LEARNING MULTIPLE FACTORS-AWARE DIFFUSION MODELS IN SOCIAL NETWORKS

Dr. L. Prasanna Kumar
Associate Professor, Department of CSE, DIET, Dadi Institute of Engineering & Technology
KONATALA LOKESH M.Tech student , Department of CSE,DIET, Dadi Institute of Engineering & Technology, lokeshnist@gmail.com

ABSTRACT

Information diffusion is a natural phenomenon occurring in social networks. The adoption behavior of a node toward an information piece in a social network can be affected by different factors, e.g. freshness and hotness. Previously, many diffusion models are proposed to consider one or several fixed factors. In fact, the factors affecting adoption decision of a node are different from one to another and may not be seen before. For a different scenario of diffusion with new factors, previous diffusion models may not model the diffusion well, or are not applicable at all. Moreover, uncertainty of information exposure intrinsically exists between two connected nodes, which causes modeling diffusion more challenge in social networks. In this work, our aim is to design a diffusion model in which factors considered are flexible to be extended and changed and the uncertainly of information exposure is explicitly tackled. Therefore, with different factors, our diffusion model can be adapted to more scenarios of diffusion without requiring the modification of the learning framework. We conduct comprehensive experiments to show that our diffusion model is effective on two important tasks of information diffusion, namely activation prediction and spread estimation.

Key Words: Social Networks, Diffusion Technique,

1. INTRODUCTION

For more than a decade, researchers have been studying information transmission in social networks. The spread of information from node to node through a social network is a natural phenomena that operates like an epidemic. Information diffusion has a wide range of uses, including more successfully marketing an idea or a product, preventing negative attitudes, identifying important nodes, and tracking information flows of subjects in a social network. Influence maximisation, suggested by Domingos and Richardson for viral marketing using word-of-mouth effects and subsequently stated as a discrete optimization problem, is a well-known issue in this field. On a particular diffusion model that simulates how information diffuses in a social network, the goal of impact maximisation is to locate a set of target nodes to be persuaded of a concept initially in order to maximise the spread size, i.e. the number of nodes accepting the idea after propagation.

A basic and significant question in information dispersion is how to represent an information item propagating in a social network. Different diffusion models have emerged as a result of various factors. The Independent Cascade (IC) model and the Linear Threshold (LT) model are two diffusion models that have been extensively used in numerous applications and have many modifications. Both models take into account the strength of a neighbor's effect. The major distinction is that IC analyses just one activated neighbor's effect with a predetermined chance for a node turning to embrace an idea, while LT incorporates the joint influence contribution from all active neighbours. Nonetheless, the actual world is so complex that it's difficult to represent it in a basic issue. Adoption decisions are likely influenced by a number of variables. For example, a concept that has been embraced by the majority of people is more likely to influence someone, but an idea that has been adopted by a smaller number of people is less likely to influence someone as time passes. Furthermore, a person's level of interest in various things may vary. Many diffusion models in social networks, on the other hand, address all of the same elements. Previous diffusion models may not accurately describe the dissemination of new components in a different setting, or they may not be relevant at all. As a result, it is common to propose a new diffusion model to better explain the dissemination of a certain situation by taking into account new parameters.

In addition to the various elements that influence adoption decisions, uncertainty of information exposure is a real-world phenomenon that should be taken into account when modelling dissemination in social networks. For example, assuming that a person in an Online Social Network (OSN) would read what her neighbours say may easily overestimate the spread size for impact maximisation. Uncertainty may be caused by a number of factors:

- In the age of information explosion, users generally have limited time and attention to deal with a flood of data.
- Different OSNs have different ranking systems, such as Facebook's news feed ranking1. To put it another way, OSNs change what we can see the most.
- A user may enjoy an information item offered by a neighbour, but she takes no specific actions, such as like, replying, or sharing on Facebook. A user who supports a political party, for example, may choose not to discuss political subjects in OSNs.
- Owing to limitations in the data collection process, such as the sample rate of Twitter Streaming API2, or because only partial data is collected for training due to restricted computing resources, only partial data is available for training models.

Designing a diffusion model for each element of adoption choice and proposing the relevant techniques, such as parameter learning and impact maximisation algorithms, may be time consuming. In this paper, we attempt to create a diffusion model that can take into account a variety of characteristics and explicitly predicts the uncertainty of information exposure between two linked nodes. A Multiple-Factors Aware Diffusion (MFAD) model, which is a two-stage propagation model, is proposed. An activated node u seeks to influence its inactive neighbour v probabilistically in the first step, known as influence transmission. If u's influence is effectively conveyed to v, the second step, known as adoption decision, occurs, in which v determines whether it will be activated based on its considerations, which are anticipated using a classification model trained on previous adoption data with many parameters.

