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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Pattern Based Topics for Document Modelling Using HLA

Padmaja .R.C[1], Venkatesan.N[2]

PG Scholar, Dept.of CSE, Sri Vidya College of Engineering & Technology, Virudhunagar, Tamil Nadu, India[1]

Assistant Professor, Dept. of CSE, Sri Vidya College of Engineering & Technology, Virudhunagar, Tamil Nadu, India[2]

ABSTRACT: Different mature term-based or pattern-based approaches have been used in the area of information filtering to generate users' information needs from a group of documents. A primary assumption for these approaches is that the documents in the collection are all about only one topic. Topic modelling is a type of text mining, a way of finding patterns in a corpus. By process the corpus and run it through a tool which groups words across the corpus into relevant 'topics'. Latent Dirichlet Allocation was proposed to generate statistical models to represent multiple topics in a collection of documents, and this has been extensively utilized in the fields of machine learning and information retrieval, and so on. But its effectiveness in information filtering has not been so well explored. Patterns are always considered to be supplementary discriminative than single terms for describing documents. However, the massive amount of revealed patterns hinders those from being effectively. Therefore, selection of the most discriminative and representative patterns from the vast amount of discovered patterns becomes essential. To compact with the above mentioned limitations, we propose some high level algorithms. The main distinctive features of the proposed model includes user information needs are generated in terms of multiple topics, each topic is represented by patterns, patterns are generated from topic models and are organized in terms of their statistical and taxonomic features.

KEYWORDS: Topic modelling, pattern mining, Text Mining, Relevant Ranking Algorithm

I. INTRODUCTION

Early on topic model was illustrated by Papadimitriou, Raghavan, Tamaki and Vempala in the year of 1998. An alternative approach called Probabilistic latent semantic indexing (PLSI), was formed by Thomas Hofmann in the year of 1999.Latent Dirichlet allocation (LDA) is one of the most common topic models currently in use. It allows documents to have a multiple topics. Other topic models are generally extensions on LDA, such as Pachinko allocation, which improves on LDA by modelling correlations between topics in addition to the word correlations which constitute topics.

Topic modelling has become one of the most popular probabilistic text modelling techniques and has been promptly accepted by machine learning and also the text mining society. It can automatically classify documents in a collection by a number of topics and represents every document. In machine learning and natural language processing, a topic model is a form of statistical model, it discovers topics that suggest itself in a collection of documents. By instinct, known set of document is about a particular topic; one would expect particular words to appear in the document more or less frequently: "apple" and "computer" will appear more often in documents about apple personal and laptop computers. A document typically may have multiple topics in different extent; thus, in a document that is 20% about apples and 80% about computers, there would probably be about 8 times more computer words than apple words. A topic model captures this perception in a mathematical structure, which allows groping a set of documents and discovers, based on the statistics of the words in each documents.

An Information filtering system is a scheme which is used to remove redundant information as well as the unwanted information in an information stream using the semi automated methods or by some computerized methods. The major goal is the supervision of the information's huge load and growth of the semantic signal-to-noise ratio. To accomplish this, the user's profile is compared to some of the reference individuality. This individuality may initiate from the information item such as content-based approach or the user's social environment such as the collaborative filtering

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

approach. While in information transmission signal processing filters are used beside syntax-disrupting noise on the bit-level, the methods engaged in information filtering perform on the semantic level. The collection of machine methods engaged building on the similar principles as those for information extraction.

At present the major problem is not discovering the best method to filter information. To achieve these systems need to gain knowledge of independently the information needs of users. Not since they automate the process of filtering but it construct and also adapt the filter. Several branches based on it, say statistics, machine learning, pattern recognition text mining and data mining, are the foundation for increasing of development of information filters that appear and adapt in base to familiarity. To let the learning process can be done; part of the information has to be pre-filtered. It says that there are positive and negative examples which we named training data, which can be created by experts, feedback from side to side of the usual users.

II. RELATED WORK

Sailaja.G, et.al (2014) proposed a methodology for frequent term measuring from the documents based on text clustering. Text clustering is the most important process in text mining .It referring to the process of grouping document with similar contents or topics into clusters. This improvises the availability as well as reliability of the mining. The main idea is to apply any obtainable frequent item finding algorithm such as a prior, Dp-tree to the initial set of text files. This will reduce the dimension of the input text files. The document feature vector is created for all the documents, then a vector is formed for all the static text input files. The algorithm outputs a set of clusters from the initial input of text files. The Proposed algorithm has the input as similarity matrix and output a set of clusters as compared to other clustering algorithms. In this work, frequent items are generated using APRIORI approach by a similar method. We can replace a priori algorithm by any frequent item Finding algorithm. The algorithm for clustering considers the set of frequent items generated from all the Documents. This gives the commonality between Document pairs. The count of frequent items serves as the Distance measure.

