Diffusion model based on shared friends-aware independent cascade

M K Alasadi ¹ and H N Almamory ²

¹, ² College of Information Technology University of Babylon, Babylon, Iraq

Email: mk3_alasadi@yahoo.com

Abstract. Due to the swift growth of social network and the large variety of its application, numerous research effort aims at predict or explain the information diffusion phenomenon and how the user behaviour changes according to the effect of social pressure. Applications like viral marketing, rumour controlling, and individual behaviour analysis are some areas where these researches are applied. Many models proposed to handle the problem of information diffusion. Such models mostly require disseminated probabilities to be assigned for each link in diffusion network. In this paper we address the problem of predicting the information diffusion cascade. The proposed model named “Shared Friends-Aware IC-Based Diffusion Model (SFA-ICBDM)” can be seen as extension to the well known diffusion model, Independent Cascade model(IC). (SFA-ICBDM) model make use of structural feature of diffusion network specifically the common friend and similarity in terms of behaviour among users as a probability of influencing among users, then predict the cascade of information dissemination. The experiment has been conducted using real world social network dataset (Meme Tracker). In sum, the proposed model has been compared with baseline model IC and related works, where it shows improvement over baseline and related works. In addition, some promising conclusions have been obtained.

Keywords: online social network; Information diffusion; Meme Tracker; predictions; common friends; dissemination probability; hybrid method.

1. Introduction

Recently, the rapid growth of Online Social Networks (OSNs), (e.g. Twitter, Google+, and Microblog) has revolutionized the way we interact with each other. Large variety of communication channel helps users widely share knowledge, ideas and beliefs, thus influencing people behaviours in daily life. The study of human online behaviour, especially with presence of huge amount of data, has drawn the attention of researchers from different disciplines, such as sociology, economy and computer scientists mainly in modelling information diffusion phenomenon [1].

Numerous studies have been conducted for answering questions like: which factor effect the diffusion process, how can accelerate spreading of information, and what direction may diffusion take, etc. All these queries and other can be answered via modelling the phenomena of information diffusion and all its related aspect in OSNs; therefore it has become an important research problem [2].

The research effort in this field can be seen as a solution for a lot of real life problems. Hence, modelling how influence disseminate in OSNs lies for helping people to better understand the
diffusion process of information and then help in better optimizing business performance (e.g., finding the most influential users), monitoring revolutionary waves, locating the source of inauthentic news and rumours, etc. [1] [3].

Literature associated with social network analysis introduces variety of diffusion models arising from the sociology communities and economics.

In this scope, two main classes of models for information diffusion can be distinguished, i.e., explanatory models and predictive models [4]. Other categories can be noticed such as graph and non-graph models depending on the nature of interaction among the social network entities.

The explanatory models such as SIS (Susceptible Infected Susceptible) model and SIR (Susceptible Infected Recovered) model infer the underlying spreading cascade, observing an order set of nodes that accept the information (infected users). Using that class of models one can answer question like, who are the most influential users, which factors play an important role in diffusion process and why does the information diffuse in the way it take. [5]

On the other hand, the predictive models predict the pattern of diffusion in OSNs by learning from previous diffusion processes, such models are used to predict the future information diffusion process in social networks based on certain factors. As example of events in which that model can be used, is when a government want to know how a situation will develop in case of revolutionary or other serious cases [4]. IC (Independent Cascade) model [6] and LT(Linear Threshold) model [7][8], are two of the well known predictive models and have been widely used for modelling information diffusion over OSNs[9].

Game Theory Model (GTM) is another approach which has been used for prediction purpose. This strategy consider number of individuals or groups with specific limitations and aim to maximize the profit. In this model a piece of information may or may not spread depending on factors like costs, benefits, and strategic choice.[4][10]

In this work, we have greatly improved the initial version of IC model, and proposing new technique to precisely and accurately predict the diffusion process. We applied the model on real world data set and the evaluation method we use indicate a great improvement in the prediction result.

2. Related Works

Recently, a lot of efforts have been made to study the information diffusion phenomenon over online social network. And due to its important; plenty of research efforts focused on developing predictive models for information propagation [11][12].