The pattern of adoptions may be captured by learning parameters of a diffusion model from data, and the diffusion model can then be used to real-world tasks such as activation prediction and spread estimate. The goal of activation prediction is to anticipate whether a node will adopt an information piece based on the adoption of one of its neighbours. Spread estimation is an essential function for impact maximisation since it estimates the spread of an information piece in the network given a set of nodes that first adopted the piece. Two situations are examined in the paper for learning using real-world diffusion data. The data including source information is the first scenario. In Twitter-like social networks, for example, a user may submit a message, known as a tweet, and repost another user's tweet. A retweet is a tweet that has been republished from another user. Not only does a retweet identify the user (follower) and the time of the post, but it also identifies the source of the rebroadcast (followee). The third case is for data in which the source is neither visible or available. In other words, we have no way of knowing who prompted an adoption. This kind of dataset may be partial data from scenario one, such as Twitter-like datasets with missing retweet origins, or data that is inherently lacking in source information, such as Flixster3, an online social network for movie ratings and reviews, or Digg4, a social website for news sharing. As a result, we provide two learning frameworks for diverse types of diffusion data in two situations in this paper. The following is a list of our contributions.

1) Because the suggested learning frameworks are independent of the variables studied, our MFAD model is adaptable in terms of extending and changing factors. This is done by using a classification strategy to anticipate a node's adoption behaviour. Furthermore, the frameworks aren't constrained by a single categorization method.
2) It is difficult to anticipate adoption behaviours due to the limitations of observation on dissemination in the actual world. In the first situation, we directly address this problem by learning nodes' classifiers for adoption choice using only positive and unlabeled examples, and in the second scenario, we use an Expectation-Maximization technique.
3) Experiments demonstrate that MFAD is successful at activation prediction and spread estimation when the uncertainty of information exposure across nodes is explicitly modelled. MFAD is superior in terms of spread estimate, in particular. The quality refers to how many nodes in a diffusion model's anticipated spread actually accept an item after propagation. Previous research has focused on the quantity component, such as spread size, while disregarding the quality aspect, resulting in estimated solution outcomes that differ significantly from genuine real-world diffusion results.

The rest of the paper is laid out as follows. In Section 2, we go through the relevant tasks. In Sections 3 and 4, we explain how to learn the MFAD model for the first case, and in Sections 5 and 6, we explain how to learn the model for the second scenario. In Section 6, we discuss how to apply the model for activation prediction and spread estimates once it has been learnt. In Section 7, we assess the performance of MFAD by conducting tests on both synthetic and actual data. Finally, in Section 8, we come to a close.

Literature survey

With the use of social influence and product attributes, we construct an improved iterative scaling approach to figure out the parameters that would optimise the probability of our unique feature-aware propagation model. It is important to understand the propagation model in order to create an effective algorithm. If the product's characteristics are fixed and we look for the set of consumers to target in the viral marketing campaign, the anticipated spread, i.e., the objective function that should be maximised, is monotone and submodular, as we demonstrate in our paper. Instead, the predicted distribution is neither submodular nor monotone when we fix the set of consumers and seek to identify the best characteristics for the product (as it is the case, in general, for product design). As a result, we create an algorithm that alternates between optimising the product's characteristics and picking the audience for the campaign. According on the results of our real-world experiments using LastFM and Flixster data, the paradigm we've provided for merging product design with viral marketing is indeed beneficial [1].

Topic modeling is how we look at social impact. New topic-aware influence-driven propagation models that we demonstrate in our tests are more accurate than the usual (i.e., topic-blind) propagation models investigated in the literature are introduced. The first topic-aware adaptations of well-known Independent Cascade and Linear Threshold models are proposed here. However, the sheer number of parameters in these propagation models raises the possibility of overfitting [2].

We believe that product adoption and impact are distinct concepts. We provide a model that takes into account a user's actual (or anticipated) experience with a product. By creating an objective function that clearly measures product adoption rather than impact, we modify the traditional Linear Threshold (LT) propagation model for our purposes. An approximation approach is possible because the NP-hard adoption maximisation problem is monotone and submodular under our paradigm. On three major social networks, we do tests that demonstrate our model can discriminate between influence and adoption and reliably forecast product uptake much better than the traditional Long-Term Influence (LTI) model. [3].

