Improving Vietnamese Web Page Classification by Combining Hybrid Feature Selection and Label Propagation with Link Information

Ngo Van Linh, Nguyen Thi Kim Anh, and Cao Manh Dat

School of Information and Communication Technology
Hanoi University of Science
and Technology
{linhnv,anhnk}@soict.hut.edu.vn, caomanhdat317@gmail.com

Abstract. Classification of web pages is essential to many information management and retrieval tasks such as maintaining web directories and focused crawling. One problem in web page classification is that, unlabeled training examples are readily available, while labeled ones are often costly to obtain. Furthermore, the uncontrolled nature of web content presents additional challenges to web page classification, whereas the interconnected characteristic of hypertext can provide useful information for the process. To address these problems, we propose a graph-based semi-supervised classification framework which combines iteratively hybrid semi-supervised feature selection and Label Propagation learning using link information to improve the Vietnamese web page classification. The experimental results show that proposed method outperforms the state-of-the art methods applying to Vietnamese web page classification.

Keywords: Feature Selection, Label Propagation, Web Classification, Web Mining.

1 Introduction

The web has become one of the most important information sources and knowledge base for science, education and research applications. With the exponential growth of information and data in the Internet, people often need to spend a huge amount of time obtaining expected the information even with the help of search engines. Meanwhile, all of the machine learning and data mining methods are aimed to provide more powerful functionalities to meet the needs of users. One way of organizing this overwhelming amount of data is to classify it into descriptive or topical taxonomies. Web page classification can help to improve the quality of the web search results which eventually saves users from a large number of unexpected web pages. Besides, the classification also plays a vital role in many information management and retrieval tasks. On the web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to the analysis of the topical structure of the web.

P.C. Vinh et al. (Eds.): ICCASA 2012, LNICST 109, pp. 324–334, 2013.

Web page classification, as a traditional supervised machine learning task, is to train a classifier with labeled examples, so as to predict the label for any new pages. But in web page classification, just as in many other practical machine learning and data mining applications, unlabeled training examples are readily available, while labeled ones are often laborious, expensive and slow to obtain. This is because labeled examples normally require much efforts and accuracy of experienced annotators. Semi-supervised learning addresses this problem by leveraging a large amount of unlabeled data, together with a small labeled dataset to build better classifiers. However, the limited size of the "labeled" dataset poses hug challenges of selecting an ideal feature subset only based on the "labeled" data.

This practical problem drives the need for "semi-supervised feature selection" to choose the best set of features that produces the most accurate classifier for a learning algorithm given both labeled and unlabeled examples.

Moreover, the uncontrolled nature of web content presents additional challenges to web page classification compared to traditional text classification. To solve this problem, we can refer to the interconnected characteristic of hypertext since the key for implementing effective web page classification is to find intrinsic relationships between web pages. For this purpose, web page content, hyperlinks and usage data (server log files) could be utilized together as important features for the system. Among them, hyperlink analysis has its own advantages, as hyperlinks convey semantics between web pages in most cases.

The motivation for our work is based on the observation that web pages on a particular topic are often linked to other pages on the same topic. In fact, with a few exceptions, the authors of web pages create links to other pages usually with an idea in mind that the linked pages are relevant to the linking pages. If a hyperlink is reasonable, it reflects human semantic judgment and this judgment is objective and independent of the synonymy and polysemy of the words in the pages. This latent semantic, once being revealed, could be employed to find higher-level relationships among the pages. Besides, the hyperlink analysis has also been proven successful in many web-related areas, such as page ranking in the search engine.

In this paper, we propose a graph-based semi-supervised classification framework combined with iteratively hybrid semi-supervised feature selection and Label Propagation learning using link information to improve the accuracy of Vietnamese web page classification. The experimental results show that our algorithm achieves promising results.

The rest of this paper is organized as follows. In section 2, we give a brief review of link-enhanced text classification and clustering methods, semi-supervised learning and semi-supervised feature selection. Section 3 introduces Gini-index-based supervised feature selection method. Section 4 represent the method of constructing integrated graph which combines content-based graph and link-based graph. Section 5 introduces the Label Propagation algorithm in graph. Section 6 proposes a graph-based semi-supervised classification framework which combines iteratively hybrid semi-supervised feature selection and Label Propagation learning using link information. Section 7 reports on the experimental results. Finally, in section 8, we make some conclusions and raise several issues for future work.

