© 2017 IJSRSET | Volume 3 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Pattern Based Filtering Approach for Big Data Application

Bhagyshri Adhau, Dr. V. T. Gaikwad

Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati, Maharashtra, India

ABSTRACT

Now a day's large Number of services are emerging on the Internet due to various social networking sites, services, cloud computing. Result of this is, service-relevant data become too big to be effectively processed by traditional approaches. Similarly growing technologies like Internet of Things is also responsible for generation of massive raw data and this is reason complexity and resource consumption increases. In this paper the system suggest the concept of pattern based filtering in which it automatically discovers new, hidden or unsuspected data from the large text collection. The propose model MPBTM consist of topic distribution describing topic preference of each collection of document and Pattern-based topic representation.

Keywords: LDA, User interest model, Pattern mining, Relevance ranking

I. INTRODUCTION

The amount of data is increasing, because of different service and cloud computing. Most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information from that dataset. Big data has gain large popularity and attracting attentions from industry and academic. Big data dealing with the tremendously high volume of data, such as for scientific projects, telecommunication companies, result of this data become more difficult to handle or processed in traditional way. It is very important to develop techniques that automatically discover new, hidden or unsuspected data from the large text collection. Large amount of data occur but user refers single topic to finding his needs However, in reality this is not necessarily the case. The user interest can be change or Diverse continuously according to situation and demand of services. Therefore, in this paper, we propose to model user's interest in multiple topics rather than a single topic, which reflects the dynamic nature of user information needs.

Topic modeling has one of the most popular text modeling techniques. It can represent any document with multiple Topics with numbers of variations in the same topic means one document can find numbers of different representation and corresponding distribution. The Patterns are generated with the word through combining topic models representation of traditional topic model MPBTM system.

MPBTM model use for information search, here system performing search on the basis of multiple topics for a particular document because in real time system user information interest can be diverse continuously. In this model pattern are organized into group called as equivalent classes based on their characteristics and statistical features in each present class pattern have same frequency and important meaning with this structured representation the most representative pattern can be recognized and this pattern used for information filtering and to get relevant document.

I I . LITERATURE REVIEW

User always want to extract important and useful data in which we are interested rather than having another unwanted raw data with obviously less complexity. The pattern mining based techniques have been used to utilize patterns to represent user's interest and have achieved some improvements in effectiveness since patterns carry more semantic meaning than terms [1]. The Information Filtering System (IF) works on such

motive that system help long term information need for particular user or group of user. The Objective of IF system is that provide accurate and effective mapping of incoming data document. A content-based filtering system select item on the basis of correlation in between content and item and also the preference of the user search. Mobile data mining systems (MDMS) play a vital role in utilization of on-board resources to uncover knowledge patterns in user locality and provision of local knowledge for immediate local utilization [2]. Test clustering method for such purpose to extract or may structure large number of data or Huge text hypertext the document but there are some problem with this technique that is very high dimensionality of the data. High dimension data simply can say goes beyond limit of text Clustering Technique. There is another one approach called Direct Discriminative Pattern Mining for Accuracy and Efficiency it is to tackle efficiency issue arising from the two-step approach. The DD mining approach perform a branch and bound search for directly mining discriminative patterns without generating a complete pattern set. Term based modeling also the technique but has many drawbacks and limitation on expressing problems of polysemy and synonymy. There people tend t extract more semantic features such as phrases and patterns to represent document in many applications.

III. PROBLEMS IN BIG DATA

The data become more complex and that is due to different v's. These v's means volume, velocity, variety, value, veracity, and variability.

• Volume: The volume refers to the size of the data that produced in big data systems. Generally, a data size which could not be easily processed by ordinary systems by conventional way is known to be volume of big data. But the terms large data size is different for different system. For example, A few GB file is a big data file for a Smartphone buy it may not be the case for PC or cloud enabled systems. Similarly, a few TB file is a big data file for a PC but a cloud enabled system may handle it easily.

