Incorporating Tweet Relationships into Topic Derivation

Robertus Nugroho*, Diego Molla-Aliod*, Jian Yang*, Youliang Zhong*, Cecile Paris† and Surya Nepal†
*Dept. of Computing, Macquarie University, Sydney, Australia
Email: robertus.nugroho@students.mq.edu.au, {diego.molla-aliod, jian.yang, youliang.zhong}@mq.edu.au
†CSIRO, Sydney, Australia
Email: {cecile.paris, surya.nepal}@csiro.au

***Abstract*—With its rapid users growth, Twitter has become an essential source of information about what events are happening in the world. It is critical to have the ability to derive the topics from Twitter messages (tweets), that is, to determine and characterize the main topics of the Twitter messages (tweets). However, tweets are very short in nature and therefore the frequency of term co-occurrences is very low. The sparsity in the relationship between tweets and terms leads to a poor characterization of the topics when only the content of the tweets is used. In this paper, we exploit the relationships between tweets and propose *intLDA*, a variant of Latent Dirichlet Allocation (LDA) that goes beyond content and directly incorporates the relationship between tweets. We have conducted experiments on a Twitter dataset and evaluated the performance in terms of both topic coherence and tweet-topic accuracy. Our experiments show that *intLDA* outperforms methods that do not use relationship information.**

***Keywords*-Topic Derivation; Twitter; Tweets Relationship;**

I. Introduction

With around 350 thousands Twitter messages (tweets) per minute at the time of writing[1], Twitter has become one of the best places on the Internet to get an understanding of what is happening in the world. With such rapidly-changing information, the ability to derive the most important topics from Twitter data is critical to provide an effective way to navigate through the data and explore the information.

In this paper we aim to determine the most important topics of a Twitter dataset by performing topic derivation. For the purposes of this paper, *topic derivation* of a collection of tweets is the process of determining the main topic of every tweet and characterizing the main topics of the collection of tweets by listing their most important words.

Unlike traditional documents with lengthy and structured content, a tweet is limited to 140 characters. Additionally, a tweet could include expressions in informal language, such as emoticons, abbreviations, and misspelled terms. Given their short-text nature, deriving topics from tweets is a challenging problem. The very low co-occurrences between terms will heavily penalize the topic derivation process. Because of this sparsity problem, conventional methods for topic derivation such as Latent Dirichlet Allocation (LDA) [1], PLSA [2] or NMF [3] do not work well in the Twitter environment, as they focus only on content.

Several studies in the literature address the sparsity problem on microblogging environments. For example, [4] and [5] presented a content expansion method based on an external document collection. Relying on external resources, this approach would become difficult to deal with a highly dynamic environment like Twitter. The study of [6] exploited the semantic features of Twitter content by building the term correlation matrix, but this still potentially suffers from the sparsity problem since the term co-occurrences in Twitter are very low.

The limitations of those methods have inspired us to go beyond content to address the sparsity problem. We investigate the possibility of incorporating the social interaction features in Twitter. Studies by [7] and [8] show that social interaction features in Twitter play an important role on both topic quality and credibility.

We propose a new method, *intLDA*, that uses the contents of tweets *and* specific relationships between tweets to perform topic derivation. In this paper, we define the relationships between tweets as the interactions based on users (*mentions*), actions (*reply* and *retweet*) and content similarity. Our analysis and experimental results show that our proposed method can significantly outperform other advanced methods and configurations in terms of topic coherence and cluster quality. The main contribution of the paper can be summarized as follows:

- We observe that tweets are related to each other through both interactions and content features. Our analysis reveals that a matrix of tweet relationships have a higher density than one based on term-to-term or tweet-to-term relationships.
- We develop a novel extension of the LDA method, *intLDA*, which incorporates the tweet relationships into topic derivation. Our proposed *intLDA* method can effectively determine and characterize the main topic of each tweet.
- We conduct comprehensive experiments on a Twitter dataset, using widely accepted topic derivation metrics. The experimental results demonstrate significant improvements over popular methods such as LDA, Plink-LDA [9] and NMF. We also discuss an implementation of a simple variation to LDA that takes into account tweets relationships (*eLDA*) and show that *intLDA* is

[1] http://www.internetlivestats.com/twitter-statistics/

still far better in comparison to this simpler method.

