A Novel Embedding Method for News Diffusion Prediction

Ruoran Liu,1,2 Qiudan Li,1 Can Wang,1,2 Lei Wang,1 Daniel Dajun Zeng1,2,3

1Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
2University of Chinese Academy of Sciences, Beijing, China
3University 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 pre-existing 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 ni, tij is used to describe the time when ni was quoted by news site uj. Specifically, tij=∞ means that uj didn’t quote ni. Moreover, a feature vector fi ∈ RM is used to characterize attributes of news ni, where M is the size of the feature space. To model the news diffusion processes in the continuous space, we first define Z={Z1, Z2, ⋯, ZN}, where Zj ∈ Rd is the location of news site uj 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 ni and its source site usi, for any pair of users (ux, uy), if tix<tiy, ux should be closer to usi in the continuous space. This goal is formally expressed as the objective function (1).

\[ \sum_{i=1}^{W} \sum_{t_{ix}<t_{iy}} \max \left( 0, 1 - \text{dist}_i(u_x) - \text{dist}_i(u_y) \right) \]  

In our model, the Euclidean distance is adopted to measure the distance between news site u and news ni’s source site usi. Meanwhile, the feature vector fi is integrated as an offset of source site usi such that the location of usi moves
Results and discussion.

Liang et al. (2016). recall@k (r@k) are adopted to evaluate the performance of baseline methods. 1) Independent Cascade Model (IC): Given a training dataset, the remaining as the testing dataset. The two parameters are updated as (3) if and only if ∥Zsi + f1W - Zx∥2 - ∥Zsi + f1W - Zy∥2 < 1:

\[ Zx \leftarrow Zx + 2\eta(Zsi + f1W - Zx) \]
\[ Zy \leftarrow Zy + 2\eta(Zsi + f1W - Zy) \]
\[ Zsi \leftarrow Zsi + 2\eta(Zx - Zy) \]
\[ W \leftarrow W + 2\eta*f1^T(Zx - Zy) \]

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>
<thead>
<tr>
<th>semantic category</th>
<th>finance</th>
<th>location of source site</th>
<th>Beijing</th>
</tr>
</thead>
<tbody>
<tr>
<td>true sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predicted sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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.

References


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.