DOI : 10.3745/JIPS.2009.5.1.005

SVD-LDA: A Combined Model for Text Classification

Nguyen Cao Truong Hai*, Kyung-Im Kim* and Hyuk-Ro Park*

Abstract: Text data has always accounted for a major portion of the world's information. As the volume of information increases exponentially, the portion of text data also increases significantly. Text classification is therefore still an important area of research. LDA is an updated, probabilistic model which has been used in many applications in many other fields. As regards text data, LDA also has many applications, which has been applied various enhancements. However, it seems that no applications take care of the input for LDA. In this paper, we suggest a way to map the input space to a reduced space, which may avoid the unreliability, ambiguity and redundancy of individual terms as descriptors. The purpose of this paper is to show that LDA can be perfectly performed in a "clean and clear" space. Experiments are conducted on *20 News Groups* data sets. The results show that the proposed method can boost the classification results when the appropriate choice of rank of the reduced space is determined.

Keywords: *Latent Dirichlet Allocation, Singular Value Decomposition, Input Filtering, Text Classification, Data Preprocessing.*

1. Introduction

The explosion in the volume of information has been accompanied by an increasing challenge of effective content indexing and summarization. Text classification, which automatically assigns predefined categories to new documents, is considered as an effective solution to this problem. It is notable, however, that the high dimensionality of the feature space makes the task much more difficult [1]. For example, the dimensionality of the original feature space, which consists of unique terms (words or phrases) in documents, can be up to hundreds of thousands even for a moderately sized text collection. This is prohibitively high for many learning algorithms [1][11]. To ease the situation, numerous information techniques have been developed one by one.

Latent Semantic Analysis (LSA) is a well-known technique that uses the linear algebra Singular Value Decomposition tool. The key idea of this technique is to map high-dimensional count vectors, such as the ones arising in vector space representations of text documents [3], on to a lower dimensional representation in a so-called latent semantic space. Hoffman's Probabilistic Latent Semantic Analysis (PLSA) marked a significant step forward when using this novel statistical technique for the analysis of two-mode and co-occurrence data. PLSA has applications in the fields of information retrieval and filtering, natural language processing, machine learning from texts, and in other related areas. It took another four years until Blei et al. (2003) suggested the Latent Dirichlet Allocation (LDA), which overcame the limitations of PLSA. LDA provides a probabilistic model at the document level. In the context of text modeling, the topic probabilities provide an explicit representation of a document [4].

However, none of these methods care much about the status of the input space of data. The input space may contain a lot of noise, ambiguity, and even imprecision (missing or incorrect values), all of which may lead to a decrease in the results of classification.

In this paper, we try to find a way to cope with this situation, based on the model suggested by Deerwester. We use the algebra matrix Singular Value Decomposition technique to map the input data space on to a rank-specified reduced space. The main purpose is to remove the noise, ambiguity and perhaps the imprecision from the input space. The results show that the proposed method can somewhat boost the classification results just in the event that the appropriate choice of rank of the reduced space can be determined.

2. Related Work and Motivation

2.1 Latent Semantic Analysis by Singular Value Decomposition

As mentioned in the introduction, the key idea of Latent Semantic Analysis (LSA) is to map documents (and by symmetry terms) on to a vector space of reduced dimensionality, namely the latent semantic space [7]. The mapping is restricted to linearity and is based on a Singular Value Decomposition (SVD) of the co-occurrence table. However, its theoretical foundation remains to a large extent unsatisfactory and incomplete in that LSA does not

Manuscript received February 2, 2009; accepted February 26, 2009.
Corresponding Author: Nguyen Cao Truong Hai
* School of Electronics and Computer Engineering, Chonnam National University, South Korea (justosue@yahoo.com, kyungim@moiza.chonnam.ac.kr, hyukro@chonnam.ac.kr)

define a properly normalized probability distribution and there is no obvious interpretation of the directions in the LSA latent space.

A matrix X of terms and documents, N x M, can be decomposed into the product of three other matrices, where $X = W_0 S_0 D_0^t$, W_0 and D_0 have ortho-normal columns ($W_0^t W_0 = D_0^t D_0 = I$), and S_0 is diagonal. W_0 and D_0 are the matrices of the *left* and *right singular vectors* and S_0 is the diagonal matrix of the *singular values* of X. This is called the Singular Value Decomposition of matrix X. The LSA approximation of X is computed by setting all but the largest L singular values in S_0 to zero, S(LxL). Thus, W_0, D_0 are also reduced to W(NxL) and D(MxL). The result is a reduced model $\hat{X} \approx X = WSD^t$. SVD can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors. Each term and document is represented by its vector of factor values. That is, the "meaning" of a particular term/word, query, or document can be expressed by L factor values, or equivalently, by the location of a vector in the L-space defined by the factors [3][7][11].

