PATTERN-BASED TOPICS FOR DOCUMENT MODELLING USING HLA

R.C.Padmaja[1] ,PG Scholar, Dept. Of. Computer Science & Engineering, Sri Vidya College of Engineering & Technology, Virudhunagar, India

N.Venkatesan[2] ,Assistant professor, Dept. Of. Computer Science & Engineering, Sri Vidya College of Engineering & Technology, Virudhunagar, India

Abstract

In data mining, pattern based approaches are employed which comes under the topic of mining. From input dataset the mining topics are collected and each document must contain at least a title. The selected title must be suitable to the document content. The process of topic selection by the humans requires more time. But the time consumption is reduced by the system since it finds a suitable title in a short period of time. Therefore some rules are developed and given to the system. According to the rules the system can choose a topic suitable to the document. For that the high level algorithm is used in the proposed system. The two things such as the dictionary and the anchor words are so essential to the high level algorithm. At first, it is necessary to consider the preprocessing operation that contains two major things like stop words and stemming. The preprocessed data is obtained by removing the stop words and the suffix from the dataset. Then the dictionary generation process is initiated and the hundreds and thousands of files are given as the input to the system. The dictionary is generated by taking input words from all the input files. Thus the dictionary comprises all the words present in all the documents. Then the anchor words are considered which are chosen on the basis of support count. In dictionary the anchor words are searched and chose the topic for all the documents. This technique requires relatively low time consumption which gives the better results when compared to the existing systems.

Key words :- *Pattern mining, High Level Algorithm, Dictionary generation and Anchor words.*

I. Introduction

Information filtering (IF) is a framework that expels repetitive or undesirable data from a data or document stream. This process is completely based on document representations that depict user's interest. On the basis of a term-based approach conventional IF models were produced which has an advantage of effective computational performance and mature theories for term weighting as well [1], [2]. The polysemy and synonymy are the two major issues that suffer the term based document representation. In order to overcome such restrictions of term based methodologies pattern mining based techniques have been utilized for information filtering and accomplished some upgrades on effectiveness [3], [4], since patterns convey more semantic significance than terms. Likewise, data mining has added to a few techniques like maximal patterns, closed patterns and master patterns in order to remove the redundant and noisy patterns [5]–[6].

Topic modeling [7] is a popular probabilistic text modeling technique that has frequently acknowledged by machine learning and text mining groups. It can consequently classify documents in an accumulation by various topics and depicts each document with numerous topics and their corresponding distribution. There are two representative methodologies like Probabilistic Latent Semantic Analysis (PLSA) [8] and LDA [9] are present. In any case, there are two issues in specifically applying topic models for information filtering. The primary issue is that the topic

distribution itself is deficient to depict documents because of its fixed number of measurements like a pre-indicated number of points.

Next issue is that the word based topic representation is restricted to particularly depict document which have diverse semantic substance since numerous words in the topic representation are incessant general words [10]. In this paper the limitation of existing system is overcome by utilizing Natural Language Processing Natural dialect preparing (NLP) that is the open English NLP 2.0 library utilized as a part of improved LDA calculation for filtering semantic implications of patterns from the accumulations of points. Representation and document significance ranking and it choose the most illustrative and discriminative patterns

Here the High Level Algorithm (HDL) is proposed for maximum matched patterns and the LDA is applied through the Gibbs sampling technique. Representation and record significance positioning and it to choose the most illustrative and discriminative examples, that are to depict topics rather than utilizing successive patterns. After establishment of this application, it is supportive for document searching inefficient and simple way from number of various documents which is accessible to download and understand the document based on user's focused area or patterns. This system is effectively discovering applicable document from accumulation of document.

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 Web 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. 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 largercollections, 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. Proposed Technique

A consideration is made on the learning problem of a class of admixture distributions which are often employed for probabilistic topic models. Latent Dirichlet allocation [11], correlated topic models [12], and Pachinko allocation [13] are some of the examples of admixture distribution. Here the number of words in the vocabulary and the number of topics is denoted by W and K respectively. Each related topic k is known as a multinomial distribution throughout the words in the vocabulary which is referred as the column vector B_k of length W. A specific prior distribution over the topic distribution is required by each topic models of a document. For instance consider the latent Dirichlet allocation (LDA) in that τ is the Dirichlet distribution and τ is denoted as a logistic Normal distribution for the correlated topic model. The generative process of a document is referred as g which is initiated by drawing the document's topic distribution $V_g \sim \tau$. Then, for each position j a topic assignment is sampled as $y_j \sim V_g$, and at last a word $v_j \sim B_{y_j}$.