•Velocity: The velocity term in big data system means the speed of incoming data streams determines the velocity property. that increases delay in the big data systems so Velocity is the main challenge in big data Big data systems handles velocity in two ways: 1) raw data is collected in central data stores for lateral analysis. In the first approach, big data systems create a delay between data acquisition and knowledge discovery. This strategy is more useful for analysis of historical data and second option is 2) online data analysis of data streams right after acquisition is performed. The second approach is more appropriate for real-time data analysis.
•Veracity: The Data that can believe trust of big data that adherence is represented by veracity property of big data. This property is based on the authenticity of data sources and correctness of data. So the effective handling of veracity property of big data improves the overall effectiveness of the system.

•Variety: The composite data sources and data formats in big data systems are describe by the variety property of big data. Data sources for big data systems vary in terms of structured and unstructured data formats increase variety in big data systems. The Big data systems handle the variety challenge effectively to uncover maximum knowledge patterns.

•Variability: The handling of clashes or incompatibilities in big data is attributed with variability. The variable data rates causes computational overhead in peak-load times therefore a proper handling of variability property increases the usefulness of big data systems.

•Value: The interestingness of uncovered knowledge patterns is represented by the value property of big data systems. The value is directly affected by other 5V's (velocity, volume, variety, veracity, and variability) therefore a proper balance between all V's brings more value in the big data system. In addition, the effective handling of all other V's is directly proportional to increased value of big data [3]. In Pattern recognition, there may be higher interest to formalize explain and visualize the patterns, while machine learning

IV. LATENT DIRICHLET ALLOCATION

The LDA model is highly modular and can therefore be easily extended. The main field of interest is modeling relations between topics. This is achieving by using another distribution on the simplex instead of the Dirichlet. The correlated Topic Mode follows this approach, including correlation structure between topics by using the logistic normal distribution instead of the dirichlet.

Latent Dirichlet allocation is a generative statistical model that allows set of observations to be explained by unobserved group means That observed why particular part of data are similar. Topic modeling algorithms are used to determine a set of hidden topics from collections of documents, where a topic is represented as a set of the words. Topic model interpret low dimension representation of documents. (I.e. with a limited and manageable number of topics). Figure1.Distribution of document over topics and words. The idea behind LDA is that each document is considered to contain multiple topics and each topic can be defined as a distribution over fixed words that appear in the documents. It can search the hidden topics in collections of documents using the words that appear in the documents. LDA that allows sets of observations to be explain why some part of the data is similar as each document is a mixture of small number of topics and each word's creation is attributable to one of the document type Generally the Topic modeling Technique is used by us to discover a set of hidden Document from a group of topic where topics are also the set of words. So LDA can discover hidden topics in a set of document [5].

Let,
Document be taken D = {d1, d2, d3, d4}
Two levels = Document Level and collection level.
At document Level,
Each document say di
contained set of topics Q(di) ={Q(di 1), Q(di 2), Q(di 3), …. ,Q (di V)}.
Where,
Number of Topics in the document = V

V. PATTERN ENHANCED LDA

The Pattern Based model contains information which is structured information which gives us an idea about the relation between the words in the document. Our goal is to mine the similar meaning pattern to define a topic and document. So there are two steps these are 1) To construct Transaction Dataset and 2) to Generate pattern Enhanced Representation.

✓ Construct Transactional Dataset

As each document is Represent by di, Let, word-topic assignment = Rdi ,Zj For topic Zj. for the number of topics say Z1 in document D1, So, Rdi,Zj= {W1,W2,……Wn}. Now instead of using that sequence of word, construct a set
of words from each word-topic assignment Rdi,Zj.
So as a result,
Set of words in Rdi,Zj = Iij.
So now Iij contains words in document di and also mentioned in the Topic Zj.
So now Iij is now transactional dataset, but make sure it has no duplicates.