2 Related Work

Exploiting link information to enhance text classification has been studied extensively in the research community [1,2,7,9]. Most of these studies fall into two frameworks. One is referred to as relaxation labeling (RL) in which the label of a document is determined by both local content and its neighbors' labels [1,7]. The other improves classification accuracy by incorporating neighbors' content information text into the local content [2,9]. However, Ghani [3] et al. discovered that neighbors' text content information could be useful only when the neighbor link structure exhibits encyclopedia regularity.

In fact, because the availability of labeled examples cannot be taken for granted for many real world applications, semi-supervised learning methods that exploit unlabeled examples in addition to labeled ones have been widely researched. Indeed, for semi-supervised learning methods, when the size of the "labeled" data is limited, it is difficult to select an ideal subset of features based only on the "labeled" data and the effectiveness of semi-supervised learning algorithm, therefore, would be downgraded. Ren et al. proposed a "wrapper-type" forward semi-supervised feature selection framework by using a mechanism of random selection on unlabeled data to extend the initial labeled training set and then, the most frequently selected feature is added to the result feature subset every iteration. Later, Zhong et al. [12] proposed a hybrid graph-based semi-supervised feature selection framework using a confidence-based sampling strategy that provides an improved accuracy over random selection. However, the "Label Propagation" algorithm in [12] use only content-based graph to predict the unlabeled data *U*. Then, the top *s*% of the unlabeled data with the highest confidence are added to the original training data along with their labels to create a new training dataset. In our opinion, the sufficiency and diversity of the new training set can be improved poorly, which in return can help to choose restrictively the most discriminative features.

The work closest to ours is an approach described in [6], which uses a semi-supervised learning algorithm on a *K*-Nearest Neighbor graph for web page classification. Their algorithm uses a similarity measure between web pages to construct a *K*-Nearest Neighbor graph. Edge weights of the graph are computed by incorporating text similarity information and link similarity information of web pages. However, using mutual information to measure the correlation between the categories of linked web pages or the class dependency could perform poorly when there isn't class label information of web pages. Furthermore, [9] does not pay attention to feature selection problem when the number of labeled web pages is small and doesn't directly exploit topic-aware characteristic of hyperlinks.

For Vietnamese web page classification, in [8], N.M.Trung et al. proposed to exploit a main content extraction method to improve performance of web classification task using SVM. Their experimental results showed that the proposed method significantly improves the precision of the Vietnamese web page classification from 71% to 80%. However, [8] doesn't deal with the problem of the size of the "labeled" data.

The motivation for our work is based on the observation that the number of labeled web pages is often small; meanwhile, web pages on a particular topic are often linked to other pages on the same topic. Therefore, we propose a graph-based semi-supervised classification framework which combines iteratively hybrid semi-supervised feature selection and Label Propagation learning using link information to improve the accuracy of Vietnamese web page classification using small training set.

3 Gini-Index-Based Supervised Feature Selection

Feature selection is an important data processing step in web page classification because the corpus of web pages is very high dimensional dataset. Traditionally, supervised feature selection methods use information from "labeled data" to find the most informative or most useful feature subsets, but the information in the "unlabeled" data is not used. At present, the feature selection methods are based on statistical theory and machine learning. Some well-known methods are information gain, expected cross entropy, the weight of evidence of text, odds ratio, term frequency, mutual information, CHI, Gini-index and so on. The experiments in [11] show that the quality of Gini-index is comparable with other text feature selection methods. However, its complexity of computing is lower and its speed is higher. Therefore, we chose Gini-index-based supervised feature selection method for Vietnamese web page classification.

The original form of Gini-index is used to measure the impurity of attributes towards categorization. For web page classification, we use *Gini-index* to measure the discriminative quantification of a given word from the set of labeled web pages.

