Improving Web Page Classification by Integrating Neighboring Pages via a Topic Model

Wongkot Sriurai, Phayung Meesad, Choochart Haruechaiyasak

Department of Information Technology
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

Department of Teacher Training in Electrical Engineering
Faculty of Technical Education
King Mongkut's University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Pathumthani 12120, Thailand

s4970290021@ kmutnb.ac.th
pym@kmutnb.ac.th
choochart.haruechaiyasak@nectec.or.th

Abstract: This paper applies a topic model to represent the feature space for learning the Web page classification model. Latent Dirichlet Allocation (LDA) algorithm is applied to generate a probabilistic topic model consisting of term features clustered into a set of latent topics. Words assigned into the same topic are semantically related. In addition, we propose a method to integrate the additional term features obtained from neighboring pages (i.e., parent and child pages) to further improve the performance of the classification model. In the experiments, we evaluated among three different feature representations: (1) applying the simple BOW model, (2) applying the topic model on current page, and (3) integrating the neighboring pages via the topic model. From the experimental results, the approach of integrating current page with the neighboring pages via the topic model yielded the best performance with the F1 measure of 84.51%; an improvement of 23.31% over the BOW model.

1 Introduction

Today, the amount of Web documents (e.g., Web pages, blogs, emails) is increasing with an explosive rate. Text categorization is a widely applied solution for managing and organizing those documents. For example, a text categorization model can be used to assist 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 [Du98], [YP97]. 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 a topic model to represent the feature space for learning the Web page classification model. Words (or terms), which are statistically dependent under the topic model concept, are clustered into the same topic. Given a set of documents *D* consisting of a set of terms (or words) *W*, a topic model generates a set of latent topics *T* based on a statistical inference on the term set *W*. In this paper, we applied the Latent Dirichlet Allocation (LDA) algorithm to generate a probabilistic topic model from a Web page collection. A topic model can help capture the hypernyms, hyponyms and synonyms of a given word. For example, the words "vehicle" (hypernym) and "automobile" (hyponym) would be clustered into the same topic. In addition, the words "automobile" (synonym) and "car" (synonym) would also be clustered into the same topic. The topic model helps 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.

In addition to the concept of topic model, our proposed method also takes an advantage of hyperlink structure of the Web. Given a Web page (denoted by *current page*), there are typically incoming links from *parent* pages and outgoing links to *child* pages. Both parent and child pages are collectively referred to as the *neighboring* pages. Using the additional terms from the neighboring pages could help increase more evidence for learning the classification model. However, the terms from current page should be weighted higher than terms from neighboring pages. Therefore, the proposed method for integrating neighboring information provides a function for varying the weight values of terms coming from the parent pages and the ones from the child pages. Using the Support Vector Machines (SVM) as the classification algorithm, the experimental results showed that by integrating the additional neighboring information via a topic model, the classification performance under the F1 measure was significantly improved over the simple BOW model.

The rest of this paper is organized as follows. In next section we provide a brief review of related works. Section 3 presents the proposed framework of feature representation via the topic model for learning the Web page classification models. Section 4 presents experiments with some discussion on the results. In Section 5, we conclude the paper and put forward the directions of our future works.

2 Related Works

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 [Du98], [YP97]. Text categorization is a well-studied research area related to information retrieval, machine learning and text mining. A number of machine learning algorithms have been introduced and applied for the task of text classification including the Support Vector Machines (SVM) [Jo98], [Va95]. The SVM has been shown to yield the best performance compared to other classification algorithm in many previous works. In this paper, we adopt the SVM in our experiments.

In the domain of the Web, text categorization has been applied for classifying Web pages. Recent works in Web page classification proposed some methods to include the information from neighboring Web pages to learn the model. The information of neighboring pages is, for example, title and surrounding text of anchor text [AGS99], [SLN02], [Zh07]. Furnkranz [Fu99] proposed a classification method using anchor text, surrounding text of anchor text that precedes the hyperlink. Shen et al. [Sh06] proposed an approach to compare of implicit and explicit links for Web page classification. The experimental results showed that the use of the implicit links is better than using explicit links in classification.

Qi and Davison [QD06] proposed a method to improve Web page classification by utilizing the class information from neighboring pages in the link graph. The categories represented by four kinds of neighbors (parents, children, siblings and spouses) are combined to help with the page in question. Experiments showed that sibling pages are the most important type of neighbor to use. Qi and Davison [QD08] proposed a method by utilizing a weighted combination of the contents of neighbors to generate a better virtual document for classification. Their experimental results showed that including a weighted value from neighboring pages helps improve the performance of Web page classification. Chen and Choi [CC08] presented an automatic genre-based Web page classification system, which can work either independently or in conjunction with other topic-based Web page classification system.