The word-topic matrix B of dimension $W \times K$ is obtained by combining the column vector B_k and each K topic. Similarly the topic document matrix V of dimension $K \times m$ is obtained by aggregating the column vectors V_g and m documents. It is denoted that V is unknown and stochastically generated and it can never required to be able to retrieve it. The word topic matrix B is found by considering the learning task. In case when τ Dirichlet (LDA), the learning of hyper parameters of is τ is shown.

NP-hard is the maximum likelihood estimation of the word-topic distributions even for two topics (Arora et al., 2012b) which results the researchers mostly apply approximate techniques. Variational expectation maximization [1] is one of the most popular techniques that optimizes a lower bound on the probability and Markov chain Monte Carlo [4] that samples the posterior distribution in a asymptotic manner.

Arora et al. [5] proposed an algorithm that can learn the parameters of a topic model for given samples from that model and it also denoted that the word-topic distributions are independant (Donoho & Stodden, 2003):

Algorithm 1: High Level Algorithm

Input: Textual corpus d, Number of topics k,
Tolerance parameters $\in_x, \in_y > 0.$
Output: Word-topic matrix B, topic-topic matrix r
$C \leftarrow$ Word Co-occurrence (d)
From $\{\overline{C}_1, \overline{C}_2, ...\overline{C}_v\}$, the normalized rows of $C.$
$S \leftarrow$ Fast Anchor Words $(\{\overline{C}_1, \overline{C}_2, ...\overline{C}_v\}, k, \in_x)$
$B, r \leftarrow$ Recover$KL(C, X, \in_y)$
Return B, r

The algorithm proposed by [5] is not virtual. It cannot solve W linear program s to recognize anchor words and it require matrix inversion to retrive B and the parameters of τ which is sensitive to noise and unstable as well. But it generates the basis of the enhanced technique i.e., Algorithm 1. Both recovery and anchor selection process takes the input of $W \times W$ matrix of word-word co-occurance counts C which is normalized. Thus, the sum of all entries is 1.

3.1 Topic Recovery Process

Based on a probabilistic model, a novel recovery technique is introduced. The actual recovery process present in [5] is given as follows. At first it interchanges the rows and columns of C matrix; therefore the the K rows and columns represent the anchor words are interchanged. Let us consider C_s as the first K rows and $C_{s,s}$ as the first K rows and the first K columns. C is set to be second order moment matrix $C = \mathrm{E}[BVV^T B^T] = B\mathrm{E}[VV^T]B^T = BRB^T$ while it is constructed from infinitely several document and the block structure of this process is given.

$$C = BRB^T = \begin{pmatrix} d \\ u \end{pmatrix} R \begin{pmatrix} d & u^T \end{pmatrix} = \begin{pmatrix} dRd & dRu^T \\ uRd & uRu^T \end{pmatrix}$$

Where d is denoted as a diagonal matrix of size $K \times K$ in that rows represent to anchor words. Then, it figures out for B and R by employing the algebraic manipulations defined in Algorithm 2.

The algorithm 2 produces substantial imprecision during matrix inversion for finite data and small negative values are present in the returned B and R matrices that necessitates a consequent projection on the simplex.

Algorithm 2: Original Recover

Input: Matrix C, Set of anchor words X
Output: Matrix B, r
Permute Rows and columns of C
Compute $\vec{p}x = C_x \vec{1}$ (equals $dr\vec{1}$
Solve for $\vec{y} : C_{x,x}\vec{y} = \vec{p}x$ (Diag $(\vec{y})$ equals d^{-1})
Solve for $B^T = C_{X,X} Diag(\vec{y}))^{-1} C_X^T$
Solve for $r = Diag(\vec{y} C_{X,X} Diag(\vec{y})$
Return B, r

An issue is that Recover algorithm employs K rows of the matrix C where C has the dimension of $W \times W$. Though neglecting the omitted $W - K$ rows is enough for theoritical analysis and the remaining rows comprise no extra information.Thus, it is not adequate for real data. Modest sample sizes may create approximates of co-occurances between the anchors

and a word inaccurate.