As a feature selection step, we compute Gini-index for each word and further remove the words that have low Gini-index. A Gini-index is regarded as low when its value is smaller than a predefined threshold. However, Gini-index can only be applied effectively when the set of labeled web pages is large since it could unintentionally remove informative and discriminative features. To alleviate the problem, as the set of labeled web pages is small, we pick a set of top *sizeFS* words which have the highest values of $G(w)$ as the features of web pages, where *sizeFS is* size of feature subset and then use them in order to construct our content-based graph.

4 Graph Construction

As with many other graph-based semi-supervised learning methods, we make the assumption that the instance graph is *homophilous* i.e., that instances belonging to the same class tend to link to each other or have higher edge weight between them. When instances are not explicitly linked to each other, usually a similarity function is applied to local features of each pair of instances to derive weighted edges between them. When instances are explicitly linked to each other, the edges simply correspond to the binary presence of a link (or are weighted by the number of links between two instances). For the corpus of web pages, hybrid approaches are used because both local features and explicit links are available.

We construct a graph based on the combination of web page content and links as follows:

1. The first phase, we build two graphs: content-based graph and link-based graph. They share the same set of nodes, but differ from on edges and their corresponding weights.
2. The second phase, we linearly combine the two graphs into an integrated graph.

Content-Based Graph Construction

We build a graph $G^1(V,E^1,W^1)$, in which V is a set of nodes, E^1 is a set of edges and W^1 is the weight matrix of E^1. Each node in V represents a webpage. The relationship between two webpage is represented by an edge in E^1. The weight W^1 could take the form of a matrix or a linked list. In this paper, W^1 is a matrix, each of its elements represent the weight of corresponding edge in the set E^1.

We represent the processed text of each webpage as a feature vector with the set of selected features of web pages based on the TF*IDF model. The weight of the edge between the two page nodes d_i and d_j based on their similarity should be calculated as follows [13]: $w_{ij}^1 = \exp(-(1-\cos(d_i, d_j))/a)$

To reduce the edge number, we replace the fully connected graph G^1 by a ε-weighted graph that has the same set of vertices with $G^1$and allows only the edges with corresponding weights greater than the given threshold ε. The employment of cutting-edge threshold has additional benefit as it could delete cross-topic edge. Although the reduction in the edge set could result in disconnected graph, this is not a problem in Label Propagation if each connected component has some labeled point.

Link-Based Graph Construction

The link-based graph $G^2(V,E^2,W^2)$ shares the same set of nodes with $G^1(V,E^1,W^1)$, and its number of edges depends on the number of hyperlinks between web pages. The reason for the employment of hyperlinks is that any two web pages should likely belong to the same topic if they are connected by a hyperlink. Unfortunately, it is common that there are a large number of noisy links in a web page such as advertisement links and navigation links. To eliminate these noisy links, we measure the similarity between linked web pages and remove the links if their similarity is lower than a threshold γ. It could also argue that the relevant pages would have low similarities due to the diversity of vocabulary used by their different authors. Fortunately, we can hardly find such cases in reality.

Graph Combination

After contracting content-based graph and link-based graph, we combine the two graphs to form a graph G(V,E,W). While the integrated graph has the same set of nodes with two former graphs. Its edge set E and weight matrix W is computed as follow: $E = E^1 \text{ U } E^2$ and $W = \alpha W^1 + (1-\alpha)W^2$

By varying the value of α, it is possible to control the relative importance of content and link in the classification process. Furthermore, thresholds ε and γ as well as similarity measure between web pages have important role in graph construction and affect later classification confidence of graph-based semi-supervised classification algorithm.

5 Label Propagation in Graph

Label Propagation is a semi-supervised classification algorithm that assigns labels for unlabelled examples based on labeled ones. The central idea of Label Propagation algorithm is that the labels of a vertex propagate to other nodes through the edges. The algorithm remains effective when the size of training set is small. In reality, it is costly and laborious to assign labels to a large amount of data so the algorithm is more preferable. The most notable advantage of Label Propagation algorithm is that its convergence is always ensured. However, the accuracy of the algorithm is largely dependent on the similarity matrix built for web pages.

Problem statement: Given a graph G(V,E,W) and a label set C of size m. Let V_l be the subset of labeled data and $V_u = V \setminus V_l$ be the subset of unlabelled data in V. The problem is to assign labels to the unlabelled data based on G and labeled data.