The rest of the paper is organized as follows. Section II presents the task of topic derivation and justifies its use for characterizing the most important topics of a collection of tweets. Section III introduces the relationships that exist between tweets based on their interactions and content. Section IV describes how to incorporate the tweet relationships into LDA. Details of the experiments and evaluation are presented in Section V. We discuss the related work in Section VI and conclude in Section VII.

II. Topic derivation of tweets

Social media in general, and Twitter in particular, are being used by a large community of people worldwide to post short pieces of information on any matters that are directly relevant to them. People might post for a wide range of reasons, such as to state someone's mood in the moment, to advertise one's business, or to report an accident or disaster. The widespread and continuous use of Twitter by such a large community makes it a desirable source for information sharing. In this paper, we aim to characterize the most salient topics being discussed in Twitter at any point in time by detecting the most important topics and listing their most representative words. This is useful for a wide range of applications. For example, in emergency relief agencies (*e.g.,* fires, floods and other disasters), detection of possible burst of epidemics by health monitoring institutions, and marketing studies to identify possible trends in large communities of potential users.

Topic modeling methods such as Latent Dirichlet Allocation (LDA) [1] model a document as a bag of words drawn from a mixture of topics. LDA has been used to determine the most likely distribution of words per topic, and the most likely distribution of topics in documents. After performing LDA, it is straightforward to determine the most salient topics in a document and the most salient words in a topic. However, since a document is considered as a mixture of topics, it is not trivial to determine the most important topics in the collection.

We have performed LDA on the first 500 tweets of our collection (see Section V-A for details of our dataset) and observed a marked predominance of one topic per tweet, as we describe below. For any tweet, let t_1 be the topic with the highest probability (p_1) and t_2 the next ranking topic (with probability p_2), as determined by performing LDA on the 500 tweets. We call the ratio of $p1/p2$ the "Prominent Factor" or PF. If t_1 is much more prominent than t_2, PF will be high. Figure 1 shows the prominent factor for each tweet, after performing LDA with 20 topics. The ratio of the prominent factor in this figure is sorted in ascending order. The values are clipped at a factor of 8, but we observed a maximum factor of 2000. Furthermore, 271 tweets (54%) have a prominent factor over 100. The figure shows that 85% of the tweets have a prominent factor of 1.4 or higher. A factor higher than 1.4 (e.g: 0.418 for the highest probability and 0.279 for the next ranking) or higher means that one topic is relatively predominant for this tweet. The larger the factor, the more predominant the topic.

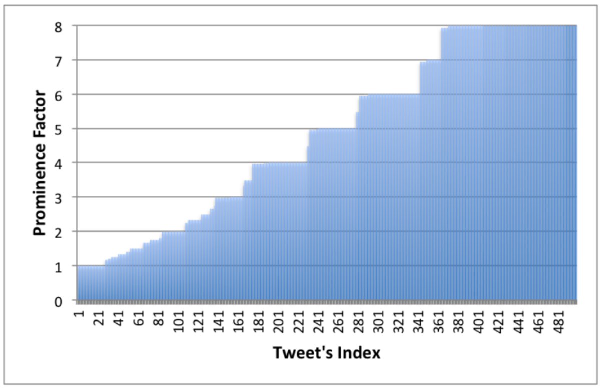

Figure 1: Topic prominence in the tweets of a collection of 500 tweets, sorted by prominence factor (ratio between the highest and the second highest topic probability for each tweet). The values are clipped at a factor of 8.

Given the marked preference of one topic in each tweet for most tweets, it is sensible to characterize a tweet by its most salient topic. By establishing this one-to-one mapping from tweets to topics, we can determine the importance of a topic in the collection of tweets by counting how many tweets are mapped to the topic. We therefore perform *topic derivation* of a collection of tweets by determining the main topic of every tweet by grouping tweets on the same topic and by characterizing the most important topics of the collection of tweets by listing their most important words.

III. Observing the relationships between tweets

Topic derivation by straight LDA suffers from the fact that the tweets are very short. Directly applying LDA on Twitter data may produce a poor characterization of the topics due to the sparse relationship between the tweets and the terms [10]. Several studies in the literature address the sparsity problem that occurs when dealing with short text. For example, [4] presented a query expansion method based on an external document collection. Relying on external resources, this approach would become difficult to predict what relevant content will be relevant to add in a highly changing environment like Twitter. Yan and his colleagues in [6] and [11] exploited the semantic features of a document content to build the term correlation matrix, but this still potentially suffers from the sparsity problem since the term co-occurrences in Twitter are very low.