Murali Krishna.S, et.al (2010) evaluated the performance of various methodologies for improving text clustering. Text clustering is the most important process in text mining. This sub-section, describes how clustering is done on the set of partitions obtained from the previous step. This step is necessary to form a sub cluster (describing sub-topic) of the partition (describing same topic) and the outlier documents can be significantly detected by the resulting cluster. In this paper, we have conducted an extensive analysis of frequent item set-based text clustering approach with different text datasets. For different text datasets, the performance of item set based text clustering approach has been evaluated with precision, recall and F-measure. The experimental results of the item set based text clustering approach are given for Reuter newsgroups and Webkb datasets. The performance study of the text clustering approach showed that it effectively groups the documents into cluster and mostly, it provides better precision for all datasets taken for experimentation.

Yuanfeng Song, et.al (2014) analyzed significant frequent patterns for the effective ranking of documents in a collection. Ranking documents in terms of their relevance to a given query is fundamental to many real-life applications such as information retrieval and recommendation systems. In this paper, we present a theoretical analysis on which frequent patterns are potentially effective for improving the performance of LTR, and then propose an efficient method that selects frequent patterns for LTR. First, we define a new criterion, namely feature significance (or simply significance). Specifically, we use each feature's value to rank the training instances, and define the ranking effectiveness in terms of a performance measure as the significance of the feature.

Xing Wei and W. Bruce Croft, (2012) proposed Latent Dirichlet Allocation Based Document Models for Ad-hoc Retrieval. An approach to building topic models based on a formal generative model of documents, Latent Dirichlet Allocation is heavily cited in the machine learning literature, but its feasibility and effectiveness in information retrieval is mostly unknown. In this paper, we study how to efficiently use LDA to improve ad-hoc retrieval. We propose an LDA-based document model within the language modelling framework, and evaluate it on several TREC collections. Gibbs sampling is employed to conduct approximate inference in LDA and the computational complexity is analyzed. We show that improvements over retrieval using cluster-based models can be obtained with reasonable efficiency.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

LDA-based retrieval can potentially be used in applications where pseudo-relevance feedback would not be possible. In summary, LDA-based retrieval is a promising method for IR, although more work needs to be done with even larger collections, such as the Web data from the TREC Terabyte track.

Sheng-Tang Wu, et.al (2006) depolyed some approaches for pattern refinement in text mining. Instead of the keyword-based approach which is typically used in this field, the pattern based model containing frequent sequential patterns is employed to perform the same concept of tasks. However, how to effectively use these discovered patterns is still a big challenge. In this study, we propose two approaches based on the use of pattern deploying strategies. The performance of the pattern deploying algorithms for text mining is investigated on the Reuters dataset RCV1 and the results show that the effectiveness is improved by using our proposed pattern refinement approaches. In this study we propose two pattern refinement methods to deploy the discovered patterns into a feature space which is used to represent the concept of documents. Our methods adopt the mining sequential pattern technique to find semantic patterns from text documents and then deploy these patterns using proposed deploying algorithms.

III. LITERATURE REVIEW

A literature study has been made on the different algorithms and various methodologies to produce a pattern based topic assigning for the documents in the collection. Every author has their own perspective for the document modelling and naming the documents. By analysing the views of authors we may get a efficient way for topic assigning process. The table following shows the literature study for pattern based topic document model.

AUTHOR & YEAR	TITLE	METHODOLOGY	DISADVANTAGES
S.Murali Krishna, et al., 2010	An Efficient Approach for Text Clustering Based on Frequent Itemsets	In the proposed research, they have invented an efficient advanced methodology for text clustering based on the item sets used frequently. The renowned method, called as priori algorithm is used for mining the item sets used frequently. The mined item sets are then used for obtaining the partition, where the documents are initially clustered without overlapping. The access to a large quantity of textual documents turns out to be effectual because of the growth of the digital libraries, web, technical documentation, medical data and more.	It does not require a pre-specified number of clusters
G. sailaja, et al., 2014	A Novel Similarity Measure for frequent Term Based Text Clustering on high dimensional data	*Sailaja.G, et.al (2014)* proposed a methodology for frequent term measuring from the documents based on text clustering. The main idea is to apply any obtainable frequent item finding algorithm such as a prior, Dp-tree to the initial set of text files. The document feature vector is created for all the documents, then a vector is formed for all the static text input files. The algorithm outputs a set of clusters from the initial input of text files. The Proposed algorithm has the	This does not satisfy the second condition of a metric. Since after all the combination of two copies is a different object from the original document.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

