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 10, October 2015

A Survey on Topics Modeling Methods over Information Filtering

V. Vishnupriya[1], S. K Somasundaram[2]

PG Scholar, Dept. of CSE, PSNA college of Engineering and Technology, Dindigul , India[1]

Assistant Professor, Dept. of CSE, PSNA college of Engineering and Technology, Dindigul, India[2]

ABSTRACT*:* In recent years, the appearance of Information filtering is a rare method of web routine data mining has become an interesting explores area. Information filtering is a system that removes surplus information from an information torrent using computerized methods prelate to staging to a human user. Information filtering is planned for formless or semi structured data, as dispute to database applications, which use very precise data. Filtering is based on beyond words of individual or group of preferences, often called profiles. Filtering also understood removal of data from an incoming stream rather than finding data in an stream; users only see the data is remove or take out and it concerned large amount of data. The goal of information filtering is to filter out the irrelevant data items. This paper contain survey on various topic modeling used for document rupture which gives the definitiveness results of these topic that have been carried out on the different types of modeling.

KEYWORDS: Information Filtering Systems, Topic Models, Methods of Topic modeling User's profile, User interest model.

I. INTRODUCTION

Information filtering system is a reputation used to illustrate a variety of measures involving the approach of information to people who need it. It is only by making that dissimilarity, however that the specific explore issues associated with filtering can be searched out and numbered. Recommendation systems are a subclass of information filtering that seeks to forecast the 'rating' or 'preference' that a user would give to an item. Content-based information filtering is based on descriptions of the item and a profile of the user's preference. In particular various applicant items are evaluated within the past rated by the users and the best matching items are favored. Information filtering extents primarily with textual Information. In fact, unstructured data is often used as a analogue for textual data.

A. Information Filtering System Architecture

Information filtering system consists of a filtering agent and user's profile.

Filtering Agent

Filtering Agent acts as an edge between the user and document system, and helps the user in finding the significant topics of a given topic through the user proxy. It decreases the user's period and effort in discovering the relevant document through the focused domain knowledge it possess. It cast include pool with the source document subsystem managing the user-profile is calculating the relevance of a document-vector to the user-profiles and communicating with the user.

The process flow is associated with the filtering agent :

1. The topic for the current document filtering session is obtained from the users.

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 10, October 2015

2. User profile vector is initialized using the current topic's title and text.
3. Document vector d is obtained from source document subsystem.
4. Current user-profile vector is reintegrated.
5. The following probabilities are calculated
   Probability of Relevance (PR) =Probability(Relevant/d;t)
   Probability of Non-Relevance(PN)=Probability(Non-Relevant/d;t)

6. If PR>PN, then for the given relevant document d :
   (a) Text of d is obtained from source document subsystem
   (b) User is informed text of d
   (c) Actual relevance judgment is obtained from the user for the document d
   (d) User-profile vector t is updated with the actual relevance judgment of d
7. Repeat from step 3.

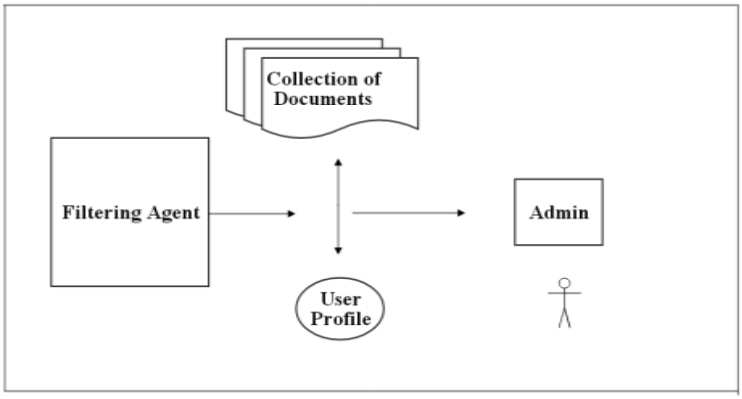

Figure 1 : Information Filtering System Architecture

User Profiling

User profile represents user details and is continuously updated in response reply. The quality of user's profiles has a major impact on the performance of information filtering systems. The user has information that can differentiate one user from a multitude of other users. Profiles normally include topics of interest but may also include topic of interest by making into account relevant and irrelevant documents.