A number of researchers have investigated the possibility of incorporating social interactions in Twitter. For example, studies by [7] and [8] show that social interaction features in Twitter play an important role on the determination of both topic quality and its credibility. In the approach presented here, we use the interactions between tweets as means to address the sparsity problem to achieve better topic coherence and higher topic quality.

Owing to the social networking nature of Twitter, there are various relationships on the Twitter platform. Twitter provides a *following-follower* mechanism to connect users, so that all followed users' tweets will be shown on a user's home page. In addition, Twitter offers several interactive features enabling users to interact with each other through tweets, such as *mention*, *reply*, *retweet* and *hashtag*. These features have made Twitter a network of not only people but also information. In this paper, we define the relationships between tweets as the interactions based on users (*mentions*), actions (*reply* and *retweet*) and content similarity.

Mention and *reply* are helpful methods for initiating or joining a conversation in Twitter. Intuitively, all tweets belonging to the same conversation have a high probability of sharing the same or similar topic even if no terms co-occur in their content. A *mention*, denoted as '@' followed by a user name, directly refers to another user. In contrast, a *reply* is used to send out a message in reply to a specific tweet. In a *reply* tweet, the user name of the original tweet's author is included in the message.

Different from the *mention* and *reply* relations, a *retweet* is a re-posting of someone else's tweet. This can be used to further disseminate a tweet, for example to ensure one's followers see it. Since a *retweet* has many words in common with the original tweet, the term co-occurrence between the two tweets (original and reteweet) will be high, and both tweets are likely to share a topic.

To capture the interactions in Twitter, we classify the interactions based on people $po(i,j)$ and on actions $act(i,j)$. Let p_i be the number of people mentioned in tweet i. Then, $po(i,j)$ uses the *mention* relationship and is defined as the number of common mentioned people in tweets i and j, normalized by the number of people involved in both tweets.

$$po(i,j) = \frac{|p_i \cap p_j|}{|p_i \cup p_j|} \ . \tag{1}$$

$act(i,j)$ is determined by the *retweet* and *reply* relations between two tweets. If tweet i is a *retweet* or *reply* of tweet j or vice-versa, or if both tweets are replying or retweeting the same tweet, the $act(i,j)$ value will become 1, otherwise 0. Generally speaking, an $act(i,j)$ value of 1 means that two tweets have a strong relationship with each other, and most likely they share the same topic.

$$act(i,j) = \begin{cases} 1, (rtp_i = j) \ or \ (i = rtp_j) \ or \ (rtp_i = rtp_j) \\ 0, \ otherwise \end{cases} \tag{2}$$

where rtp_i stands for the *retweet* or *reply* information in a tweet i.

Table I: Comparison of the density between the relationships of tweet-to-tweet (*R*), term-to-term (*T*), and tweet-to-term (*W*)

# of tweets	# of terms	*R*	*T*	*W*
5K	6119	32.93%	0.37%	0.13%
10K	9103	32.07%	0.29%	0.09%
15K	11973	32.88%	0.24%	0.07%
20K	14283	32.67%	0.22%	0.06%
25K	16121	32.64%	0.21%	0.05%

There are many *self-contained* tweets in the Twitter platform, where a tweet does not have any references (*mention*, *reply* or *retweet* relation) to another tweet [12]. We thus also include content based interactions in the relationship between tweets for the purposes of topic derivation. We use *content similarity* ($sim(i,j)$) between two tweets i and j to measure the content based interaction. In this paper we will simply use the word overlap between i and j. Thus, if W_i denotes the set of words of tweet i after preprocessing the text as described in Section V-A, then:

$$sim(i,j) = |D_i \cap D_j|. \tag{3}$$

We can now formalize the relationship between tweets i and j (R_{ij}) based on their interactions (based on people, actions and content), as shown in Equation 4:

$$R_{ij} = \begin{cases} 1 & \text{if } po(i,j) > 0 \text{ or } act(i,j) > 0 \\ & \text{or } sim(i,j) > 0 \\ 0 & \text{otherwise .} \end{cases} \tag{4}$$

Table I compares the density between the relationships of tweet-to-tweet (*R*), term-to-term (T), and tweet-to-term (W) from a Twitter dataset. The table shows that the relationship between tweets has the highest density by a large margin. Adding information about tweet relationships can thus dramatically decrease the sparsity of information.

