Sparse Representation Based Query Classification Using LDA Topic Modeling

Indrani Bhattacharya and Jaya Sil

Abstract In recent years, tremendous growth of documents provides scope and challenges to the interdisciplinary research community in text processing for retrieving information. Text analytics reveals high-quality information by identifying patterns and its trends using statistical methods. In this paper, we propose a novel approach to classify user query in a reduced search space by considering the query as a collection of words distributed over different topics. Latent Dirichlet allocation (LDA) has been used for topic modeling and a collection of topics containing words are obtained following Dirichlet distribution. We construct a sparse matrix called topic-vocabulary matrix (TVM) using probability distribution of words appearing in the topics. Finally, sparse representation based classifier (SRC) has been applied for classifying query using TVM consisting of training patterns. Here, we have analyzed the effect of number of patterns in classifying the queries and achieved 90.4 % accuracy.

Keywords Topic modeling · LDA · Sparse classifier · Statistical methods

1 Introduction

Huge collection of electronic documents posed new challenges to the researchers for developing automatic techniques to visualize, analyze and summarizing the documents in order to retrieve information accurately and computationally efficient manner. Topic models identify patterns which reflect underlying semantic embedded in the documents [1], needed to classify the documents based on the requirement of the users. Topic modeling is applied to index the documents using relevant terms whereas a topic is a probability distribution over words [2].

Indrani Bhattacharya (✉) · Jaya Sil
Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
e-mail: Indrani.84@hotmail.com

Jaya Sil
e-mail: jayaiiests@gmail.com

S.C. Satapathy et al. (eds.), *Proceedings of the International Conference on Data Engineering and Communication Technology*, Advances in Intelligent Systems and Computing 469, DOI 10.1007/978-981-10-1678-3_59

In classical document clustering approaches [3] documents are represented by bag-of-word (BOW) model based on raw term frequency and therefore, poor in capturing the semantics. Topic models are able to combine similar semantics into the same group, called topic.

Term frequency inverse document frequency (tf-idf) method [4] can be able to identify discriminative words for a document, but pays little attention to inter- or intra-document statistical structure [5]. To address the problem, first latent semantic indexing (LSI) [6] and latter a generative probabilistic model [7] of text corpora were proposed. A significant step toward probabilistic modeling of text is probabilistic latent semantic analysis (PLSA) reported in [8]. LDA model has been developed to improve the process of forming the mixture models by capturing the exchangeability of both words and documents previously not explored in PLSA and LSA [9]. There are many LDA-based models including temporal text mining, author-topic analysis, supervised topic models, latent Dirichlet co-clustering, and LDA-based bioinformatics [10]. Several improvements have been proposed on LDA, such as the hierarchical topic models [11] and the correlated topic models [12].

A query consisting of several keywords can be viewed as distribution of words with probability over topics. Challenge is to develop more efficient retrieval mechanism for searching related topics from the corpus similar to the query submitted by the user. In this paper, we use LDA method to extract the topics from a large corpus of documents [2]. Then we propose a sparse representation-based classifier [13] for classifying the query, which is distribution of words with probability among the topics. A term vocabulary (TRV) has been constructed using unique terms in the topic corpus, representing the document repository. Since the number of terms present in the query is very specific, the query vector is highly sparse with respect to the TRV. In Sparse representation based classifier (SRC), query is represented in an overcomplete dictionary whose base elements are the training samples. In this paper, the topics are encoded using TRV and considered as training samples in topic-vocabulary matrix (TVM). Finally, we apply SRC to classify the query vector in a reduced search space.

This paper is divided into four sections. Section 2 describes detailed methodology using statistical topic modeling and sparse representation based classifier while results are summarized in Sect. 3 and conclusions are arrived in Sect. 4.

2 Methodology

In the proposed method, first we obtain the topics and the word distribution among the topics using traditional statistical topic models (LDA model). In the second part, we apply SRC to classify the users' query.

2.1 Statistical Topic Modeling

LDA is a basic probabilistic model that describes a generative topic model for a large corpus of documents. In this model, each document is defined as a mixture of words with a given probability. The most dominant topics in the document are with highest probabilities.

Basic LDA model is built on certain assumptions. It assumes that (i) k number of topics is in the corpus, (ii) each document has topic proportions of θ, (iii) a word w is generated from a topic z, and (iv) each topic is defined as the word proportions β over the number of existing words.

