ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 4, No.16 , July 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Automatic Classification for Vietnamese News

Phan Thi Ha[1], Nguyen Quynh Chi[2]

[1] Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
hapt@ptit.edu.vn

[2] Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
chinq@ptit.edu.vn

Abstract

This paper proposes an automatic framework to classify Vietnamese news from news sites on the Internet. In this proposed framework, the extracted main content of Vietnamese news is performed automatically by applying the improved performance extraction method from [1]. This information will be classified by using two machine learning methods: Support vector machine and naïve bayesian method. Our experiments implemented with Vietnamese news extracted from some sites showed that the proposed classification framework give acceptable results with a rather high accuracy, leading to applying it to real information systems.

Keywords: *news classification; automatic extraction; support vector machine, naïve bayesian networks*

1. Introduction

In the modern life, the need to update and use of information is very essential for human's activities. Also we can see clearly the role of information in work, education, business, research to modern life. In Vietnam, with the explosion of information technology in recent years, the demand for reading newspapers, searching for information on the Internet has become a routine of each person. Because of many advantages of Vietnamese documents on the Internet such as compact and long-time storage, handy in exchange especially through the Internet, easy modification, the number of document has been increasing dramatically. On the other hand, the communication via books has been gradually obsolete and the storage time of document can be limited.

From that fact, the requirement of building a system to store electronic documents to meet the needs of academic research based on the Vietnamese rich data sources on the site. However, to use and search the massive amounts of data and to filter the text or a part of the text containing the data without losing the complexity of natural language, we cannot manually classify text by reading and sorting of each topic. An urgent need to solve the issue is how can automatically classify the document on the Vietnamese sites. Basically, the sites will contain pure text information and processing and classifying documents by topic has been interested and researched on the worldwide [2]. Therefore, they build the methods of text classification to strongly support for finding information of Internet users.

This paper proposes an automatic framework to classify Vietnamese news from electronic newspaper on the Internet under Technology, Education, Business, Law, Sports fields to build archives which serve the construction of internet electronic library of Vietnam. In this proposed framework, the extracted main content of Vietnamese news is performed automatically by applying the improved performance extraction method from [1]. This information will be classified by using two machine learning methods: Support vector machine and naïve bayesian method. Our experiments implemented with Vietnamese news extracted from some sites showed that the proposed classification framework gives an acceptable result with a rather high accuracy, leading to applying it to real information systems.

The rest of the paper is presented as the followings. In section 2, the methods of news classification based on automatically extracted contents of Web pages on the Internet are considered. The main content of the automatic classification method is presented in section 3. Our experiments and the results are analyzed and evaluated in section 4. The conclusions and references are the last section.

2. Related works and motivation

In recent years, natural language processing and document content classification have had a lot of works with encouraging results of the research community inside and outside Vietnam.

The relevant works outside Vietnam have been published a lot. In [3], they used clustering algorithm to generate the sample data. They focused on optimizing for active machine learning. An author at University of Dortmund

Germany presented that the use and improvement of support vector machine (SVM) technique has been highly effective in text classification [4]. The stages in the a text classification system including indexing text documents using Latent semantic Indexing (LSI), learning text classification using SVM, boosting and evaluating text categorization have been shown in [5]. "Text categorization based on regularized linear classification methods" [6] focused in methods based on linear least squares techniques fit, logistic regression, support Vector Machine (SVM). Most researchers have focused on processing for machine learning and foreign language, English in particularly. In the case of applying for Vietnamese documents, the results may not get the desired accuracy.

The work of Vietnamese text categorization can be mentioned by Pham Tran Vu et al. They talked about how to compute the similarity of text based on three aspects: the text, user and the association with any other person or not [7]. The authors applied this technique to compute the similarity of text compared with training dataset. Their subsequent work referred matching method with profiles based on semantic analysis (LSA). The method presented in [8] was without the use of ontology but still had the ability to compare relations on semantics based on the statistical methods. The research works in Vietnam mentioned have certain advantages but the scope of their text processing is too wide, barely dedicated for a particular kind of text. Moreover, the document content from Internet is not extracted automatically by the method proposed in [1]. Therefore, the precision of classification is not consistent and difficult to evaluate in real settings.

To extract document content on the Internet, we must mention to the field of natural language processing – a key field in science and technology. This field includes series of Internet-related applications such as: extracting information on the Web, text mining, semantic web, text summarization, text classification... Effective exploitation of information sources on the Web has spurred the development of applications in the natural language processing. The majority of the sites is encoded in the format of Hyper Text Mark-up Language (HTML), in which each Web page's HTML file contains a lot of extra information apart from main content such as pop-up advertisements, links to other pages sponsors, developers, copy right notices, warnings…Cleaning the text input here is considered as the process of determining content of the sites and removing parts not related. This process is also known as dissection or web content extraction (WCE). However, structured websites change frequently leading to extracting content from the sites becomes more and more difficult [9]. There are a lot of works for web content extraction, which have been published with many different applications [10, 11, 12, 13, 14, 15].