		input as similarity matrix and output a set of clusters as compared to other clustering algorithms. In this work, frequent items are generated using APRIORI approach by a similar method. We can replace a priori algorithm by any frequent item Finding algorithm. The algorithm for clustering considers the set of frequent items generated from all the Documents. This gives the commonality between Document pairs.	
Florian, et al. 2002	Frequent Term-Based Text Clustering	This paper introduced a new approach which uses frequent item (term) sets for text clustering. Such frequent sets can be efficiently discovered using algorithms for association rule mining. By measuring the mutual overlap of the frequent sets with respect to the sets of supporting documents. It can be used to structure large sets of text as well as hypertext documents.	Very high dimensionality of the data, this requires the ability to deal with sparse data spaces or a method of dimensionality reduction.
Anton Bakalov, et al. 2012	Topic Models for Taxonomies	This paper introduces two semi supervised topic models that automatically enlarge a given classification with numerous supplementary keywords by leveraging a corpus of multi-labeled documents. The models provide a better information rate compared to Labeled Latent Dirichlet Allocation.	The concept node names often do not de-scribe the concept in sufficient detail for unfamiliar users to fully understand the topics a node is intended to capture.
Wei Li, et al. 2008	Pachinko Allocation: Scalable Mixture Models of Topic Correlations	This paper proposed the pachinko allocation model (PAM). This confines arbitrary topic correlations using a directed acyclic graph. The leaves of the directed acyclic graph represent individual words in the vocabulary. While every interior node represents a correlation among its children, which may be words or other topics. Latent Dirichlet Allocation's performance peaks at 40 topics and decreases as the number of topics increases.	It is not feasible to perform exact inference in this model.
Christopher, et al. 2011	Modeling the Ownership of Source Code Topics	This paper combines both software repository mining and topic modeling. This helps to measure the ownership of linguistic topics in source code.	It does not model correlations among topics.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

		They conducted an exploratory study of the relationship between linguistic topics and ownership in source code using 10 open source Java systems. pachinko allocation model allows an arbitrary DAG to model topic correlations.	
Haiping Maa, et al. 2012	Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning	Multi-label learning approach on text data, they applied the proposed model for inferring the training data. The standard Four-Level Pachinko Allocation Model for the test data. In addition to that they proposed pruned Gibbs Sampling algorithm in the test stage, which is used to reduce the inference time. In this approach the computational efficiency is high compare to other high-performing multi-label learning methods.	Latent Dirichlet Allocation does not perform very well in supervised settings
Ivan Titov, et al. 2008	Modeling Online Reviews with Multi-grain Topic Models	In this paper authors' presented a new framework for extracting the ratable aspects of objects from online user reviews. All the models are based on extensions to standard topic modeling methods such as Latent Dirichlet Allocation. The performance will improve considerably upon standard topic models.	It does not model the appropriate aspects of user reviews.
S.Durga Bhavani , et al. 2014	Performance Evaluation of an Efficient Frequent Item sets-Based Text Clustering Approach	A great deal of attention in research community has been received by the use of such common item sets for text clustering. Since the dimensionality of the documents is severely reduced by the mined recurrent item sets. Based on recurrent item sets, an efficient approach for text clustering has been developed. It can be engaged to look through a set of documents or to arrange the results given by a search engine in answer depends upon the users query.	Clusters are not required by this text clustering approach.
Hong Cheng, et al. 2007	Discriminative frequent Pattern Analysis for Effective Classification	In this paper, They conducted a systematic exploration of frequent pattern based classification. It provides hard reasons sustaining this tactic. It is useful for classification by mapping data to a higher dimensional space.	It is computationally intractable to enumerate them when the number of single features is large
Mohammed	CHARM: An Efficient	In this paper, they presented	It cannot be extended at the same

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

J, et al. 2015	Algorithm for Closed Itemset Mining	CHARM. That is a well-organized algorithm for mining all frequent closed item sets. It catalogues the closed sets using a dual itemset and tidset search tree. By means of an efficient mixture search that bounce many levels. It also uses a method named as diffsets to reduce the memory footprint of intermediate computations. Computing the supports is simpler and faster. Only intersections on tidsets is required, which are also well-supported by current databases.	time

Table 1 Literature Study

IV. PROPOSED WORK

In proposed system we are going to introduce some of the high level algorithms. This helps to assign the most relevant topics to the documents in the collections.