The generative process is described below:

1. For each document d in the corpus
   - (a) Generate $\theta_d \sim Dir\,(\alpha)$
   - (b) For each word
     - i. Draw a topic $z_{d,n} \sim Mult\,(\theta_d)$
     - ii. Draw a word $w_{d,n} \sim Mult\,(\beta_{z_{d,n}})$

where n is the number of words and α is Dirichlet prior vector for θ and β is the topic probability. The joint distribution of topic mixture θ for given parameter α and β, a set of N topics zd and a set of n words wd is given by Eq. (1):

$$p(\theta, z_d, w_d \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta).p(w_{d,n} \mid z_{d,n}, \beta) \tag{1}$$

where $p(z_{d,n} \mid \theta)$ is $\theta_{d,i}$ for the unique i such that $z_n^i = 1$.

Equation (2) shows the marginal distribution of a document:

$$p(w_{d,n} \mid \alpha, \beta) = \int p(\theta \mid \alpha).(\prod_{n=1}^{N} \sum p(z_{d,n} \mid \theta).p(w_{d,n} \mid z_{d,n}, \beta))\, d\theta \tag{2}$$

Finally, we obtain the probability of a corpus as given in Eq. (3):

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha)\, (\prod_{n=1}^{N_d} \sum p(z_{dn} \mid \theta_d) p(w_{dn}, \beta))\, d\theta_d \tag{3}$$

In the learning method, latent variables z and θ are searched using LDA with an objective to maximize log-likelihood of the data and this problem is NP hard. Several approximate inference algorithms include Gibbs sampling [14] and variation inference [15] have been used for learning purpose. Figure 1 shows graphical representation of LDA model where each node is represented as a random variable and labeled according to the generative process.

In our experiment, we apply LDA considering variable number of topics. We set initial value of parameter α in the range [0, 1] and obtain its effect on distribution of number of topics. We choose 10^{-2} as threshold of difference in α varied over number of topics for the proposed retrieval method. We choose 25 topics as threshold because no significant change in α has been observed further.

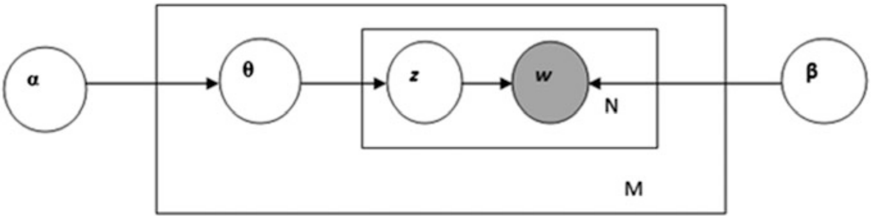

Fig. 1 Graphical representation of LDA model

2.2 Sparse Classifier

We obtain n topics containing words using LDA from the corpus of documents and a vocabulary TRV is prepared using unique terms present in different topics. Let us assume the number of unique words in n no. of topics is k, so the dimension of TRV is 1 × k. On the basis of TRV a feature vector (FV) of dimension 1 × k for each topic is defined in Eq. (4):

$$\begin{aligned}\mathbf{FV}[i] &= P(w_i \mid T) \cdot P(w_i), \quad \text{if } w_i \text{ presents in the topic} \\ &= P(w_i), \quad \text{otherwise}\end{aligned} \tag{4}$$

where $P(w_i|T)$ denotes the probability of word w_i in topic T and $P(w_i)$ gives the probability distribution of the word in the corpus.

Let us assume that there are c known pattern classes. Let Ai be the matrix obtained using the training samples of class i, i.e., $\boldsymbol{A}_i = [\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{\boldsymbol{iRi}}] \in \mathbb{R}^{d \times Si}$ where M_i is the number of training samples of class i.

Let us define a matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots \boldsymbol{A}_c] \in \mathbb{R}^{d \times S}$, where $S = \sum_{i=1}^{c} S_i$. The matrix $\boldsymbol{A}$ is built for the entire training samples. Given a query test sample $\boldsymbol{y}$, we represent $\boldsymbol{y}$ in an overcomplete dictionary whose basis are training samples, so $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ If the system of linear equation is underdetermined (P < S), this representation is naturally sparse.

