Predicting the knowledge flow of social networks based on machine learning

Muhi Saadi Radhi1*, Mohammadreza Soltan Aghaei2

1,2 Department of Computer Engineering, Isfahan Branch (Khorasgan), Islamic Azad University, Isfahan, Iran.

mohisaadi2018@gmail.com

Abstract. Predicting the release of information deals with the path of publishing a news or information or topic in a structural data such as a graph. Researchers in this field seek to solve the following questions by providing methods for predicting the path of publication: 1- Which information or topics that are most common, 2- Why, how and in what Information path will be published and will be published in the future? And 3- Which network member has an important role in the dissemination process? Machine learning is an area which has been very helpful recently in order to answer such questions. Machine learning as an artificial intelligence subset presents so acceptable to predict the dissemination of information. Since predicting novel users who are in information flow is the process of diagnostic, the issue is able to be solved by the algorithms of Machine learning.

1. Introduction

Online social networks allow hundreds of millions of Internet users around the world to produce and publish content. People on these networks have access to a vast source of information on an incredible scale. Online social networks play an important role in disseminating this information by increasing the dissemination of new information and various perspectives [1]. The issue of information dissemination on small world social networks was addressed by Mark Granwater in his 1973 paper The Strength of Weak Ties, and has been one of the areas of academic study active for the past three decades. Facebook's significant role during the Arab Spring 2012, as well as the important role of Twitter in the 2008 US presidential election, are examples of the influence of these networks. According to online social networks impact on society, recent concentration is on extracting worthy information from the large data amount [3]. During the Greater Japan Tsunami in 2011, many people used social media such as Twitter to access important information, but despite all the significant impact that social media has had on the lives of human societies, it has had devastating effects. They can have, including the dissemination of misinformation that can cause chaos and confusion among people in critical situations such as floods, earthquakes, tsunamis, etc. Analysis of online community analysis is the forerunner of extensive research into the impact of information dissemination on online social networking communities. There are different methods and methods of analysis for disseminating information in virtual networks, the main purpose is to disseminate information dissemination information in a virtual network with the least calculations (path length) [4]. Events, issues, interests, etc., when they happen, evolve very quickly in social networks and their absorption, understanding, visualization and prediction create important expectations from experts and researchers. This is because understanding the dynamics of these networks may lead to better follow-up of events (eg revolutionary wave analysis), problem
solving (such as preventing terrorist attacks, predicting natural hazards), optimizing business performance (For example, optimizing social marketing campaigns. For this reason, researchers in recent years have introduced a variety of techniques and models for 1- compiling information on social networks, 2- analysing, 3- extracting knowledge, and 4- predicting their dissemination. [3] One of the most important dynamic processes in social networks and complex networks in general is the process of information dissemination analysis. In general, anything that can be moved physically or virtually between nodes can be considered in this diffusion process. From the spread of viral or socially contagious diseases to the anticipation of disasters, from rumors to the spread of a belief or religion, from the promotion of a product to the spread of technology, from the spread of news. Until the vote is published, everyone will be in this category. Such publications may not be obvious, and we can only conclude from the available evidence. For example, when a contagious virus breaks out, we only see people getting sick one after another. But we do not know how the disease reached them. The main value of the results of this paper is to find hidden flows of information that by observing various events will eventually reach a proposed graph as an information bus that can be useful in solving all the problems mentioned.

The primary purpose of this paper is to map the problem of predicting the path of information dissemination at the level of social networks to a machine learning problem. Since predicting new users who are exposed to information is a diagnostic process, it can be solved by machine learning algorithms. Based on what is stated in the results section of the article, machine learning algorithms work well in solving prediction problems based on graph network data.

This paper remaining parts are grouped as below. Part II defines related works. Part III defines various proposed method elements. Part IV defines procedure of testing also analyzes results. Part V provides final conclusions.

2. Related works

Understanding, modelling and predicting rapid developments and events in network is one of the most essential network systems tasks like social networks. Basically, it is because of reality recognition named structure discovery of network, which predicts patterns of social events such as their shape, size and growth as information dissemination [5-7]. Many researchers have examined different techniques and methods for modelling information in homogeneous and heterogeneous networks such as social networks [6-10]. These techniques seek to find a solution to how to predict the publishing process in a data network, from a temporal or spatial perspective by learning from past publications. Formally, the network of information with graph (G = (V, E)) with a V nodes set as well as an edges E set is homogeneous if and only if nodes and edges are of one kind. If various nodes kinds as well as the relationships are present at edges, this is heterogeneous. Different papers have been done on homogeneous networks such as link prediction and semantic analysis [11-13]. Recently, heterogeneous networks have been taken into consideration as the attractive research domain because of more natural heterogeneous networks assumption in a lot of real events in the world. [14] investigated network structure and threshold value role in dissemination of information. However, the dissemination of information in [15] has been studied as a special issue in this regard using different main paths in heterogeneous networks. In [16] with the recognition of power in the transmission of surrounding information for different types of main routes is proposed. As in [17], a weight is considered between the two nodes through which the predictions are made.

