Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Information Credibility Evaluation on Social Media

Shu Wu, Qiang Liu, Yong Liu, Liang Wang, Tieniu Tan
Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, China
{shu.wu, qiang.liu, liuyong, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

With the growth of social media, rumors are spread fast and viewed by more and more people on the Internet. Rumors bring significant harm to daily life and public security. It is crucial to evaluate the credibility of information and detect the rumors on social media automatically. In this work, we establish a Network Information Credibility Evaluation (NICE) platform, which collects a database of rumors that have been verified on Sina Weibo and automatically evaluates the information which is generated by users on social media but has not been verified. Users can use a query to search related information. If the according information appears in our database, users can identify it is a rumor immediately. Otherwise, NICE will show users with real-time results crawled automatically from social media and can calculate credibility of a specific result with our algorithm. Our algorithm learns dynamic representations for information on social media based on behavior information, dynamic information, user information and comment information. Then, we use an ordinary logistic regression to classify information into rumors and non-rumors. Based on our algorithm, NICE system achieves satisfactory performance on evaluating information credibility and detecting rumors on social media.

Introduction

Nowadays, online social media, such as Facebook, Twitter and Sina Weibo, is developing rapidly all over the world. Users are free to share information and publish their comments con the Internet. On these platforms, information can be generated and spread more rapidly, while rumors or misinformation can also be spread and viewed by more people. A rumor is an unverified and instrumentally relevant statement of information spread among people (DiFonzo and Bordia 2007). Rumors are showing their harm on our daily life or even public security. For instance, the lost of MH370 draw world-wide attention, great amount of rumors have been generated and spread on social media, e.g., MH370 has landed in China, the lost MH370 is caused by terrorists and Russian jets are related to the lost of MH370. These rumors may lead public opinion to a wrong direction and delay the search of MH370. Rumors are exploding on social media. Up to September 3th, 2015, on China's biggest microblog website Sina Weibo[1], 28374 rumors have been reported and collected on its misinformation management center[2]. Accordingly, it is crucial and urgent to evaluate the credibility of information and detect the rumors on social media.

Extensive works have been done for automatically evaluating the credibility of information on social media. Existing methods are mainly based on feature engineering, most of which are based on content information and user credibility. Some methods evaluate credibility at the message level (Castillo, Mendoza, and Poblete 2011), while some evaluate credibility at the event level (Kwon et al. 2013; Zhao, Resnick, and Mei 2015; Ma et al. 2015). And some research studies the credibility aggregation from message level to event level (Jin et al. 2014). Moreover, considering dynamic information, temporal features based on prorogation properties over time have been incorporated (Kwon et al. 2013). And some works train a model with features generated in different time periods (Ma et al. 2015). There are also some works which consider the user feedback and suspicious tweets (Zhao, Resnick, and Mei 2015). Based on feature engineering, conventional methods can not reveal underlying properties of data and require great labor in designing features. Besides, features in conventional methods are all based on global statics, which have difficulty in revealing the correlation among different aspects of information.

In this work, we plan to evaluate the credibility of the event, which is composed several messages posted and reposted by users on social media. Rather than designing artificial features, for credibility evaluation, we plan to learn a representation of each message to model the dynamic behaviors in spreading. In our algorithm, each user is represented as a vector, while time interval, user behavior and user comment opinion are represented as matrices respectively. The aggregation of these representations generates the representation of a piece of message. Finally, after aggregating all the dynamic behavioral representations of messages, we can generate the credibility representation of an event.

