ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1559

Topic-Level Influencers Identification in the Microblog Sphere

Yakun Wang and **Zhongbao Zhang** and **Sen Su** and **Cheng Chang** and **Muhammad Azam Zia**[1]

Abstract. This paper studies the problem of identifying influencers on specific topics in the microblog sphere. Prior works usually use the cumulative number of social links to measure users' topic-level influence, which ignores the dynamics of influence. As a result, they usually find faded influencers. To address the limitations of prior methods, we propose a novel probabilistic generative model to capture the variation of influence over time. Then a influence decay method is proposed to measure users' current topic-level influence.

1 Introduction

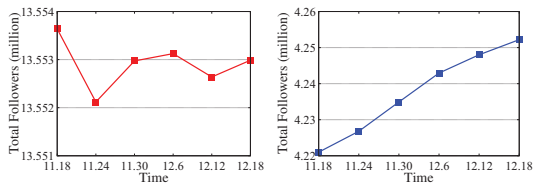
Researchers have focused on topic-level influence analysis [1, 6]. In general, people want to find recent influencers rather than outdated ones. In these prior studies, it is a common way to utilize the cumulative number of social links (e.g., followship, reposts and mentions) to identify the topic-level influencers. However, we observe that the links are created over time. For measuring users' influence, it is critical to incorporate the variation trend of influence. A real example about two famous basketball players Jianlian Yi and Jeremy Lin in Sina Weibo is illustrated in Fig. 1. We can see that although Yi has more followers than Lin, the number of Yi's followers no longer increases, while Lin gets more and more followers along with time. Accordingly, we can not simply assume Yi has more influence than Lin just because Yi owns more followers. However, if assuming both Yi and Lin are followed for basketball, all prior methods will select Yi as the key influencer rather than Lin, which leads to inaccurate models. This example conveys that the learned influence by the cumulative number of links is far from adequate, since users' influence is dynamic and rises or falls over time [3].

In this paper, we intend to identify recent popular and influential users on specific topics rather than faded ones. To address this problem, we firstly propose a novel probabilistic generative model, which we refer to as Topic-level Influence over Time (TIT), for capturing the temporal aspect of influence on specific topics. Then we design an exponential decay method that works on the learned temporal influence to compute the influence of each user on specific topics, which takes both quantity and trend of influence into consideration. Through extensive experiments on real-world dataset, we demonstrate the effectiveness of our approach.

2 Topic-Level Influence Analysis

2.1 TIT Model

Firstly, we intend to model the topic-level influence over time, which can better help us capture the dynamics of influence. We propose a Topic-level Influence over Time (TIT) model jointly over text, links and time based on the LDA model [2]. It uncovers the latent topics and users' topic-level temporal influence in a unified way. The plate notation is given in Fig. 2. Specifically, there are two components in this model: the user-word component in the right part and the user-(link, time) component in the left part of Fig. 2.



(a) Jianlian Yi (b) Jeremy Lin

Figure 1. The Number of Total Followers over Time (Year 2015)

The user-word component is to model user u's words. We aggregate the words w posted by u into an integrated document from which we use LDA model to discover the latent topics. As a result, each user has a Multinomial distribution θ over topics and each topic has a Multinomial distribution φ over words. The user-(link, time) component is to model the u's links (e.g., followship) and the corresponding generation time in the microblog network. We discretize the time by dividing the entire time span of all links into T time slices. We consider the network as a document corpus and each user u is represented by a document where user f that u communicated with and the corresponding time t pairs form the words in this document. Note that this component consists of two levels of mixtures: an upper-level Bernoulli mixture μ and two underneath-level multinomial mixture parts σ and π. μ is for deciding whether the link creation is based on u's topics or not. If topic based, we model the topic x (generated by θ) over (f, t) by a multinomial distribution σ. Otherwise, we use a global multinomial distribution π to model (f, t). Benefiting from the learning results of σ, we can generate the influence trend line over time of each user like Fig. 1, and this can greatly help us to identify the key topic-level influencers on microblogs.

2.2 Parameter Estimation

We use Gibbs sampling [5] to obtain samples of the hidden variable assignment and to estimate the model parameters from these samples. Let $x_{\neg i}$ denote the set of all hidden variables of topics except x_i and $n^{(\cdot)}_{\cdot,\neg i}$ denote the count that the element i is excluded from the corresponding topic or user. Here, we only give the sampling formula of links. For a link f_i and the corresponding time t_i with index $i = (u, l)$, we jointly sample y_i and x_i from the conditional as the following two equations:

$$p(x_i, y_i = 1 | f, t, x_{\neg i}, y_{\neg i}, z, \alpha, \gamma, \rho) \quad (1)$$

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China, email: {wangyakun, zhongbaozb, susen, changwow, zia}@bupt.edu.cn. Corresponding author of this paper is Prof. Sen Su.

$$\propto \frac{n^{(f,*)}_{k,\neg i}+\gamma}{\sum_{f=1}^{U} n^{(f,*)}_{k,\neg i}+U\gamma}(n^{(y=1)}_{u,\neg i}+\rho_1)(n^{(k)}_{u(w)}+n^{(k)}_{u(f),\neg i}+\alpha)$$

$$p(x_i, y_i = 0|f, t, x_{\neg i}, y_{\neg i}, z, \alpha, \epsilon, \rho) \quad (2)$$

$$\propto \frac{n_{(f,*),\neg i}+\epsilon}{\sum_{f=1}^{U} n_{(f,*),\neg i}+U\epsilon}(n^{(y=0)}_{u,\neg i}+\rho_0)(n^{(k)}_{u(w)}+n^{(k)}_{u(f),\neg i}+\alpha),$$

where $n^{(f,*)}_{k,\neg i}$ denotes the number of times that user f occurs in topic k, $n_{(f,*),\neg i}$ denotes the number of times that user f occurs without any topic, $*$ represents an aggregation on time dimension, $n^{(k)}_{u(w)}$ denotes the number of times that topic k has been observed with a word w of user u, and $n^{(k)}_{u(f),\neg i}$ denotes the number of times that topic k has been observed with a link f of u, $n^{(y=1)}_{u,\neg i}$ and $n^{(y=0)}_{u,\neg i}$ denote the number of times the links created by u is related to topics or regardless of topics, respectively.

