Advancement from Topic based to Information based Model: A Survey

Jyoti Dua
Computer Science and Engineering
United Institute of Technology, Allahabad
India

Prashant Shukla
Computer Science and Engineering
United Institute of Technology, Allahabad
India

ABSTRACT

Online Social Network played a vital role in diffusing information at a very large scale, lot of work has been done in this area to explain and understand this phenomenon, classifying from predominant topic detection to information diffusion modeling, containing prominent diffuser's identification. This paper presents a survey of illustrative procedures that deal with these topics and propose a classification that reviews the state-of-art. The aim of this paper is to provide an extensive analysis of prevailing efforts around information diffusion in social networks is the aim of the paper. This survey is projected to assist scholars to understand it quickly and bring the enhancement in the present work.

General Terms

Social Network

Keywords

frequency metric, linear influence model, online social network, temporal variation.

1. INTRODUCTION

Individuals can simultaneously exchange information with multiple users via Online Social Networking technologies. In order to quantify the underlying effect of these modes on the propagation of information entails identifying the influences, and whether peers would still disseminate information when the social signals about that information are absent. Social influence plays an essential role in a range of behavioral phenomena, from the diffusion of information to the acceptance of political views and knowledge [15] [16], which are progressively facilitated through online systems. Presently Online Social Networking technologies focus on to extract significant information from the large amount of data. In social networks: the issues, events, interests, etc. that occur, grow rapidly, therefore their apprehension, understanding, conception, and estimation are becoming critical outlooks from both end users and investigators. The understanding of network dynamics may help in following events, resolving issues, refining business performance, etc. in an efficient way. Hence various techniques and models have been developed to elicit and analyze information dissemination in social networks and obtain knowledge from it and forecast it. Since a very large amount of data is available on online social networks, therefore identifying peer influence is a challenge. Time and time again, it is impossible to determine own material from observational data whether an association between two persons' activities exists because they are alike or because one individual has been influenced by the behavior of the other, as persons tend to involve in alike actions as their peers. The main objective of this paper is to survey the developments related to these issues so as to present a clear view of the arena. In regard to this subject, the strengths and weaknesses of prevailing techniques have been identified and structured in taxonomy. The study has been designed to provide guidelines for researchers who are interested in developing novel techniques in this field.

2. MODELS ON TOPIC DETECTION TECHNIQUE

Topic detection techniques were established earlier for static corpora and not reformed to the stream of messages generated by O.S.N. Therefore, it has been recommended to focus on burst.

Leskovec *et al.* [3] developed a model for tracking and identifying popular topics and ideas on social media. They developed a mathematical model for analyzing temporal variation to provide coherent representation on the basis of abrupt spikes in the contents that spreads with great extent of popularity and then weakens over time. Fig. 1 represents the temporal variation of popular topics. Thereafter, based on this concept, many researches are performed on the basis of these successive burst where all the approaches are performed on discrete data, therefore, they must be discretized by converting continuous data into sequence of messages circulated during equal slice of time. Fig. 2 represents discretization of stream of messages.

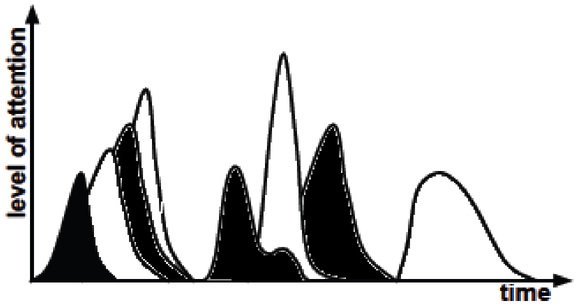

Fig 1: Temporal Variation of popular topics [3]

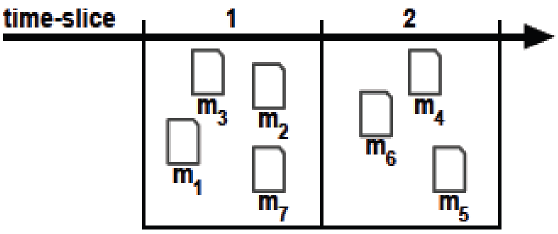

Fig 2: Discretization of stream messages

Shamma *et al* [5] demonstrated two metrics for identifying temporal topics. These two metrics are peaky topic and persistence conversation. Peaky topics are the term of interests and persistence conversation are less salient term. To extract these texts the authors used well-known model known as tf-idf model [20] for scoring. The model uses normalized term frequency metric ($ntf_{t,i} = {tf_{t,i}}\Big/{cf_t}$, where $tf_{t,i}$is the frequency of term t at the i^{th} time slice and cf_t is the frequency of term t in the whole message stream) to extract table of content by finding peaky topics. The method score each term on the basis of its level of interest and then rank all the term according to the score.