In this paper, we also apply the neighboring information for improving the classification model. However, we adopt the topic model to represent the feature space. There have been many studies on discovering latent topics from text collections [SG06]. Recently, the Latent Dirichlet Allocation (LDA) has been introduced as a generative probabilistic model for a set of documents [BNJ03]. The basic idea behind 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 classification and collaborative filtering [BNJ03].

The process to generate a topic model can be explained as follows. The input data for the LDA algorithm consists of an article collection which is a set of *m* documents denoted by $D = \{D_0, \ldots, D_{m-1}\}$. The LDA algorithm generates a set of *n* topics denoted by $T = \{T_0, \ldots, T_{n-1}\}$. Each topic is a probability distribution over *p* words denoted by $T_i = [w^i_0, \ldots, w^i_{p-1}]$, where w^i_j is a probabilistic value of word *j* assigned to topic *i*. Based on this topic model, each document can be represented as a probability distribution over the topic set *T*, i.e., $D_i = [t^i_0, \ldots, t^i_{n-1}]$, where t^i_j is a probabilistic value of topic *j* assigned to document *i* [HD08].

3 Feature Representation via the Topic Model

Figure 1 illustrates the proposed framework of feature representations for learning the Web page classification models. In our proposed framework, we evaluated among three different feature representations: (1) applying the simple BOW model on current page, (2) applying the topic model on current page, and (3) integrating the neighboring pages via the topic model. Each approach is described in details as follows.

Approach 1 (BOW): Given a Web page collection, the process of text processing is applied to extract terms. The set of terms is then filtered by using the feature selection technique, information gain (IG) [DL97]. Once the term features are obtained, we apply the Support Vector Machines (SVM) to learn the classification model. The model is then used to evaluate the performance of category prediction.

Approach 2 (TOPIC_CURRENT): Given a Web page collection, the process of text processing is applied to extract terms. The set of terms is then generated by using the topic model based on the LDA algorithm. The output from this step is the topic probability representation for each article. The Support Vector Machines (SVM) is also used to learn the classification model.

Approach 3 (TOPIC_INTEGRATED): The main difference of this approach from Approach 2 is we integrate the additional term features obtained from the neighboring pages to improve the performance of Web page classification. The process of integrating the neighboring pages is explained as follows.

