www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 12 Dec. 2016, Page No. 19548-19551

Topic Digging Over Asynchronous Text Sequences

Monali S Patil[1], Sandip S Kankal[2]

[1] Department of Computer Science and Technology, Maharashtra Institute of Technology,
Beed By Pass Road, Satara Parisar, Aurangabad, Maharashtra
monali.patil80@gmail.com

[2]Department of Computer Science and Technology, Maharashtra Institute of Technology,
Beed By Pass Road, Satara Parisar, Aurangabad, Maharashtra
Sandipkankal25@gmail.com

Abstract: ***Nowadays use of internet is growing rapidly. Most of the information share and use by end users on the web. In any format data or information access by user, most of the available information is in text format. All such text documents available have common content between them with different time stamp. The mutual action between common topics may deduce valuable knowledge but as they are not aligned properly they are not arranged in index fashion. The absolute purpose of this paper is to extract common topic with the help of topic model using appropriate time period. Generative topic model follows two alternative steps. First is retrieve common topic from all the documents with their adjusted time stamp and second is adjusting the time stamp with respect to time period dispersion of common topic being formely generated.***

Keywords: Text Mining, Topic Mining, Asynchronous Sequences, Time Stamp

1. Introduction

The development of internet in present is enhancing speedily. The data point present is in unstructured, semi structured and structure format. The electronic information sources are applied to receive information which is introduced on web. According to text data survey 80% of information is available in unstructured text format. We require few methods to summarize, analyze and discover useful information from such unstructured data. Text mining is the computational technique use to retrieve high quality information from text. Text mining technique used to extract and discover covered knowledge in the text. Document Categorization, Document Organization, Summarization, Visualization and numeric prediction these are certain steps followed by text mining .To cure the trouble of overloaded text information, the automated text extractor and summarizer is needed. To detect prior knowledge from text sequence, the initial stage is to extract topic with both semantic and temporal information. The key supposal while extracting a topic is that all text sequences have same time stamp.which means that common topics from different sequences shared same time period. On the contrary, most of the topic from different text sequences shares different time stamp.For illustration. In case of news feeds,There are cases of publishing of same news by different agencies on different time. There might be hours of delay, days of delay or delays of weeks likewise, In case of research paper archives There are lots of research articles on same topic by different authors on different time stamp. In such cases where extracting common information on different time period is necessary in that case generative model for topic mining came into picture.

A topic model is designed to automatically extract topics from a corpus of text documents. Topic models were originally developed as a means of automatically indexing, searching, clustering, and structuring large corpora of unstructured and unlabeled documents. Using topic models, documents can be represented by the topics within them, and thus the entire corpus can be indexed and organized in terms of this discovered semantic structure.

Topic Mining is one of two very important steps in the process of summarizing a text; the second step is summary text generation. To discover valuable knowledge from a text sequence, the first step is usually to extract topics from the sequence with both semantic and temporal information, which are described by two distributions, respectively: a word distribution describing the semantics of the topic and a time distribution describing the topic’s intensity over time. In many real-world applications, we are facing multiple text sequences that are correlated with each other by sharing common topics.

The method proposed therein relied on a fundamental assumption that different sequences are always synchronous in time, or in their own term Coordinated, which means that the common topics share the same time distribution over different sequences. Rather, asynchronism among multiple sequences, i.e., documents from different sequences on the same topic have different time stamps, is actually very common in practice. For instance, in news feeds, there is no guarantee that news articles covering the same topic are indexed by the same time stamps. There can be hours of delay for news agencies, days for newspapers, and even weeks for periodicals, because some sources try to provide first-hand flashes shortly after the incidents, while others provide more comprehensive reviews afterward. Another example is research paper archives, where the latest research topics are closely followed by newsletters and communications within weeks or months, then the full versions may appear in conference proceedings, which are usually published annually and at last in journals, which may sometimes take more than a year to appear after submission. To visualize it, we have the relative frequency of the occurrences of two terms warehouse and mining.

We do not assume that given text sequences are always synchronous. Instead, we deal with text sequences that share common topics yet are temporally asynchronous.

2. Literature Review

In Information Filtering, Information Retrieval Vector Space model is used. For representation of text document a vector space algebraic model plays important role. There are three stages need to follow in vector space model .Those are Document Indexing, Term Weighting and Similarity Coefficient.