IV. INCORPORATING TWEET RELATIONSHIPS INTO LDA

In this section, we discuss our method of incorporating the tweet relationships into the LDA process. We present two LDA implementations which directly incorporate the relationships. We first discuss the basic LDA method, then a simple method we call *eLDA*, our naïve way of expanding the tweet content by adding the new content from the related tweets. We then present our proposed method *intLDA*, another variant of LDA that directly incorporates the relationships between tweets.

A. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was presented by Blei et al. [1]. This method is used to automatically discover the topics from a collection of documents, with the intuition that every document exhibits multiple topics. LDA models the words of a document as generated randomly from a mixture of topics where each topic has a latent distribution of word probabilities. The documents and their words are generated according to the following generative process:

1) For each document d, draw a topic distribution θ_d, which is randomly sampled from a Dirichlet distribution with hyperparameter α. ($\theta_d \sim Dir(\alpha)$)
2) For each topic z, draw a word distribution ϕ_z, which is randomly sampled from a Dirichlet distribution with hyperparameter β. ($\phi_z \sim Dir(\beta)$)
3) For each word n in document d:
    a) Choose a topic z_n sampled from the topic distribution θ_d. ($z_n \sim Cat(\theta_d)$)
    b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n. ($w_n \sim Cat(\phi_{z_n})$)

B. eLDA: expanding tweet content based on tweet relationship

From the generative process shown in previous subsection, we can see that LDA works solely on the tweet content, without incorporating the relationships that may exist between tweets. It has a ”bag of words” assumption where the order of the words in the documents does not have any effect on the topic derivation process. When dealing with short texts such as tweets, term co-occurrences amongst tweets can be low, which hurts the topic derivation process. A naïve way of improving the LDA method is to augment the tweet content to increase the term co-occurrences. While expanding the content of the tweets using external documents seems to be ideal [4], the method would become difficult to deal with Twitter's highly dynamic environment, as already mentioned. Furthermore, the language used in tweets is mostly informal, and therefore the words occurring in a tweet may not easily match those terms in external corpora.

A simple, intuitive use of the tweet-relationship matrix R consists in expanding the tweet content by adding the words from the related tweets (tweets with the observed tweet relationships discussed in section III). In this approach, we add only words that are not already occurring in the original tweet. Our implementation of this content expansion is denoted as *eLDA* in this paper. A possible drawback of this method is that the added words might not be related to the tweet, therefore introducing noise.

C. intLDA: incorporating the tweet relationship to improve the tweet-topic distributions

In LDA, each tweet i defines a multinomial distribution θ_i of topics. The global tweet-topic distribution θ can be learned based on the observed words present in each tweet through a Markov Chain Monte-Carlo algorithm such as Gibbs sampling [13].

Since working only on content makes LDA suffer from the sparsity problem, we extend the model to directly incorporate the observed relationships between tweets R in the process of learning θ. We use R to add an additional constraint to the θ distributions, so that if two tweets are related, then the θ of those two tweets will be simultaneously adjusted based on the sampled topic.

The difference between LDA and *intLDA* is in the process of sampling the tweet-topic distribution using Gibbs sampling. In each iteration of Gibbs sampling, LDA updates the document-topic counts of each tweet i independently of each other. In contrast, *intLDA* updates the document-topic counts of tweet i, as in LDA, but in addition it updates the document-topic counts for the sampled topic z of all tweets j that are related to i as defined by $R_{i,j}$. In other words, the estimation of the document-topic distribution θ_i for tweet i is affected by information from related tweets.

The posterior probability used to estimate the parameters in the Gibbs sampling is shown in equation 5.

$$P(z_{(d,t)}|z_{-(d,t)}, W, R, \alpha, \beta) = \frac{P(z_{(d,t)}, z_{-(d,t)}, W, R, \alpha, \beta)}{P(Z_{-(d,t)}, W, R, \alpha, \beta)} \tag{5}$$

where $z(d,t)$ denotes the z hidden topic of the n^{th} word token in the d^{th} tweet, W is the vocabulary, and R denotes the relationship between tweets. In Algorithm 1, the difference between LDA and *intLDA* is the addition of lines 14 to 16.