Zhu et al. studied the problem of identifying the main paths of diffusion in online social networks. Where the proposed method consists of three steps: measurement of tie strength, identification of influential users, and identification of main paths. By calculates the weights of links based on historical interaction frequency between pairs of nodes and adopt the influential node ranking algorithm (Leader Rank algorithm) to identify the influential nodes. In other words, they consider the main paths connecting influential nodes are relatively important paths by which the information diffusion is promoted. In general, their main concern is finding the most important paths to avoid trace the information in a entire network. [13]

Lagnier et al. propose a new family of probabilistic models to predict how content diffuses in a network. In fact, the content of the piece of information diffused, user’s profile and willingness to diffuse a given piece of information, are additional dimensions used in their work.
In User-Based Approach the thematic interest of each user in the content diffused can be modelled as a proximity between user profiles (describing their interests) and the content diffused. Cosine similarity has been used to identify a similarity between content diffused and user profile. In essence, the probabilistic model allow each user has a certain probability of diffuse the given content. When the thematic interest of the user is high, or when willingness to diffuse or social pressure is high, $P(n^i, c^k, t)$ the probability that user $n^i$ diffuses content $c^k$ at time $t$, should be high [14].

Another related research area study the volume to which information could be propagate over social media. In [15], they propose an information diffusion model to predict whether a post is going to be forwarded or not (if a tweet is going to be re-tweeted). Moreover, they aim at predicting how much it is going to be diffused (how popular it will be). In detail, their model based on machine learning techniques and make use of three types of features: user-based, time-based and content-based to represent tweets and perform prediction process. In fact, there are a total of 29 features, like: Number of people who follow the user (user-based), whether the tweet is created on public holiday (time-based) and the tweet contains a location name(content-based). They conclude that some introduced features are of highest importance in prediction retweet ability like the numbers of followers and number of communities a user belongs to.

In [16] the authors seek to answer the following question: Given information initiated from a seed node(s), what is the fraction of user that could be influenced by (adopt) that piece of information after a period of time?

In like manner of the work that presented by [15], however, they suggest to use different technique to solve the research problem. Note here the study did not pay attention to answer question like 'how influence home in social network', in other words their target was not to predict the information cascade which identify the direction of dissemination and the source and distension of influence at each time step. The question which we focuses our effort answering it in this paper.

In their work, non-linear differential equations have been used to study the temporal patterns of information diffusion process in an online social platform on the assumption that the social carrying capacity is dynamic. By minimizing the Mean Absolute Error (MAE) and make use of Genetic Algorithm with random initial guess for the minimization, they conclude to find the volume of diffused information.

In article [17] the author use a crawler topic based sampling approach for acquiring twitter data set related to specific event during short duration of time. Their prediction was manly base on the linguistic and emotional features of the tweet as well as the user communication behaviour that created the initial tweet.

For a given subject they examine the frequencies of all possible diffusions and extract some simple and basic patterns of diffusion. Then, by using linguistic analysis and machine learning techniques they associate the content of tweets with the propagation patterns. they argue that the content of the message and the type of network diffusion are the two main dimensions that lead to robust cascade prediction.

Bourigault et al. embed users of social network in a continuous latent space in which representation learning techniques was proposed, where the relative positions of users in latent space were used to define their content transmission likelihood rather than those defined by classical graphical learning approaches on a discrete structure.

The hypothesis was the following: influence is binary, the diffusion network is unobserved, influence relationships do not depend on the disseminated information as well as the infection probabilities between users do not vary in time [18].
3. Modelling Diffusion through Networks

3.1. Independent Cascade model (IC)

The social networks are generally defined as a graph $G = (V,E)$ where $V$ is the set of nodes or users and $E$ is the set of links (edges). Each link is denoted by $e = (v,w) \in E$ and $v \neq w$, which refers to the existence of a directed link from node $v$ to node $w$.

The IC model [Talk of the network, Information diffusion through Blogspace, Maximizing the spread of influence through a social network] is mainly used for prediction and influence research. In the IC model, a real value $P_{v,w}$ should be assigned in advance for each directed edge $e$. Where $P_{v,w}$ represents the diffusion probability through the edge $(v,w)$ and $0 \leq P_{v,w} \leq 1$.

The diffusion process begins from a given initial active set $A(0)$ in the following way. Suppose a node $v \in A(t)$, firstly becomes active at time-step $t$, it has only one chance to activate (infect) all its inactive child node $w$ with probability $P_{v,w}$. Subsequently, $w$ becomes active at time-step $t + 1$, i.e., $w \in A(t + 1)$. If multiple active nodes at time-step $t$ try to activate the same child node, then their activation attempts are sequenced in an arbitrary order at time-step $t$. Generally, in all next steps node $v$ will never try to activate $w$, no matter whether or not $v$ succeeds in previous time step. The diffusion process stops if no more activation is possible.

