A SURVEY ON LDA APPROACH IN PREDICTING LINK BEHAVIOR IN A SOCIAL NETWORK

MD ABDUL NAVEED MASTAN
M.Tech, Department of Computer Science Engineering
Velgapudi Ramakrishna Siddhartha Engineering College (Autonomous), Affiliated to JNTUK
Vijayawada, AP, India
m.a.naveedmastan@gmail.com

S.RAVI KISHAN
Associate Professor, Department of Computer Science Engineering
Velgapudi Ramakrishna Siddhartha Engineering College (Autonomous), Affiliated to JNTUK
Vijayawada, AP, India
suraki_in@yahoo.com

***Abstract*—Social network sites (SNS) in recent times are focusing mainly on user interactions. These SNS are attracting the attention of academic and industry researchers who are intrigued by their accordance and reach rapidly. Mainly data mining techniques have been very effective in using the content and graph structure that was available to solve various problems such as friendship link prediction, estimating the percentage of their friendship…etc. Topic models are one among the most effective approaches to discover latent topic analysis and text data mining. One desirable feature of a social network is to be capable to suggest potential friends to its existing users and the approach must be proved to be effective in improving the predictions. Topic modeling approach provides an easy way to analyze large volume of data and the topic modeling techniques like Latent Dirichlet Allocation (LDA) to uncover latent structure in user interests which have to be explored is going to be implemented. By using LDA, the users are predicting their friends and with the how much amount of percentage ratio they are becoming friends. In this review, it has been identified that LDA has a limitation of topic correlation modeling which can be overcome by using CTM (correlated topic model) and it can work better than Arm (Association Rule Mining) for list of 4 or more communities while the tagging can be effectively done when both LDA and association rules are used together.**

Keywords: Social Networks, Link Prediction, Learning, Topic Modeling.

I.INTRODUCTION:

Social network sites (SNS) in recent times are focusing mainly on user interactions. These SNS are attracting the attention of academic and industry researchers who are intrigued by their accordance and reach rapidly. Mainly data mining techniques have been very effective in using the content and graph structure that was available to solve various problems such as friendship link prediction, estimating the percentage of their friendship…etc. Topic models are one among the most effective approaches to discover latent topic analysis and text data mining. One desirable feature of a social network is to be capable to suggest potential friends to its existing users and the approach must be proved to be effective in improving the predictions. Topic modeling approach provides an easy way to analyze large volume of data and the topic modeling techniques like Latent Dirichlet Allocation (LDA) to uncover latent structure in user interests which have to be explored is going to be implemented. By using LDA, the users are predicting their friends and with the how much amount of percentage ratio they are becoming friends. In this review, it has been identified that LDA has a limitation of topic correlation modeling which can be overcome by using CTM (correlated topic model) and it can work better than Arm (Association Rule Mining) for list of 4 or more communities while the tagging can be effectively done when both LDA and association rules are used together.

II.RELATED WORK:

Large documents collections are readily available online and are widely accessed by diverse communities. The advent of new tools for browsing, searching and allowing the productive use of such archives is thus an important technological challenge and provides new opportunities for statistical modeling. In [3] topic models have been considered by which latent variable models of documents that exploit the correlations among the words and latent semantic themes. Topic models can extract interpretable and useful structure without any explicit understanding of the language by computer. In [3] paper, the correlated topic model (CTM) explicitly

models the correlation between the latent topics in the collection and enables the topic graphs construction and document browsers which allow a user to navigate the collection in a topic guided manner.

The correlated topic model builds on the earlier latent dirichlet allocation (LDA) model of [3] which is an instance of a general family of mixed membership models for decomposing data into multiple latent components. LDA mainly assumes that the words of each document arise from a mixture of topics where each topic is a multinomial over a fixed word vocabulary. The topics are shared by documents in the collection but the topic proportions change stochastically across documents as they are randomly drawn from a Dirichlet distribution. A recent work in [3] has used LDA as a building block in more sophisticated topic models. They fail to directly model the correlation between topics in the document. In most of the text corpora, it is natural to expect subsets of the underlying latent topics will be highly correlated. Consider an example for instance in science, an article about genetics may be likely to be about health and disease but unlikely to be about X-ray astronomy. For the LDA model, this limitation stems from the independence assumptions implicit in the dirichlet distribution on the topic proportions.

