IMPROVING TEXT CATEGORIZATION BY USING A TOPIC MODEL

Wongkot Sriurai

Department of Mathematics Statistics and Computer, Faculty of Science,
Ubon Ratchathani University, Thailand
scwongsr@ubu.ac.th

ABSTRACT

Most text categorization algorithms represent a document collection as a Bag of Words (BOW).The BOW representation is unable to recognize synonyms from a given term set and unable to recognize semantic relationships between terms. In this paper, we apply the topic-model approach to cluster the words into a set of topics. Words assigned into the same topic are semantically related. Our main goal is to compare between the feature processing techniques of BOW and the topic model. We also apply and compare between two feature selection techniques: Information Gain (IG) and Chi Squared (CHI). Three text categorization algorithms: Naïve Bayes (NB), Support Vector Machines (SVM) and Decision tree, are used for evaluation. The experimental results showed that the topic-model approach for representing the documents yielded the best performance based on F_1 measure equal to 79% under the SVM algorithm with the IG feature selection technique.

KEYWORDS

Text categorization , Bag of Words,Topic-Model

1. INTRODUCTION

Today, the amount of text documents is increasing with an explosive rate. Text categorization has become one of the key techniques for managing and organizing those documents and also assists the information retrieval process in filtering the documents for a specific topic. Text categorization process usually adopts the supervised machine learning algorithms for learning the classification model [1-2]. To prepare the term feature set, the bag of words (BOW) is usually applied to represent the feature space. Under the BOW model, each document is represented by a vector of weight values calculated from, for example, the term frequency–inverse document frequency (TF-IDF), of a term occurring in the document. The BOW is very simple to create, however, this approach discards the semantic information of the terms (i.e., synonym). Therefore, different terms whose meanings are similar or the same would be represented as different features. As a result, the performance of a classification model learned by using the BOW model could become deteriorated.

In this paper, we apply the topic model approach to cluster the words into a set of topics. The concept of topic model, the words (or terms) are clustered into the same topics. Given *D* is a set of documents composed of a set of words (or terms) *W*, *T* is a set of latent topics, that are created based on a statistical inference on the term set *W*. In the paper, the topic model is applied based on the Latent Dirichlet Allocation (LDA) [3] algorithm to produce a probabilistic topic model from a web page dataset. The topic model can help capture the synonyms, hypernyms and hyponyms of a given word such as the words “human” (hypernym) and “John” (hyponym) would be clustered into the same topic.

DOI : 10.5121/acij.2011.2603

In addition, the words “film” (synonym) and “movie” (synonym) would also be clustered into the same topic. The topic model improve the performance of a classification model by (1) reducing the number of features or dimensions and (2) mapping the semantically related terms into the same feature dimension. Our main goal is to compare between the feature processing techniques of BOW and the topic model. We also apply and compare between two feature selection techniques: Information Gain (IG) and Chi Squared (CHI). Three text categorization algorithms: Naïve Bayes (NB), Decision tree (Dtree) and Support Vector Machines (SVM), are used for evaluation.

The remainder of this paper is organized as follows. In the next section we provide a brief review of related works. Section 3 presents the proposed framework. In the section 4, the experimental result is shown with the discussion. In Section 5, we conclude the paper and put forward the directions of our future works.

2. BACKGROUND AND RELATED WORKS

In this section, we review text categorization and feature selection and description of the topic model based on the Latent Dirichlet Allocation algorithm.

2.1. Text Categorization and Feature Selection

Text categorization (also known as *document classification*) is a supervised learning task, concerning the assigning of category labels to new documents based on the information learned from a labeled training data [1-2],[4]. Text categorization is a well-studied research area related to information retrieval, machine learning and text mining. Yang and Liu [1] presented a comparative study on five text categorization algorithms including Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Neural Network (NNet), Linear Least-Squares Fit (LLSF) and Naive Bayes (NB). Based on the thorough evaluation, the SVM has been shown to yield the best performance compared to other classification algorithm.

Most known text categorization algorithms represent a document collection as BOW. Using the BOW usually leads to an explosion in the number of features, so that even tens of thousands of features. The major problem of this representation is the high dimensions of feature space and information loss of the original texts. Feature selection (FS) [5] is one technique to deal with such problems. The main idea of feature selection is to select a subset of terms occurring in the training set and using only this subset as features in text categorization. There are several previous research works, which proposed and compared among many feature selection methods. For example, Dash and Liu [6] gave a survey of feature selection methods for classification. Yang and Pedersen [5] compare state of the art five feature selection methods, such as document frequency (DF), information gain (IF), mutual information (MI), chi-square (CHI) and term strength (TS). The study results found IG and CHI to be the most effective.