Topic mining has been extensively studied in the literature, starting with the Topic Detection and Tracking, which aimed to find and track topics (events) in news sequences with clustering-based techniques. In many real applications, text collections carry generic temporal information and, thus, can be considered as text sequences .

2.1 Methods of Topic Digging

I. Latent Semantic Analysis

Latent semantic analysis (LSA) is a method or a technique in the area of Natural language processing (NLP). The main goal of Latent semantic analysis (LSA) is to create vector based representation for texts to make semantic content. By vector representation (LSA) computes the similarity between texts to pick the heist efficient related words.

B. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is an approach that has been release after LSA method to fix some disadvantages that have found into LSA. Jan Puzicha and Thomas Hofmann introduced it in 1999. PLSA is a method that can be automated document indexing which is based on a statistical latent class model for factor analysis of count data, and also this method tries to improve the Latent Semantic Analysis (LSA) in a probabilistic sense by using a generative model. The main goal of PLSA is that identifying and distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus.

C. Correlated Topic Model

Correlated Topic Model (CTM) is a kind of statistical model used in natural language processing and machine learning. Correlated Topic Model (CTM) used to discover the topics that shown in a group of documents. The key for CTM is the logistic normal distribution. Correlated Topic Models (CTM) is depending on LDA .

D. Latent Dirichlet Allocation

The reason of appearance of latent Dirichlet allocation (LDA) model is to improve the way of mixture models that capture the exchangeability of both words and documents from the old way by PLSA and LSA. This was happening In 1990, so the classic representation theorem lays down that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture .

3. Proposed Work

Model For Topic Mining: Latent Dirichlet Allocation (LDA) is a model used at highest degree for text mining. This Algorithm apply on the basis of statistical inferences. a LDA is a productive framework which means it efforts to mimes what the writing action is. So it attempts to produce a text file containing topic. It can also be employ to different format of data. There are Number of LDA based models admitting: temporal text mining, author- topic analysis, supervised topic models, latent Dirichelet co-clustering and LDA based bio-informatics.

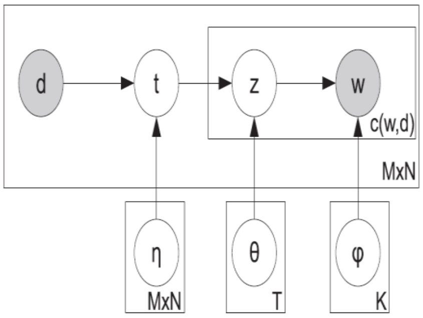


Figure No.4.6 An illustrative Generative Model

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Topic models are also referred to as probabilistic topic models, which refer to statistic algorithms for discovering the latent semantic structures of an extensive text body. In the age of information, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models now has been used to detect instructive structures in data such genetic information, images and networks, they also have applications in other fields such as bioinformatics.

Topic Mining using Latent Dirichelet Allocation is useful to find out the top results by time synchronization and topic extraction. Performance of system is evaluated to find out the Topic. We have implemented our method in research paper repository and the experimental results show that approach achieves high search efficiency.

The Data flow diagram shows actual flow of proposed system as shown in following figure. It is an iterative process and graphical representation of workflows of stepwise activities and actions.

Data flow diagram shows overall flow of control of system.