We first transform the problem above into the problem of finding a probability matrix Y of size mxn in which m and n are the number of labels and vertices, respectively. The ith row of the matrix Y represents the probability distribution of vertex i over the label set C. Specifically, the value Y_{ic} corresponds to the probability of vertex i having label c. We could derive label Y_i for each vertex i from matrix Y as follows: $Y_i = \arg\max_c(Y_{ic})$. We first initialize the matrix Y^0: Y_{ic}^0=1 if vexter i has label c, 0 otherwise.

```
Algothrim 1: Label Propagation
Input: $Y^0$, G(V,E,W).
Output: Y
$P = D^{-1}W$, where D is the diagonal matrix with D(i,i) equal
to the sum of the i-th row of W
$Y \leftarrow Y^0$
$t \leftarrow 1$
repeat
  $Y^t \leftarrow PY^{t-1}$
until convergence to $Y^*$
$Y \leftarrow Y^*$
return Y
```

The convergence of Y is reached when the algorithm executes a fixed number of iterations, *Iteration_prop,* which is usually small compared to the size of dataset.

6 Combining Iteratively Hybrid Semi-Supervised Feature Selection and Label Propagation

In fact, for web page classification, the number of labeled web pages is often small which makes it difficult to generate an ideal subset of features based only on this small labeled dataset. The Label Propagation learning algorithm, as a result, would perform poorly due to the unqualified set of features. Meanwhile, web pages on a particular topic are often linked to other pages on the same topic. Therefore, a hybrid semi-supervised feature selection method combined iteratively with the Label Propagation algorithm using link information would be a reasonable solution.

The framework takes advantages of both the supervised and semi-supervised feature selection paradigms, while can alleviate their deficiencies. Concretely, we first perform the supervised feature selection on the labeled data to obtain an initial "seed" feature subset which is used to represent the processed text of each web page, to measure the content-based similarity between web pages and to construct our content-based graph. Then, to predict the unlabeled data, we apply the Label Propagation learning on the integrated graph which combines content-based graph and link-based graph. The link-based graph is used in label propagation process to directly exploit topic-aware characteristic of hyperlinks. A new training set is built by unifying both the labeled examples and those unlabeled examples whose predicted labels are likely to be correct. After that, feature selection will be carried out on the new training set again, and the selected features will be employed to construct a new graph model for the next iteration. Therefore, during each iteration, we improve this feature subset using the unlabeled data and hence, enhance the performance of the Label Propagation learning. The proposed approach is summarized in Algorithm 2.

```
Algorithm 2: IterHybridFS& LabelPropagation
Input: L, U, $Y^0$, sizeFS, s, Iterations
Output: Y
FS = Gini-index-based Feature Selection (L), where |FS| = sizeFS
newL = L; newAvgCon f  = 0
G = GraphConstruction(L + U, FS )
$Y^{\Pr edict}$ = Label Propagation (G, $Y^0$) through algorithm 1
For i = 1 to Iterations do
      UwithLabel defined through $Y^{\Pr edict}$
      PU= top s% from UwithLabel with highest confidence
      AvgCon f = average prediction confidence on PU
      If  AvgCon f >newAvgCon f then
         newAvgCon f = AvgCon f
        newL = L + PU
         FS = Gini-index-based Feature Selection (newL), where |FS | = sizeFS
      Else
         Break
      End if
      G = GraphConstruction(L + U, FS )
      $Y^{\Pr edict}$ = Label Propagation (G, $Y^0$) through algorithm 1
End for
Y =$Y^{\Pr edict}$
Result Y
```

The classification confidence of vertex i being labeled using the Label Propagation algorithm is $Y_i = \arg\max_c (Y_{ic}^{\Pr edict})$.

Following a similar formal performance analysis as [12], the feature selection process depends on the sampled unlabeled data whereas the classification confidence by the Label Propagation algorithm is estimated based on weights of edges or similarity measures between web pages. Therefore, our framework should benefit from a better similarity measure obtained by using a confidence-based sampling strategy that provides an improved accuracy over random selection. Our experimental results corroborate the analysis.