The CTM replaces the dirichlet by the more flexible logistic normal distribution which incorporates a covariance structure among the components. This gives more realistic model of the latent topic structure where the presence of one latent topic may be correlated with the presence of another a hierarchical topic model of documents that replaces dirichlet distribution of per document topic proportions with a logistic normal. This allows the model to capture correlations between the occurrences of latent topics. The resulting correlated topic model gives better predictive performance and uncovers interesting descriptive statistics for facilitating browsing and search. Use of the logistic normal may have benefit in the many applications of dirichlet-based mixed membership models. Much information is readily accessible online still we don't have means for processing all of it. To help users overcome the information overload problem and sift through huge amounts of information efficiently and easily, recommender systems have been developed to generate suggestions based on user preferences.

In [4], focus is on applying CF to community recommendation. Investigating which notions of similarity are most useful for this task, we examine two approaches from different fields. First, the association rule mining (ARM) is a data mining algorithm that finds association rules based on frequently co-occurring sets of communities and then it makes the recommendations based on the rules. ARM can discover the explicit relations between communities based on their co-occurrences across the multiple users. Second, LDA is a machine learning algorithm that models user community co-occurrences using the latent aspects and makes recommendations based on the learned model parameters. Unlike ARM, LDA models the implicit relations between communities through the set of latent aspects present. Comparison of ARM and LDA for the community recommendation task was made and evaluated their performances using the top-k recommendations metric. LDA performs consistently better than ARM for the community recommendation task when recommending a list of 4 or more communities. However, for recommendation lists of up to 3 communities, ARM is still a bit better.

Ontology can be defined as an explicit formal specification of the terms and relations among terms in a domain. It can be achieved by a systematic grouping of domain concepts may be user interests based on their definitions in machine interpretable form. Although the ontology constructed in [1] has proven helpful for improving the predictions of friendship relationships the use of WordNet-Online, IMDB and AWS for a semantic understanding of user interests is cumbersome and this may not always give complete and accurate definitions of interests. Work in [5] explores different ontology engineering approaches and more comprehensive knowledge bases to address the limitations mentioned in [1]. In first approach, we obtain the definitions of interests from Wikipedia and use the technique of latent semantic analysis (LSA) to measure the similar behavior between interests. While this approach produces more sensible ontology than the one produced by the approach in [1], this ontology is still a binary tree and it consists of internal clusters labeled based on the child information. Our second and third approaches explore reuse of knowledge from existing hierarchies such as the Wikipedia Category Graph (WCG) and Directory Mozilla (DMoz) to group interests.

Three approaches are explored to the problem of building ontology over the interests specified by the users in a social network. The first and third approaches produce usable hierarchies although the Wikipedia/LSA hierarchy has some limitations. While the second approach didn't produce a useful ontology, it served as a bridge between the Wikipedia/LSA approach and DMoz approach. Moreover, it motivated the reuse of knowledge from existing hierarchies in the ontology engineering process. Extensive exploration of the usefulness of both Wikipedia/LSA and DMoz based interest hierarchies for the predicting friendship links.

Tagging systems have now become major infrastructures on the web. They allow users to create tags which annotate and categorize the content and share them with other users that are very helpful in particular for searching multimedia content. Tagging is not constrained by a controlled vocabulary and annotation guidelines. These tags tend to be noisy and sparse. In [6] an approach based on Latent Dirichlet Allocation (LDA) for recommending tags of resources in order to improve search was introduced. The goal of the approach presented

in [6] is to overcome the cold start problem for tagging new resources. In [6], we use Latent Dirichlet Allocation (LDA) to elicit latent topics from resources with a fairly stable and complete tag set to recommend topics for new resources with only a few tags.