The sparsest solution can be obtained by solving the following L_1 optimization problem given in Eq. (5),

$$(L_1)\ \widehat{\boldsymbol{w}_1} = \arg\min \|\boldsymbol{w}\|_1, \quad \text{subject to } \boldsymbol{A}\boldsymbol{w} = \boldsymbol{y} \tag{5}$$

This problem can be solved in polynomial time by standard linear programming algorithm [16]. After the sparsest solution $\widehat{\boldsymbol{w}_1}$ is obtained, the SRC can be done in the following way [17].

For each class i, let $\partial_i : \mathbb{R}^S \to \mathbb{R}^S$ be the characteristic function that selects the coefficient associated with the ith class.

For $\boldsymbol{w} \in \mathbb{R}^S$, $\partial_i(\boldsymbol{w})$ is a vector whose nonzero entries are in w associated with class i. Using only the coefficient associated with the ith class, reconstruction can be done on a given test sample y as $\boldsymbol{v}^i = \boldsymbol{A}\partial_i\widehat{w_1}$; $\boldsymbol{v_i}$ is called the prototype of class i with respect to the sample $\boldsymbol{y}$. Equation (6) shows the residual distance between $\boldsymbol{y}$ and its prototype $\boldsymbol{v_i}$ of class i,

$$r_i(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{v}^i\|_2 = \left\|\boldsymbol{y} - \boldsymbol{A}\partial_i(\widehat{\boldsymbol{w}_1}).\boldsymbol{v}^i\right\|_2 \tag{6}$$

The SRC decision rule is
if $r_l(\boldsymbol{y}) = \min_{\,i}\, r_i(\boldsymbol{y})$, $\boldsymbol{y}$ is assigned to class l.

In the experiment, we consider the **TVM** as the training set. **TVM** has been built by considering the feature vectors $\mathbf{FV}_{\boldsymbol{i}}$ for each topic i as described below.

$$\mathbf{TVM} = [\mathbf{FV_1},\ \mathbf{FV_2}\ \ldots\ \mathbf{FV}_n]^{\mathrm{T}}$$

The dimension of **TVM** is $n \times k$. Now, we consider a user query q as a test sample and convert it into feature vector $\mathbf{FV_q}$ of dimension $1 \times k$ as described in Sect. 2.2.

We apply SRC on **TVM** for reconstruction of $\mathbf{FV_q}$ and assigned nearest topic to $\mathbf{FV_q}$. The procedure is described in Algorithm 1.

Algorithm 1 Query classification using SRC

```
Input: Set of topics T , query Q, Set of unique keywords TRV, Number of topics $n$
Output: Topic related to the query,
    Topic-SRC (T, Q, TRV, $n$)
    TVM ← Φ; FV ← Φ   // Topic-vocabulary matrix & Feature vector
    For each $t$ ϵ T
        FV = Feature-Vector (T, TRV)
    $i$ ← 0
    For $i < n$
        TVM[$i$] = FV [$i$]$^{T}$
    FV$_q$ = Feature-Vector (Q, TRV)
    $t$-Class = SRC (FV$_q$, TVM, $n$)   // Topic of the query
    Return $t$-Class
    Procedure1: Feature-Vector (T, TRV)
    For each $w$ ϵ T
        Calculate P($w$ | $T$)
    FV ← Φ; $i$ ← 0
    For $i$ < length (TRV)
        If T[$i$] = = TRV[$i$]
            FV[$i$] ← P($w$ | $T$) .P($w$)
        Else
            FV[$i$] ← P($w$)
    Return FV
    Procedure 2: SRC (FV$_q$, TVM, $n$)
    W ← Φ; D ← Φ        //Sparse co-efficient vector & Set of distance
    W = pinv(TVM) * FV$_q^{T}$ // Construction of W
    Find sparsest solution W$_s$ for W by equation (5)
    $i$ ← 0
    For each $i < n$
        (FV$_q$)$^{i}_{new}$ = TVM * W$_s$   // Reconstruction of sample vector
        D[$i$]  ← Norm (FV$_q$ , (FV$_q$)$^{i}_{new}$ )
    $t$-Class = min(D)        //Finding minimum residue distance
    Return $t$-Class //Returning nearest topic class of Query
```

3 Results and Discussion

In the experiment, we used a subset of TREC AP (Academy Press) corpus containing 2246 news articles with 10473 unique terms. After preprocessing the document, we apply LDA and obtain 25 topics that are optimum for this experiment. Each topic is visualized as a set of words where each element of the set is assigned with the posterior topic measures. Table 1 shows five different topics with most frequent 10 keywords. Test samples as users' query are classified by executing algorithm1.