Time complexity of IterHybridFS& LabelPropagation can be obtained as follows. *k* is the number of features, *n* is the number of web pages, m is the number of labels and *Iterations_prop* is the number of iterations in label propagation. Gini-index-based feature selection is bounded by *O*(*k***n*+*klogk*). The construction of graph *G* is bounded by $O(k*\mathrm{n}^2)$, while the Label Propagation is bounded by $O(\mathrm{m}*\mathrm{n}^2*Iterations\_prop)$. The costs for calculating *PU*, *FS*, and *AvgConf* are bounded by *O*(*n*log*n*), $O(k*\mathrm{n}^2)$, and *O*(*n*), respectively. These operations need to be performed *Iterations* times. Therefore, the overall time complexity of IteraGraph FS is bounded by $O((k*\mathrm{n}^2+\mathrm{m}*\mathrm{n}^2*Iterations\_prop)*Iterations)$.

7 Experiments and Results

Currently, a standard dataset for Vietnamese web page classification research does not exist. Therefore, to evaluate the performance of our graph-based Vietnamese web Page classification method, we use web pages dataset collected from 11 websites in [8]. The dataset consists of 2500 web pages in *6* categories with nearly 25500 links between pages. These categories can be used as class labels for later evaluation. Since this dataset contains the general categories, it can be used for evaluating the overall performance of classification across websites.

In reality, web pages, especially online web pages, often contains a large amount of redundant and noise information including navigational elements, templates, and advertisements. We use the algorithm devised by Kohlschütter et al. [4] which was reported to remove noise and extract main content of web pages with a high accuracy. After obtaining main content, we perform the word segmentation on the web pages based on the probabilistic model suggested by Phuong [5]. The model could attain the accuracy of 97%.

In natural language, there are many words that convey no or little meaning in a sentence, such as linking words, preposition, conjunctions, etc. These words has no contribution to, even degrade, the classification process. They are called stopwords and we use a stopword list to remove them from the content of web pages. In the first experiment, we perform the Label Propagation algorithm as described in Section 5 to categorize the Vietnamese web pages with a= 0.1, ε= 0.009, γ= 0.001, α=0.9.

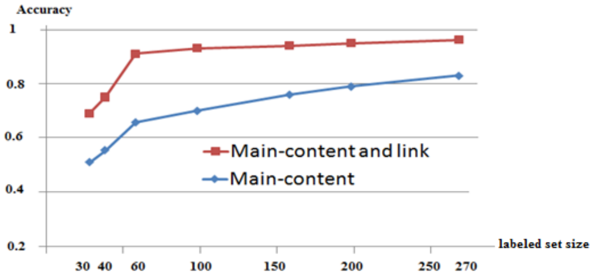

Fig. 1.

Figure 1 depicts the curves of the average accuracy of Label Propagation algorithm on dataset by using two different graph constructions: one is only to consider the content similarity of the web pages; the other is to combine the content similarity with link information. We test this two algorithms on different labeled set sizes. For each labeled set size, we perform 6 trials. In each trial, we randomly sample labeled web pages from the entire dataset, and use the rest of dataset as unlabeled web pages. Figure 1 shows that the classification accuracies are higher when both content similarity and link information are involved. It demonstrates that the exploiting link information allows improving the accuracy of Vietnamese web page classification significantly.

In the second experiment, we perform the IterHybridFS& LabelPropagation algorithm as described in Section 6 to catego-rize the Vietnamese web pages with *sizeFS =3000, %s=30%, Iteration_prop = 20, Iterations = 10.*

The performance of the IterHybridFS& LabelPropagation algorithm is compared with the two Label Propagation algorithms in the integrated graphs in which content-based graph construction uses full text and Gini-index based supervised selection feature. Figure 2 depicts the curves of the average accuracy of classify-cation on the dataset by using three different algorithms and confirms our remark.

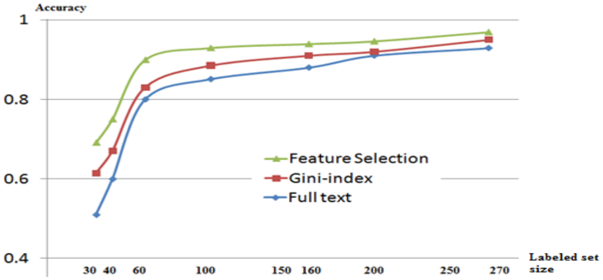

Fig. 2.