Based on this, other tags belonging to the recommended topics can be recommended. The use of LDA for collective tag recommendation has been explored in [6]. Compared to association rules, LDA achieves better accuracy and in particular it recommends more specific tags which are more useful for search. In general, our LDA based approach is able to elicit a shared topical structure from the collaborative tagging effort of multiple users more over the association rules are more focused on simple terminology expansion. However, both approaches succeed to some degree in overcoming the idiosyncrasies of individual tagging practices. The main contribution of latent topic models is to reduce the sparsity of the tag space. This gives rise to several interesting lines of research which will investigate mapping resources to their latent topics that may result in more robust resource recommendation.

The problem of Named Entity Recognition in Query (NERQ) involves detection of the named entity in a given query and classification of the named entity into predefined classes. NERQ is potentially useful in various applications in web search. The [7] approach proposes taking a probabilistic approach to the task using query log data and Latent Dirichlet Allocation. In this task for a given query the named entity has to be detected within the query and identify the most likely classes of the named entity. In the LDA model, the contexts of a named entity is represented as words of a document where as classes of the named entity are represented as topics of the model. The alignment between model topics and predefined classes needs to be guaranteed. To address this problem, a weakly supervised learning method referred to as WS-LDA (Weakly Supervised Latent Dirichlet Allocation) which can leverage the weak supervision from humans. Experimental results indicate that the proposed approach can accurately perform NERQ and outperforms other baseline methods.

III.TOPIC MODELS AND LATENT DIRICHLET ALLOCATION

With the growth of data on the web in the form of web sites, articles, social networking sites, news etc., there is an increased need to process the data that has to extract hidden patterns and information from them. Data mining techniques like vector space model were used in the past to extract the patterns from text. Vector space model is an algebraic model for representing the text documents as vectors of identifiers. It uses the bag of words representation the documents are seen as vectors in the word space to represent each document in the document corpus.

Topic modeling is the method for analyzing large quantities of the data that is unlabeled. A topic model provides an easy and simple way to analyze large volumes of unlabeled data. A topic in the model consists of a word clusters that occur frequently together. By the usage of the contextual clues, topic models can connect the words that have similar meanings and distinguish between words uses that have multiple meanings. Topic models extract topics from texts which represent a family of computer programs. Topic to computer is a list of words that exist in statistically meaningful ways. The text can be an email, a blog post, a book chapter, a journal article...etc any kind of unstructured text. By unstructured means that there exists no computer readable annotations that tell the computer about the semantic meaning of the words in text available. In general, the topic modeling programs doesn’t have any idea about the word meanings in a text. Instead, they make an assumption that any piece of text is composed by selecting words from possible word baskets where each basket corresponds to topic. If that results true, then it becomes easy to mathematically decompose a text into probable baskets from where the words first came. The tool goes through this process again and again until it settles on the most likely word distributions into baskets which probably called as topics.

Figure 1 illustrates the topic modeling approach as a generative model. Probabilistic Latent Semantic Analysis (pLSA) is one such generative model that is used to model the documents. It is reported that pLSA has many over fitting problems. The number of parameters grows linearly with the number of documents. Even though pLSA is the generative model of the documents in the collection used to estimate the model but it is not a generative model of new documents. The idea is that use of a probabilistic mixture of a number of models is that to explain some observed data. Each observed point of data is assumed to have come from one of the existing models in the mixture but which one it has come out is unknown. The latent parameter that specifies which model each point of data came from.