2.2. A Review of Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for a set of documents [3], [7]. The concept of this approach is that documents are represented as random mixtures over latent topics. Each topic is represented by a probability distribution over the terms. Each article is represented by a probability distribution over the topics. LDA has also been applied for identification of topics in a number of different areas such as collaborative filtering [3], content-based filtering [7],[8] and classification [9-13]. In the generation process of LDA, a word w is generated from combination of topics z in a document d and sampling a word from topic-word distribution. The process is generated using Equation (1)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \tag{1}$$

where $P(z_i = j)$ represents the probability of topic j was sampled for word token w_i in document; $P(w_i \mid z_i = j)$ represents the probability of w_i under topic j ,and T is the number of topics. To simplify notation, let $\phi^{(j)} = P(w \mid z = j)$ refer to multinomial distribution over words for topic j , and $\theta^{d} = P(z)$ refer to multinomial distribution over topics for document d . The estimated parameters ϕ and θ are the basis for latent-semantic representation of words and documents. The LDA model is shown in Figure 1 [3]. The model has three levels to the representation. For the generating corpus process, the variables α and β are the corpus-level parameters, that are supposed to be sampled once.

The α and β are hyperparameters for the document-topic and topic-word Dirichlet distributions, respectively. The variable θ is a document-level variable, sampled once per document. Finally, the variable z is a word-level variable, and the variable w is a word-level variable, sampled once for each word in each document. The variable N_d is the number of word tokens in a document d and variable D is the number of documents.

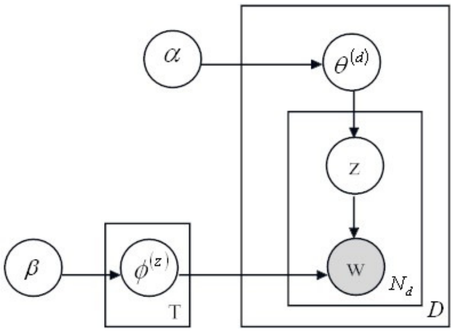

Figure 1. The Latent Dirichlet Allocation (LDA) model [3]

In the LDA model, there are many algorithms for estimating parameters, such as variation Bayes [3], Gibbs sampling [7] and expectation propagation [14]. This paper, we choose Gibbs sampling algorithm. Gibbs sampling algorithm is easy to implement and requires little memory. The procedure for LDA model uses Gibbs sampling to estimate topics from the document collection as well as estimate the word-topic and topic-document probability distributions. The Gibbs sampler is used to compute the conditional probability in Equation (2).

$$P(z_i = j \mid z_{-i}, w_i, d_i, .) \propto \frac{C^{WT}_{w_{ij}} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W.\beta} . \frac{C^{DT}_{d_{ij}} + \alpha}{\sum_{t=1}^{T} C^{DT}_{dt} + T.\alpha} \tag{2}$$

where $z_i = j$ represents the topic assignment of word token w_i to the topic j , z_{-i} refers to the topic assignments of all other word tokens, and "." refers to all other known or observed information. W is the number of word tokens, D is the number of documents, and T is the number of the topics. The C^{WT} and C^{DT} are matrices of counts with dimensions a word-topic matrix and a topic-document matrix, respectively. C_{wj}^{WT} is composed of the number of times word w_i is assigned to topic j, not including the current word token w_i and C_{dj}^{DT} is composed of the number of times topic j is assigned to some word token in document d, not including the current word token w_i. The suitable values for α and β are discussed in Steyvers and Griffiths [6]. Next the sampling process, we can estimate parameters ϕ and θ with equations (3) and (4).

$$\hat{\phi}_i^{(j)} = \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W \cdot \beta} \quad (3)$$

$$\hat{\theta}_j^{(d)} = \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dt}^{DT} + T \cdot \alpha} \quad (4)$$

3. PROPOSED FRAMEWORK

Figure 2 illustrates the proposed framework of feature representations for learning the classification models. In our proposed framework, we evaluated among two different feature representations: (1) applying the simple BOW model and (2) applying the topic model. Each approach is described in details as follows.