The greedy technique is put to the test against a variety of heuristics. As shown in the trials, degree centrality and other low-cost heuristics outperform the greedy strategy in most circumstances. When there is a lack of data, we also look at the issue of impact restriction, where the present states of nodes in the network can only be predicted with a limited degree of certainty. We present a random spanning tree-based prediction system and analyze its effectiveness. As a result of these trials, we found that the prediction algorithm can tolerate up to 90% missing data before the system's performance begins to decrease, and even with a high quantity of missing data, it can still perform as well as with full data [4].

Recent research on the spread of information through social networks has mostly focused on models based on the effect of local friends. To simulate user behaviour, social scientists have used ideas about diffusion of innovations to test the generalizability of this technique. To this goal, we investigate alternative diffusion models in Digg and Twitter, two popular online social networks.. Two sample local impact models are evaluated first, and we demonstrate that most social network users' behaviours are not reflected by these local models. Using ideas from the diffusion of innovations study we offer a new diffusion model called Gaussian Logit Curve Model (GLCM) that simulates user behaviour in relation to the general population. In particular for Digg, our research reveals that GLCM better captures user behaviour than local models. We develop a variety of hybrid models and use statistical approaches to assess their effectiveness in capturing both local and global data. The complexity of human behaviour is captured by our technique, which analyses each user individually, automatically deciding which users are motivated by their local relationships and which individuals are best described by adopter categories [5].

We provide two improvements to this approach in this paper. PNB has been further developed by presenting an approach to building more complicated Bayesian classifiers without the use of negative examples. Our PTAN algorithm (positive tree augmented naive Bayes, or PTAN) is a novel method for constructing tree-augmented Bayes models in the positive unlabeled domain. With regard to dealing with the prior probability of being in the positive class, we suggest a novel Bayesian strategy based on a Beta distribution to explain the uncertainty of this parameter. As a consequence, two new algorithms, PNB and PTAN, are created using this strategy. Real and synthetic datasets are used to test the effectiveness of the four methods. Extending PNB to larger complicated networks improves classification accuracy in cases where predicting variables are not conditionally independent given a particular class. Using our Bayesian method to the prior probability of the positive class, they demonstrate that our PNB and PTAN findings may be improved [6].

The heuristic technique we provide in this study can be readily scaled to millions of nodes and edges in our tests, and this is what we demonstrate in this work. A simple configurable parameter in our method allows users to fine-tune the system's execution duration and effect spread. On numerous real-world and synthetic networks, our thorough simulations show that our approach is presently the most scalable solution to the challenge of maximising influence: A) Our algorithm can handle graphs with millions of nodes, and (b) in all size ranges, it outperforms other heuristics by as much as 100 percent to 260 percent in terms of influence spread — it is always among the best algorithms, and in most cases it significantly outperforms all other scalable algorithms [7].

According to Goffman and Newill1, there is a parallel between the transmission of infectious illness and the dissemination of knowledge. Using mathematical epidemiology, we recently explored how rumours spread and have a few quick things to say about it before it's detailed elsewhere2. A mathematical model for rumour spreading may be developed in a variety ways, depending on the mechanism proposed to explain the development and decay of rumour spreading itself. Models like this one may be used to study epidemics, however the findings may not be as predicted because of the formal resemblance between epidemics and mathematical epidemiology procedures [8].

A customer's network value should be modelled as well, including her potential effect on the sales of other customers she may inspire to purchase, as well as the consumers those customers influence. Markov random fields are used to describe the market as a social network rather than a collection of isolated entities. Using a collaborative filtering database, we demonstrate the benefits of this technique. As a marketing strategy, viral marketing may be incredibly powerful yet remains a mystery to many. It may be seen as a step toward establishing a more solid basis for it, taking use of the enormous relevant datasets that are available [9].