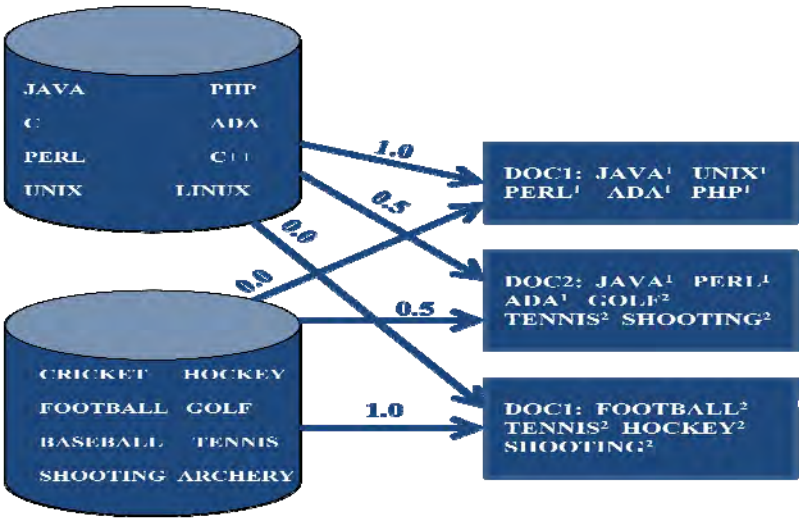

Fig. 1. Illustration of Probabilistic Generative Process

LDA is another popular topic model in application and it is also the simplest one. It solves the problems with over fitting and increased number of parameters. Latent Dirichlet Allocation (LDA) is a probabilistic generative model for the collection of discrete data like text corpora [3]. LDA was focused at solving the disadvantages exhibited by the probabilistic LSA model. LDA is almost similar to pLSA difference is that in LDA distribution of topic is assumed to have the Dirichlet prior.

LDA is a true generative model that means it have the ability to generate documents and it allows sets of observations that are explained by unobserved groups which explain why some parts of the data are similar. LDA represents the documents in given corpus as topic mixture that spit out words with some probabilities. The goal of LDA is to determine the degree to which a document explains various topics. The LDA method can also be applied to data collections other than text documents but the terminology of natural language processing provides an intuitive way to describe the algorithm. The LDA model is Bayesian model where each document is represented as a topic mixture while each topic is a discrete probability distribution mostly by an array that defines how common each word is in each topic. In general, everyone normally think of a document as a sequence of words but LDA sees a document as merely a collection of weighted topics from which words can be generated.

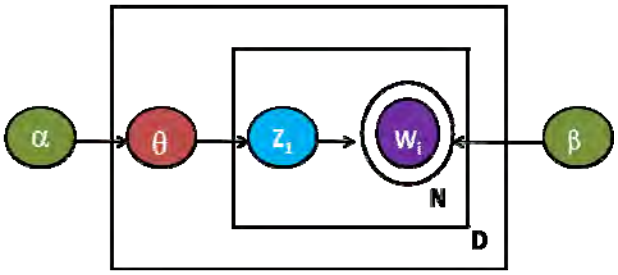

Fig. 2.
Plane Notation Of LDA Algorithm.

LDA assumes the following generative process can be described as follows [3]

For each document in a corpus D:

1. Choose topic distribution

   $\theta_i \sim$ Dirichlet(α) where i {1…M} and Dirichlet (α) is the Dirichlet distribution for parameter α

2. For each of the words w_{ij} in the document, where j {1,….,N_i}

   (a) Choose a specific topic

   $z_{ij} \sim$ Multi (θ_i)

   Where multi () is a multinomial

   (b) Choose a word $w_{ij} \sim \beta z_{ij}$

Here,

- w represents a words

- z represents a vector of topics,
- β is a k X V word-probability matrix for each topic (row) and each term column),

  Where $\beta_{ij} = p(w_j = 1|z^i = 1)$

  Plane notation is shown in figure1

  Step (a) reflects that each document contains topics in different proportion

For instance consider one document may contain a many words taken from the topic on climate and no words taken from the topic about diseases, while a different document may have an equal number of words drawn from both topics. Step (ii) reflects that each individual word in the document is drawn from one of the K topics in proportion to the document's distribution over topics as determined in Step (i). The selection of each word depends on the distribution over the V words in vocabulary as determined by the selected topic j. The generative model does not make any assumptions about how the orders of the words in the documents are present. This is known as the bag-of-words assumption. The central goal of topic modeling is to automatically discover topics from a collection of documents.

In Fig. 3, α and β parameters constitute to the outermost level of the model, parameter θ_d forms the middle level and parameters Z_{dn}, W_{dn} are at the innermost level of the model. Parameters α and β are corpus level parameters. These corpus level parameters are assumed to be sampled once in the process of corpus generation. The variables θ_d are document-level variables sampled once per document and the variables Z_{dn} and W_{dn} are at the word level.