Examples of User profiling
Name : John
User Name : John
Password : **********
E-mail id : John@gmail.com
Topic of Interest : DataWarehouse

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 10, October 2015

II. THE VARIOUS METHODS OF TOPIC MODELING

In this survey [10] we will dispute about some of the topic modeling's method that deals with words, documents and topics. Also, these methods involve in many application so it will have a brief thought in what applications that can see these approaches works with.

A. Latent Semantic Analysis

Latent Semantic Analysis [20][21]is a method or technique in the divisional of natural language processing. The main goal of Latent Semantic Analysis is to produce a vector based representation for text's to make linguistic content. By vector representation computes the similarity between the text's to pick the highest proficient related words. In the past latent semantic analysis was named as latent semantic indexing but perfected for information retrieval tasking. So, finding recent documents that doubt to the query that given from many documents. Latent Semantic Analysis should have many aspects to given approach such as keywords matching and vector representation depends on occurrences of words in documents. Latent semantic analysis uses a singular value decay to rearrange the data. It is a method that uses a matrix to reconfigure and calculate all the diminutions of vector space. By investigating about words that have a high rate of similarity will be occurred if that words have similar vector.

B .Probabilistic Latent Semantic Analysis

Probabilistic [3] is based on a statistical model that is referred to as an aspect model. An aspect model is a latent variable model for co-occurrence data, which associates unobserved class variable. Probabilistic Latent Semantic Analysis is a method that can be preset document indexing which is based on statistical latent class model for influence analysis of calculation data and also this method to improve the latent semantic analysis in a probabilistic sense by using a procreative model. The goal of probabilistic latent semantic analysis is that identifying and distinguishing between various contexts of words usage without alternative to a dictionary. It is based on two important implications: First one, it is allows to demonstrate the polysemy, Second one, is reveal the topic parallels by grouping stable words that shared a common vocabulary. Probabilistic latent semantic analysis method comes to improve the method of latent semantic analysis. It has been successfully developed in many real-world applications, including computer vision, and recommender systems. Probabilistic Latent Semantic Analysis suffers from over fitting problems.

In probabilistic model it introduces a latent variable zk € { z1,z2,….zk},which corresponds to a potential semantic layer. The full model p(di) on support of the document in the data set the probability; p(wj|zk) zk representatives as clear semantics, the connected term of the opportunities are many;p(zk|di) represents a semantic document distribution. Using this distribution these model will generate a new data by the following steps:

1. Select a document di with probability p(di)
2. Pick a latent class zk with probability p(zk|di)
3. Generate a word wj with probability p(wj|zk).

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 10, October 2015

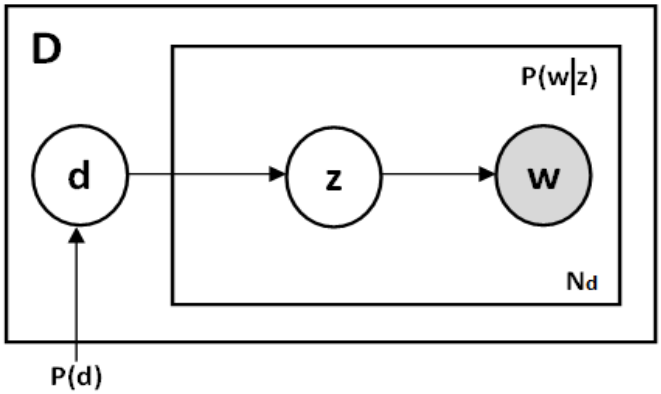

Figure 2 : High Level View of Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis has two different formulations that can instant this method. The first formulation is symmetric formulation, which will help to get the word and document from the latent class in similar ways. The second formulation is asymmetric that is each document d a latent class is selected conditionally to the document according to the word can be generated from the class. These two formulations have rules and algorithms which have been used for different utilities This method is based on Recursive probabilistic Analysis it is an extension of probabilistic latent semantic analysis.