The result showed in figure 2 indicates that the mean of the classification accuracies of IterHybridFS& LabelPropagation algorithm are higher than that of two others. In the algorithm based on full text, the high dimensionality of the feature space makes the similarity between any two web pages approximately the same which puts the accuracy of algorithm at the lowest position. Besides, when the number of labeled web pages is small, Gini-index-base supervised feature selection techniques often fails due to the sample selection bias or the unrepresentative sample problem. Experimental results demonstrate that the proposed IterHybridFS& LabelPropagation algorithm outperforms two other algorithms by at least 8% in accuracy.

8 Conclusion

Classification plays a vital role in many information management and retrieval tasks. On the web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to analysis of the topical structure of the web. Web page classification can also help improve the quality of web search.

This paper proposed a graph-based semi-supervised classification framework which combines iteratively hybrid semi-supervised feature selection and Label Propagation

The framework has the following main advantage:

(1) It performs supervised feature selection before predicting on unlabeled data, thereby maintaining the most critical features.
(2) It uses Label Propagation learning with link information, thereby providing a better prediction confidence.
(3) It uses confidence-based sampling strategy through Label Propagation learning with link information, thereby producing better new training dataset. Furthermore, the sufficiency and diversity of the new training set can be improved, which in return helps to choose the most discriminative features.

As the result, our experiments show that the proposed classification method outperforms the state-of-the art methods applying to Vietnamese web page classification.

While our results are encouraging, there are still much improvements to be made. The current training set is still using electronic newspapers. Therefore, it can not represent all types of web page. It is necessary to build a standard Vietnamese web page data set in the future. Another problem is that we can not perform this method on a web page without text such as a flash website or a picture website, so another solution should be used in this case.

References

1. Angelova, R., Weikum, G.: Graph-based text classification: learn from your neighbors. In: SIGIR 2006 (2006)
2. Chakrabarti, S., Dom, B.E., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In: SIGMOD 1998, pp. 307–318 (1998)
3. Ghani, R., Slattery, S., Yang, Y.: Hypertext Categorization using Hyperlink Patterns and Meta Data. In: ICML 2001 (2001)

4. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate Detection using Shallow Text Features. In: WSDM 2010 – The Third ACM International Conference on Web Search and Data Mining, New York, City, USA (2010)
5. Hông Phuong, L., Thi Minh Huyên, N., Roussanaly, A., Vinh, H.T.: A Hybrid Approach to Word Segmentation of Vietnamese Texts. In: Martín-Vide, C., Otto, F., Fernau, H. (eds.) LATA 2008. LNCS, vol. 5196, pp. 240–249. Springer, Heidelberg (2008)
6. Liu, R., Zhou, J., Liu, M.: A Graph-based Semi-supervised Learning Algorithm for Web Page Classification. In: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, ISDA 2006 (2006)
7. Lu, Q., Getoor, L.: Link-based classification. In: ICML (2003)
8. Trung, N.M., Tam, N.D., Phuong, N.H.: Using main content extraction to improve performance of Vietnamese web page classification. In: SoICT 2011, Hanoi, Vietnam, October 13-14 (2011)
9. Oh, H.J., Myaeng, S.H., Lee, M.H.: A practical hypertext categorization method using links and incrementally available class information. In: SIGIR, pp. 264–271 (2000)
10. Ren, J., Qiu, Z., Fan, W., Cheng, H., Yu, P.S.: Forward Semi-supervised Feature Selection. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 970–976. Springer, Heidelberg (2008)
11. Shang, W., Huang, H., Zhu, H.: A Novel feature selection algorithm for text categorization. Expert System with Application 33, 1–5 (2007)
12. Zhong, E., Xie, S., Fan, W., Ren, J., Peng, J., Zhang, K.: Graph-based Iterative Hybrid Feature Selection. In: Proceeding ICDM 2008 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (2008)
13. Strehl, A., Ghosh, J., Mooney, R.J.: Impact of similarity measures on web-page clustering. In: AAAI Workshop (2000)