ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Topic Model for Graph Structured Data

Gauri Arunrao Shete[1], Prashant Borkar[2]

M. Tech Student, Dept. of CSE., G.H. Raisoni College of Engineering, Nagpur, India[1]

Assistant Professor, Dept. of CSE., G.H. Raisoni College of Engineering, Nagpur, India[2]

ABSTRACT: Many types of data can be represented as graphs such as documentary data, Chemical molecular structures and images. There is one issue in these graphs that they cannot find the hidden data topics. Topic model can solve this problem successfully. To address this problem a Graph Topic Model (GTM) can be used. A Bernoulli distribution may used to model edges between nodes in the graph. It will make edges in graph to find latent topic discovery and then accuracy of supervised and unsupervised learning can be improved.

KEYWORDS: Graph mining, Latent Dirichlet Allocation (LDA), Topic model

I. INTRODUCTION

Graph is a combination of nodes and edges, where these edges are connected by links nothing but edges. Many structured and unstructured data can be represented as graphs. Hence the research about this graph structured data belongs to graph mining area. In graph mining, edges contribute in classification or clustering of data. In Text mining, a document is made by some words denoted as nodes and relation between words, such as association relation or some semantic relations denoted as edges of nodes of graphs. By these classification of document graphs the accuracy of document can be improved.

On the other hand the research about topic detection in text mining and video processing is done, but the issue about topic detection in graph mining is not well solved. Hence the question arises, how do we discover the hidden topics in graph structured data? In order to solve this issue, a Topic Model for graph structured data (GTM) is proposed. Because of "bag-of-word" assumption a standard topic model cannot directly applied on graph structured data. Hence one assumption can be made that if an edge is present in between two nodes then these two nodes possess same content. To detect existence of an edge parameterised by topics of two linked nodes, a Bernoulli distribution is used in GTM. GTM makes the edges put up to latent topic discovery by directly measuring the edges. Latent DirichletAllocation(LDA) method is used to detect the hidden topics of the graph. A GTM for graph structured data is built by modelling the edges in graph to discover latent topics and LDA for detecting hidden topics. Also one Author Topic (AT) model is may also used to discover the latent topics of graphs. Hence the model LDA is used for Graph topic model i.e. GTM.

II. LITERATURE SURVEY

Sr. no	Title	Author	Description
1.	Topic model for Graph Mining	Junyu Xuan, Jie Lu, Guangquan Zhang and Xiangfeng Luo	In this paper a GTM model is proposed which uses Bernoulli distributions is to model edges between the edges. And further applies MCMC inference algorithm to propose the GTM. Also comparison between LDA and MCMC is made.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

2.	Hashtag graph based topic model for tweet mining	Yuan Wang, Jie Liu, Jishi Qu and Xia Feng	In this paper a Novel topic Model is proposed to handle Semi-structured tweets, denoted as Hashtag Graph based Topic Model (HGTM). By utilizing the relation between hashtags in these hashtag graphs it builds word semantic relation, even if they haven't occurred in that specific tweet.
3.	External Evolution of Topic Models: A Graph mining Approach	Hau Chan and Lemon Akoglu	In this paper, a graph mining and machine learning approach is developed for external evolution of topic models. On the basis of graph centric features, the extraction of topic words from projection is done on the Wikipedia page links graph.
4	Topic models and advanced algorithms for profiling of knowledge in scientific papers	V.Jelisavcic, B. Furlan, J. Protic, and V. Milutinovic	Models introducing in-domain knowledge are sparse comparing to others, but with increasing need for semantically richer topics and applications, incorporating in-domain knowledge is gaining on popularity.
5	Graph Mining: A Survey of Graph Mining Techniques	Saif Ur Rehman, AsmatUllah Khan, Simon Fong	Summary information of different graph mining techniques is given in this paper. These graph mining techniques are based on the classification, clustering, decision tree approaches which are the data mining fundamentals.
6	Probabilistic Topic Modelling for Genomic data Interpretation	Xin Chen, Xiaohua Hu, Xiajiong Shen, and Gail Rosen	A probabilistic topic modelling is introduced to find the genome level comparison of DNA sequences, in which the concurrence pattern of N-mer feature across the whole genome set are modelled as latent topics. The proposed probabilistic topic modelling is capable of characterising the core and distributed genes within a species.
7	Topic Model Allocation of Conversational Dialogue records by Latent Dirichlet Allocation	Jui-Feng Yeh , Chen-Hsien Lee, Yi-Shiuan Tan, and Liang-Chih Yu	The dialogs records can deviate the topics, event has the sentences which are not connect with the topics. Here LDA is utilized which uses the Bag-of-word, hence it is essential to select the word it frequently selects the similar vocabulary for each topic.
8	LDA based Model for Topic Evolution on Text	Qingqiang Wu, Caidong Zhang, Xiang Deng, Changlong Jiang	This paper analyses the topic of the evolution model, and then achieves the LDA model, and then finally results shows that the evolution of model in topics has good result.

III. PROPOSED METHODOLOGY

Here in this project LDA method is used. LDA is a probabilistic model used to collect discrete data. In this project there are some basic terminologies used in LDA method as shown below:-

1. **Words:-**Defined as item from a vocabulary as shown as {1,….N}. These words are represented as unit-basis vectors having single component equal to one and all other component equal to zero.
2. **Document:-** Document is sequence of n words denoted as D = $\{W_1,W_2,\ldots W_n\}$, where W_n is the n^{th} word in the sequence.
3. **Corpus:-** It is collection of m documents such as $\{D_1,D_2,\ldots D_m\}$.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

The aim of this Topic model is to find out the topics from the corpus. It assumes that document is made up of number of topics with different weights, and topic is composed by number of keywords with different weights called keyword distribution.