AlSumait *et al.* [6] proposed online topic model, known as a non-Markov on-line Latent Dirichlet Allocation Gibbs sampler topic model (O.L.D.A.). The O.L.D.A. approach is based on L.D.A. generation process with few enhancements. The L.D.A. is a statistical generation model that depends on hierarchical Bayesian Network that relates messages and words via latent topic. The idea behind O.L.D.A. is to acquire evolution of the topic, build the evolutionary matrix and then permit to detect bursty topic. This approach, incrementally updates the topic model at each time slice using the previously generated model as a prior and the corresponding message collection monitor the learning of the new generative process.

Cataldi et al. [1] presented a work on Temporal and Social Term Evaluation that considers both temporal and social attributes. The objective of authors' topic detection approach is to retrieve real-time topics or the most emerging term in the community. They defined term as emerging which are frequent term but rare in past. In order to determine the authority of active users, the author analyzed social relationship using PageRank algorithm. Lastly, they connect emerging terms with the semantically related keywords using a topic graph. The model follows 5-step process. In the first step, the user-generated content are extracted and formalized as term vector with relative frequencies. Secondly, based on social relationship directed graph of active users are defined and PageRank algorithm is used to calculate the users' authority. Designing of life cycle model is done in the third step for each term according to aging theory. In fourth step, PageRank algorithm is applied to select emerging term depending on their status or energy value. At last, topic graph is designed which connects emerging term with semantically term to achieve a set of emerging term as output. The authors in their theory defined semantically related term as single argument.

Rong *et al.* [7] developed the method for predicting trends and reason analysis of new emerging topics on Twitter. The method is based on Moving Average Convergence Divergence (M.A.C.D) indicator. This M.A.C.D indicator is used in the analysis of stocks. Therefore, based on the characteristics of emerging topics authors used the indicator accordingly.to define trend momentum and also to predict trends of emerging topics. To offer best trend following and momentum, the authors provide M.A.C.D two trend-following turns indicator i.e. long moving average and short moving average. The trend momentum of new topic in [1] is obtained by subtracting longer moving average from the shorter moving average. Now, the resulting trend momentum is used to predict trends of new topic in near future. At last, authors also proposed two circumstances for trend variations. The first situation, when trend rises, this is due to the involvement of influential users and the second situation when trend falls is due to the popularity of other topic that attracts the attention of users on social network.

The above method discussed to predict emerging new topics are conventional approaches which are based on term-frequencies. These approaches are not well suited as the information shared on social network is not only text but images, videos and URLs. Thus, Takahashi *et al.* [8] studied aspects of social network where links are formed dynamically. So, they proposed a probability model o predict the emergence of new topic from anomaly measure. They combine the anomaly score with the recent change point detection technique that is based on Sequentially Discounting Normalized Maximum Likelihood or with Kleinberg's burst model to detect emerging topic based on reply/mention relationships under probability model.

3. INFORMATION DIFFUSION MODELS

Diffusion process is more challenging in a dynamic network as it grows and change very quickly. The two factors are used to categorize diffusion process. They are structure and temporal dynamic. The structure describes the topology in which who is influenced by whom. The temporal dynamic is the number of nodes that accepts the part of information over time called as diffusion rate. Communication and Social Network substrate are the two factors on which today's social media operate. Most of the social network models relies on either the structural growth of the system that deals with the dynamics of the network or information diffusion processes that deals with the dynamics on the network. Mainly, the models developed in this regard of O.S.Ns assume that people only influenced by activities performed by their connection. That is why path followed by the part of information in the network is known as spreading cascade that provide knowledge where and when a part of information disseminated but not why and how. Therefore, to capture and predict underlying hidden mechanism information diffusion models are used. There are two commonly used models i.e. Explanatory Model and Predictive Model.

3.1 Explanatory based Models

This model takes into account underlying sociological factors that causes to establish links that helps understand how information is propagated and allows to follow the path that is taken by the part of information for propagation. In contrast to structural models, explanatory models are more detailed and specific to the network which is being considered. There were various models that were proposed under this category.

