Volume 6, Issue 5, May 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Pattern Enhanced Topic Model for Information Filtering

[1]B. S. Jadhav, [2]Dr. D. S. Bhosale, [3]Dr. D. S. Jadhav

[1] PG-Student, Dept. of Computer Science & Engineering, Ashokrao Mane Group of Institutions, Kolhapur, India

[2] PG-Coordinator, Dept. of Computer Science & Engineering, Ashokrao Mane Group of Institutions, Kolhapur, India

[3] Head, Faculty of Management Studies, Ashokrao Mane Group of Institutions, Kolhapur, India

Abstract— Traditionally term-based & word-based approaches used for information filtering. Topic model has used for discovering unseen topics in a set of credentials. In topic model most commonly used LDA which generates a structural model to represent multiple topics. Term-based & Word-based approaches have disadvantages which are polysemy & synonymy. The pattern mining technique used in field of topic modeling generates model for finding out more meaningful & discriminative topics from collection of documents. The pattern enhanced topic model used in the field of information filtering for finding out most appropriate user interested data from a collection of documents.

Index Terms— Information filtering, LDA, Pattern mining, Pattern enhanced topic model.

I. INTRODUCTION

Information filtering is a classification that eliminates redundant & superfluous information from a set of credentials. It is usually applied to input data which based on user interested data. If it is a method that is managing large information flows. Pattern based techniques have been operating in the vicinity of information filtering & realized a quantity of enhancements on efficiency [1]. Information system gets user interest or user information needs. The user always needs information from large amounts of data. This set of documents which consist of more useful addict's desired information is created by using information filtering system.

Topic modeling techniques it is a probabilistic model for collection of discrete data. Discrete data it is a collection of text & it has approved, by text mining & machine learning communities. It represents probabilistic distributions which break relationships between associated words. Topic model it can classify documents in a set of topics & every document produces with multiple topics & their related distributions [2]. LDA it is a generative probabilistic topic modeling techniques used for discovering topic.

The pattern enhanced topic model representing topics using patterns which make it possible to interpret the topics with semantic meanings. The pattern enhanced topic model can be used in the area of information filtering for constructing a content-based user interest modeling. A pattern mining technique utilized in a topic model which carry more semantic meaning in topic patterns in topic model. Pattern can use to represent document more accurate & meaningful.

II. TOPIC MODEL

Topic model it is probabilistic text modeling technique. Topic modeling used to discover number of topics of document where each topic is consist as a number of words. Topic model gives an interpretable representation of the document with a manageable number of topics.

Two approaches used in topic model which is Latent Dirchilet Allocation (LDA) & probabilistic latent semantic analysis (pLSA). A Probabilistic LSA model which is a generative data model can provide a solid statistical foundation. LDA it is a mixture of a tiny quantity of topics & topic & that every statement configuration is endorsed to one of the credentials topics.

LDA is a one of the examples of a topic model and it also was first offered as a graphical model for topic discovery [3].The LDA technique attracted due to its robust & interpretable topic representation. The idea behind LDA first assumes the same number of topics which are distributed over words.

In LDA D is a collection of credentials $\{d_1,d_2,\ldots.,d_m\}$ & m is total number of credentials. The representation of the LDA topic model results at three levels. At the first level is documented level in that each document d_i represented by topic distribution $\theta_{d_i} = (\vartheta_{d_{i,1}}, \vartheta_{d_{i,2}}, \ldots..\vartheta_{d_{i,v}})$. v is the hold value of number of topics. Second is collection level, D is set which consists collection of topics. Each topic consists of a probability distribution over words ϕ_j for topic j. Third is word level in that considered word occurrence into the topics.

III. PATTERN ENHANCED TOPIC MODEL

Pattern mining techniques carry semantic meaning & more understandable rather than only words. Pattern enhanced topics representation is more accurate & more meaningful rather than word based topic representations. Pattern carries more identifiable meaning.

The pattern Enhanced topic model technique is to use repeated patterns created from each transaction datasets Γ_j to represent Z_j for particular minimum support threshold σ, an itemset X which in Γ_j is considered as frequent if supp(X)>=σ, where supp(X) is indicate support of X which is the number of transaction in Γ_j that consists X. The frequency of item set X is defined $\frac{supp\,(x)}{|\Gamma_j|}$.

IV. PATTERN ENHANCED TOPIC MODEL FOR INFORMATION FILTERING

Information filtering system that removes unwanted or unnecessary information from an incoming information stream by use of automated or computerized methods [4]. Pattern enhanced topic model use in areas of the IF system to extract relevant data which based on user interest.

The idea behind pattern enhanced topic model is first find out number of topics v from a collection of documents D. After applying LDA to D construct transactional datasets which is $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_v\}$. After generating transactional datasets, find out PBTM representation for set of U= $\{X_{z1}, X_{z2} \ldots X_{zv}\}$ & each X_{zi} is collection of repeated pattern generate from transactional dataset Γ_i. After pattern enhanced topic representations find Equivalence classes which denoted as $E(Z_i)$ which is set of $\{EC_{i1}, \ldots\ldots, EC_{ini}\}$ for topic Z_i .

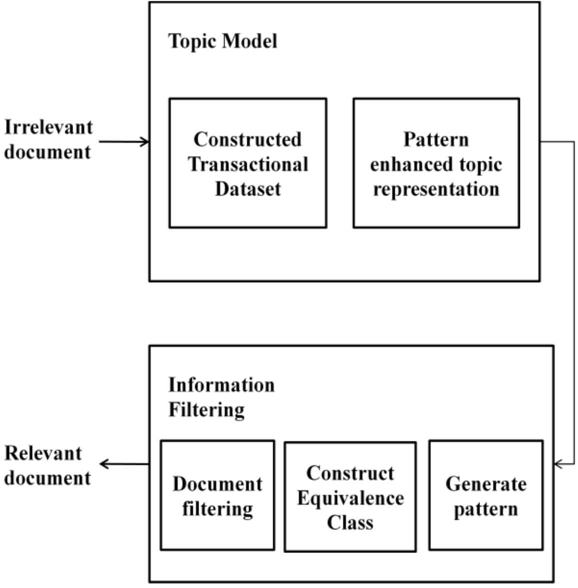

Fig. Structure of pattern enhanced topic model for information filtering.