In our work, we adapt this model to perform the prediction process. Our model combines a spiritual aspect of the IC model and the community of friends.

3.2. Proposed Method

Numerous methods have been proposed to determine the probability $P_{v,w}$. Where it can be assigned based on frequency of interactions, geographic proximity, or historical infection traces, and many other techniques [19][20].

In this paper, we intended to use different scenarios to learn the diffusion probability. As the IC model assumes that the diffusion process mainly depends on the edge probability, from this point of view we pay much attention to find a new learning method that accurately finds these probabilities.

We argue that using feature extracted from the structure of the network which represents the asocial graph. In fact, motivation of using this feature is that friends affect each other. Furthermore, user behaviour in social media can be used as the other feature.

3.2.1. Structural – Based Feature.

Many previous studies conclude that similar users tend to largely influence each other, and other work states that such users are more likely to be part of the same community, which is densely connected to each other and loosely connected to other parts of the network[21][22]. Therefore and instead of relying mainly on the frequency of interaction between users to model the diffusion probabilities and starting from the facts mentioned above, we think to use the structural information and believe that the number of common friends between two nodes (user) in the network will be of great usefulness, if considered as an indicator to the level to which one node can influence another node that is directly connected with it.

The suggested method is using the relation available between nodes, where for each directly connected pair of nodes $(v$ and $w)$ with link goes from $v$ to $w$, one can calculate the tie strength (influence probability) as follow:
Let $S(v)$ and $S(w)$ are two disjoint sets containing all the neighbours of node $v$ and $w$ respectively. By considering the size of intersection to size of union (Jaccard similarity) of these two sets the probability value will be $0 \leq I(v, w) \leq 1$ as in equation (1).

$$I(v, w) = \frac{S(v) \cdot S(w)}{S(v) + S(w)}$$

3.2.2. **User Behaviour – Based Feature.**

"Birds of the same feather flock together" is an ecological phenomenon which drowns the attention of large number of social network analysts. From this point of view and based on the fact which state that 'influence between similar people occurs at a higher rate than among dissimilar people'. We use in this paper the similarity among users based on their social behaviour, as an indicator to the strength at which users can affect each other activates. The user profile extraction is the module where the user’s activity and the content diffused in social media are analysed during fixed time period in the past.

Depending on the nature of social network and available dataset, user profile can be constricted representing their interest in the content diffused so far in social network. These profiles can be seen as vectors of the same length for all users, defined on the same feature space. Entries of such vector represent a specific topic trend or field of interest that can be detected by analysing the diffused contents. The values in each profile (vector) entry reflect the tendency of a user to publish in this topic or trend of information.

We proposed two methods for modelling user interest and user-user profile similarity, one can use either of them based on application under hand and relative dataset.

In first method user interest represented by binary value (0 and 1), where the vector entry carry 1 if specific user have ever participate in that topic trend or participate for number of times exceed specified threshold, otherwise it carry 0.

In this case users profile similarity can be calculated using Jaccard Coefficient (asymmetric attributes) as illustrated in equation (2).

$$J(v, w) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Where for example, $M_{11}$ is the number of attributes where $(v)$ was 1 and $(w)$ was 1, and $M_{01}$ is the number of attributes where $(v)$ was 0 and $(w)$ was 1, and so on.

The second method assumes that, user interest in each topic or trend of diffused information is described by integer value. This value is nothing else the number of times the user has diffused such information in the period of data collection process.

In contrast to first method, we propose to use pearson correlation equation (3) as a metric to measure similarity among user profile which represented by real values.

$$P(v, w) = \frac{COV(v, w)}{\sigma_v \cdot \sigma_w}$$

Where $P(v, w)$ is a pearson correlation value, $v$ and $w$ represent two user profiles. COV is the covariance and $(\sigma_v, \sigma_w)$ are standard deviation for $v$ and $w$ respectively.