The performance of the classifier is evaluated using different statistical measures [18] considering 10 queries for each 25 topics, i.e., 10 × 25 = 250 queries. We considered varied length of query up to five keywords and no significant change is

Table 1 Topics with top 10 keywords

Economy	Administration	Judiciary	Healthcare	Aviation
Oil	Police	Court	Aids	Air
Cents	People	Trail	Health	Space
Price	Killed	Case	Hospital	Flight
Futures	Authorities	Charges	Medical	Plane
Cent	Army	Attorney	Disease	Two
Lower	City	Prison	Drug	Aircraft
Market	Man	Judge	Patients	Planes
Higher	Government	Two	Care	Accident
Million	Officials	Guilty	Federal	Navy
Farmers	Reported	Years	Doctors	Ship

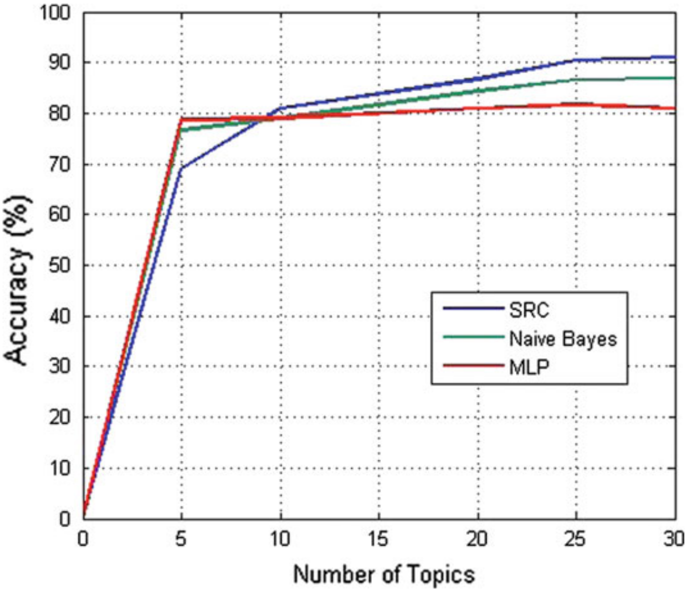

Fig. 2 Accuracy versus number of topics

Table 2 Comparison with different classifiers

Classifiers	Statistical measures							
	Accuracy (%)	Misclassification (%)	TP rate	FP rate	Precision	Recall	F-measure	Specificity
Naive-Bayes	86.6	13.3	0.86	0.015	0.888	0.867	0.868	0.985
Multilayer perceptron	81.66	18.33	0.82	0.02	0.823	0.817	0.819	0.98
SRC	90.4	9.6	0.9	0.09	0.978	0.9	0.94	0.91

reported in the retrieved information. It has been observed statistically that maximum five keywords are provided by most of the users for their query [19]. For instance, the query vectors [judge trial charges guilty]^T, [oil higher market price]^T, [government authorities reported official]^T and [aids patients doctors care]^T of length 4 are classified as ‘Judiciary’, ‘Economy’, ‘Administration’, and ‘Healthcare’.

Accuracy has been improved with the number of topics indicating appropriateness of applying sparse classifier in the proposed method. Figure 2 shows improvement in accuracy with increasing number of topics, not remarkable for other classifiers unlike SRC. High accuracy and high TP rate ensures good precision and recall for retrieval method that guarantees lower misclassification too. Detailed comparison of different classifiers using statistical measures summarized in Table 2 and ROC curve for SRC is shown in Fig. 3.