T	TDT	TDT	TDT
1	{w1,w3,w8,w9}	{w4,w5,w6}	{w5,w10}
2	{w1,w2,w3}	{w1,w2,w3}	{w2,w7,w10}
3	{w2,w6}	{w1,w5}	{w1, w12, w11}
	T1	T2	T3

Table1. Transactional Datasets generation

As in the table we get different sets of words that we can called it as Transactional Datasets of the given document.

✓ Generate Pattern Enhanced Representation

The pattern-based is simply for use frequent pattern generated from each transactional dataset say Gj to represent zj.

S= minimal support threshold,
If supp(X)>=d,
Where supp(x) is support of X. and X is item set in Gj.

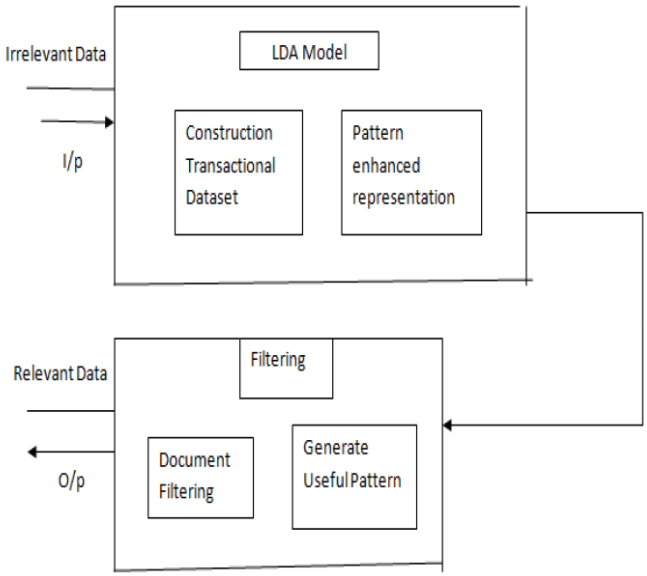

Figure 2. Block Diagram for Document Filtering by generating Patterns

Suppose you have the following set of sentences:

- ✓ I ate a banana and spinach smoothie
- ✓ for breakfast.
- ✓ I like to eat broccoli and bananas.
- ✓ Chinchillas and kittens are cute.
- ✓ My sister adopted a kitten yesterday.
- ✓ Look at this cute hamster munching on
- ✓ a piece of broccoli.

Latent Dirichlet allocation is a way of automatically discovering topics that these sentences contain. For example, given these sentences and asked for 2 topics, LDA might produce something like

- ✓ Sentences 1 and 2: 100% Topic A
- ✓ Sentences 3 and 4: 100% Topic B
- ✓ Sentence 5: 60% Topic A, 40% Topic B
- ✓ Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, (at which point, you could interpret topic A to be about food)
- ✓ Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, (at which point, you could interpret topic B to be about cute animals)

The question, of course, is: how does LDA perform this discovery?

LDA Model

In more detail, LDA represents documents as **mixtures of topics** that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you

- ✓ Decide on the number of words N the document will have (say, according to a Poisson distribution).
- ✓ Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.
- ✓ Generate each word in the document by:
- ✓ ....First picking a topic (according to the multinomial distribution that you sampled above; for example, user might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability).
- ✓ ....Then using the topic to generate the word itself (according to the topic's multinomial distribution). For instance, the food topic might output the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

Example

Lets make an example According to the above process, when generating some particular document D, it might

- ✓ Decide that D will be 1/2 about food and 1/2 about cute animals.
- ✓ Pick 5 to be the number of words in D.
- ✓ Pick the first word to come from the food topic, which then gives you the word "broccoli".
- ✓ Pick the second word to come from the cute animal's topic, which gives you "panda".
- ✓ Pick the third word to come from the cute animal's topic, giving you "adorable".
- ✓ Pick the fourth word to come from the food topic, giving you "cherries".
- ✓ Pick the fifth word to come from the food topic, giving you "eating".