Gomez *et al.* [17] [18] proposed an iterative algorithm called N.E.T.I.N.F. which relies on sub-modular function optimization. This algorithm is used to find the spreading cascade that maximizes the likelihood of the observed data. Later, the authors extended the algorithm and developed the diffusion process model in which temporal processes occur at different rate. The probability density function is used to calculate the likelihood of a node infecting other at a given interval of time. The extended N.E.T.I.N.F. algorithm assumes pair wise transmission rates and the graph of diffusion by designing and calculating the convex maximum likelihood problem [19].

Choudhary *et al.* [2] after performing experiments on Twitter data analyzed that sampling method that consider both network topology and users' attribute such as activity and localization permits to acquire information diffusion with minor error than naïve approaches like random or activity only-based sampling.

Sadikov *et al*. [9] formulated k-tree model of cascades to address the problem of missing data and analyzed its properties for missing data. Using Information Propagation Cascade model they evaluated their methodology on Twitter and found that the properties of complete cascade C such as size or depth can be accurately estimated even when 90% of data is missing.

3.2 Predictive based Models

According to [21], the main idea behind this model is to predict that how particular diffusion process would describe itself in the given snapshot of a network by learning from past diffusion approaches from temporal or spatial point of view.

The predictive based model can be classified in two forms, Non-graph method and graph method.

3.2.1. Approaches based on Graph

Graph based method is divided into two models, Independent Cascade which is sender-centric and Linear Threshold which is receiver-centric. The diffusion probability in the Independent Cascade model is to be associated to each edge whereas, in Linear Threshold an influence degree is to be associated to each edge and an influence threshold to each node. The diffusion process for both models continues iteratively along a discrete time-axis and synchronously, beginning from a set of initially activated nodes, often named as early adopters [10]. In the case of I.C., at each iteration the newly activated node try only once to activate their neighbor with the probability associated to edges joining them. In the case of L.T. model, at every iteration, if their own influence threshold is exceeded by the sum of influence degrees, then their activated neighbors activate the inactive nodes. The process terminates either, when no fresh transmission is possible or, when no neighboring node can be communicated [21].

Galuba et al. [11] proposed the L.T. oriented model. The model depends on parameters such as pair wise user's degree of influence, information virility, and user likelihood of accepting any information. The L.T. model is applied on data describing the diffusion process by first optimizing the parameter using gradient ascent method. But the realistic temporal dynamics cannot be reproduced by L.T.

Saito *et al.* [12] proposed asynchronous extension of L.T. and I.C. called Asynchronous Linear Threshold (As.L.T.) and Asynchronous Independent Cascade (As.I.C.). The model needs same parameters as synchronous I.C. and L.T. comprising of a time delay parameter on each edge of a graph and continue iteration along continuous time axis. The parameters are defined in parametric mode. The authors considered the process as maximum likelihood estimation method and also guarantee the convergence. However, the limitation of this model is that it is limited to synthetic data and does not have any practical explanation.

3.2.2. Approaches based on Non-Graph

The non-graph based approach does not consider structure of a graph. However, the approach classifies nodes in numerous classes. There are two important models in this approach, Susceptible Infected and Recovered and Susceptible Infected Susceptible. In both the cases, the nodes in the class 'S' switch to the class 'I' with a fixed probability β. In case of, S.I.S., nodes in the 'I' class switch to the class 'S' with the probability γ, whereas in the S.I.R case the nodes permanently move to R class. The differential equations are used to express percentage of nodes in each class. Both model assume that the probability of every node to be connected to another is same. The connection inside the population is random [21].

Leskovec *et al.* [4] proposed a model based on S.I.S. that considers only single parameter β. In this model, every node has the same probability β to capture the information. At the succeeding time step the nodes that have adopted the information become vulnerable. There is no even distribution of influence among all nodes and this characteristic is necessary to consider in the social network.

Yang *et al.* [13] developed a Linear Influence Model that focus on global influence of the node that has on the diffusion rate. The total count of freshly infected nodes of which other nodes got infected in the past is termed as the influence function. This influence function is non-parametric formulation and is estimated by solving a non-negative least squared method. They demonstrated that this model accurately simulates influence of the nodes and foresees the temporal dynamics efficiently. Fig. 3 represents Linear Influence forecast model.