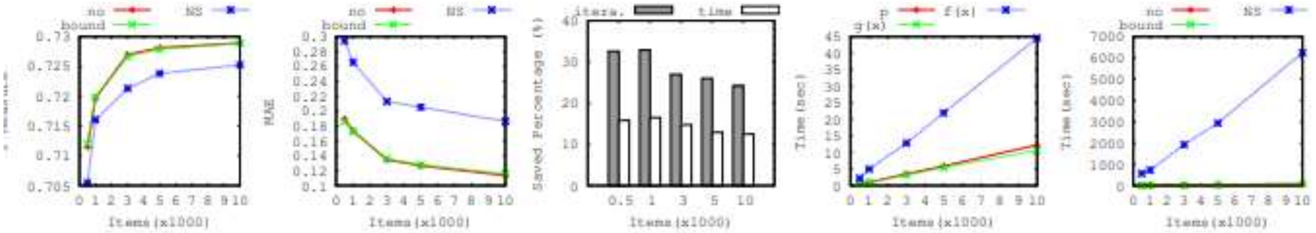
Cascades are the only method we have to expose the hidden network. Due to the diverse nature of networks, a simple parametric model cannot adequately capture the structure finding challenge. As a result, we have developed a kernel-based technique that can capture a wide variety of influences without making any previous assumptions. Using both synthetic and actual data, we demonstrate that our approach can better reconstruct the underlying diffusion network and greatly enhance the estimate of transmission functions across networked entities [10].

Result

For the purposes of activation prediction and spread estimation, we run a series of experiments on both synthetic and actual data in this section. In this section, we begin by describing the setup.

Results on Synthetic Datasets

To test the performance of MFAD, we utilize synthetic datasets to analyze the two situations of learning with source information (WS) and learning with no source information (NS). Figure 1 shows the findings of MFAD and NS on five synthetic datasets with T = 10 for predicting WS transmission probability and activation prediction. For learning transmission probabilities, the series "no" was generated from learning with the derived limits in Propositions 1, 2, and 3 whereas the series "bound" was derived from learning with the native bounds in WS, namely 0 and 1. Both the F-Measure and the MAE can be shown in Figures 1(a) and 1(b) to be quite close to each other. Since learning with derived boundaries requires less iteration and runs more quickly in Figure 1(c), this is the default setting for WS in our test. Notably, the amount of time saved by employing two derived limits isn't nearly as substantial as the amount of iterations saved, since bounds need additional calculations, particularly for the upper bound in Proposition 2's estimation of |N (D q,s)|. The time spent learning each MFAD component for WS is also shown in Fig 1. (d). When using positive and unlabeled learning [11], it takes the longest to learn the classic classifier fv(x) for each node v because the positive training examples include extra copies from weighted unlabeled instances, as explained in Section 4.1.2. In Fig 1(a), 1(b), and 1(c), the performance of NS is lower than that of WS because of the incompleteness of data (e). To further understand time frame T.



(a) F-Measure

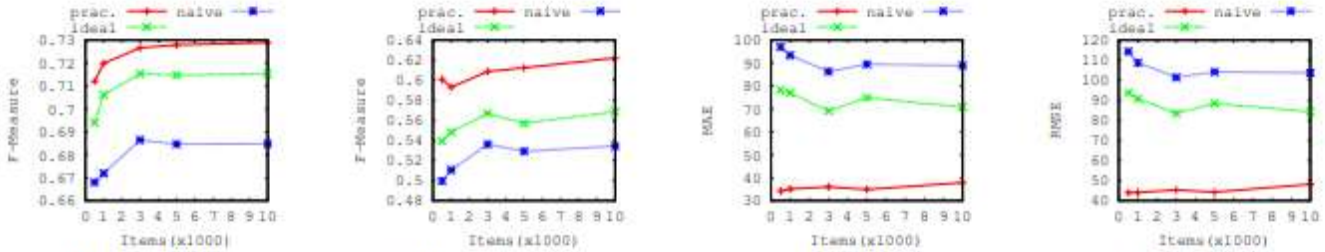
(b) MAE of transmission probabilities

(c) Saved percentage of average iterations and running time for estimating all $p_{u,v}$

(d) Running time of three components of MFAD (bound)

(e) Total running time

Fig. 1. MFAD activation prediction for WS and NS on five synthetic datasets (T = 10). (NS=learning without source information; WS=learning with source information)



(a) F-Measure (AP)

(b) F-Measure (SE)

(c) MAE (SE)

(d) RMSE (SE)

Fig. 2. Three AP and SE decision boundaries (100 runs) using MFAD (T = 10) source data.

In order to accurately estimate the spread, a good diffusion model must meet two requirements.