Different methods for different problems mentioned above are introduced in the field of information propagation path in heterogeneous networks. Recently, machine learning methods have been widely used to solve such problems. Since most of the problems in this field are related to diagnosis and prediction, machine learning and classification algorithms are used in this field. For example, the problem of predicting the propagation path of different topics. The problem in which active and inactive users (whether they have received the information) have been mapped to a classification problem by several articles and by algorithms. Different classifications are solved. Various methods for this problem have been introduced based on classical classification algorithms such as decision trees, neural networks, nearest neighbour k, and so on.
In addition to classical machine learning methods, deep machine learning algorithms have recently been introduced to solve information propagation path problems. Much attention has been focused on deep learning in the dissemination of information of heterogeneous networks [18]. For example, in [19] a Deep Learning with Deep Learning Framework (CDL) has been explored to address the problem of heterogeneous VIS-NIR compliance through network dissemination. In [20] a deep architecture for heterogeneous information network (HIN) is proposed. Recently, a new HIN algorithm has been proposed that describes the main types of network propagation paths in HIN.

3. Proposed method

One of the extensive dynamic processes is Information Dissemination in networks study which has potential applications in different domains. Information like innovation, news, malware or virus begins from a nodes and spreads set across the network. There is a rich literature (study) on information dissemination in complex networks, in various models and their interaction with network topology are studied. In the last study, mainly heterogeneous networks have been studied. The network of information \( G = (V, E) \) that \( V \) as nodes set and \( E \) as edges set is a homogeneous network if nodes and edges are of one kind. Networks with nodes or edges of more than one type of network are called heterogeneous networks. For example, in the Facebook network, which is an important provider of bibliography in social humanity, nodes are users. In this network, edges can be the user-user relationship when the user (first user) is in contact with the second user and attends a group. The flowchart of the proposed method is shown in Figure 1. Input is a data set on social networks. The learning algorithm performs the detection process until it is completed.

Here first are the details of Graph. Then in the next section, this algorithm will be used to learn the data graph in this paper. A data graph can be decomposed into two main elements: the vertices of the \( v_{ij} \) graph and the edges of the \( a_{ij} \) graph. A graph can be represented as a triad.

\[
G = (V, A)
\]  

Where \( V \in \mathbb{R}N \times \)’s signal matrix is the descriptor of the N edge of each with the property \( f \). \( A \in \mathbb{R}N \times \) Shows the proximity matrix that encodes edge information as in Section I. Each entry \( A \) is defined as follows.

\[
a_{ij} = \begin{cases} 
   d_{ij}, & \text{if there is an edge between } i \text{ and } j \\
   0, & \text{otherwise}
\end{cases}
\]  

Figure 1. One sample of Graph
Figure 2. Flowchart of Proposed Method

A representation of the graph and matrix of vertices V and the adjacency matrix A is shown in Figure 2 and Figure 3.

|     | 0 | 1 | 2 | 3 | 4 |
|-----|---|---|---|---|---|
| 0   | 0 | 1 | 0 | 0 | 1 |
| 1   | 1 | 0 | 1 | 1 | 1 |
| 2   | 1 | 0 | 0 | 1 | 0 |
| 3   | 0 | 1 | 1 | 0 | 1 |
| 4   | 1 | 1 | 0 | 1 | 0 |

Figure 3. Adjacency Matrix of Graph in Fig 2
The algorithm 1 shows an overview of the proposed algorithm for this paper. The general framework of the proposed method consists of two main steps: 1- Designing a learning machine for the forecasting process and 2- Evaluating the accuracy of the machine designed to predict the flow of test information in the data set used. In the next step, a learning machine is trained whose input data is collected from the data collected from the information network graph and its output is labelled "yes" or "no". The output specifies whether the input node will be selected as the next path for information dissemination. The purpose of this machine is to create a regression function that best maps between input data and output labels.

In the second step, the test data will be tested using the test data, which is also selected from the selected data collection set. For the test phase and the accuracy of the operation of this machine are selected for classification operations once randomly and once with the designed machine. Finally, the quality of the selected nodes will be compared using these two methods.