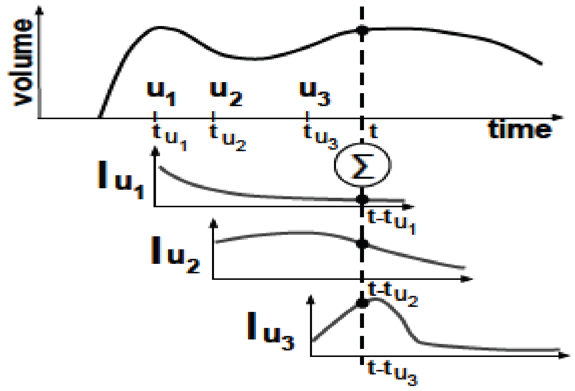

Fig 3: Linear Influence Forecast Model [13]

Weng *et al.* [14] developed a model based on partial differential to predict diffusion of information injected by a given node in the network. Logistic diffusive equation is used to predict topological and temporal dynamic, diffusive logistic equation model is used. The topology is considered in the term of distance from each node to source node. The density of influenced nodes at the given distance source at a given time is represented by logistic equation. The logistic equation is used to represent process dynamics. The cubic spline interpolation method is used to estimate the parameter in the model.

4. CONCLUSION

Social network is a vast source and a tunnel through which information propagate worldwide. Many information model have been developed to focus on which information has the highest popularity, through which path the information should be diffused and how it can be diffused so that it can reach to a large extent, who influence whom in the social network so as to deliver best and relevant information as possible. In short, the information models have been developed for optimizing and for accelerating diffusion process. Diffusion process is more challenging task in a dynamic network. Spreading of information, data routing, propagating new ideas are some of the example of diffusion process. Many models have been developed by the researchers to accelerate the diffusion process and to predict hot topics or new emerging topics on social network. Therefore, this paper tried to focus on the researches that were performed to diffuse information on the social network in more efficient way so that large number of users may be benefitted from the information. Even today, the researchers are contributing a lot in this area and different approaches are made to be included in the topic such as, data mining, information retrieval.

5. REFERENCES

[1] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation", MDMKDD '10, pages 4–13, 2010.

[2] M.D. Choudhury, Y.R. Lin, , H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, " How does the data sampling strategy impact the discovery of information diffusion in social media?" ICWSM '10, pages 34-41, 2010.

[3] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", KDD '09, pages 497-506, 2009.

[4] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst, " Cascading behavior in large blog graphs", SDM '07, pages 551-556, (short paper) 2007.

[5] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In CSCW '11, pages 355–358, (short paper) 2011.

[6] L. AlSumait, D. Barbar´a, and C. Domeniconi., "Online lda: Adaptive topic models for mining text streams with applications to topic detection and tracking", ICDM '08, pages 3–12, 2008.

[7] L. Rong and Y. Qing, "Trends analysis of news topics on Twitter. International Journal of Machine Learning and Computing ", 2(3):327–332, 2012.

[8] T. Takahashi, R. Tomioka, and K. Yamanishi, ". Discovering emerging topics in social streams via link anomaly detection", ICDM '11, pages 1230-1235, 2011.

[9] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina, "Correcting for missing data in information cascades", WSDM '11, pages 55–64, 2011.

[10] E. M. Rogers. Diffusion of Innovations, 5th Edition. Free Press, 5th edition, August, 2003.

[11] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers - predicting information cascades in microblogs", WOSN '10, pages 3–11, 2010.

[12] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda,"Learning diffusion probability based on node attributes in social networks", ISMIS '11, pages 153-162, 2011.

[13] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In ICDM '10, pages 599 608, 2010.

[14] F. Wang, H. Wang, and K. Xu., "Diffusive logistic model towards predicting information diffusion in online social networks", ICDCS '12 Workshops, pages 133–139, 2012.

[15] M. S. Granovetter, "Threshold models of collective behavior", Am. J. Sociol., 83(6):1420{1443), 1978.

[16] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world' networks", Nature, 393(6684):440{442), June 1998.

[17] M. Gomez Rodriguez, J. Leskovec, and A. Krause,. "Inferring networks of diffusion and influence", KDD '10, pages 1019–1028, 2010.

[18] M. Gomez-Rodriguez, J. Leskovec, and B. Schokopf. Structure and dynamics of information pathways in online media. In WSDM '13, pages 23–32, 2013.

[19] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September, 2012.

[20] G. Salton and M.J. McGill. "Introduction to Modern Information Retrieval", McGraw-Hill, 1986

[21] A Guille,, H Hacid,, C Favre, D.A Zighed, " Information Diffusion in Online Social Networks: A Survey", ACM SIGMOD, Vol. 42, No. 2, June 2013.