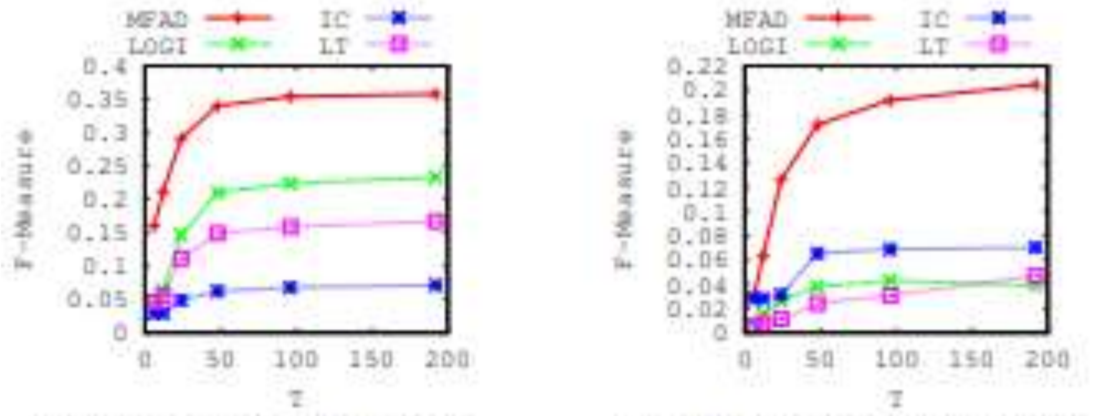
1) There is a component of quantity. Each item's projected spread size is quite similar to the genuine spread size as determined by MAE and RMSE.
2) Secondly, Quality. Each node predicted by a diffusion model is active in the genuine spread, which is determined by F-Measure.

Only the spread size is taken into account in the first condition, which is often employed in earlier research. For influence maximization issues, the number does not indicate whether a node adopts an item in the projected spread. A diffusion model can't forecast how many people will use an item based on how far it spreads. To maximize the spread of their product, many companies focus only on the quality of their product, which might result in a significant loss of income. When using a diffusion model, it's important to take into account both the first and second conditions.

For activation prediction in Fig 2(a) and spread estimate in Fig 2(b) through 3, MFAD with the practical decision boundary has the best results (d). Better than the other two WS setups, especially in terms of estimating spread. Figure 8 in Appendix C shows comparable findings for NS activation prediction and MAE estimate. In this way, the experiment serves as a decision-making guideline for both situations of MFAD.

Results on Real Datasets

F-measure activation prediction findings on Weibo datasets are given in Fig 3(a) and 3(b) for two situations (b). The activation source is known in Fig. 2a, where models are trained using data. It's not possible to train a model in Fig. 3(b) since there's no way to tell where the data came from. In all datasets, IC and LT's prediction results are low because real-world adoption variables are complex, but the factors addressed by IC and LT are simplistic. F-Measure improves for all models as the number of positive unlabeled cases decreases, allowing for a larger number of positive examples to be used for training. Fig. 3(b) shows that when T rises, F-Measure typically improves as more potential influences are taken into account. For the two scenarios, Appendix C provides or analyses the training time for all models, the training size for WS, the convergence analysis of EM for MFAD, initial-value strategies for learning transmission probabilities, single-stage improvement of MFAD, and another scalable method [15] for training IC without source information.



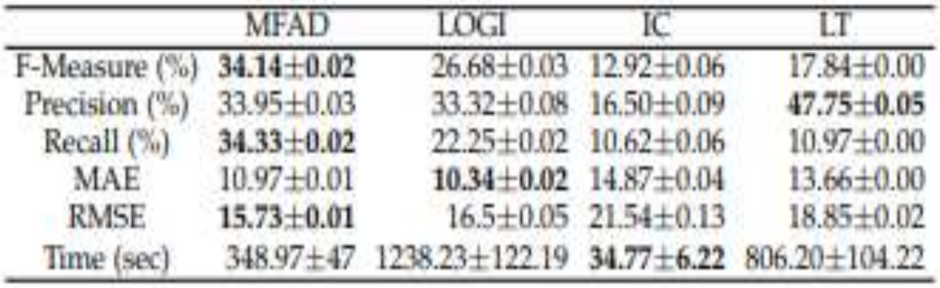
(a) With source information

(b) Without source information

Fig. 3. Activation prediction in Weibo for two learning scenarios with different observation time window T. (Models in 4(b) are trained with only 1 EM due to very long running time.)