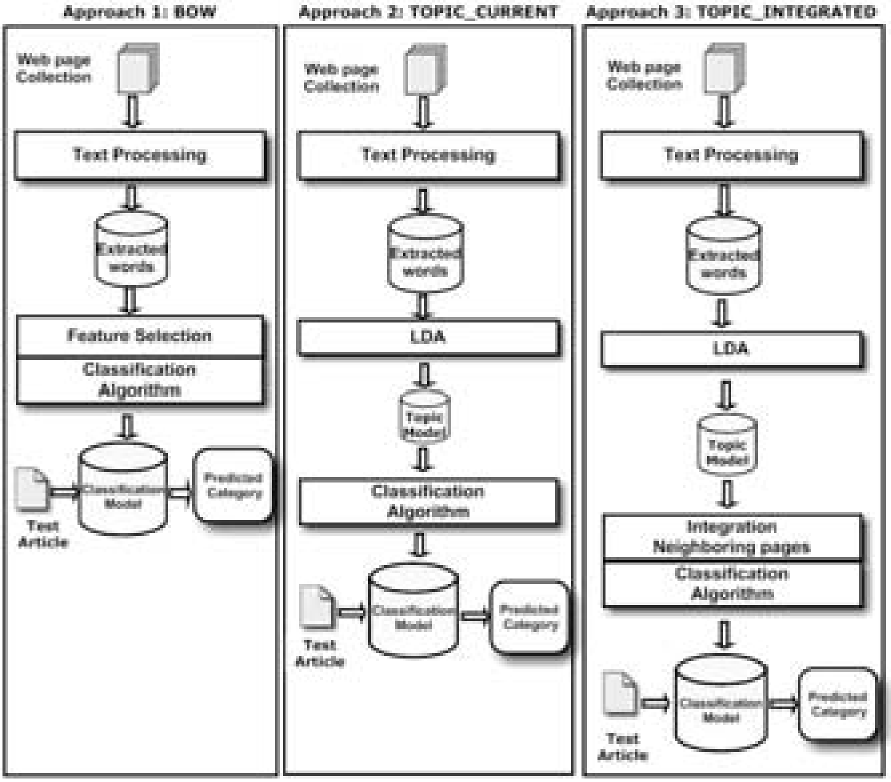

Figure 1: The proposed feature representation framework

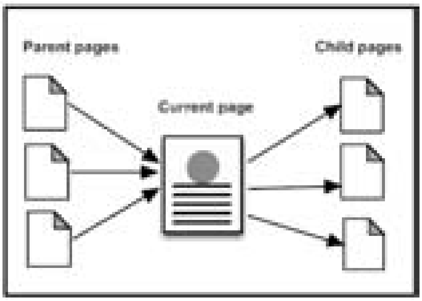

Figure 2: A Current Web page with two types of neighboring pages

Figure 2 shows two types of neighboring pages, parent and child pages. Given a Web page (denoted by *current page*), there are typically incoming links from *parent* pages and outgoing links to *child* pages. Both parent and child pages are collectively referred to as the *neighboring* pages. Using the additional terms from the neighboring pages could help increase more evidence for learning the classification model.

In this paper, we vary a weight value of neighboring pages from zero to one. A weight value equals to zero means the neighboring page is not included in the feature representation. Under this approach, terms from different page types (i.e., current, parent and child) are first transformed into a set of n topics denoted by $T=\{T_0,...,T_{n-1}\}$ by using the LDA algorithm. The weight values from 0 to 1 are then multiplied to the topic dimension T_i of parent and child pages. The combined topic feature vector by integrating the neighboring topic vectors with adjusted weight values can be computed by using the following equation:

$$T(Integrated) = T(current\ page) + w_p \times T(parent\ pages) + w_c \times T(child\ pages) \qquad (1)$$

where $T(Integrated)$, $T(current\ page)$, $T(parent\ pages)$ and $T(child\ pages)$ are topic sets of integrated page, current page, parent pages and child pages, respectively. w_p and w_c are the weights of parent pages and child pages, respectively. The values of w_p and w_c are varied from 0.0 to 1.0 with 0.1 interval.

4 Experiments and Discussion

In our experiments, we use a collection of articles obtained the Wikipedia Selection for Schools, which is available from the SOS Children's Villages Web site[1]. There are 15 categories: art, business studies, citizenship, countries, design and technology, everyday life, geography, history, IT, language and literature, mathematics, music, people, religion and science. The total number of articles is 4,625.

We used the LDA algorithm provided by the linguistic analysis tool called LingPipe[2] to run our experiments. LingPipe is a suite of Java tools designed to perform linguistic analysis on natural language data. LingPipe tools include a statistical named-entity detector, text classification and clustering. In this experiment, we apply the LDA algorithm provided under the LingPipe API and set the number of topics equal to 200 and the number of epochs to 2,000.

For text classification process, we used WEKA[3], an open-source machine learning tool, to perform the experiments. The standard performance metrics for evaluating the text classification used in the experiments are precision, recall and F1 measure [Du98]. We tested all algorithms based on the 10-*fold* cross validation.

[1] SOS Children's Villages Web site. http://www.soschildrensvillages.org.uk/charity-news/wikipedia-for-schools.htm

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

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

We start by evaluating the weight values of neighboring pages under Approach 3. Table 1 shows the results of weight value adjustment on parent pages and child pages based on Eq. (1). The best weight value of parent pages is equal to 0.4 with the F1 measure of 0.8463. For the child pages, the maximum value of F1 measure is 0.8463 with the weight value of 0.2. The results showed that using information from parent pages is more effective than child pages for improving the performance of a classification model. The reason is due to the parent pages often provide terms, such as in the anchor texts, which provide additional descriptive information of the current page.

Neighbors	w_p	P	R	F1
Parent pages	0.1	0.8587	0.8319	0.8440
	0.2	0.8575	0.8321	0.8435
	0.3	0.8575	0.8323	0.8439
	0.4	**0.8569**	**0.8313**	**0.8463**
	0.5	0.8566	0.8318	0.8428
	0.6	0.8533	0.8279	0.8391
	0.7	0.8541	0.8274	0.8391
	0.8	0.8543	0.8251	0.8378
	0.9	0.8527	0.8225	0.8358
	1	0.8524	0.8225	0.8355

Neighbors	w_c	P	R	F1
Child pages	0.1	0.8590	0.8343	0.8455
	0.2	**0.8577**	**0.8335**	**0.8463**
	0.3	0.8576	0.8332	0.8442
	0.4	0.8548	0.8305	0.8416
	0.5	0.8560	0.8333	0.8437
	0.6	0.8576	0.8331	0.8444
	0.7	0.8549	0.8285	0.8404
	0.8	0.8539	0.8286	0.8400
	0.9	0.8548	0.8293	0.8407
	1	0.8547	0.8289	0.8403

Table 1: Weight value adjustment of parent pages and child pages under Approach 3

The experimental results of three feature representation approaches are summarized in Table 2. From the table, the approach of integrating current page with neighboring pages via the topic model (TOPIC_INTEGRATED) yielded a higher performance compared to applying the topic model on current page (TOPIC_CURRENT) and application of the BOW model. Applying the TOPIC_INTEGRATED approach yielded the best performance with the F1 measure of 84.51%; an improvement of 6.11% over the TOPIC_CURRENT approach and an improvement of 23.31% over the BOW model. Applying the TOPIC_CURRENT approach helped improve the performance over the BOW by 17.2% based on the F1 measure. Thus, integrating the additional neighboring information, especially from the parent pages and child pages, via a topic model could significantly improve the performance of a classification model.

