The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

A Novel Embedding Method for News Diffusion Prediction

Ruoran Liu,[12] **Qiudan Li,**[1] **Can Wang,**[12] **Lei Wang,**[1] **Daniel Dajun Zeng**[123]
[1]Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]University of Arizona, Tucson, Arizona, USA
{liuruoran2016, qiudan.li, wangcan2015, l.wang, dajun.zeng}@ia.ac.cn

Abstract

News diffusion prediction aims to predict a sequence of news sites which will quote a particular piece of news. Most of previous propagation models make efforts to estimate propagation probabilities along observed links and ignore the characteristics of news diffusion processes, and they fail to capture the implicit relationships between news sites. In this paper, we propose an algorithm to model the news diffusion processes in a continuous space and take the attributes of news into account. Experiments performed on a real-world news dataset show that our model can take advantage of news' attributes and predict news diffusion accurately.

Introduction

News media is increasingly becoming an important platform for information diffusion. Understanding news diffusion mechanism can help management department maximize the news influence such that more people are informed of the hot topics. Most of the existing propagation models are graph-based and focus on estimating the propagation probabilities along observed links (Guille et al. 2012; Li et al. 2014). Since these probabilistic models assume that information only diffuses along observed links, their ability to predict future diffusions is limited. Recently, instead of modeling information diffusion processes on discrete preexisting graph structure, Bourigault et al. (2014) and Gao et al. (2017) proposed to model diffusion processes in a latent continuous space by learning nodes' embedding from observed diffusion processes directly.

During the processes of news diffusion, news's semantic category and the location category of its source site are key factors that determine which news sites will quote this news. For example, financial news is more likely to be quoted by professional financial news sites than sports news sites, and news about local events is usually spread among local medias. Thus, we propose a news diffusion model which maps news sites into a continuous space based on the observed diffusion processes in the training dataset directly. Moreover, the attributes of news being propagated are integrated as the offset of source site's location in the latent space. Finally, the news quoting sequence is obtained based on the distances to the source site of news in the latent space. To the best of our knowledge, this work is a first step towards utilizing an embedding method to analyze news diffusion processes among news sites.

Proposed News Diffusion Model

Given news's attributes and its source site, news diffusion model aims to predict a sequence of news sites which will quote this news. Suppose that there are N news sites and W pieces of news. For news $\mathrm{n_i}$, $\mathrm{t_{ij}}$ is used to describe the time when $\mathrm{n_i}$ was quoted by news site $\mathrm{u_j}$. Specifically, $\mathrm{t_{ij}}=\infty$ means that $\mathrm{u_j}$ didn't quote $\mathrm{n_i}$. Moreover, a feature vector $\mathrm{f_i} \in \mathrm{R^M}$ is used to characterize attributes of news $\mathrm{n_i}$, where M is the size of the feature space. To model the news diffusion processes in the continuous space, we first define $\mathrm{Z}=\{\mathrm{Z_1}, \mathrm{Z_2}, \cdots, \mathrm{Z_N}\}$, where $\mathrm{Z_j} \in \mathrm{R^d}$ denotes the location of news site u_j in the continuous space and d is the size of the space. Then a constraint is introduced to learn news sites' embedding such that sites quoting the news earlier are closer to the source site in the space. Given news $\mathrm{n_i}$ and its source site $\mathrm{u_{s_i}}$, for any pair of users $(\mathrm{u_x}, \mathrm{u_y})$, if $\mathrm{t_{ix}}<\mathrm{t_{iy}}$, $\mathrm{u_x}$ should be closer to $\mathrm{u_{s_i}}$ in the continuous space. This goal is formally expressed as the objective function (1).

$$\sum_{\mathrm{i=1}}^{\mathrm{W}} \sum_{\mathrm{t_{ix}<t_{iy}}} \max\left(0,\ 1\text{-}\ \mathrm{dist_i}(\mathrm{u_y}) + \mathrm{dist_i}(\mathrm{u_x})\right) \tag{1}$$

In our model, the Euclidean distance is adopted to measure the distance between news site $\mathrm{u_j}$ and news $\mathrm{n_i}$'s source site $\mathrm{u_{s_i}}$. Meanwhile, the feature vector $\mathrm{f_i}$ is integrated as an offset of source site $\mathrm{u_{s_i}}$ such that the location of $\mathrm{u_{s_i}}$ moves from Z_{s_i} to $Z_{s_i} + f_i W$, where $W \in R^{M\times d}$ is a matrix to transform feature vector f_i into the same continuous space as news sites. Therefore, the final distance function is defined as (2):

$$\mathrm{dist}_i\left(u_j\right)=\left\|Z_{s_i} + f_i W - Z_j\right\|^2 \qquad (2)$$