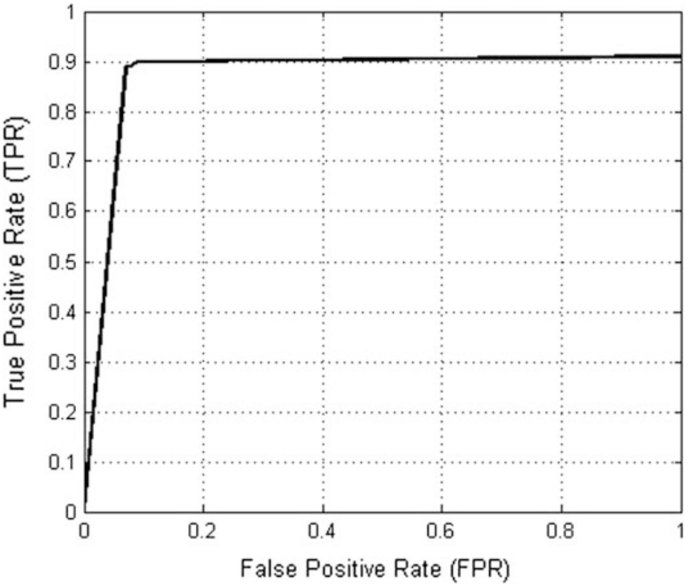

Fig. 3 ROC curve for SRC

4 Conclusion

LDA is a basic and generative model for topic modeling used in the paper for initial latent topic identification. We consider any query as a distribution of words obtained from topics which is sparse with respect to high dimensional input space represented using vocabulary. Proposed approach gives satisfactory results in terms of statistical measures and improves with the size of the training set. For Naïve Bayes classifier performance is better for less number of topic, however, proposed approach using SRC outperforms when size of the training set increases, shown in Fig. 2.

Acknowledgments This research was partially supported by grants from Information Technology Research Academy (ITRA), under the Department of Electronics and Information Technology (DeitY), Government of India.

References

1. Baeza-Yates, R., and Ribeiro-Neto, B.: Modern information retrieval (Vol. 463). ACM press, New York (1999).
2. Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation. The Journal of machine Learning research, Vol. 3, 993–1022 (2003).
3. Manning, C. D., Raghavan, P., and Schütze, H.: Introduction to information retrieval (Vol. 1). Cambridge university press, Cambridge (2008).
4. Salton, G., Singhal, A., Mitra, M., and Buckley, C.: Automatic text structuring and summarization. In: Information Processing & Management, Vol. 33(2), 193–207 (1997).
5. Salton, G., and McGill, M.: Introduction to modern information retrieval. McGraw Hill Book Co, New York (1983).
6. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A.: Indexing by latent semantic analysis. JAsIs, Vol. 41(6), 391–407 (1990).
7. Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala, S.: Latent semantic indexing: A probabilistic analysis. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM, 159–168 (1998).
8. Hofmann, T., and Puzicha, J.: Latent class models for collaborative filtering. In: IJCAI, Vol. 99, 688–693 (1999).
9. Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W.: Joint latent topic models for text and citations. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 542–550 (2008).
10. Shen, Z. Y., Sun, J., and Shen, Y. D.: Collective latent Dirichlet allocation. In: Eighth IEEE International Conference on Data Mining, ICDM'08. IEEE, 1019–1024 (2008).
11. Griffiths, D. M. B. T. L., and Tenenbaum, M. I. J. J. B.: Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems, Vol. 16(17) (2004).
12. Blei, D., and Lafferty, J.: Correlated topic models. Advances in neural information processing systems, Vol. 18(147) (2006).
13. Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A.: Face recognition: A literature survey. ACM computing surveys (CSUR), Vol. 35(4), 399–458 (2003).

14. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M.: Fast collapsed gibbs sampling for latent dirichlet allocation. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 569–577 (2008).
15. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K.: An introduction to variational methods for graphical models. Machine learning, Vol. 37(2), 183–233 (1999).
16. Chen, S. S., Donoho, D. L., and Saunders, M. A.: Atomic decomposition by basis pursuit. SIAM journal on scientific computing, Vol. 20(1), 33–61 (1998).
17. Yang, J., Chu, D., Zhang, L., Xu, Y., and Yang, J.: Sparse representation classifier steered discriminative projection with applications to face recognition. Neural Networks and Learning Systems, IEEE Transactions on, Vol. 24(7), 1023–1035 (2013).
18. Davis, J., and Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. ACM, 233–240 (2006.).
19. http://www.keyworddiscovery.com/keyword-stats.html.