Algorithm 1. Proposed Machine Learning Algorithm

1. Input: A_Data // Adjacency Matrix of Graph, Label
2. ML_Algorithm = {'DecisionTree', 'NaiveBayes', 'K Nearest Neighbors', 'NeuralNetwork'};
3. For Each ML_Algorithm
   4. if (GP.operators == DecisionTree)
      5. tree = ClassificationTree.fit(A_Data, Label);
      6. precision = precision + kfoldLoss(cvtree);
      7. precision_counter = precision_counter + 1;
   8. elseif (GP.operators == NaiveBayes)
      9. NB = fitcnb(A_Data, Label);
     10. precision = precision + kfoldLoss(cvnb);
     11. precision_counter = precision_counter + 1;
   12. elseif (GP.operators == K Nearest Neighbors)
     13. knn = ClassificationKNN.fit(A_Data, Label);
     14. precision = precision + kfoldLoss(cvknn);
     15. precision_counter = precision_counter + 1;
   16. elseif (GP.operators == NeuralNetwork)
     17. knn = ClassificationNN.fit(A_Data, Label);
     18. precision = precision + kfoldLoss(cvknn);
     19. precision_counter = precision_counter + 1;
   20. End
4. End
5. Result = precision / precision_counter;
4. Conclusion

The study shows method of ML in which the passive node is chosen by the active neighbours of it on social networks. Actually it is predicting information propagation path in which active nodes activate inactive nodes. Since predicting new users who are in the flow of information is a diagnostic process, it can therefore be solved by machine learning algorithms. Experimental conclusions present that in this field proposed method is more effective rather than the other methods.

References

[1] Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. 2012. The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web, pp. 519-528.
[2] Granovetter, M. S. 1977. The strength of weak ties Social networks, pp. 347-367.
[3] Hu, Y., Song, R. J., & Chen, M. 2017. Modeling for information diffusion in online social networks via hydrodynamics. IEEE Access, 5, pp. 128-135.
[4] Ikeda, K., Okada, Y., Toriumi, F., Sakaki, T., Kazama, K., Noda, I., ... & Kurihara, S. 2014. Multi-agent information diffusion model for twitter. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE, Vol. 1, pp. 21-26.
[5] Yang, R., Wang, B. H., Ren, J., Bai, W. J., Shi, Z. W., Wang, W. X., & Zhou, T. 2007. Epidemic spreading on heterogeneous networks with identical infectivity. Physics Letters A, 364(3-4), pp. 189-193.
[6] Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., & Montesi, D. 2015. Spreading processes in multilayer networks. IEEE Transactions on Network Science and Engineering, 2(2), pp. 65-83.
[7] Wang, L., & Dai, G. Z. 2008. Global stability of virus spreading in complex heterogeneous networks. SIAM Journal on Applied Mathematics, 68(5), pp. 1495-1502.
[8] Nadini, M., Sun, K., Ubaldi, E., Starnini, M., Rizzo, A., & Perra, N. 2018. Epidemic spreading in modular time-varying networks. Scientific reports, 8(1), pp. 1-11.
[9] Demirel, G., Barter, E., & Gross, T. 2017. Dynamics of epidemic diseases on a growing adaptive network. Scientific reports, 7(1), pp. 1-15.
[10] Serrpezi, P., & Spyropoulos, T. 2013. Information diffusion in heterogeneous networks: The configuration model approach. In 2013 Proceedings IEEE INFOCOM. IEEE, pp. 3261-3266.
[11] Zhou, Y., & Liu, L. 2013. Social influence based clustering of heterogeneous information networks. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 338-346.
[12] Molaei, S., Babaei, S., Salehi, M., & Jalili, M. 2018. Information spread and topic diffusion in heterogeneous information networks. Scientific reports, 8(1), pp. 1-14.
[13] Li, S., & Jin, Z. 2014. Modeling and analysis of new products diffusion on heterogeneous networks. Journal of Applied Mathematics.
[14] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., ... & Zanin, M. 2014. The structure and dynamics of multilayer networks. Physics reports, 544(1), pp. 1-122.
[15] Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv, pp. 1301-3781.
[16] Cheng, W., Greaves, C., & Warren, M. 2006. From n-gram to skipgram to concgram. International journal of corpus linguistics, 11(4), pp. 411-433.
[17] Perozzi, B., Al-Rfou, R., & Skiena, S. 2014. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710.
[18] Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), pp. 1735-1780.
[19] Cao, Q., Shen, H., Cen, K., Ouyang, W., & Cheng, X. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1149-1158.

[20] LeCun, Y., & Bengio, Y. 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10).

[21] Ordóñez, F. J., & Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors, 16(1), pp. 115.

[22] Citation Network Dataset. URL http://konect.uni-koblenz.de/networks/subelj_cora.

[23] Zahra Arbabi et al 2020 J. Phys.: Conf. Ser. 1530 012111.