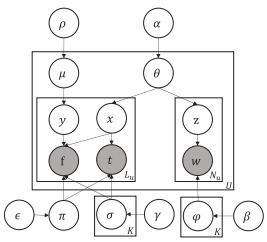


Figure 2. Plate diagram of TIT

2.3 Measuring Users' Influence

Given the topic-level influence trend lines over time derived from σ, users who get lots of attention from others and have a upward trend of influence can be easily found as the key influencers on the corresponding topics. However, for some cases like the example Yi and Lin in Fig. 1, we can not easily identify who exhibits more influence, since Yi has more followers than Lin, while Lin has a better growing trend of influence than Yi. Intuitively, links generated long time ago have little contribution to users' influence. It means the more closer of the links generated in time, the more important they are to users' influence. Hence, we utilize the exponential decay function to model the influence decay. Specifically, σ is a distribution of topics over a set of 2-tuples $\{(f,t)\}$. That is, σ is a $U \times T \times K$ matrix in the procedure of sampling recording the number of times (f,t) has been assigned to topic k, denoted as $n^{(f,t)}_k$, plus prior parameter γ, i.e., $\sigma_{u,t,k} = n^{(f,t)}_k + \gamma$. Thus, we can use the following equation to measure the influence of user f on topic k till time T:

$$Influence(f)@(k,T) = \gamma + \sum_{t=1}^{T} n^{(f,t)}_k \times e^{-\frac{T-t}{\lambda}} \quad \lambda{>}0. \quad (3)$$

Here, λ is a parameter controlling the decay rate of influence.

3 Experiment

Dataset: We crawl the followship network from Sina Weibo[2]. Since Sina Weibo does not release the information about when a user follows another, we periodically crawl the follow list of all users in our seed set, monitor their changes and then label the new generated links with timestamps. We also crawl their recent 100 messages. Finally, after preprocessing, there are 0.4M users, 207M words, 46M links with 7M time-tagged and 24 time slices with each nearly 1.5 days in our dataset. We empirically set the values of hyperparameters of TIT in line with other topic modelling work [1]. We set $\lambda = 11$ through minimizing held-out perplexity on a validation set.

[2] http://weibo.com

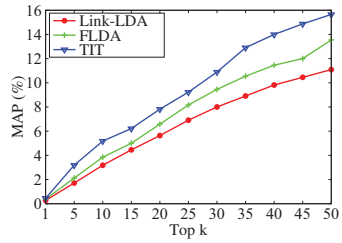


Figure 3. All Categories

Precision: We evaluate our approach by comparing it with Link-LDA [4] and Followship-LDA (FLDA) [1]. For the ground truth, Sina Weibo gives the lists of popular users or organizations about 36 categories such as sports and music. Each category list contains 100 ranked users. It is clear that these popular users or organizations are some kind of the key influencers on the corresponding topics. Sina Weibo states that these lists are updated by month. Intuitively, TIT considering the temporal dynamics of influence should produce more precise results. Although these rankings do not necessarily have 100% precision, they give us enough information to facilitate relative comparisons across different approaches. Fig. 3 shows the results of Mean Average Precision (MAP) across all categories. It is clear that TIT significantly outperforms the competitors.

4 Conclusion

This paper studies the problem of analyzing the topic-level temporal influence of users for the finding recent influencers on specific topics in microblog sphere. To achieve this, we first propose the TIT (Topic-level Influence over Time) model, a novel probabilistic generative model jointly over text, links and time. Then, we design an influence decay based approach to measure users' topic-level influence from the learned temporal influence. We compare our approach with Link-LDA and FLDA on a real dataset crawled from Sina Weibo. Experimental results demonstrate the effectiveness of our approach.

5 Acknowledgements

This work is supported in part by the following funding agencies of China: National Natural Science Foundation under Grant 61170274 and U1534201, and the Fundamental Research Funds for the Central Universities (2015RC21).

REFERENCES

[1] Bin Bi, Yuanyuan Tian, Yannis Sismanis, Andrey Balmin, and Junghoo Cho, 'Scalable topic-specific influence analysis on microblogs', in *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 513–522. ACM, (2014).

[2] David M Blei, Andrew Y Ng, and Michael I Jordan, 'Latent dirichlet allocation', *the Journal of machine Learning research*, **3**, 993–1022, (2003).

[3] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi, 'Measuring user influence in twitter: The million follower fallacy.', *ICWSM*, **10**(10-17), 30, (2010).

[4] Elena Erosheva, Stephen Fienberg, and John Lafferty, 'Mixed-membership models of scientific publications', *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5220–5227, (2004).

[5] Thomas L Griffiths and Mark Steyvers, 'Finding scientific topics', *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5228–5235, (2004).

[6] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He, 'Twitterrank: finding topic-sensitive influential twitterers', in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270. ACM, (2010).