So the document generated under the LDA model will be relevant.

VI. MAXIMUM MATCHED PATTERNS

The documents are the set of multiple topics so furthermore the topic should be also organize in the different classes so that it will be easy to discovered correct topics or document and also in very less amount of time.

- ✓ The propose model MPBTM consist of topic distribution describing Topic Preference of each collection of document and Pattern-based topic representation which representing semantic meaning of each for each Topic.
- ✓ We propose structural Patterns based Topic Representation in which the different patterns are classified into different classes called Equivalence classes based on their different taxonomic and their statistical features.
- ✓ Now when searching particular Topic the pattern which is more representative to that searching one topic can be identified which will more benefit to relevance ranking.
- ✓ The system then introduced new document ranking method to determine relevance of new incoming document based on the structured Patterned-Based representation model.

Patterns are used to represent documents, which not only can solve the synonymy problem, but also can deal with the low frequency problem of phrases [4].

A. Topic-based Document Relevance Ranking

In the filtering step, we get Relevance Document by removing irrelevant data. That is we first to be find maximum match pattern which user has more interest and then estimate relevance of document but significant of these pattern is based on size. The longer size is more specific suppose For example Filtering is the -ing form of Filter where as 'Document Filtering' Represent more specific but 'Relevance Document Filtering' is even more specific. Let d = document, Zj = topic in the user interest model, ECj1; . . . ; ECjnj=pattern equivalence classes of Zj, Then a pattern in d is considered a maximum matched pattern to equivalence class ECjk, denoted as MCdjk, if the following conditions are satisfied: [⊆ and ∈] here {∄ h h ∈ , ⊆ ⊆ }. The maximized matched pattern MCjk to equivalence class ECjk must be the largest pattern in ECjk which is contained in d and all the patterns in ECjk that are contained in d must be covered by MCdjk. Therefore, the maximum matched patterns MCd jk, where k=1; . . . ; nj are considered the most significant patterns in d which can represent the topic Zj. For an incoming document d, we propose to estimate the relevance of d to the user interest based on the topic significance and topic distribution. The Document relevance is estimated using the following equation: Rank ( ) =Σ (,) =1× , Higher Rank (,) = {1, f ∈00}.

VII. CONCLUSION

The MPBTM and PBTM have the same computational complexity as frequent closed pattern mining. On the other hand, the MPBTM and the PBTM generate patterns from very small transactional datasets compared with the datasets used in general data mining tasks, because the transactional datasets used in the MPBTM and the PBTM are generated from the topic representations produced by the LDA model rather than the original document collections. The MPBTM model comparatively achieves better performance than other modeling system. It generates pattern enhanced topic representations to model user's interests across multiple topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modeling and the specificity as well as the statistical significance from the most representative patterns. In comparison with the state-of-the-art models, the proposed model demonstrates excellent strength on document modeling and relevance ranking.

VIII. REFERENCES

[1]. F. Beil, M. Ester, and X. Xu, "Frequent termbased text clustering," inProc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 436–442.

[2]. Rehman, M.H., Liew, C.S., and Wah, T.Y.: 'Frequent pattern mining in mobile devices: A feasibility study', in Editor (Ed.)^(Eds.): 'Book Frequent pattern mining in mobile devices: A feasibility study' (IEEE, 2014, edn.), pp. 351-356.

[3]. Gani, A., Siddiqa, A., Shamshirband, S., and Hanum, F.: 'A survey on indexing techniques for big data: taxonomy and performance evaluation', Knowledge and Information Systems, 2015, pp. 1- 44.
[4]. Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for information filtering," in Proc. Int. Conf. Data Min. Workshop SENTIRE, 2013, pp. 921–928.
[5]. X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178–185..
[6]. H. M. Wallach, "Topic modeling: Beyond bag-of-words," in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 977–984.
[7]. Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000.