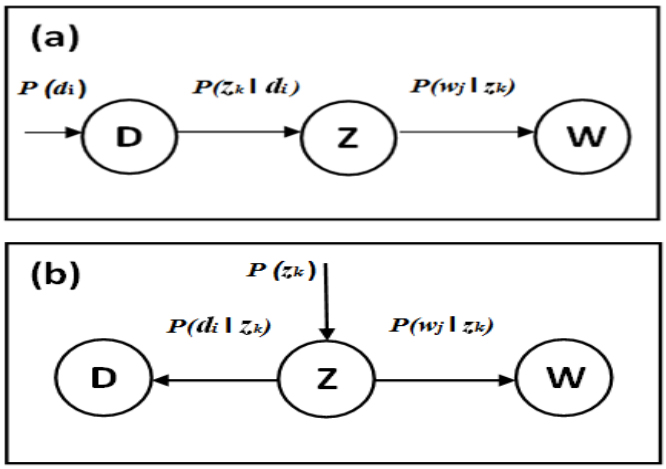

Figure 3: (a) Symmetric (b) Anti Symmetric

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation [8][12]is an algorithm for text mining that is located on analytical topic models and it is widely used. Latent Dirichlet is a generative model that means it efforts mimics so it to generate a document based on given topic. It can also be applied to others types of data. There are different types of latent dirichlet allocation models including temporal topic model, supervised topic model, latent dirichlet co-clustering and latent dirichlet based bioinformatics. In this model each document described as a mixture of topics and each topic described as a probability distribution that defines how likely each word is given to a topic. Here a "document" is represented as a "bag of words" with no structure further

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 10, October 2015

than the topic and word statistics. Latent Dirichlet model each of D documents as a combination over K latent topics, each of describes a multinomial distribution over a W word vocabulary

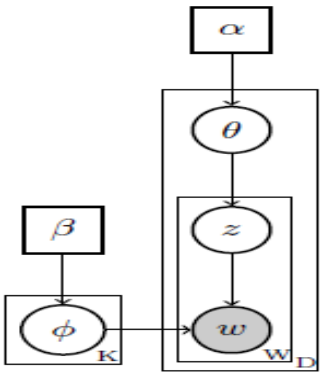

Figure 4 : A graphical model representation of LDA

Example of Latent Dirichlet Allocation

It is used to provide an descriptive example of the use of an LDA model on actual data. By using the subset of TREC AP Corpus containing 16,000 documents. In first it take away the stop words in TREC AP Corpus before continuously topic modeling and then after the use of EM algorithm to search the dirichlet and provisional multinomial parameters for a hundred topic .The top words from some of the resulting multinomial distributions are illustrated.
.

"ARTS"	"BUDGET"	"CHIDREN"
New	Million	Children
Film	Program	Women
Play	Federal	Families
Actress	Governement	Care
First	State	Family
Musical	Year	Work

Figure 5 : Mostly likely words from four topics in LDA from the AP Corpus

D. Correlated Topic Model

Correlated topic model [4][7]is a type of statistical model used in machine learning, it is used to discover the topics and it is advertised in the group of documents. Logistic normal distribution is the keynote for correlated topic model and it is

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 10, October 2015

depends on latent dirichlet allocation and simplex that allows for a general design of variability between the workings by transforming a multivariate normal random variable .The random variables takes the specific value and exactly one component equal to one. The mapping between the mean parameterization and the natural parameterization is calculated .The Strong independence assumption imposed by the Dirichlet in LDA is not realistic when analyzing document collections, where one may find strong correlations between topics.

III . METHODS ABOUT TOPIC EVOLUTION MODELS

A. Overview of Topic Evolution Model

This is top-up important to model topic evolution,[3] so user can search topics within the situation and see how topics obtain over time. This section appraisal various important papers that model topic evolution. These papers model topic evolution by using different models, but all of them consider the most important factor time when model topics. For example ,probabilistic time series models are used to handle the issue in paper "dynamic topic models" and non-homogeneous Poisson process and multiscale analysis with Haar wavelets are employed in paper "multiscale topic tomography" to model topic evolution.