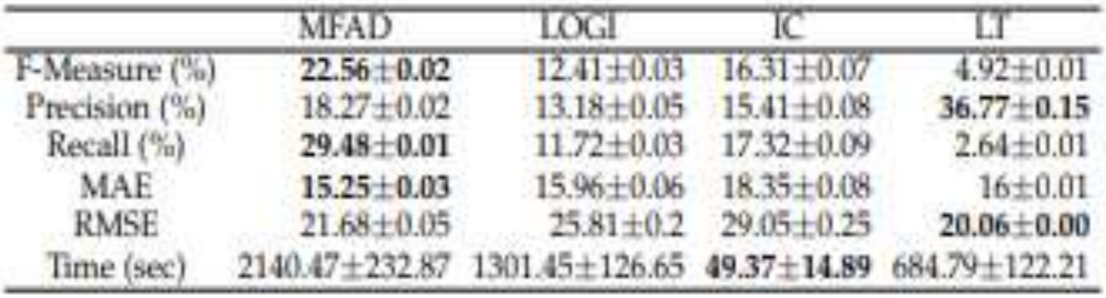
Tables 1 and 2 describe the findings of the Weibo spread estimate. T = 192 was used to train the models shown in Table 1. Models trained using 10 EMs and 192 T are shown in Table 2. When it comes to F-Measure predictions, MFAD is substantially better than LOGI, IC, and LTI. Tables 1 and 2 indicate that four models are equivalent for MAE and RSME. When poor Recall prevents the observation of source information, LT's F-Measure performance suffers. IC is able to swiftly estimate the spread since it doesn't take into account any characteristics at the node level while making adoption decisions. Additionally, in Appendix D, MFAD adoption choice experiments are undertaken to compare alternative categorization systems.

TABLE 1. Spread estimation with source information in Weibo



	MFAD	LOGI	IC	LT
F-Measure (%)	**34.14±0.02**	26.68±0.03	12.92±0.06	17.84±0.00
Precision (%)	33.95±0.03	33.32±0.08	16.50±0.09	**47.75±0.05**
Recall (%)	**34.33±0.02**	22.25±0.02	10.62±0.06	10.97±0.00
MAE	10.97±0.01	**10.34±0.02**	14.87±0.04	13.66±0.00
RMSE	**15.73±0.01**	16.5±0.05	21.54±0.13	18.85±0.02
Time (sec)	348.97±47	1238.23±122.19	**34.77±6.22**	806.20±104.22

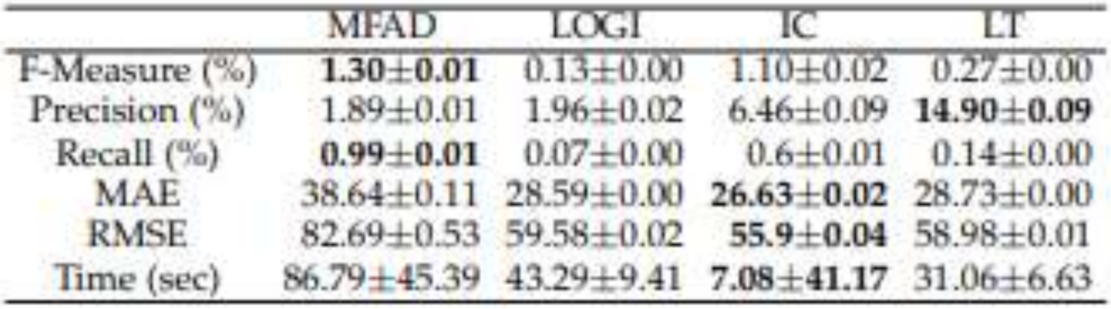
TABLE 2. Spread estimation without source information in Weibo.



	MFAD	LOGI	IC	LT
F-Measure (%)	**22.56±0.02**	12.41±0.03	16.31±0.07	4.92±0.01
Precision (%)	18.27±0.02	13.18±0.05	15.41±0.08	**36.77±0.15**
Recall (%)	**29.48±0.01**	11.72±0.03	17.32±0.09	2.64±0.01
MAE	**15.25±0.03**	15.96±0.06	18.35±0.08	16±0.01
RMSE	21.68±0.05	25.81±0.2	29.05±0.25	**20.06±0.00**
Time (sec)	2140.47±232.87	1301.45±126.65	**49.37±14.89**	684.79±122.21