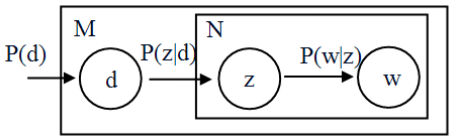


Fig. 1 Graphical model representation of PLSA

2.2 Probabilistic Latent Semantic Analysis

We have M documents containing terms from a vocabulary of size N. The corpus of text documents is summarized in the N by M co-occurrence table X, where $X = (x(w_i, d_j))_{ij}$ stores the number of occurrences of a term w_i in document d_j . This is known as the bag of words model. In addition, there is a hidden (latent) topic variable z_k associated with each occurrence of a word w_i in a document d_j.

The joint probability $P(w_i, d_j , z_k)$ is assumed to have the form of the graphical model shown in Figures 1 and 2. Marginalizing over topics z_k determines the conditional probability $P(w_i|d_j)$:

$$P(w_i|d_j) = \sum_{z_k} P(w_i|z_k) P(z_k|d_j) \quad (1)$$

where $P(z_k|d_j)$ is the probability of topic z_k occurring in document d_j, and $P(w_i|z_k)$ is the probability of word w_i occurring in a particular topic z_k.

However, in the PLSA space, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with over-fitting, and (2) it is not clear how to assign probability to a document outside of the training set [1, 4, 8]. Although PLSA represents a significant step towards probabilistic modeling of textual data, it provides no probabilistic model at the level of documents [7].

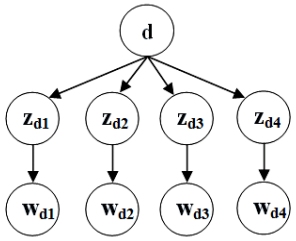


Fig. 1 Another graphical model presentation of PLSA

Examples:

Assume that there is a document d containing 5 characters A, B, C, D and E; and that there is a set of corresponding topics z, including 4 topics z1, z2, z3 and z4. Then we have:

d={A, B, C, D, E} and z={z1, z2, z3, z4}, in which:

- A, D is known to have belonged to z1.
- C, E is known to have belonged to z2.
- B is known to have belonged to z4.

Thus, we can present d as a mixture of portions of z1, z2 and z4.

2.3 Latent Dirichlet Allocation

Unlike PLSA, LDA (Figure 3) treats the multinomial weights P(z|d) over topics as latent random variables. The LDA model is extended by sampling those weights from a Dirichlet distribution, the conjugate prior to the multinomial distribution. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, thus reducing over- fitting [13]. LDA assumes the following generative process for each document in a corpus:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir(α).
3. For each of the N words w_n:

(a) Choose a topic z_k Multinomial(θ).

(b) Choose a word w_n from $p(w_n|z_k,\beta)$, a multinomial probability conditioned on the topic z_k.

By using some generative variables to control the objects of interest (documents, words, and topics), LDA can overcome the limitations of local observation and the problem of the linear increase in the number of parameters in PLSA. Variable α will control the documents, while β will control the words and θ will control the topics. Figure 3 illustrates this idea.

Given the Dirichlet parameters α and β, with a topic mixture θ, a set of K topics $\{z_k\}$, and a set of N words $\{w_n\}$, we have the marginal distribution over topics $\{z_k\}$ of a document:

$$P(d|\alpha,\beta) = \int P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_k} P(z_k|\theta)P(w_n|z_k,\beta)\right)d\theta. \quad (2)$$

Each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. This three-level hierarchical probabilistic model gives LDA the strength to overcome the problems of local training set indexing and parameter over-fitting.