In this paper, we proposed a novel technique based on Baye's rule. Let us consider w_1 and w_2 are any two words and y_1 and y_2 are the topic assignments. $B_{j,k}$ is used as the index to the matrix of word-topic distribution, ie., $B_{j,k} = p(v_1 = j \mid y_1 = k) = p(v_2 = j \mid y_2 = k)$. For infinite data, the factors of C matrix can be represented as $C_{i,j} = p(w_1 = i, w_2 = j)$. $\overline{C}$ is denoted as the row-normalized C matrix that acts a role in both recovery process and searching the anchor words. It can be taken as a conditional propability $\overline{C}_{i,j} = p(v_2 = j \mid v_1 = i)$.

Algorithm 3: RecoverKL

Input: Matrix C, Set of anchor words X, toleranceParameters $\in$.
Output: Matrix B, r
Normalize the rows of C to form $\overline{C}$
Store the normalization constants $\vec{p}_v = C\vec{1}$
$\overline{C}_{xk}$ is the row of $\overline{C}$ for the K^{th} anchor word.
For $j = 1,....,W$ do
Solve $Q_j = \arg\min_{\overline{C}_j} d_{kL}\left(\overline{C}_j \| \sum_{K \in X} Q_{j,K}\overline{C}_{xK}\right)$
Subject to: $\sum_K Q_{j,K} = 1$ and $Q_{j,K} \geq 0$
With tolerance: $\in$
end for

Consider the exponents of the anchor words as $X = \{x_1, x_2, ..., x_k\}$ and the rows guided by factors of X are peculiar in which every other row of $\overline{C}$ rests in the convex hull of the matrix rows guided by the anchor words. For the anchor word x_k,

$$\overline{C}_{x_k,j} = \sum_{k'} p(y_1 = k' \mid v_1 = s_k)\, p(v_2 = j \mid y_1 = k') \quad (1)$$

$$= p(v_2 = j \mid v_1 = k), \quad (2)$$

Where eqn(1) emplys the fact present in an admixture model $v_2 \perp v_1 \mid x_1$, and eqn(2) is because $p(x_1 = k \mid v_1 = x_k) = 1$. For some other word i, we have

$$\overline{C}_{i,j} = \sum_k p(x_1 = k \mid v_1 = i)\, p(v_2 = j \mid x_1 = k).$$

Representing the probability $p(x_1 = k \mid v_1 = i)$ as $A_{i,k}$, we have $\overline{C}_{i,j} = \sum_k A_{i,k} \overline{C}_{x_k,j}$. Because A is non-negative and $\sum_k A_{i,k} = 1$, thus we have that any row of $\overline{C}$ rests in the convex hull of the rows representing the anchor words. $p(x_1 \mid v_1 = i)$ is the mixing weights. Applying this weights together with $p(v_1 = i)$, we can recover the matrix B simply by employing Baye's rule.

$$p(v_1 = i \mid x_1 = k) = \frac{p(x_1 = k \mid v_1 = i)\, p(v_1 = i)}{\sum_{i'} p(x_1 = k \mid v_1 = i')\, p(v_1 = i')}$$

At last, we discover that $p(v_1 = i)$ is simple to solve for since $\sum_j C_{i,j} = \sum_j p(v_1 = i, v_2 = j) = p(v_1 = i)$.

The proposed novel algorithm discovers each row of the actual row normalized co-occurance matrix $\overline{C}_i$ and the vector of non-negative co-efficients $p(x_1 \mid v_1 = i)$ which is best and reconstruct it as a covex combination of the rows which represent the anchor words. Here, the objective function is used to measure the "best" which permits this step to be resolved frequently in parallel of each word by employing the exponentiated gradient algorithm. Once we find $p(x_1 \mid v_1)$ then we can recover the matrix B by applying Baye's rule. The complete algorithm employing KL divergence is as the target which is found in Algorithm 3. Thus our proposed non-negative recovery algorithm provides better results on a wide range of performance metrices rather than the original recover.