Approaches	P	R	F1
1. BOW	0.6000	0.6610	**0.6120**
2. TOPIC_CURRENT	0.7960	0.7710	**0.7840**
3. TOPIC_INTEGRATED	0.8571	0.8299	**0.8451**

Table 2: Evaluation of different feature representation approaches

5 Conclusions and Future Works

To improve the performance of Web page classification with the bag of words feature representation, we proposed a method based on a topic model to integrate the additional term features obtained from the neighboring pages. We applied the topic model approach based on the Latent Dirichlet Allocation algorithm to cluster the term features into a set of latent topics. Words assigned into the same topic are semantically related. From the experimental results, the approach of integrating current page with the neighboring pages via the topic model yielded the best performance with the F1 measure of 84.51%; an improvement of 6.11% over applying the topic model on current page approach and an improvement of 23.31% over the BOW model. Thus, integrating the additional neighboring information, especially from the parent pages and child pages, via a topic model could significantly improve the performance of a classification model.

In our future work, we plan to evaluate our proposed method on other interesting data sets such as social media contents. Other idea is to include other types of neighboring pages such as sibling and spouse pages, in addition to the parent and child pages.

References

[AGS99] Attardi, G.; Gulli, A.; Sebastiani,F.: Automatic Web page categorization by link and context analysis: Proc. of European Symposium on Telematics, Hypermedia and Artificial Intelligence (THAI), 1999, pp. 105-119.

[BNJ03] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, 2003. vol. 3, no. 5, pp. 993-1022.

[CC08] Chen, G.; Choi, B.: Web page genre classification : Proc. of 2008 ACM symposium on Applied computing , 2008, pp. 2353-2357.

[DL97] Dash, M.; Liu, H.: Feature Selection for Classification. Intelligent Data Analysis, 1997, vol. 1, no. 1–4, pp. 131-156.

[Du98] Dumais, S. et al.: Inductive Learning Algorithms and Representations for Text Categorization: Proc. of CIKM1998, 1998, pp. 148-155.

[Fu99] Furnkranz, J.: Exploiting structural information for text classification on the WWW: Proc. of the 3rd Symp. On Intelligent Data Analysis (IDA), Springer, 1999, vol. 1642, pp. 487-497.

[HD08] Haruechaiyasak, C; Damrongrat C.: Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools: Proc. of the 11th International Conference on Asian Digital Libraries**,** 2008, pp. 339-342.

[Jo98] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features: Proc. of the European Conference on Machine Learning (ECML), Berlin, 1998, pp. 137-142.

[QD06] Qi, X.; Davison, B. D.: Knowing a Web Page by the Company It Keeps: Proc. of the 15th ACM Conference on Information and Knowledge Management (CIKM), Arlington, Virginia, USA, 2006, pp. 228-237.

[QD08] Qi, X.; Davison, B. D.: Classifiers Without Borders: Incorporating Fielded Text From Neighboring Web Pages: Proc. of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, Singapore, 2008, pp. 643-650.

[SG06] Steyvers, M.; Griffiths, T.: Probabilistic topic models, In T., Landauer, D., McNamara, S., Dennis, and W., Kintsch, (eds), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2006.

[Sh06] Shen, D. et al.: A comparison of implicit and explicit links for web page classification: Proc. of the 15th international conference on World Wide Web table of contents, 2006, pp. 643-650.

[SLN02] Sun, A.; Lim, E.-P.; Ng, W.-K.: Web classification using support vector machine: Proc. of the 4th Int'l Workshop on Web Information and Data Management (WIDM), ACM Press, 2002, pp. 96-99.

[Va95] Vapnik, V.: The Nature of Statistical Learning Theory. In Springer-Verlag, New York, 1995.

[YP97] Yang, Y.; Pederson, J.P.: A comparative Study on Feature Selection in Text Categorization: Proc. of the 14th International Conference on Machine Learning, 1997.

[Zh07] Zhu, S. et al.: Combining content and link for classification using matrix factorization: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval, 2007, pp. 487-494.