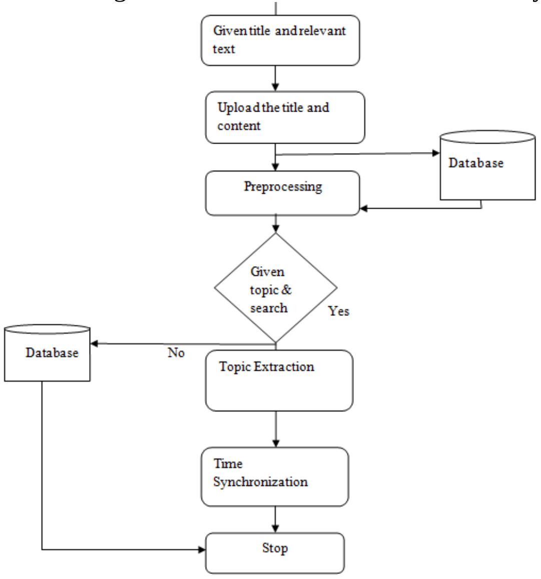


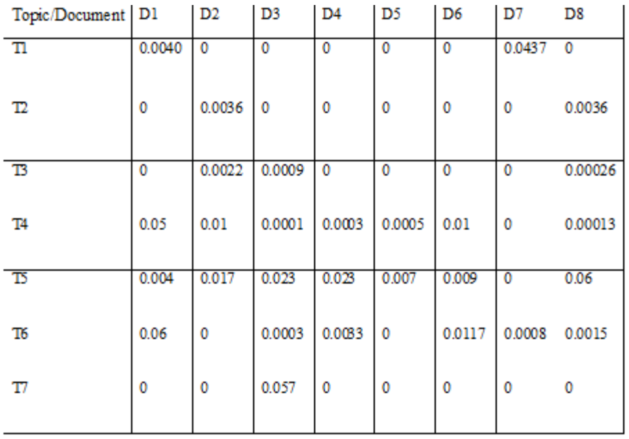
Figure No. 4.7 Data Flow Diagram

A. Result Analysis

We have performed Topic Digging over Research Article Repository of different research areas. such as, Computer Network, Software testing, database management

DOI: 10.18535/ijecs/v5i12.52

system, operating system, Network security etc. In the result analysis evaluation of efficiency and response time of the system. A systematic examination and evaluation of data or information, by breaking it into its component parts to uncover their interrelationships is the analysis.



Topic/Document	D1	D2	D3	D4	D5	D6	D7	D8
T1	0.0040	0	0	0	0	0	0.0437	0
T2	0	0.0036	0	0	0	0	0	0.0036
T3	0	0.0022	0.0009	0	0	0	0	0.00026
T4	0.05	0.01	0.0001	0.0003	0.0005	0.01	0	0.00013
T5	0.004	0.017	0.023	0.023	0.007	0.009	0	0.06
T6	0.06	0	0.0003	0.0033	0	0.0117	0.0008	0.0015
T7	0	0	0.057	0	0	0	0	0

Figure No. 5.1 TF-IDF Table for Topic Ranking

B. Precision and Recall

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

Suppose a computer program for recognizing dogs in scenes from a video identifies 7 dogs in a scene containing 9 dogs and some cats. If 4 of the identifications are correct, but 3 are actually cats, the program's precision is 4/7 while its recall is 4/9. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are".

4. Data Extraction

The proposed method is used by utilizing correlation between the semantic and temporal information in the sequences. It performs topic extraction and time synchronization alternatively to optimize a unified objective function. A local optimum is guaranteed. Preventing duplications in text sequences considering similarities. According to temporal analysis is a constraint proceeds further.

1) The method is able to find meaningful and discriminative topics from asynchronous text sequences;
2) The performance of our method is robust and stable against different parameter settings and random initialization.

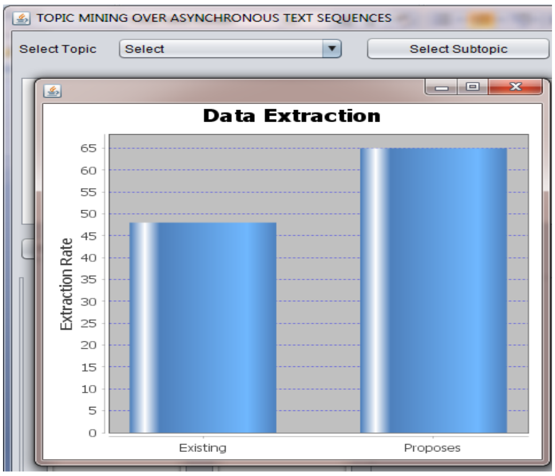


Figure No. 5.2 Data Extraction Rate

5. Data Reduction

Duplicate Words or Topic removes by using Tokenizer in word net database, so that we can get burst free topic and the efficiency of the system can be improved. The Data reduction rate of our proposed system is as follows.

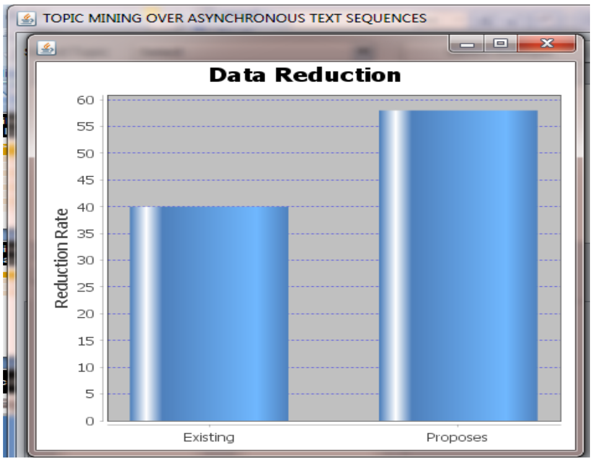


Figure No.5.3 Duplicate Topic reduction rate

6. Conclusion

The proposed system combines the Topic Models and Probabilities. It is based on Latent Dirichelet Allocation (LDA) Topic Model. Most of the time Multiple Asynchronous Text Sequences shares a common topic between them. The proposed Approach achieves the probabilities of common topic and topic synonyms with their respective time stamp in synchronize manner.