IV. Performance evaluation

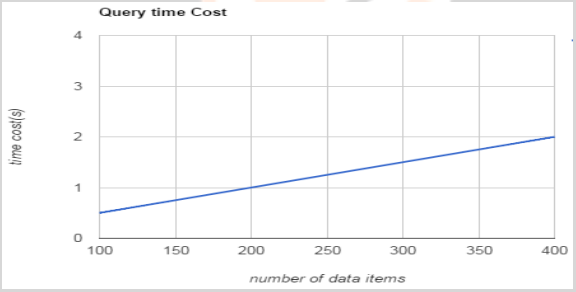

Figure 1: Query Time Cost

In Figure 1, Quary time cost of the proposed HDL algorithm is calculated which is the ratio of number of data items and time cost.

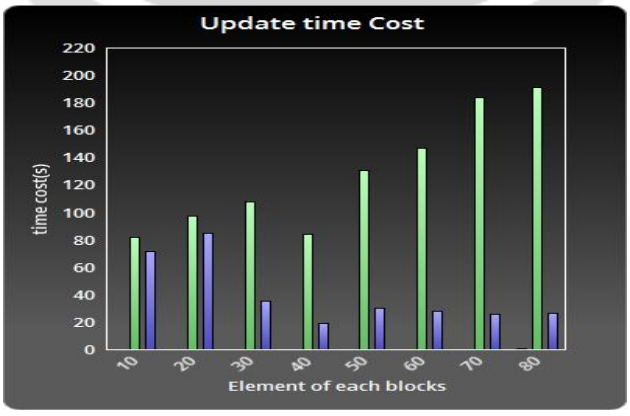

Figure 2: Update Time Cost

In Figure 2, the performance of the proposed HDL and existing Gibbs algorithms are compared in which the blue line indicates the HDL algorithm and green line indicates the Gibbs algorithm. Update time cost is the ratio of elements of each blocks and time cost. Our proposed HDL algorithm proides better performance with reduced time cost.

Figure 3: Verify Time Cost

In Figure 3, verify time cost of the Gibbs, RecoverKL and the proposed HDL algorithms are compared. Green line indicates the performance of Gibbs algorithm, blue line indicates the performance of RecoverKL algorithm and red line indicates the performance of the proposed HDL algorithm. The verify time cost is the ratio of number of data blocks and time cost. Through the graph we can show that the proposed HDL algorithm can provide better performance of reduced verify time cost.

V. Conclusion

Thus we proposed novel algorithms for topic modeling which is so efficient and it needs simple process to implement and keep up provable guarantees. Based on the size of the corpus, the running time of the proposed algorithms are effectively independent. For further optimization we tried to utilize the output of the proposed algorithms as initialization process but unfortunately we have not yet discovered hybrid which performs either technique by itself.

References

[1] S. Robertson, H. Zaragoza, and M. Taylor, “Simple BM25 extension to multiple weighted fields,” in Proc. 13th ACM Int. Conf. Inform. Knowl. Manag., 2004, pp. 42–49.

[2] F. Beil, M. Ester, and X. Xu, “Frequent term-based text clustering,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 436–442.

[3] 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.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, “Discriminative frequent pattern analysis for effective classification,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725.

[5] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowl. Discov., vol. 15, no. 1, pp. 55–86, 2007.

[7] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining." in Proc. SDM, vol. 2, 2002, pp. 457–473.

[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data Knowl. Eng., vol. 70, no. 6, pp. 555–575, 2011.

[9] 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.

[10] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011, pp. 448–456.

[11] Blei, D. Introduction to probabilistic topic models. Communications of the ACM, pp. 77{84, 2012. 1

[12] Blei, D. and La erty, J. A correlated topic model of science. Annals of Applied Statistics, pp. 17{35, 2007.

[13] Li, W. and McCallum, A. Pachinko allocation: Dagstructured mixture models of topic correlations. In ICML, pp. 633{640, 2007.

[14] McCallum, A.K. Mallet: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu

[15] Arora, S., Ge, R., and Moitra, A. Learning topic models { going beyond svd. In FOCS, 2012b.