[1] http://weibo.com

[2] http://service.account.weibo.com/?type=5&status=4

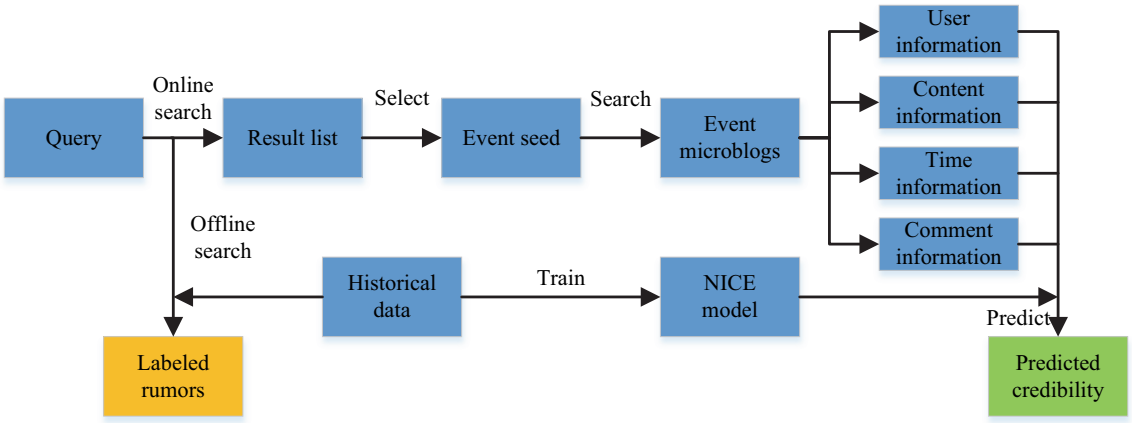

Figure 1: Overview of NICE system.

Algorithm

In this work, our task is to predict the credibility of an event, which contains several messages (posting or reposting). For an event e_i containing n_{e_i} messages (denoted as a set M^{e_i}), we can learn event representation as:

$$\mathbf{r}^{e_i} = \frac{1}{n_{e_i}} \sum_{m_j^{e_i} \in M^{e_i}} (\mathbf{T}_t^{e_i} + \mathbf{C}_j^{e_i} + \mathbf{B}_j^{e_i}) \mathbf{u}_j^{e_i} , \qquad (1)$$

where $m_j^{e_i}$ is a message of the event e_i, $\mathbf{u}_j^{e_i}$ is latent vector of the corresponding user posted or reposted the message $m_j^{e_i}$, $\mathbf{B}_j^{e_i}$ is operating matrix of his or her behavior (posting or reposting), $\mathbf{C}_j^{e_i}$ means the operating matrix of comment opinion (whether the message is suspicious or nor), latent matrix $\mathbf{T}_j^{e_i}$ denotes operation of the time interval since the beginning of the event.

Then, after generating the credibility representations of events, we apply Logistic Regression (LR) for learning of the model, where the prediction whether an event e_i is a rumor can be calculated as:

$$y^{e_i} = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{r}^{e_i}}} , \qquad (2)$$

where $\mathbf{w}$ is a latent vector denoting the weights of regression. Based on this model, using the related information of an event, we can evaluate the credibility of the event.

System

The NCIE system is constructed based on our algorithm and historical dataset collected from Sina Weibo. In order to train the model, we first construct a database containing rumors and non-rumors from Sina Weibo. To crawl rumors, we first collected some rumor seeds from misinformation management center of Sina Weibo, then extracted key words from the rumor seeds and retrieved microblogs based on these key words. We crawled all the matching microblogs, and for each microblog, we collected its reposts, comments, dynamic information and the corresponding user's profile. To crawl non-rumors, we collected some hot topics on Sina Weibo and used similar process of crawling rumors to collect corresponding information. Finally, this database contains the verified rumors and non-rumors, which are used to train our model.

Figure 1 illustrates the flow chart of the NICE system. The user can input a query to retrieval the related information using the system. If a user's query matching the rumor in the database, users can identify the rumor immediately. Otherwise, NICE will crawl real-time information from social media, and user can select an event to evaluate the information credibility based on our model. Based on the selected microblog, the system will crawl the related microblogs from Weibo, and collect related information including content information, time information, comment information and corresponding user profile. Based on these information and the trained model, NICE can evaluate the credibility of the related information and provide a predicted score of the event.

Acknowledgments

This work is jointly supported by National Basic Research Program of China (2012CB316300), and National Natural Science Foundation of China (61403390, U1435221).

References

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *WWW*, 675–684.

DiFonzo, N., and Bordia, P. 2007. Rumor, gossip, and urban legend. *Diogenes* 54(1):19–35.

Jin, Z.; Cao, J.; Jiang, Y.-G.; and Zhang, Y. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *ICDM*, 230–239.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *ICDM*, 1103–1108.

Ma, J.; Srivastava, M.; Toniolo, A.; and Norman, T. J. 2015. Detect rumors using time series of social context information on microblogging websites. In *CIKM*.

Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *WWW*, 1395–1405.