Finally, stochastic gradient descent algorithm is adopted to learn the parameters Z and W simultaneously from the news quoting sequences in the training dataset. The two parameters are updated as (3) if and only if $\left\|Z_{s_i} + f_i W - Z_y\right\|^2 - \left\|Z_{s_i} + f_i W - Z_x\right\|^2 < 1$:

$$\begin{gathered} Z_x \leftarrow Z_x + 2\eta(Z_{s_i} + \ f_i W - Z_x) \\ Z_y \leftarrow Z_y - 2\eta(Z_{s_i} + f_i W - Z_y) \\ Z_{s_i} \leftarrow Z_{s_i} + 2\eta(Z_x - Z_y) \\ W \leftarrow W + 2\eta \times f_i^T(Z_x - Z_y) \end{gathered} \qquad (3)$$

where η is the learning rate.

Experiments and Results

Datasets. We carried out experiments on a news dataset which consists of 298307 pieces of news and 5521 news sites. 80% of news and their quoting sequences are used as the training dataset, the remaining as the testing dataset.

Baseline methods. 1) Independent Cascade Model(IC) (Bourigault et al. 2014): In this graph-based model, the news is propagated through the pre-existing links and the probabilities along links are learnt by EM algorithm from quoting actions in the training dataset. 2) APP: The proportion of news published by a site in the training dataset is used as the likelihood of quoting news for this site in the test phase.

Evaluation Measures. Mean average precision (map) and recall@k (r@k) are adopted to evaluate the performance (Liang et al. 2016).

Results and discussion. We empirically set d=500*,* η=0.01 for all news diffusion models and the results of different models are shown in Table 1.

Table 1: The results of different models

	news diffusion model				IC	APP
	O	S	L	S + L		
map	0.4924	0.5017	0.5104	**0.5117**	0.3767	0.1991
r@5	0.4982	0.5179	0.5229	**0.5239**	0.4122	0.2213
r@10	0.5297	0.5425	0.5513	**0.5554**	0.4737	0.2936
r@15	0.5967	0.6067	0.6165	**0.6211**	0.5338	0.3654
r@20	0.6543	0.6612	0.6666	**0.6713**	0.5792	0.4168

We first predict news diffusion with no attributes of news (O, for short) and this model outperforms baseline methods. This reflects that our news diffusion model based on continuous space can capture the implicit relationships which are unobserved in the training dataset.

Furthermore, there are 15 news semantic categories (S, for short) and 35 site's locations (L, for short) in our dataset. We take into account the two attributes respectively. Both models perform better than original model without any news attributes, which suggests that news's attributes can help learn more detailed propagating patterns.

Finally, we predict news diffusion by the model with both semantic category and location attributes and obtain best performance, which shows that both attributes should be considered to model news diffusion comprehensively.

Case study. An intuitive example is shown in Table 2 to illustrate the influence of news attributes. The predicted sequence is obtained by model with both attributes and only top 5 sites are shown in the table for space reason. These examples explain that financial news is usually quoted by professional financial sites. Moreover, the predicted sequences show that our model can map sites with similar attributes into close locations in the continuous space and capture the propagating patterns of news with different attributes.

Table 2: The examples for illustrating the importance of news attributes. All of these sites are professional finance sites.

semantic category	finance	location of source site	Beijing
true sequence	www.cs.com.cn,www.jrj.com.cn,www.stockstar.com, www.ce.cn, www.cnfol.com		
predicted sequence	**www.cs.com.cn**, **www.jrj.com.cn**, **www.cnfol.com**, www.eastmoney.com, **www.stockstar.com**		

Acknowledgements

This research is supported by the Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3; National Natural Science Foundation of China under Grant No. 71621002, 61671450, 61402123.

References

Bourigault S, Lagnier C, Lamprier S, et al. Learning social network embeddings for predicting information diffusion. *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014: 393-402.

Gao S, Pang H, Gallinari P, et al. A Novel Embedding Method for Information Diffusion Prediction in Social Network Big Data. *IEEE Transactions on Industrial Informatics*, 2017.

Guille A, Hacid H. A predictive model for the temporal dynamics of information diffusion in online social networks. *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012: 1145-1152.

Li H, Cao T, Li Z. Learning the information diffusion probabilities by using variance regularized em algorithm. *Advances in Social Networks Analysis and Mining*, *2014 IEEE/ACM International Conference on*. IEEE, 2014: 273-280.

Liang D, Altosaar J, Charlin L, et al. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. *Proceedings of the 10th ACM conference on recommender systems*. ACM, 2016: 59-66.