant drops. This result shows that *tNMijF* is reliable for different numbers of topics in terms of the topic coherence.

The above results show that introducing a time factor on the interaction features (in particular the *mention*) when performing topic derivation with a non-negative joint matrix factorization process greatly improves the accuracy of tweet clustering and the coherence of topics. This improvement is consistent for any number of derived topics.

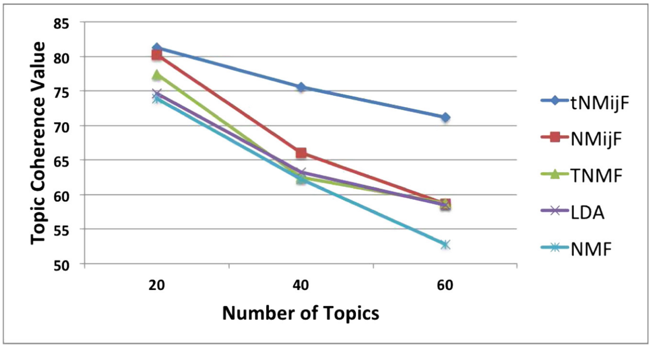


Fig. 7. Topic Coherence

5 Related Work

The short-in-content nature of Twitter presents a challenging problem for deriving the topics of a tweet collection. The very limited length for each tweet renders the frequency of co-occurences between terms extremely low. This sparsity heavily penalizes the performance of the state of the art topic derivation methods such as LDA [3], PLSA [4] and NMF [5], as they generally work solely on content features.

A lot of studies have been conducted to extend those popular methods to handle the sparsity issues. [10] proposed a variant of *labeled-LDA* to work on Twitter environment with the hashtag and other content features as labels for a partially supervised topic learning process. Albakour et al. [9] and Vosecky et al. [11] addressed the problem by expanding the content with the help of external documents collections. However, relying on external documents brings an extra burden when dealing with highly dynamic environments like Twitter. The approach reported in [6,7] exploits the term co-occurrence patterns to improve the topic learning process in a short text environments. Unfortunately, in Twitter environment, the relationship between terms is very sparse and it only provides a small improvement with respect to density in comparison with the original tweet-to-term relationships [2].

To deal with the dynamic nature of the Twitter environment, several methods have been proposed by including temporal features. The proposed method in [11] uses a temporal weight function for the recency sensitivity of the tweet content based on hashtags and urls. [13] proposed a temporal based regularization in NMF method to learn the topics in social media. The study in [14] introduced the content aging theory to mine the emerging topics from Twitter stream. Stilo et al. [15] proposed *Symbolix Aggregate approximation* (SAX) to discretize the temporal series of terms to discover the events from Twitter content. All these studies still focus on contents and overlook the social features available in the Twitter environment. As a result, they still suffer from the sparsity issue.

Different from the topic derivation work that only takes content into account, [12] incorporated the relationships between tweets to deal with the sparsity problem in the Twitter environment and showed improvements in performance. The work presented in this paper builds on this foundation, adding a time dimension to the interactions. To the best of our knowledge, our proposed method is the first one to incorporate temporal features, social interactions and content in a unified model to derive topics from a collection of tweets.

6 Conclusions

In this paper, we investigate the effect of time on user interactions for topic derivation in Twitter. We propose a new topic derivation method that includes this time factor. It can simultaneously achieve the clustering of the tweets based on topics and the identification of the representative terms for each topic. We conducted a set of experiments on a set of tweets collected over a period of one month.

Our results show that incorporating of a time aspect on the interaction features improves the results of the topic derivation process. In particular, the proposed method results in a consistent improvement in the accuracy of the tweet clusterings and topic coherence for different numbers of topics over both well-known baseline methods and our prior method, which was not time-sensitive. Currently, the method works for a static Twitter dataset. We are developing the incremental model of the proposed method to work with the stream based Twitter messages.

Acknowledgment. This work is supported by the Indonesian Directorate General of Higher Education (DGHE), Macquarie University, CSIRO and Australian Research Council Linkage Project (LP120200231).

References

1. Wan, S., Paris, C.: Improving government services with social media feedback. In: Proceedings of the 19th International Conference on Intelligent User Interfaces. IUI 2014, New York, NY, USA, pp. 27–36. ACM (2014)
2. Nugroho, R., Molla-Aliod, D., Yang, J., Paris, C., Nepal, S.: Incorporating tweet relationships into topic derivation. In: Proceedings of the 2015 Conference of the Pacific Association for Computational Linguistics, PACLING (2015)
3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
5. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2000)
6. Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y.: Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: Proceedings of the SIAM International Conference on Data Mining. SIAM (2013)

7. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 1445–1456 (2013)
8. Hu, Y., John, A., Wang, F., Kambhampati, S.: Et-lda: joint topic modeling for aligning events and their twitter feedback. AAAI **12**, 59–65 (2012)
9. Albakour, M., Macdonald, C., Ounis, I., et al.: On sparsity and drift for effective real-time filtering in microblogs. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 419–428. ACM (2013)
10. Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. ICWSM **10**, 1–1 (2010)
11. Vosecky, J., Jiang, D., Leung, K.W.T., Xing, K., Ng, W.: Integrating social and auxiliary semantics for multifaceted topic modeling in twitter. ACM Trans. Internet Technol. (TOIT) **14**, 27 (2014)
12. Nugroho, R., Zhong, Y., Yang, J., Paris, C., Nepal, S.: Matrix inter-joint factorization - a new approach for topic derivation in twitter. In: Proceedings of the 4th IEEE International Congress on Big Data. IEEE Services Computing (2015)
13. Saha, A., Sindhwani, V.: Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 693–702. ACM (2012)
14. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, p. 4. ACM (2010)
15. Stilo, G., Velardi, P.: Time makes sense: Event discovery in twitter using temporal similarity. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, pp. 186–193. IEEE Computer Society (2014)
16. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
17. Von Seggern, D.H.: CRC Standard Curves and Surfaces with Mathematica. CRC Press, Boca Raton (2006)
18. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
19. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 262–272 (2011)