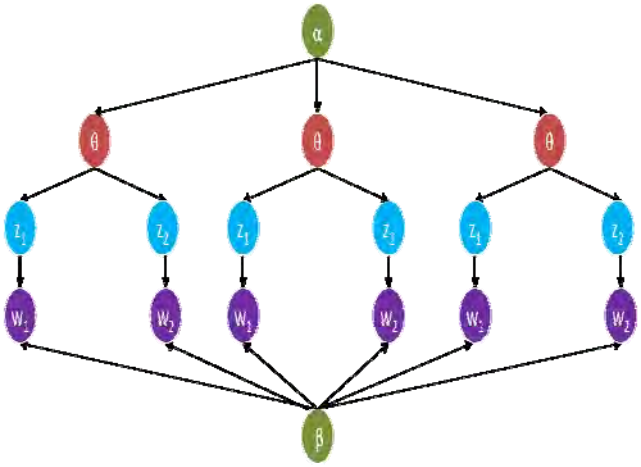

Fig. 3. Graphical Notation Representing LDA Model.

III.CONCLUSION

From all the above discussed topics, it has been identified that LDA has a limitation of topic correlation modeling which can be overcome by using CTM (correlated topic model) and it can work better than Arm (Association Rule Mining) for list of 4 or more communities while the tagging can be effectively done when both LDA and association rules are used together. Instead of LDA, WS-LDA can work better for recognition of query processing. By using this information, the friendship links has to be predicted from the user dataset in a social network with help of LDA.

IV.ACKNOWLEDGEMENTS

Authors acknowledge the help received from the scholars whose articles have been cited and included in references of this manuscript. The authors are also grateful to authors/editors /publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed. Authors are grateful to IJCSE editorial board members and IJCSE team of reviewers who have helped to bring quality to this manuscript.

V.REFERENCES

[1] Bahirwani, V. 2008. Ontology engineering and feature construction for predicting friendship links and users' interests in the live journal social network. Master's thesis, Kansas State University, 2008.

[2] Nowell, D.L. and Jon Kleinberg. 2003. The link prediction problem for social networks, 2003. Proc. of the Twelfth Annual ACM Int. Conf. on Information and Knowledge Management (CIKM'03), November 2003, pp.556-559.

[3] Blei, D. and Laerty, D.J. 2008. A correlated topic model of science. *Ann. Appl. Stat*. 1: 17-35.

[4] Chen, W., Chu, J., Luan, J., Bai, H., Wang, Y. and Chang, Y.E. 2009. Collaborative filtering for orkut communities: Discovery of user latent behavior. In Proc. of Int. World Wide Web Conf. 2009.

[5] Haridas, M 2009. Exploring knowledge bases for engineering a user interest's hierarchy for social network applications. Master's thesis, Kansas State University, 2010.
[6] Krestel, R., Fankhauser, P. and Nejdl, W. 2009. Latent dirichlet allocation for tag recommendation. In Proc. of RecSys'09, New York, USA, 2009.
[7] Guo, J., Xu, G., Cheng, X. and Li, H. Named entity recognition in query. In Proc. of SIGIR'09, Boston, USA, 2009.
[8] Blei, D., Ng, Y.A. and Jordan, I.M. 2003. Latent Dirichlet Allocation. *J. Machine Learning Res*. 3: 993-1022.
[9] Castillo, C., Donato, D., Gionis, A., Murdock, V. and Silvestri, F.2007. Know your neighbors: Web spam detection using the web topology. In Proc. of SIGIR'07, Amsterdam, Netherlands, 2007.
[10] Steyvers, M. and Griffiths, T. 2007. Probabilistic topic models, 2007. In Handbook of latent semantic analysis
[11] Reed, C. 2012. Latent Dirichlet Allocation: Towards a deeper understanding, 2012.
[12] McCallam, K.A. 2002. Mallet: machine learning for language toolkit, 2002. Retrieved from http://mallet.cs.umass.edu.
[13] Steyvers, M. and Griffiths, T. 2004. Finding scientific topics. In Proc. of Nat. Acad. Sciences, U.S.A, 2004.