Algorithm 1 *intLDA* Gibbs Sampling

INPUT: tweets t, number of tweets D, number of topics K
OUTPUT: topic assignments z and counts cdt, cwt and ct

```
1:  randomly initialize $z$ and increment counters
2:  for $i = 1 \to D$ do
3:      for $l = 1 \to N_i$ do
4:          $w \leftarrow t_{i,l}$
5:          $topic \leftarrow z_{i,l}$
6:          $cdt_{i,topic}- = 1; cwt_{w,topic}- = 1; ct_{topic}- = 1$
7:          for $k = 1 \to K$ do
8:              $p_k = (cdt_{i,k} + \alpha_k)\frac{cwt_{k,w}+\beta_w}{ct_k+\beta\times W}$
9:          $n\_topic \leftarrow$ sample from $p$
10:         $z_{i,l} \leftarrow n\_topic$
11:         $cdt_{i,n\_topic}+ = 1;$
12:         $cwt_{w,n\_topic}+ = 1;$
13:         $ct_{n\_topic}+ = 1$
14:         foreach $j$ such that $R_{ij} == 1$ do
15:             $cdt_{j,topic}- = 1$
16:             $cdt_{j,n\_topic}+ = 1$
17: return $z, cdt, cwt, ct$
```

V. Experiments

In this section, we discuss the details of our experiments, including the experimental dataset, the baseline methods and the evaluation metrics, and the results.

A. Dataset

For the experiments, we use Twitter messages collected from 03 March 2014 to 07 March 2014 using the Twitter public stream API [2]. Our experiments deal with only English tweets. Our data set includes 729,334 tweets involving 509,713 users, 12,221 reply tweets and 101,272 retweets.

A preprocessing step was performed against the test dataset by removing all irrelevant characters (e.g., emoticons, punctuations) and stop words, and performing spelling correction and lemmatization using NLTK python packages. For the purposes of evaluation, around 20% of the tweets were manually labeled as the training set (one label/topic for every tweet).

B. Baseline Methods

We evaluate *eLDA* and *intLDA* against the following alternatives.

- *LDA*. This is a straight use of LDA [1].
- *Plink-LDA*. This is a variant of LDA that uses relationships between documents as prior information for topic derivation [9]. This variant of LDA is thus closest to our approach. However, the implementation of the prior information in the topic sampling process seems to have no direct impact on the document topic distributions, as the sparse relationship between content and vocabulary still has a higher negative effect on the quality of the topics. For the purpose of this evaluation, we use our observed tweet relationships as the link information between tweets.
- *NMF*. This is a popular algorithm of Non Negative Matrix Factorization that factorizes a tweet-term matrix into tweet-topic and topic-term matrices [3].

C. Evaluation Metrics

We evaluated both the quality of tweet-topic distributions and the coherence of words in the topics.

As mentioned in Section II, for each tweet we chose the topic with highest value in the topic distribution. We subsequently clustered the tweets by their chosen topic and compared the clusters against the clusters generated by our manually labeled training set. We used pairwise *F Measure* and Normalized Mutual Information (*NMI*) metrics to compare the clusters with the annotations.

The pairwise *F-Measure* [14] computes the harmonic mean of both precision p and recall r.

$$F = 2 \times \frac{p \times r}{p + r} \ . \tag{6}$$

where precision p is calculated 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.

NMI [15] measures the mutual information shared between tweet-topic clusters and the training set $I(K;C)$, normalized by the entropy of the clusters $H(K)$ and training set $H(C)$. The value of *NMI* ranges between 0 and 1 (higher is better).

$$NMI(K,C) = \frac{I(K;C)}{[H(K) + H(C)]/2} \ . \tag{7}$$

To measure the coherence between words in a topic, we adopt the metric defined in equation 8, in which $Co(k, W)$ is the measurement of *topic coherence* for a topic k described by its topic-terms in W [16].

$$Co(k,W) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{T(w_m, w_l) + 1}{T(w_l)} \tag{8}$$

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 terms respectively; M is the size of the set W of topic-terms.

D. Discussion

We have conducted experiments on several possible setups for all the methods. We set the number of the topics starting from 20 (k = 20, 40, 60, 80, 100) to assess the performance of the methods for a different number of topics. For every value of k, we ran the algorithms over the dataset 30 times and noted the mean of each evaluation metric. In each experiment, we retrieved the 10 words with highest values in the topic probability distribution as the representative words for the topic.