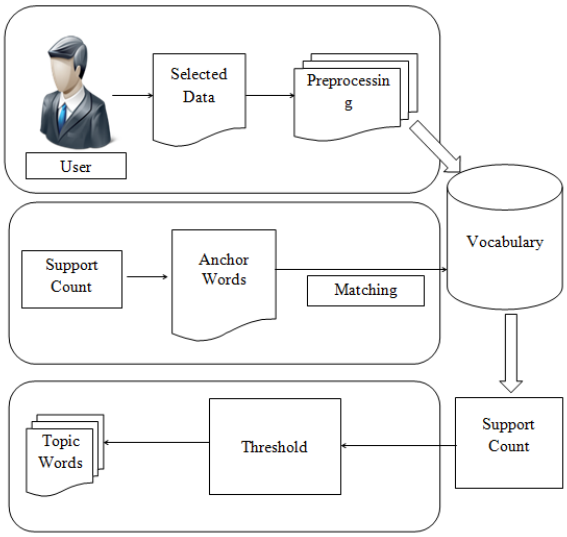

Fig. 1. Architecture of a proposed system

The dataset is given as the input to the system. All the words from the documents are gathered is called the vocabulary. Then find the support value for all the words in the list. From this the anchor words are calculated. The anchor words are the high frequency words in the document. The anchor words are getting from the list and find support from the vocabulary. From this the topics are generated.

V. CONCLUSION

This paper proposed High level algorithm for topic modelling. The proposed model consists of topic distributions describing topic preferences of documents or a document collection. And structured pattern-based topic representations representing the semantic meaning of topics in a document. The proposed new model automatically generates discriminative and semantic rich representations for modelling topics. And documents by combining statistical topic modelling techniques and data mining techniques.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

ACKNOWLEDGMENT

My deepest gratitude is to my advisor, Mr.N.Venkatesan M.Tech., I have been amazingly fortunate to have advisors who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. They taught me how to question thoughts and express ideas. Their patience and support helped me overcome many crisis situations and finish this work.

REFERENCES

[1] S.Murali Krishna and S.Durga Bhavani," An Efficient Approach for Text Clustering Based on Frequent Itemsets", European Journal of Scientific Research., ISSN 1450-216X Vol.42 No.3 (2010), pp.385-396.
[2] G. sailajaP and B.PrajnaP," A Novel Similarity Measure for frequent Term Based Text Clustering on high dimensional data ", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.
[3] Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey, "A Text Mining Technique Using Association Rules Extraction", International Journal of Computational Intelligence, Vol. 4; No. 1, 2008.
[4] Hong Cheng, Xifeng Yan, Jiawei Han, Chih-Wei Hsu," Discriminative Frequent Pattern Analysis for Effective Classification", U.S. National Science Foundation NSF IIS-05-13678/06-42771 and NSF BDI-05-15813.
[5] I. H. Witten and E. Frank," *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann", 2nd edition, 2005.
[6] Yuanfeng Song, Wilfred Ng, Kenneth Leung, and Qiong Fang," SFP-Rank: Signi_cant Frequent Pattern Analysis for E_ective Ranking", Department of Computer Science and Engineering ,The Hong Kong University of Science and Technology, Hong Kong.
[7] Mohammed J. Zaki and Ching-Jui Hsiao," CHARM: An Efficient Algorithm for Closed Itemset Mining", Computer Science Department, Rensselaer Polytechnic Institute, Troy NY 12180.
[8] Yue Xu, Yuefeng Li, Gavin Shaw," A Reliable Basis for Approximate Association Rules", IEEE Intelligent Informatics Bulletin November 2008 Vol.9 No.1.
[9] Xing Wei and W. Bruce Croft," LDA-Based Document Models for Ad-hoc Retrieval", Computer Science Department, University of Massachusetts Amherst, MA 01003,@cs.umass.edu
[10] Florian Beil,Martin Ester,Xiaowei Xu," Frequent Term-Based Text Clustering", SIGKDD 02 Edmonton, Alberta, Canada Copyright 2002 ACM 1-58113-567-X/02/0007.
[11] Mittar Vishav, Ruchika Yadav, Deepika Sirohi," Mining Frequent Patterns with Counting Inference at Multiple Levels", International Journal of Computer Applications (0975 – 8887) Volume 3 – No.10, July 2010.
[12] Sheng-Tang, Wu Yuefeng, Li Yue Xu," Deploying Approaches for Pattern Refinement in Text Mining",IEEE computing society, Proceedings of the Sixth International Conference on Data Mining (ICDM'06) 0-7695-2701-9/06.
[13] Saman Hina, Eric Atwell, Owen Johnson," Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard", IJICR, Volume 2, Issues 1/2/3/4, Mar/Jun/Sept/Dec 2011