The graphical representation of LDA is as shown below:

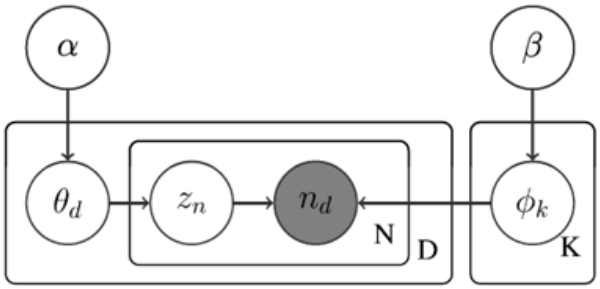


Fig 1. Graphical representation of LDA

Θ_d= Topic distribution per document,
Φ_k = Word distribution per topic,
Z_n = Topic assignment of node n,

Dirichlet priors α and β are conjugate priors of the parameters of the multinomial distribution over topics/words.

The process of LDA is as follows:

1. Draw Φ_k from Dir(β) for each topic.
2. Draw Θ_d from Dir(α) for each document.
3. For all keywords in document: calculates posterior distribution
   a) Draw topic Zfrom Multi(Θ_d) for each keyword.
   b) Draw a word $n_d$from Multi(Φ_k) for each keyword.

By giving the parameters α & β, the joint distribution of topic mixture Θ, a set of n topics Z, and set of N words W is given by:-

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta),$$

Where,
P(Θ|α) is the probabilistic dirichlet distribution,
P(Z_n|Θ) is simply Θ_i

Integrating Θ & summing over Z, we get marginal distribution of a document as :

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta) \right) d\theta.$$

Finally take the product of probabilities of a single document, the probability of corpus is find as :

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

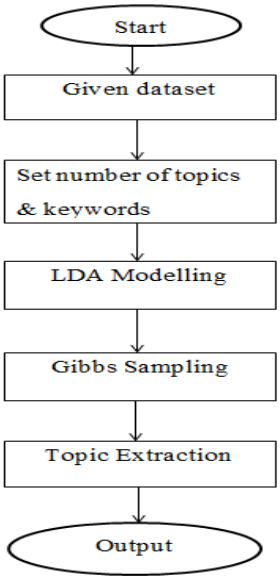
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Hence by this the latent topics can be detected.
By applying Gibbs Sampling on the above equations we can detect the latent topics for a given dataset and do the topics discovery.

System Flow:-



IV. CONCLUSION

In this paper, a LDA method is used for to find the latent topics from a given data set. A GTM model is proposed and hidden topics get detected by using LDA. Also a Bernoulli distribution has been used to find the latent topics of the dataset. The result shows that the latent topics can be find by using LDA in Graph Topic Model i.e. GTM.

REFERENCES

1. J. Paisley, C. Wang, D. Blei, and M. Jordan, "Nested hierarchical Dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 256–270, Feb. 2015.
2. S. Wang, J. Wang, and F. L. Chung, "Kernel density estimation, kernel methods, and fast learning in large data sets," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 1–20, Jan. 2014.
3. Junyu Xuan, Jie Lu, *Senior Member, IEEE*, Guangquan Zhang, and Xiangfeng Luo, *Member, IEEE*. "Topic Model for Graph Mining", IEEE transactions on cybernetics, vol.45, No. 12, Dec 2015.
4. Hau Chan and Leman Akoglu.," External Evaluation of Topic Models: A Graph Mining Approach", IEEE International Conference on Data Mining,DOI 10.1109/ICDM.2014.60,2014.
5. V. Jelisavcic, B. Furlan, J. Protic, and V. Milutinovic, "Topic models and advanced algorithms for profiling of knowledge in scientific papers," in *Proc. 35th Int. Conv. (MIPRO)*, Opatija, Croatia, pp. 1030–1035,2012.
6. J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sequential topic model for mining recurrent activities from long term video logs," *Int. J.Comput. Vis.*, vol. 103, no. 1, pp. 100–126, 2013.
7. K. Amailef and J. Lu, "Ontology-supported case-based reasoning approach for intelligent m-Government emergency response services," *Decis. Support Syst.*, vol. 55, no. 1, pp. 79–97, 2013.
8. Q. Liu, H. Huang, and C. Feng, "Micro-blog post topic drift detection based on LDA model," in *Behavior and Social Computing* (Lecture Notes in Computer Science 8178), L. Cao *et al.*, Eds. Cham, Switzerland: Springer , pp. 106–118,2013.
9. Jui-Feng Yeh , Chen-Hsien Lee, Yi-Shiuan Tan, and Liang-Chih Yu, "Topic Model Allocation of Conversational Dialogue Records by Latent Dirichlet Allocation"*Vol*. 20, *pp*. 273-297, 1995.
10. Hau Chan and Leman Akoglu, "External Evaluation of Topic Models: A Graph Mining Approach",IEEE 13th International Conference on Data Mining, DOI 10.1109/ICDM.2013.112,*vol*. 3, *pp*. 993–1022,2013.
11. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation,"Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
12. Y. Chen, Y. Yang, H. Zhang, H. Zhu, F. Tian, "A topic detection method based on Semantic Dependency Distance and PLSA," IEEE 16th InternationalConference on Computer Supported Cooperative Work in Design(CSCED), pp. 703-708 May. 2012.