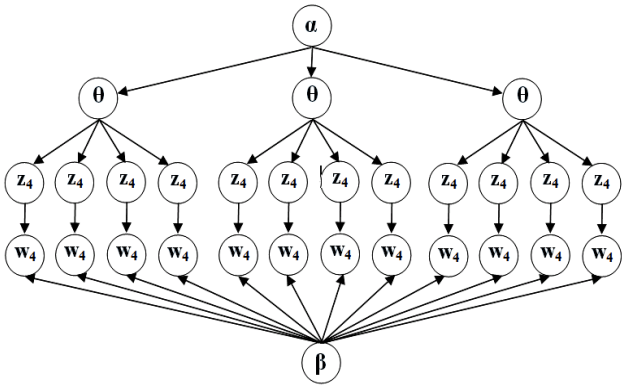


Fig. 2. LDA is an updated PLSA

This method is the most advanced of the three; however, it does not take care of the input space, which may contain a lot of noise, ambiguity and even imprecision, all of which may lead to a decrease in the classification results.

2.4 Motivation

Singular Value Decomposition (SVD) is an interesting algebra tool which has been included in many applications. These employ SVD to compute the least squares fitting of data and to determine the rank, range and null space of a matrix. Many other works use SVD as an effective technique in image compressing/decompressing, image decomposition, and circuit signal filtering or clustering. For text data, SVD has just been significantly used in LSA. The LSA technique led to the idea of using SVD to create a reduced space where documents and terms will be presented with the inference vectors of factor values. The reduced space is significant in that it can avoid the unreliability, ambiguity and redundancy of individual terms as descriptors [3].

LDA, as mentioned above, is one of the best among the infinite mixture of probability models. This line of thinking has led to the key idea of my work. SVD is included in LSA, which is neither well-modeled nor equipped with any random probabilistic models. But, SVD alone can create the reduced space without unreliability, ambiguity and redundancy. This advantage should not be wasted. At the same time, LDA, an updated model of which constitutes a well-designed, three-level hierarchical probabilistic model, can model and classify the data very well. That being the case, why should we not use these strong elements to create a new procedure? The new combined method may have the ability to outperform the old ones. This work has been executed as an answer to the foregoing question.

3. The Combined Model

As mentioned above, the combined model wants to use the strong points of both LSA and LDA. LSA helps to map the original feature description (documents by symmetry terms) on to a vector space of reduced dimensionality, the latent semantic space, while LDA has the exceptional ability to represent and infer the latent relationship among terms/words, documents and topics. To take advantage of these two methods, the combined model tries to perform LDA on the latent semantic space generated by LSA. In this semantic space, the terms, documents and queries are already represented in a way that can avoid the unreliability, ambiguity and redundancy of individual terms as descriptors [3]. The hope here is that LDA can perfectly perform in a "clean and clear" compact space.

Assume that we have M documents of size N_m ($m \in [1, M]$) containing terms from a vocabulary V. The corpus of text documents is then summarized in an M by N co-occurrence table/matrix X, where matrix $X=(x(m, n)_{mn}$ ($n \in [1, N_m]$) stores the number of occurrences of a particular term for the term placeholder [m, n]. The decomposition of X in the way of SVD is the product of three other matrices, $X = U_0S_0V_0^T$, where U_0 and V_0 have ortho-normal columns ($U_0^TU_0 = V_0^TV_0 = I$) and S_0 is diagonal.

The reduce space $\hat{X}$ is reconstructed by setting all but the largest L singular values in S_0 to zero. Since zeros were introduced into S_0, the representation can be simplified by deleting the zero rows and columns of S_0 to obtain a new diagonal matrix S, and then by deleting the corresponding columns of U_0 and V_0 to obtain U, V and S, respectively. The result is a reduced model $X \approx \hat{X} = USV^T$. The new space now has the same size as X but the content is *stressed*. From now on, the later LDA will be performed on this reduced space instead of on X in the hope that within the clean and clear space, free from unreliability, ambiguity and redundancy, LDA can show only the best results.

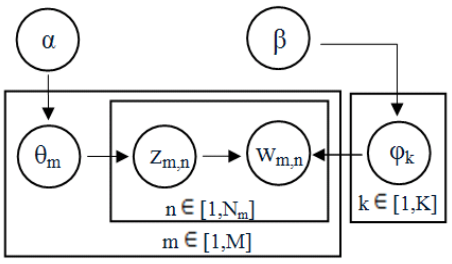


Fig. 3 An LDA generative graphical model

In order to create a reduced space/matrix, we have to choose a value L that is known to be the rank of the new

space. This is not a simple problem. The choice of rank L critically affects our work. Ideally, we want a value of L that is large enough to fit all the real structures in the data, but small enough so that we do not also fit the sampling error or unimportant details. The proper way to make such choices remains an open issue in the factor analytic literature. In this work, we assume that a value of L which yields good retrieval performance is chosen.