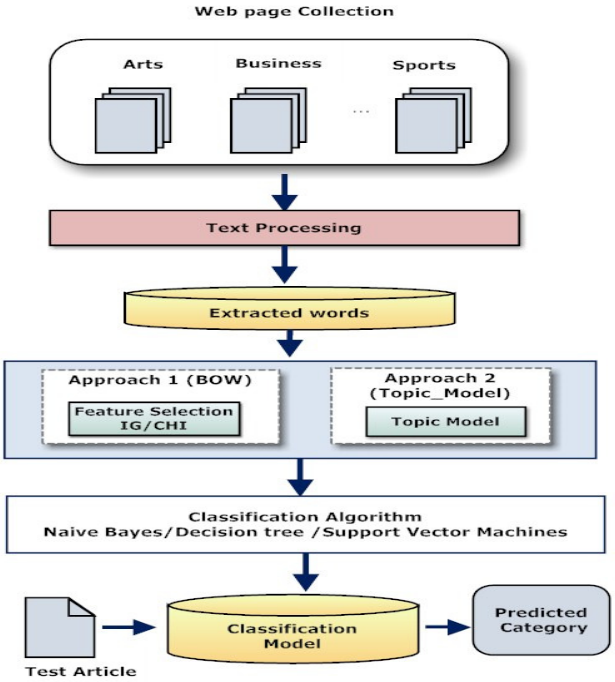

Figure 2. The proposed framework

Approach 1 (BOW): Given a web page collection, the process of text processing is applied to extract terms. For filtering the set of term process, two feature selection techniques are compared such as Information Gain (IG) and Chi Squared (CHI) [4].We apply the Naïve Bayes (NB), Decision tree (Dtree) and Support Vector Machines (SVM) to learn the classification model. Three models are investigated to evaluate the performance of category prediction.

Approach 2 (Topic_Model): Given a web page collection, the process of text processing is applied to extract terms. The set of term is originated by applying the topic model based on the LDA algorithm. The result is the topic probability representation for each article.We apply the Naïve Bayes (NB), Decision tree (Dtree) and Support Vector Machines (SVM) to learn the classification model. Three models are used to estimate the performance of category prediction**.**

4. EXPERIMENTS

4.1. Web page collection

In our experiments, we use a collection of documents obtained the dmoz collection[1]. We selected 13 categories : Arts , Business, Computers, Games, Health, Home, News, Recreation, Reference, Science, Shopping, Society and Sports. We selected 600 documents for each category. The total number of documents is 7,800.

4.2. Experiments

For running our experiments, we use the linguistic analysis tool called LingPipe[2] based on LDA algorithm. LingPipe is a Java API for the linguistic analysis on natural language data. The LDA algorithm is provided under the LingPipe API to investigate the experimental result. The number of topic is set equal to 200 and the number of epochs to 2,000. We used WEKA[3] tool in text classification process. WEKA is a java based on open-source machine learning tool.

4.3. Evaluation Metrics

For the experiment, the precision, recall and F_1 measure is used [2] for evaluating the text classification. All algorithms are examined by using the 10-fold cross validation.

Precision (P) is the percentage of the predicted documents for a given category that are classified correctly, defined as:

$$precision = \frac{(categories\ found \quad and \quad correct)}{(total \quad categories \quad found)} \tag{5}$$

Recall (R) is the percentage of the documents for a given category that are classified correctly, defined as:

$$recall = \frac{(categoriesfound \quad and \quad correct)}{(total \quad categories \quad correct)} \tag{6}$$

F_1 measure is a measure that combines precision and recall. F_1 measure ranges from 0 to 1 and lower the poor. F_1 measure can be defined as follows:

[1] http://www.dmoz.org

[2] http://alias-i.com/lingpipe

[3] http://www.cs.waikato.ac.nz/ml/weka/

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

4.4. Experimental results

Table 1 shows the results in terms of precision, recall and F_1 are the averaged values calculated across all *10-fold* cross validation experiments. From this table, the results of classification model based on two models between the BOW model and Topic_Model, the approach of Topic_Model yielded a higher performance compared to applying the BOW model. The IG feature selection method is suitable for Naïve Bayes (NB), Decision tree (Dtree) and Support Vector Machines (SVM) algorithms. The approach of feature representations with the topic model and using Support Vector Machines (Topic_Model_SVM) to learn the classification model, however, yielded a higher performance compared to using Naïve Bayes (Topic_Model_ NB) and using Decision tree (Dtree). The experimental results showed that the Topic-model approach for representing the documents yielded the best performance based on F_1 measure equal to 79% under the SVM algorithm with the IG feature selection technique; improvement of 11.1% over the BOW model (BOW_SVM).