3.2.3. **Shared Friends-Aware IC- Based Diffusion Model (SFA-ICBDM).**

The integration of probabilities that constructed based on structural feature equation (1), along with probabilities obtained by analysing user interest equation (2) or (3) can be used to predict precisely the information diffusion cascade as follow:

$$H(v, w) = \lambda_1 I(v, w) + \lambda_2 P(v, w)$$

The $\lambda_1$ and $\lambda_2$ are the importance given for each parameter (T & P).
4. Experiment and Evaluation

We conduct our experiment using Memetracker dataset to predict the information diffusion cascade, and for evaluation purpose we use the median time difference values between content mentions by different users in dataset.

4.1. Dataset Description

The MemeTracker dataset described in [23] contains millions of blog posts and news articles. Each website or blogger stands as a user. The data constricted by analysing around 900,000 news stories and blog posts per day from 1 million online sources, ranging from mass media to personal blogs. There are two version of this data set. We make use both of them:

i. Raw MemeTracker phrase data
Data contains phrases and hyper-links extracted from each article/blog post. For each article, the timestamp, phrases and hyper-links are extracted.

ii. MemeTracker phrase cluster data
Data contains phrase clusters. For each phrase cluster the data contains all the phrases in the cluster and a list of URLs where the phrases appeared. extracted by [18].

4.2. Define diffusion probabilities

From the information available in dataset and by using the technique illustrated in [dynamic of] we construct diffusion network and for each edge in the network we find four different diffusion probabilities.

Firstly, from the structural information extracted from network (fraction of similar common friends) and for each edge \( e(v,w) \), we find the edge strength \( T(v,w) \) by using equation (1).

Second we exploit the behavioral information provided by MemeTracker phrase cluster dataset, we calculate the user profile which represented as a vector of the length equal to the number of the most largest clusters (in our case is 5000). By applying the equation (2) and (3) we uncover the second and third type of probability (\( P(v,w) \)). Finally we use (eq. 4) to find the proposed hybrid probability\( (H(v,w)) \) for all network edges.

4.3. Results and Discussion

In our experiment and by using diffusion probability defined by equation 4), we have made several trail to find optimal values for \( \lambda_1 \) and \( \lambda_2 \) which represent the importance of the fraction of common friend parameter and profile similarity parameter respectively. We find that \( \lambda_1=0.3 \) and \( \lambda_2=0.7 \) give significant improvement in term of F-measure score. In figure (1) we have compared the result obtained from these values versus assigning same value of importance \( (\lambda_1 =\lambda_2=0.5) \) for both parameters.
We have separately applied independent cascade (IC) model on the previously mentioned probability values by considering in every cascade, different nodes as a seed nodes that information diffusion process start with.

We find the average F-measure over all the training cascades for each probability schema. The experiment show that the F-measure of result obtained from IC model using hybrid-base diffusion probabilities (SFA-ICBDM) was in general greater than the F-measure scored by using other probabilities values. Table (1) details the average precision and average F-measure observed by conducting the experiment based on different baseline diffusion probabilities.

### Table 1. Comparison of proposed method with different baseline techniques.

| Approach                  | Precision | F-measure |
|---------------------------|-----------|-----------|
| IC model                  | 0.331     | 0.363     |
| User behavior based features | 0.776     | 0.758     |
| Shared friend based features | 0.776     | 0.571     |
| SFA-ICBDM                 | **0.781** | **0.805** |

In table (2) we shows the performance of the proposed model versus related works.

### Table 2. Precision values on MemeTracker datasets.

| Cascade length | 1          | 2          | 3          | >= 4         |
|----------------|------------|------------|------------|-------------|
| RUC            | 0.63       | 0.50       | 0.63       | 0.67        |
| DRUC           | 0.63       | 0.50       | 0.62       | 0.68        |
| Our Method (Hybrid Based) | 0.830    | 0.731      | 0.757      | 0.781       |

RUC (Reinforced User-Centric) and DRUC (Decaying Reinforced User-Centric) are two related modules proposed in [14].
As detailed above our method yield significant improvement in term of precision where it is about 16 % over both DRUC and RUC models.

We show in figure (2), the values of precision, recall and the F- measure metrics obtained by Applying the experiment on MemeTracker datasets using Shared Friends-Aware IC- Based Diffusion Model.