Figure 2 shows that *intLDA* presents a significant improvement of *F-measure* in comparison to the other methods for every evaluation setup. The method of expanding the tweet content (*eLDA*) also provides an improvement over the straight *LDA* method, *Plink-LDA* and *NMF*. However, the performance of *eLDA* remains below that of *intLDA*. This suggests that incorporating the observed tweet relationships directly in the Gibbs sampling process is more robust to noise than introducing words from the related tweets.

The noise from expanded content on the *eLDA* method has a big impact on the entropy. As shown by Figure 3, the *eLDA* method has the worst performance due to a higher entropy of information. In the *NMI* evaluation, our proposed method *intLDA* gives the best result over the other baseline methods. *Plink-LDA* is the next best method, showing that incorporating the relationships between tweets can produce higher mutual information than straight LDA.

Our evaluation of the topic coherence for each method (Table II) confirms the results of the F-measure of cluster quality. A higher topic coherence value means that the topic is more readable [16]. Table II shows that *intLDA* always performs best on any number of topics. The expanded *eLDA*

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

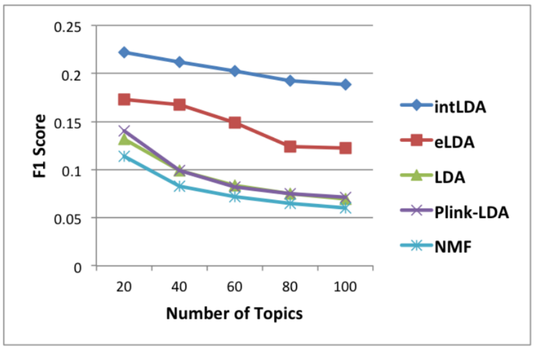

Figure 2: Experiment results using *F-Measure* metric

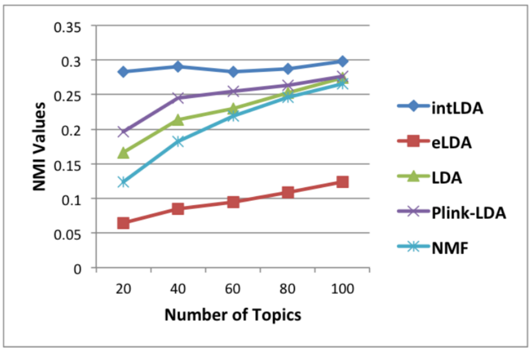

Figure 3: Experiment results using *NMI* metric

method shows only a small improvement over the original LDA.

The improvement of *intLDA* over the original LDA method for topic derivation in Twitter shows that incorporating social interactions is useful to improve topic quality. Our model tries to introduce additional information directly into the original LDA process, which previously worked solely on content. By having the ability to incorporate additional information on LDA, this method could potentially be extended for different tasks in Twitter, such as recommendation systems or collaborative filtering.

VI. Related Work

Popular topic modeling methods, such as PLSA [2], LDA [1] and NMF [3], exploit the document content to infer topics of documents. However, as already mentioned, the short-text nature of Twitter provides very low term co-occurrence which heavily penalizes the qualities of topics. In order to work on Twitter, certain extensions of these methods were proposed, e.g., [17], [18], [19], [6]; however, they are still suffering from the sparsity problem caused by the short-text nature of Twitter.

Table II: Comparison of topic-coherence values

Methods	*K=20*	*K=40*	*K=60*	*K=80*	*K=100*
intLDA	**59.12**	**48.97**	**45.69**	**42.30**	**41.27**
eLDA	58.51	47.93	43.96	41.79	40.00
LDA	58.39	47.52	43.75	41.52	38.39
Plink-LDA	55.42	46.34	43.78	41.13	38.68
NMF	54.04	44.48	43.72	40.43	37.82

The study of [4] tackled the short-text issue by exploiting external document collections. However, this brings the extra burden of identifying relevant corpora to augment the documents. In a rapidly changing environment such as Twitter, this is problematic. In addition, the language used in the tweets might not match that of the external corpora, due to the frequent informal language used on Twitter. Likewise, the study of [6] built a term-correlation matrix from the content of the documents, then jointly use document-term and term-correlation matrices to address the sparsity problem in short-text environments. However, as shown in Table I, the term-to-term relationships as the term-correlation matrix only provides a small improvement with respect to density in comparison with the original tweet-to-term relationships.