A set of K topics is assumed to be already known and fixed. A document W_m is generated by first picking a distribution over topics θ_m from a Dirichlet distribution Dir(α), which determines the topic assignment for words in that document. Then, the topic assignment for each word placeholder [m,n] is performed by sampling a particular topic $z_{m,n}$ from a multinomial distribution Mult(θ_m). And, finally, a particular word $w_{m,n}$ is generated for the word placeholder [m, n] by sampling from the multinomial distribution Mult(φ_k). Such a generative process is shown in Figure 4 and described as a pseudo code in the table below.

Table 1 A pseudo code for the LDA generation process

for *all topics k in [1,K] do*

sample mixture components k $\sim Dir(\beta)$

end for

for *all documents m in [1,M] do*

sample mixture proportion m $\sim Dir(\alpha)$

sample document length $N_m \sim Poiss(\xi)$

for *all words n in [1,* N_m*] do*

sample topic index $z_{m,n} \sim Mult(\theta_m)$

sample term for word $w_{m,n} \sim Mult(\varphi_k)$

end for

end for

Given the Dirichlet parameters α and β, with a topic mixture θ, a set of K topics, the likelihood of a document W_m can be computed by integrating over θ_m, φ_k and summing over Z_m as below.

$$P(W_m|\alpha,\beta) = \iint P(\theta_m|\alpha)p(\varphi_k|\beta)\prod_n^{N_m} P(W_{m,n}|\theta_m\varphi_k)d\varphi_k d\theta_m \quad (3)$$

By calculating the product of the likelihood of the whole data collection $\{W_m\}$, we can obtain the likelihood of all documents:

$$P(W|\alpha,\beta) = \prod_{m-1}^{M} P(W_m|\alpha,\beta) \quad (4)$$

4. Experiments

Experiments are set up using the data set "20 Newsgroups", which is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his paper titled *Newsweeder: Learning to filter netnews*, though he does not explicitly mention this collection. The '20 newsgroups' collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. The version chosen ("by date" version) is sorted by date into training (60%) and test (40%) sets, and does not include cross-posts (duplicates) or newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date).

The experiments are performed using 3 particular methods: PLSA, LDA, and the method proposed in this paper. We will conduct experiments on 405,628 data entries divided into train data (208,994 entries) and test data (196,634 entries). This amount of data contains 5 topics. Because the cost of computing SVD is extremely high, we decided to use a subset of the data set. The Matlab smoothed code for LDA and PLSA is from Jakob Verbeek (LEAR team). All options are set to default, which means estimates for gammas and betas are used. Gammas values will show the probability of one document being assigned to a topic. At the same time, the beta values show the probability of one word being assigned to a topic.

We check the accuracy in a situation where there are T=5 topics, and with SVD-LDA the rank L varies among 100, 300, 500, 700, 900 (L=3000 is the maximum with this amount of data).

With (T,L)=(5,100) or (T,L)=(5,300), SVD-LDA is worse than PLSA and LDA. But when the rank L is slightly increased to L=500, we can see that the SVD-LDA method has already passed PLSA. Then, at the point where (T,L)=(5,700), SVD-LDA is the winner. Compared to PLSA, this represents an improvement of 1.89%, while compared to LDA, an improvement of 0.71%. However, a decrease with L=900 was also noticed. From the experiments, we can tell that our method is capable of boosting the results of the LDA method with the right choice of rank L. However, L=700 may not be the best choice: the best choice must be somewhere between 500 and 900.

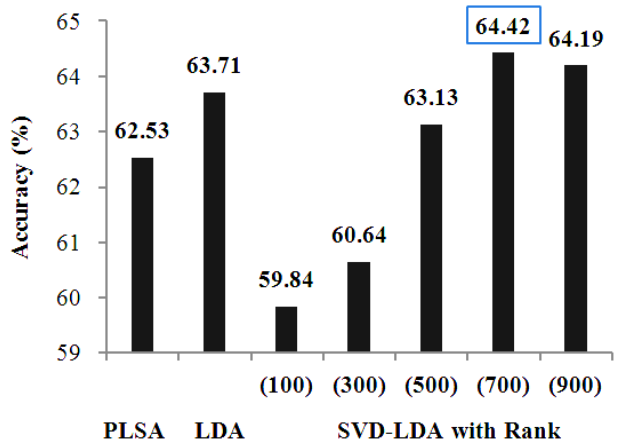


Fig. 5 Accuracy changes over the ranks (T=5).