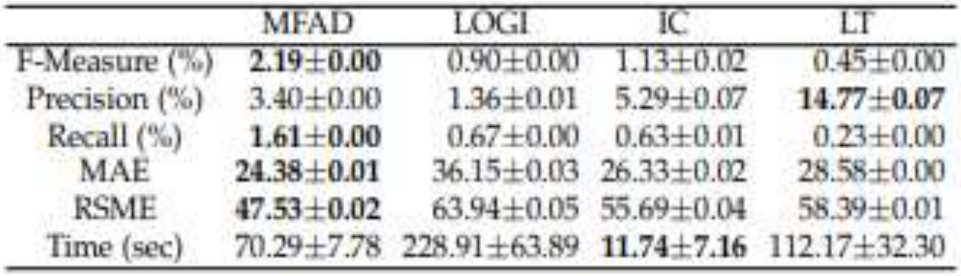
In order to estimate the spread for Douban, we run two kinds of simulations: one without external impact, the other with it. Assuming that information only spreads online, this implies that a node is only aware of information from its in-neighbors unless it is a seed. The second kind believes that a node may be influenced by sources other than its immediate surroundings. However, we don't know just how much of an impact external influences have. For the second sort of simulation, the environment is built up as follows. Every diffusion item has a global pseudo node inserted as a seed, and all nodes follow the pseudo node at the same time as the pseudo node adopts the item. There is a significant time frame T = 1,536 in which models are trained ahead of time owing to the sluggish dissemination of item adoptions, which is described in Tables 3 and 4. Weibo's prediction results are substantially poorer than MFAD's F-Measure, yet MFAD still outperforms the other baselines in Table 3 and 4. Douban's dissemination is significantly slower than Weibo's, and more diffusion is likely to spread outside of Douban. Predicting the spread outcome is substantially more difficult. F-Measure, MAE, and RSME prediction outcomes tend to improve when a global pseudo node is introduced to all models.

TABLE 3. Spread estimation in Douban



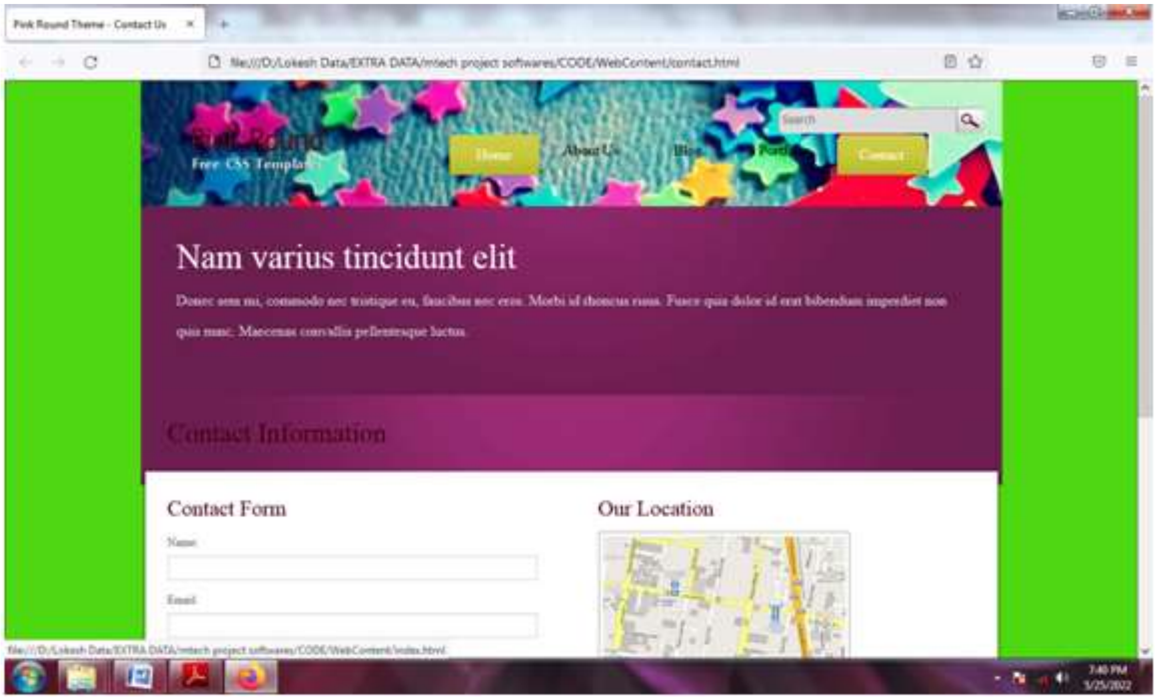
	MFAD	LOGI	IC	LT
F-Measure (%)	**1.30±0.01**	0.13±0.00	1.10±0.02	0.27±0.00
Precision (%)	1.89±0.01	1.96±0.02	6.46±0.09	**14.90±0.09**
Recall (%)	**0.99±0.01**	0.07±0.00	0.6±0.01	0.14±0.00
MAE	38.64±0.11	28.59±0.00	**26.63±0.02**	28.73±0.00
RMSE	82.69±0.53	59.58±0.02	**55.9±0.04**	58.98±0.01
Time (sec)	86.79±45.39	43.29±9.41	**7.08±41.17**	31.06±6.63