Table 1. Classification results base on two feature representations approach

Methods	**IG**			**CHI**		
	P	**R**	**F_1**	**P**	**R**	**F_1**
BOW_NB	0.623	0.616	**0.619**	0.594	0.607	**0.600**
BOW_Dtree	0.658	0.661	**0.660**	0.600	0.661	**0.629**
BOW_SVM	0.677	0.682	**0.679**	0.669	0.666	**0.667**
Topic_Model_ NB	0.676	0.696	**0.686**	0.676	0.680	**0.678**
Topic_Model_ Dtree	0.725	0.712	**0.719**	0.693	0.697	**0.695**
Topic_Model_ SVM	0.800	0.780	**0.790**	0.796	0.771	**0.783**

5. CONCLUSIONS AND FUTURE WORKS

To improve the performance of text categorization based on the bag of words feature representation. This research, the Latent Dirichlet Allocation algorithm is used to cluster the term features into a set of latent topics. We find, the same topic will has terms that are semantically related.We compared between the feature processing techniques of BOW model and the topic model. In the research, two feature selection techniques are compared, Information Gain (IG) and Chi Squared (CHI). Three text categorization algorithms: Naïve Bayes (NB), Support Vector Machines (SVM) and Decision tree, are used for evaluation. From the experimental results, the approach of feature representation with the topic model and using Support Vector Machines to learn the classification model, yielded the best performance with the F_1 measure of 79%; an improvement of 11.1% over the BOW model.

In our future work, we plan to evaluate our proposed method on more interesting web 2.0 contents such as web blogs. Neighboring web pages may be included into classification approach for further improvement.

REFERENCES

[1] Yang, Y.& Liu, X. (1999) "Are-examination of text categorization Methods". In Processing of the ACM SIGIR, pp. 42–49.

[2] Sebastiani, F. (2002) "Machine learning in automated text Categorization". ACM Computing

Surveys 34(1), pp. 1–47.
[3] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) "Latent Dirichlet Allocation". Journal of Machine Learning Research, vol 3, pp. 993-1022.
[4] Kumar D. M. (2010) " Automatic induction of rule based text categorization". International Journal of Computer Science & Information Technology (IJCSIT), vol 2, pp.163-172.
[5] Yang, Y. & Pederson, J.O. (1997) "A comparative Study on Feature Selection in Text Categorization", In Proceedings of the 14th International Conference on Machine Learning, pp. 412-420.
[6] Dash, M. & Liu, H. (1997) "Feature Selection for Classification". Intelligent Data Analysis, Vol.1, no.3, pp. 131-156.
[7] Steyvers, M. & Griffiths, T. (2006) "Probabilistic topic models". In: T., Landauer, D., McNamara, S., Dennis, and W., Kintsch, (eds), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum
[8] Haruechaiyasak, C. & Damrongrat, C. (2008) "Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools". The Eleventh International Conference on Asian Digital Libraries (ICADL 2008), pp.339-342.
[9] Bıro, I. & Szabo, J. (2009) "Latent Dirichlet Allocation for Automatic Document Categorization". In Proceedings of the 19th European Conference on Machine Learning and 12th Principles of Knowledge Discovery in Databases.
[10] Bıro, I. & Szabo, J. (2010) "Large scale link based latent Dirichlet allocation for web document classification".
[11] Tasc, S. & Güngör, T. (2009) "LDA-based Keyword Selection in Text Categorization", 24th International Symposium on Computer and Information Sciences (ISCIS 2009).
[12] Viet, Ha-Thuc & Renders, J.M. (2011) "Large-Scale Hierarchical Text Classification without Labelled Data". In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 685-694.
[13] Liu, Z. Li, M., Liu, Y. & Ponraj, M. (2011) "Performance Evaluation of Latent Dirichlet Allocation in Text Mining". International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp.2695-2698.
[14] Minka, T. & Lafferty, J. (2002) "Expectation propagation for the generative aspect mode". In UAI

Authors

Wongkot Sriurai received B.Sc. degree in Computer Science, M.S. degree in Information Technology from Ubon Ratchathani University and Ph.D. degree in Information Technology from King Mongkut's University of Technology North Bangkok. Her current research interests Web Mining, Information filtering and Recommender system.