![Figure 2. Average Precision, Recall and the F- measure metrics for different influence threshold.](image)

5. Conclusion

In this study, we addressed the problem of predicting information diffusion process by observing the user behaviour in online social network along with information diffused throw this network. The key point here is to efficiently learn the diffusion probabilities associated with all the edges running from one node to other node in social network. We use for this purpose a new hybrid method, which make use of both user profile and properties of network structure. The experiment on real world social network show that, exploiting the similarity between the behaviours of two users along with ratio of common friends of these users, give notable accuracy in prediction the cascades of information dissemination.

We have compared the performance of diffusion model based on proposed learning method with several baseline as well as result from related work. The comparison outcome confirmed that our proposed technique highly improved the result of all other method in term of F-measure.

6. References

[1 ] Xu X, Chen X and Eun D 2015 IEEE Modelling time-sensitive information diffusion in online social networks ( Hong Kong, China) Computer Communications Workshops. pp 408-413
[2 ] Easley D and Kleinberg J 2010 Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Ch.19 Cascading Behavior in Networks(Cambridge University Press) pp 573-619
[3 ] Hu Y, Song R and Chen M 2017 IEEE Modeling for information diffusion in online social networks via hydrodynamics Vol.5 pp 128 - 135
[4 ] Li M , Wang X, Gao K and Zhang S 2017 A survey on information diffusion in online social networks: models and methods. Vol. 8
[5] Guille A, Hacid H, Favre C and Zighed D 2013 ACM Information diffusion in online social networks: a survey Vol.42

[6] Goldenberg J, Libai B and Muller E 2001 Talk of the network: a complex systems look at the underlying process of word-of-mouth Vol.12, pp 211-223

[7] Lang J and Sterck H 2016 J.of Comp. Net. IEEE. A hierarchy of linear threshold models for the spread of political revolutions on social networks Vol.4, pp 426 – 456

[8] Michalski R 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) Linear threshold model in temporal networks — seed selection for social influence pp 922 – 923

[9] Zhang H, Mishra S and Thai M 2014. Opportunistic Mobile Social Networks Recent. Chapter: Advances in information diffusion and influence maximization of complex social networks. pp 37-69

[10] Zinoviev D and Duong V 2011 IGI Global. Game theoretical approach to broadcast information diffusion in Social Networks pp.146-163

[11] Chang B, Xu T, Liu Q and Chen E 2018 Inte. J. of Aut. and Com. Springer. Study on information diffusion analysis in social networks and its applications. Vol.15. pp 377-401

[12] Zhu Y, Li D, Guo H and Pamula R. 2014 10th International Conference on Mobile Ad-hoc and Sensor Networks IEEE(USA). New competitive influence propagation models in social networks

[13] Zhu H, Yin X, Ma J and Hu W 2016 Elsevier. Identifying the main paths of information diffusion in online social networks. Vol.452, pp 320-328

[14] Lagnier C, Denoyer L, Gaussier E and Gallinari P 2013 HAL. Predicting information diffusion in social networks using content and user’s profiles. book series LNCS. Vol.7814. pp.74-85

[15] Bich T, Hoang N and Mothe J2018 Elsevier Predicting information diffusion on twitter – analysis of predictive features. Vol 28, pp 257-264

[16] Davoudi A and Chatterjee M 2016 IEEE International Conference on Big Data. Prediction of information diffusion in social networks using dynamic carrying capacity

[17] Kafeza E, Kanavos A, Makris C and Vikatos P 2014 Springer. International Conference on Artificial Intelligence Applications and Innovations. Predicting information diffusion patterns in twitter. book series vol .436 pp 79-89

[18] Bourigault S, Lamprier S and Gallinari P 2016 WSDM. International Conference on Web Search and Data Mining. Representation learning for information diffusion through social networks: an Embedded Cascade Model. pp 573-582

[19] Phuke P and Rathod S 2017 DLINE. Journal of Networking Technology. Learning Influence Probabilities in Social Networks Vol. 8

[20] Goyal A, Bonchi F and Lakshmanan L 2010 ACM international conference on Web search and data mining. Learning influence probabilities in social networks pp 241-250

[21] Xie J, Kelley S and Szymanski B 2013 ACM Overlapping community detection in networks: the state-of-the-art and comparative study. Vol. 45

[22] Riaz F and Ali K 2011 IEEE Applications of graph theory in computer science. Vol. 2. Pp. 4610-4621

[23] Leskovec J Backstrom L and Kleinberg J 2009 Meme-tracking and the dynamics of the news cycle. pp 497-506