B. Non-Markov continuous –Time Method

In this method [17][18]a topic is considered being associated with a continuous distribution over time. In Topic Over Time, for each document multinomial distribution over topics is exampled from dirichlet words, are generated from multinomial of each topic generates the document time stamp's. A Topic Over Time, a topic model that explicitly models time mutually with word co-occurrence patterns. Significantly, and unlike some recent work with similar goals and this model does not discretize time. When a powerful word co-occurrence pattern appears for a brief moment in time then disappears, topic over time will create a topic with a narrow time distributions. It discovers topics with both time-localization and word-clarity improvements over latent dirichlet allocation and it are turned not only by the word co-occurrences, but also temporal information.

C. Dynamic Topic Models

Dynamic topic models[1][2]estimates topic distribution at different epochs. It uses Gaussian prior for the topic parameters instead of dirichlet prior ,and capture the topic evolution over time slices. Dynamic Topic models are classifies with the continuous time dynamic topic model. This model is based on the frequent processes that uses a Brownian motion to model the latent topics through a sequential collection of documents, where a "topic" is a pattern of word use that we expect to evolve over the course of the collection. Continuous time dynamic topic model which is an extension of discrete dynamic topic model.Topic models such as latent dirichlet allocation and the more general discrete component analysis posit that a small number of distributions over words called topics can be used top explain the observed collection.

D. Multiscale Topic Tomography

In this method the document collection[22] is stored in the ascending order and that the document collection is grouped into equal-sized chunks, each of which represents the document one epoch. Each document in a epoch is represented by a word-count vector ,and each epoch is associated with its poisson generate parameters, each of which represented the expected words count from a topic. It is used to model the evolution of topics with time of powered value in automatic summarization and analysis of huge document collections. It employs non-homogeneous poission process to model generation of word-counts. The evolution of topics modeled through a multi-scale analysis using Haar wavlets. One of the recent features of the model is its modeling the evolution of topics at various time-scales of resolution, allowing the user to zoom in and out of the scales.

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 10, October 2015

E. Summary of Topic Evolution

This paper summarizes the main characteristics of topic evolution models as mentioned in the topic modeling and topic evolution models.

Comparison of Two Categories

The main difference between the two categories is that the first category model topics with considering time and methods in the first category mainly involve model words. The methods in second category model topic in considering time, by using continuous-time, (or) by discretizing time or by combining time discretization relationship. Due to the different characteristics of these two categories, the methods in the first category are more accurate in terms of topic discovery.

IV. CONCLUSION

The topic model contains cluster of words with similar meanings and text,it contains different terms of topic modeling. In this paper we have discussed the general idea about the four types of topic modeling with their advantages and disadvantages over the other. It also includes model topics with taking into account time based on user interest model and it will cofound the topic discovery. Furthermore it has been mentioned in the some of the applications that have been involved in these methods. It has been study on the various topic modeling by analyzing the accuracy of latent dirichlet model. We have seen that the further performance has been verified through the various four types of topic models. The comparison of different topic models features is essential to design a new proposal for information filtering based on user interest model. All of these models discussed in this paper considers the time as a most vital factor.