The study of [17] and [5] exploited content based social features such as *hashtag* and *url* to improve the quality of the topics. The user's *following-follower* mechanism has also been investigated [20] for determining the popularity of authors to refine the topic learning process in Twitter. However, analyzing the relationships based on *following-follower* suffers from scalability issues in the Twitter's streaming environment, since user details information needs to be queried apart from the dataset itself.

Plink-LDA [9] is a variant of LDA that is close to our approach as it uses relationship information. This approach has been developed to analyze a collection of publications and their links via citations. It uses the link between documents as prior information. In contrast, we work on much shorter documents, and we integrate the link between tweets in the Gibbs sampling algorithm. As discussed in Section V, our approach outperformed *Plink-LDA* in the Twitter data.

Within the domain of social media, [7] applied user context to topic modeling. This approach takes into consideration only conversation patterns, ignoring the tweet contents. The study of [8] suggested that the topics discussed through interactions on social networks had higher credibility than those specified by content-based extraction methods. These studies are aligned with our experiments with respect to the impact of interactions, in their case on topic qualities. However, our research discovered that deriving topics from only the socially connected tweets will lose a great number of important topics in the Twitter environment, as the self-contained tweets occupy the majority of the total tweets. Taking this research into account, *intLDA* effectively incorporates both social interactions and content similarities in the topic derivation process to achieve high quality results.

VII. Conclusion

In this paper, we present a method that incorporates information about tweets relationship for topic derivation. *intLDA* is an extension of LDA that incorporates the relationship information directly in the Gibbs sampling process.

We have conducted several experiments of topic derivation on a Twitter dataset. Our experiments demonstrate that the defined relationships between tweets are helpful to improve the quality of the topic derivation result. Our evaluation results show that *intLDA* consistently outperforms *eLDA*, *Plink-LDA* and other methods that do not incorporate relationship information.

The relationships *intLDA* takes into account are based on the interactions of people, actions and content similarity between tweets. We are currently investigating more complex social features to observe their effects on topic derivation. Having achieved an improvement over *LDA*, *Plink-LDA* and *NMF*, we will also extend the study to incorporate the tweet-relationships for topic derivation in a real-time situation using an online and incremental version.

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] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[3] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2000, pp. 556–562.

[4] M. Albakour, C. Macdonald, I. Ounis *et al.*, "On sparsity and drift for effective real-time filtering in microblogs," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 419–428.

[5] J. Vosecky, D. Jiang, K. W.-T. Leung, K. Xing, and W. Ng, "Integrating social and auxiliary semantics for multifaceted topic modeling in twitter," *ACM Transactions on Internet Technology (TOIT)*, vol. 14, no. 4, p. 27, 2014.

[6] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, "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] R. Pochampally and V. Varma, "User context as a source of topic retrieval in twitter," in *Workshop on Enriching Information Retrieval (with ACM SIGIR)*, 2011, pp. 1–3.

[8] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on twitter," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 2012, pp. 179–188.

[9] H. Xia, J. Li, J. Tang, and M.-F. Moens, "Plink-lda: Using link as prior information in topic modeling," in *Database systems for advanced applications*. Springer, 2012, pp. 213–227.

[10] K. Erk, "Vector space models of word meaning and phrase meaning: A survey," *Language and Linguistics Compass*, vol. 6, no. 10, pp. 635–653, 2012.

[11] X. Yan, J. Guo, Y. Lan, and X. Cheng, "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, 2013, pp. 1445–1456.

[12] A. de Moor, "Conversations in context: a twitter case for social media systems design," in *Proceedings of the 6th International Conference on Semantic Systems*. ACM, 2010, p. 29.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[14] C. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[15] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2003. [Online]. Available: http://dx.doi.org/10.1162/153244303321897735

[16] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2011, pp. 262–272.

[17] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models." *ICWSM*, vol. 10, pp. 1–1, 2010.

[18] Y. Hu, A. John, F. Wang, and S. Kambhampati, "Et-lda: Joint topic modeling for aligning events and their twitter feedback." in *AAAI*, vol. 12, 2012, pp. 59–65.

[19] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 1992–2001, 2013.

[20] Y. Cha, B. Bi, C.-C. Hsieh, and J. Cho, "Incorporating popularity in topic models for social network analysis," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 223–232.