Information filtering it is system first given irrelevant document as input. Then topic model in that firstly applies the LDA technique to the documents. Then create a new transactional dataset from a set of document. The next step is a creating pattern from transactional datasets. After generating pattern find out the most useful pattern from a set of patterns. Then most appropriate data extracted from the document. This is relevant data which is based on user interest.

V. CONCLUSION

Pattern enhanced topic model use in the area of the Information Filtering system to produce more semantic & a discriminative representation for topics. Each topic to represent by a word distribution & also document represented with topic distributions. The pattern enhanced topic model effectively increased performance of information filtering system.

REFERENCES

[1] J.Han, X.Yan, C.W.Hsu, and H.Cheng, "Discriminative frequent pattern analysis for effective classification," IEEE transaction on Data Eng., vol.7, pp. 716-725, 2007.

[2] Ms Manjiri M. More 1, Prof. Mrs Archana S. Vaidya. "A Survey Paper On Modeling Methods For Information Filtering And Relevance Ranking Of Documents". International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 09 | Dec-2015 e-ISSN: 2395 -0056, p-ISSN: 2395-0072.

[3] Andrew Y., Jordan Michael I., Lafferty John, Blei David M., "Latent Dirichlet Allocation". Journal of Machine Learning Res. 3 (4–5): pp.993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

[4] Shapira B., Hanani U., Shoval P. "Information filtering: Overview of issues, research and systems, User Modeling and User-Adapted Interaction." 11, pp. 203–259.

[5] Pallavy Nath. S , Annie George. "Semantic Pattern-Based Topics Filtering for Document Modeling", International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2015. DOI: 10.15680/IJIRCCE.2015. 0311189.

[6] Chinnu C. George, Abdul Ali. "Information Filtering Model Based on Topic Pattern for Document Modeling". International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-5 Issue-5, October 2015.

[7] Vasudevan, V.Sharmila, G.Tholkappia Arasu. "Innovative Pattern Mining For Information Filtering Systems". International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 4, October 2012.

[8] Tincy Chinnu Varghese, Smitha C Thomas. “Pattern Enhanced Topic Model”. International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 4, Issue 1, pp: (189-194), Month: January - March 2016, Available at: www.researchpublish.com.

[9] Ezeife C. I., Mabroukeh N. R.,"A taxonomy of sequential pattern mining algorithms". ACM Com. Surv. 43: 1–41. doi:10.1145/ 1824795. 1824798.

[10] H.M.Wallach, “Topic modeling: Beyond bag-of- words,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 977–984.

[11] S.T. Wu, Y. Li, and N. Zhong, “Effective pattern discovery for text mining,” IEEE Transaction on Knowledge Data Eng., vol. 24, no. 1, pp. 30–44, Jan. 2012.

[12] S. T. Wu, Y. Li, and Y. Xu, “Deploying approaches for pattern refinement in text mining,” in Proc. 6th Int. Conf. Data Min., 2006, pp. 1157–1161.

[13] Y. Xu, and Y. Li, “Deploying approaches for pattern refinement in text mining,” in Proc. 6th International Conference on Data Mining, 2006, pp. 1157–1161.

[14] T. Minka, J. Callan, and Y. Zhang, “Novelty and redundancy detection in adaptive filtering,” in Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2002, pp. 81–88.

AUTHOR'S BIOGRAPHY

B. S. Jadhav was born in India, Maharashtra, in 1992. He completed a Bachelor of Engineering in Computer science & engineering from Shivaji university, Kolhapur (MS) in academic year 2011-2014. He is pursuing a Master of Engineering in Computer science & engineering from Ashokrao Mane Group of Institutions, Kolhapur (MS).

D. S. Bhosale he was received Master of Technology in Computer science & engineering. Also completed PhD in Computer science. He was worked as a Professor at Anna Dange College of Engineering & Technology, Sangli,(MS). At present he is working as Professor & Head of PG studies of computer science department in A.M.G.O.I.,Kolhapur (MS). He has published more than 20 papers in International and National journals and conference Proceedings. He is member of various National & International Professional Bodies and member of Editorial / Reviewer of various International Journals. His research interest includes Data Mining, Software Testing.

D. S. Jadhav was born in India, Maharashtra, in 1979. He received the PhD, MCA, MBA, MTech (CSE) degrees from Solapur University, Solapur, University of Pune (MS) and NIMS University, Rajasthan respectively. From 2008-2010 he was worked as lecturer Smt. K. W. College, Sangli. From 2010 to 2013 he was worked as Asst. Professor at Bharat Ratna Indira Gandhi College of Engineering, Solapur (MS) and Sinhgad Institute of Computer Sciences, Pandharpur(MS) respectively. From Oct 2013 to 2015 he was working as I/C Director at Ideal Institute of Management(IIMK), Kondigre – Ichalkaranji (MS). From 2015 to till date he is working as Head, Faculty of Management Studies at Ashokrao Mane Group of Instituttions, Vathar Tarf Vadgaon, Kolhapur (MS). He has published more than 30 papers in International and National journals and conference Proceedings. He is member of various National & International Professional Bodies and member of Editorial / Reviewer of various International Journals. His research interest includes Cyber Crime, Cyber / Computer Forensic, Information Security, Business Management.