REFERENCES

[1] Blei,D.M and Lafferty,J.D " Dynamic Topic Models",proceedings of the 23rd international conference on machine learning ,Pittsburgh,PA 2006.
[2] Chong.W and Blei,D.Heckerman, "Continuous Dynamic Topic Models", in Computer Science Department in Princeton University Princeton NJ 08540.
[3] Liu.S., Xia.C.,and Jiang,X., "Efficient Probabilistic Latent Semantic Analysis with Sparsity control",IEEE International Conference on Data Mining,2010,p.no 905-910.
[4] Blei M.David "Probabilistic Topic Models", Communications of the ACM,vol.55 no.4,pp.77-84.
[5] Blei,D.M and Lafferty,J., " A Correlated Topic Model of Science",Ann,Application,stat.,Sep 2007,pp 17-35
[6] Rosen-Zvi,.M,Chemudugunta,C. Griffiths,T.Symth.P., and Steyvers,.M.2010. Learning author topic models from text corpora ACM Transactions on Information Systems.pp.1-38.
[7] Rosen-Zvi,.M., Griffiths,T.,Steyvers.M,and Symth.p.2004, " The author topic models for authors and documents". In U'04: Proceedings of the 20th Conference on uncertainity in artificial intelligences,pp.484-494,2004.
[8] Wallach.H. David, "Topic modeling : Beyond bag of words" in proceedings of the 23rd international conference on Machine learning(2006).
[9] Jelisavcic.V,Furlan.B,.Protic ,J.Milutinovic.V. " Topic models and advanced algorithms for profiling knowledge" in scientific papers.MIrPRO,2012 proceedings of the 35th international convention vol.no 25.pp.1030-1035 may 2102.
[10] I.Sato and H.Nakagawa. "Bayesian Document Generative Model with Explicit Multiple Topics" In Proc of joint conference on empirical methods in natural language processing and computational natural learning.pp 421-429,2007.
[11] D.Gildea,and T.Hoffman. "Topic-Based Language Modeling using em",In Proceedings of the 34th European Conference on Speech Communication and Technology(EuroSpeech)2009.
[12] T.P.Minka and J.Lafferty, "Expectation-propogation for the generative aspect model",in uncertainty in Artifical intelligence 2010.
[13] Wei.X and Croft,W.B., "Lda-based document models for the ad-hoc retrieval",In SIGIR'09 proceedings of the 2009 international ACM SIGIR Confernce on research and development in Information Retrieval,Newyork,NY,USA,ACM,pp 178-185.
[14] Griffiths.T.L,and Steyvers.M, "Finding Scientific topics", in proceeding of the Natl Acad Science USA,101,pp.5228-5235.
[15] Zhai.C and Lafferty.J. " A study of Smoothing methods for language models applied to adhoc information retrieval".in proceedings of the ACM SIGIR,pp.334-342.2001
[16] Deerweste,.S,Dumais.S.T.Furnas,G.W.Landauer,T.K and Harshman.R., " Indexing by latent Semantic Analysis." Journal of the American society for information science ,pp 391-407.
[17] Steinbach M,.Karypis G.,Kumar.V, " A Comparison of Topic Clustering Techniques" in proc.Text Mining Workshop,KDD 2000.
[18] Wang,X, and A.McCallum, Topics over time: A Non-Markov Continuous-time model of topical trends.In SIGKDD,2006.
[19] Wang,.C,Meek Christopher and Thiesson.B, "Markov Topic Models" in Princeton University at the department of Computer Science Engineering.pp.145-157.

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 10, October 2015

[20] Landauer.T,K.,Folkz P,W.,& Laham,D., " Introduction to Latent Semanitc Analysis",disclosure processs,vol.25,pp.259-284.
[21] Darrell.L.,Thomas,K.L.Thomas,Peter Foltz, " Latent Semantic Indexing",at the university of Colorado,boulder state university.pp.1-37.
[22] Ramesh Nallapati.William.C.Susn Ditmore, "Multiscale Topic Tomography",at the university of Carnege Mellon,pp.36-65.

BIOGRAPHY

VishnuPriya Vijaya Raghavan is a PG Scholar in the Department of Computer Science and Engineering in PSNA College of Engineering and Technology, Anna University. She received Bachelor of Engineering (BE) degree in 2014 from Anna University ,Chennai, India.

S K Somasundaram received his B.E. degree in Computer Science and Engineering from PSNA College of Engineering and Technology, Dindigul in 2003 and M.E. degree in Computer and Communication from PSNA College of Engineering and Technology, Dindigul in 2007 and pursuing Ph.D. at Anna University, Chennai. In 2003, he joined the Department of Computer Science and Engineering, as a Lecturer, and in 2009 promoted as Assistant Professor. His current research interests include image processing, Bio-informatics and soft computing. He is a Member of Institute of Electrical and Electronics Engineers (IEEE) and International Association of Computer Science and Information Technology. He published papers in more than 20 international journals and National journals.