The results show that the proposed method is slightly better than LDA and PLSA, having the appropriate rank L of the reduced space. This once again stresses the critical role of the choice of rank L, which still requires further research before it can be used effectively.

In Fig. 2, the original clean images are located in the first column, while the second column contains the noisy images and the third one contains the de-noised images of our proposed algorithm.

5. Conclusions

We have presented a combined model SVD-LDA which can be used for text classification. This approach attempts to deal with noise, ambiguity and perhaps imprecision in the input space. To ease the situation, we suggest the idea of using the algebra matrix Singular Value Decomposition technique as a filter. The SVD filter will map the input data space on to a rank-specified reduced space. Then, LDA will be performed on this new space.

The experimental results show that the proposed method can somewhat boost the classification results in the event that the appropriate choice of rank of the reduced space can be determined. This also proves that the classification results can be improved with this combination, which may lead us to (an) other effective one(s).

References

[1] Zhiwei Zhang, Xuan-Hieu Phan, Susumu Horiguchi, "An Efficient Feature Selection using Hidden Topics in Text Categorization," *22nd International Conference on Advanced Information Networking and Application*, 2008.

[2] A. Berger, A. D. Pietra, and J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, Vol.22, no.1, 1996, pp.39-71.

[3] S. Deerwester, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Info*, Science, Vol.41, No.6, 1990, pp.391-407.

[4] D. M. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *JMLR*, Vol.3, 2003, pp.993-1022.

[5] Ramesh Nallapati and William Cohen, "Link-PLSA-LDA: A new unsupervised model for topics and the influence of blogs," *AAAI*, 2008.

[6] G. Heinrich, "Parameter estimation for text analysis," *Technical report - University of Leipzig, Germany*, 2005.

[7] T. Hofmann, "Probabilistic latent semantic indexing," *Proceedings of SIGIR'99*, 1999.

[8] Tuomo Kakkonen, Niko Myller, and Erkki Sutinen, "Applying Latent Dirichlet Allocation to Automatic Essay Grading," *FinTAL 2006*, LNAI 4139, pp. 110–120, 2006.

[9] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol.34, no.1, 2002, pp.1-47.

[10] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp.412-420.

[11] C. Andrieu, N. D. Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, Vol.50, 2003, pp. 5–43.

[12] T. Hofmann, J. Puzicha, and M. I. Jordan, "Unsupervised learning from dyadic data," *Advances in Neural Information Processing Systems*, Volume 11. MIT Press, 1999.

[13] B.C. Russell, A.A. Efros, J. Sivic, W.T. Freeman, and A. Zisserman, "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections," *Proceedings of CVPR*, June, 2006.

[14] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM TOIS*, Vol.22, no.1, 2004, pp.89-115.

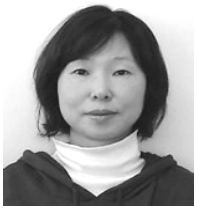
[15] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," *Proc. UAI*, 2002.

[16] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol.34, no.1, 2002, pp.1-47.

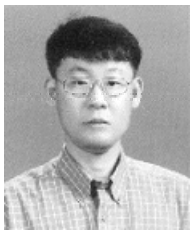
[17] http://www.puffinwarellc.com/p3b.htm.

[18] http://en.wikipedia.org/wiki/Information_retrieval.

Nguyen Cao Truong Hai
He received a BS degree in Computer Science from the University of Sciences, Ho Chi Minh City, Vietnam and an MS degree in Computer Engineering from Chonnam National University in 2006 and 2009, respectively. He is currently undertaking a doctorate course as a member of the Information Retrieval Lab at Chonnam University. His research interests include Text Classification, Content- based Image Search, Cryptography, Object Recognition, Natural Language Processing, and Data Mining.



Kyung-Im Kim
She received an MS degree from Dongguk University. She is now undertaking a doctorate degree course as a member of the Information Retrieval Lab at Chonnam National University. Her research interests are in the areas of Information Retrieval, Image Retrieval, Ontology, and the Semantic Web.



Hyuk-Ro Park
He received a BS degree from Seoul National University; and MS and Ph.D. degrees from KAIST. He has been a professor at Chonnam National University since 1999. He is now in charge of the Information Retrieval Lab at Chonnam University. His research interests are in the areas of Information Retrieval, Natural Language Processing, Database System, and Data Mining.