The K-L Divergence similarity measure is used to find out similarity between two topics. Log likelihood function and TF-IDF performance measure evaluates the most common topic among all text sequences. The experimental results shows that the proposed method performs better than existing system as it results in data extraction rate 65 and data redundancy 58 rate in comparison with existing system is 48 and 50 respectively.

References

[1]Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen, “Topic Mining over Asynchronous Text Sequences,” IEEE Transactions On Knowledge And Data Engineering, 24(1), JANUARY 2012, pp. 156–169.

[2] Sujata B. Sanap and Prof. Vivek P. Kshirsagar, “Topic Quarrying Over Same Time Period Text Documents”, IJARCCE , 5(2), FEBRUARY 2016, pp. 115-118.

[3] Uma, S. Shanawaz Basha and G. Sesha Phaneendra Babu, “Meaningful Data Extraction Using Data Correlation Technique,” International Journal of Engineering Research & Technology (IJERT), 2(4), APRIL 2013, pp. 1604-1607.

[4]Thomas Hofman, “Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning, 42, 177–196, 2001 Kluwer Academic Publishers. Manufacture, Netherlands , pp. 177-196.

[5]G. Kou, Y. Peng, ”An Application of Latent Semantic Analysis for Text Categorization,” International Journal of Computers Communication & Control ISSN 1841-9836, 10(3):357-369, June, 2015 , pp. 357-368.

[6]L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, “Xrank: Ranked Keyword Search over XML Documents,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2003, pp. 16-27.

[7]Russell Swan and David Jensen, ”TimeMines: Constructing Timelines with Statistical Models of word usage,” Department of Computer Science University of Massachusetts Amherst, Massachusetts USA.

[[8]Sheema Khan and Zafar Ul Hasan, “Text Mining: (Asynchronous Sequences),” Int. Journal of Engineering Research and Applications 2248-9622, 4(12) December 2014, pp.55-59.

[[9]David M. Blei and John D. Lafferty,”A Correlated Topic Model of Science,” Princeto University and Carnegie Mellon University, The Annals of Applied Statistics, 1(1) DOI: 10.1214/07-AOAS114 Institute of Mathematical Statistics, 2007, pp.17–35.

[1 [10]David M. Blei,Andrew Y. Ng and Michael I. Jordan,”Latent Dirichlet Allocation”,Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03, pp. 993-1022.

[[11]Zhiwei Li, Bin Wang and Mingjing Li, Wei-Ying Ma “A p probalistic Model for Retrospective News Event Detection”.

[12] G. Li, J. Feng, and L. Zhou, “Retune: Retrieving and Materializing Tuple Units for Effective Keyword Search over Relational Databases,” Proc. Int’l Conf. Conceptual (E Modelling, 2008, pp. 469-483.

[[13]Tom Magerman , Bart Van Looy , Bart Baesens Koenraad D ebackere , “Assessment of Latent Semantic Analysis (LSA) mitext mining algorithms for large scale mapping of patent and sciscientific publication documents” October 2011.

[1[14]Liangjie Hong and Brian D. Davison,”Empirical Study of Topic Modeling in Twitter” Dept. of Computer Science and , L Engineering Lehigh University Bethlehem, PA 18015 US.

[1[15]Andrew McCallum, Andres Corrada-Emmanuel, Xuerui ,” Wang, “Topic Modelling and Role Discovery in Social DeDepartment of Computer Science University of Massachusetts A Amherst, MA 01003 USA”

[[16]Chong Wang and David M Blei,”Collaborative Topic M Modelling for Recommending Scientific Articles,” Computer Sc Science Department Princeton University, Princeton, NJ, 08 540,USA.

Patil Monali S, Pursuing M.Tech from MIT college of Engineering and Technology,Aurangabad,
She has also completed Diploma from Government Polytechnic, Aurangabad. Her domain of pursuit is Data Mining as well Big Data, Text mining, Natural Language Processing