TABLE 4. Spread estimation with a global pseudo node in Douban



	MFAD	LOGI	IC	LT
F-Measure (%)	**2.19±0.00**	0.90±0.00	1.13±0.02	0.45±0.00
Precision (%)	3.40±0.00	1.36±0.01	5.29±0.07	**14.77±0.07**
Recall (%)	**1.61±0.00**	0.67±0.00	0.63±0.01	0.23±0.00
MAE	**24.38±0.01**	36.15±0.03	26.33±0.02	28.58±0.00
RSME	**47.53±0.02**	63.94±0.05	55.69±0.04	58.39±0.01
Time (sec)	70.29±7.78	228.91±63.89	**11.74±7.16**	112.17±32.30

To summarise, MFAD is extremely successful for activation prediction and spread estimate across four models, regardless of whether source information is provided or not. According to F-Measure, MFAD estimates the most accurate spread. As a result, for diffusion-based applications like impact maximisation, MFAD spread estimate will better match real-world diffusion. Furthermore, MFAD allows for a wide range of adoption variables to be easily changed without requiring a major overhaul of the learning frameworks.



Pink Round Theme - Contact Us
file:///D:/Lokesh Data/EXTRA DATA/mtech project softwares/CODE/WebContent/contact.html
Search
Pink Round
Free CSS Template
Home
About Us
Blog
Portfolio
Contact

Nam varius tincidunt elit

Donec sem mi, commodo nec tristique eu, faucibus nec eros. Morbi id rhoncus risus. Fusce quis dolor id erat bibendum imperdiet non quis nunc. Maecenas convallis pellentesque luctus.

Contact Information

Contact Form

Name:

Email:

Our Location

file:///D:/Lokesh Data/EXTRA DATA/mtech project softwares/CODE/WebContent/index.html
7:40 PM
3/25/2022

Conclusion:

The multiple-factors conscious diffusion model, which explicitly incorporates the effect of transmission and adoption decisions, is presented in this paper. Adoption choice criteria and classification algorithms have little effect on the MFAD learning frameworks. As a result, MFAD is more versatile and may be used in a variety of situations for a variety of reasons. Online learning and impact maximisation algorithms, as well as their assessment on a wide range of datasets, are some of the next steps. A model built on past data should be updated for an online system to represent dynamic changes, such as user interests, node join and exit. Diffusion records appear streamingly. MFAD's learning frameworks are independent of factors and classification algorithms provided in this study. However, there is still a lot of work to be done in the area of impact maximization and its efficient algorithms. Furthermore, we are interested in the spread of information through online social networks. Different types of diffusion data, such as epidemics, lend themselves well to the use of MFAD.

Reference:

[1] Barbieri, N., Bonchi, F.: Influence maximization with viral product design. In: SDM, pp. 55–63 (2014)

[2] Barbieri, N., Bonchi, F., Manco, G.: Topic-aware social influence propagation models. KAIS 37(3), 555–584 (2013)

[3] Bhagat, S., Goyal, A., Lakshmanan, L.V.: Maximizing product adoption in social networks. In: WSDM, pp. 603–612 (2012)

[4] Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: WWW, pp. 665–674 (2011)

[5] Budak, C., Agrawal, D., El Abbadi, A.: Diffusion of information in social networks: Is it all local? In: ICDM, pp. 121–130 (2012)

[6] Calvo, B., Larraaga, P., Lozano, J.A.: Learning bayesian classifiers from positive and unlabeled examples. Pattern Recognition Letters 28(16), 2375–2384 (2007)

[7] Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: KDD, pp. 1029–1038 (2010)

[8] Daley, D.J., Kendall, D.G.: Epidemics and Rumours. Nature 204 (1964)

[9] Domingos, P., Richardson, M.: Mining the network value of customers. In: KDD, pp. 57–66 (2001)

[10] Du, N., Song, L., Yuan, M., Smola, A.J.: Learning networks of heterogeneous influence. In: NIPS, pp. 2780–2788 (2012)