In recent years, extracting contents of the site have been researched by many groups in countries and their results were rather good [16, 17, 18, 19, 1]. These approaches include: HTML code analysis; pattern framework comparison; natural language processing. The method of pattern framework extracts information from two sites. This information is then aligned together based on the foundation of pattern recognition applied by Tran Nhat Quang [18]. This author extracted content on web sites aiming to provide information on administration web sites. The method of natural language processing considers the dependence of syntax and semantics to identify relevant information and extract needed information for other processing steps. This method is used for extracting information on the web page containing text following rules of grammar. HTML method accesses directly content of the web page displayed as HTML then performs the dissection based on two ways. The first is based on Document Object Model tree structure (DOM) of each HTML page, data specification is then built automatically based on the dissected content. The second is based on statistical density in web documents. Then, dissect the content, data obtained will become independent from the source sites, it is stored and reused for different purposes.

To automatically extract text content from the web with various sources, across multiple sites with different layouts, the authors [1] studied a method to extract web pages content based on HTML tags density statistics. With the current research stated above, we would like to propose a framework for automatic classification of news including Technology, Education, business, Law, Sports fields. We use the method in [1] which is presented in the next section.

3. Automatic News Classification

3.1 Vietnamsese web content extraction for classification

The authors have automatically collected news sites under 5 fields from the Internet and used content dissection method based on word density and tag density statistics of the site. The extracting text algorithm was improved from the algorithm proposed by Aidan Finn [11] and the results were rather good.

Aidan Finn proposed the main idea of BTE algorithm as follows: *Identify two points i, j such that some HTML tag-tokens under i and on j is maximum and the signs of text-tokens between i and j is maximum. The extraction result is the text signs between interval [i, j] which are separated.*

Aidan Fin did experiments by using BTE algorithm to extract text content for textual content classification in

digital libraries, mainly collecting new articles in the field of sports and politics in the news website. This algorithm has the advantage that dissection does not depend on the given threshold or language but the algorithm is not suitable for some Vietnamese news sites containing some advanced HTML tags.

By observing some different Vietnamese news sites, the paper [1] showed that the news sites in general have a main characteristic: in each page's HTML code, text body part contains fewer tags and many signs of text. The authors have improved algorithm BTE (by adding step 0) to extract text body from Vietnamese new sites to build Vietnamese vocabulary research corpus.

Construction algorithm: The experimental observations show that the text body of Web pages always belong to a parent tag that is located in pair (<body> … </ body>) in which HTML tags like or scripts is embedded in tags like <img> <input> <select> <option>… In addition, some content is not related to the text body but in some advanced HTML tags like (<style> … </ style> <script> … </ Script>, <a>….,…). Therefore, initial step should be removing the HTML code which certainly does not contain the content of the web page (Step 0). Then binary encoding of remaining content (HTML tags corresponding to 1, text signals corresponding -1) is performed then total of identical adjacent value is computed. Next, extract segments which have the most negative values (-1) and the least positive values (1). The complexity of this algorithm is O (n^2).

Here are main steps of the algorithm:

Step 0: Each site corresponds to one HTML file format. Clean HTML codes by removing tags, HTML codes do not contain information relating to contents such as tags: <*input*>, <*script*>, <*img*>, <style>, <*marquee*>,<!--...-->, <a>… and contents outside the HTML tags <body>,</body> of each web page. HTML tags library is collected from web site address [22, 23].

Step 1: For the remaining part of the web sites, build two arrays that are binary_tokens[] and tokens[]. Binary_tokens[] include 1 and -1:

- Binary_tokens[i] = 1 corresponds to the i^{th} element which is an HTML tag. This tag includes the beginning tags: <?...>, example: <html>, and end tags: </?...>, example: </html>, .
- Binary_tokens[i]= -1 corresponds to the i^{th} element which is a sign of text.

  Tokens[] is an array of elements including value of text signs or tags corresponding to elements in the binary_tokens[]. Example, at position 23, binary_tokens [23] = 1, tokens [23] = <td…>.

  Merge adjacent elements which have the same value in the binary_tokens[] array to make an element in encode[] array, which significantly reduces the size of binary_tokens[] array. The complexity of this algorithm is O(n)

Step 2: Locate two points i, j from binary_tokens[] array recently obtained in step 1 so that the total number of elements which have value -1 between [i, j] and 1 outside [i, j] is the largest. Perform data dissection in the scope [i, j] and remove HTML tags. The complexity of this algorithm is O(n^3).

The BTE-improved algorithm is tested and compared with original algorithm proposed by Aidan Finn with the same number of sites in the test set. The experiments and results are as follows:

First time: run BTE algorithm of Aidan Finn on HTML file obtained respectively from the URL.

Second time: run improved BTE on HTML file obtained respectively from the URL.

The ratio of text body needed over the total extraction text of 3 types of sites which are interested by many users is shown in table 1, in which each type of collected sites contains 100 files.

Table 1. Comparing ratio of text body needed to take/total of extraction text

Type of site	***Improved algorithm***	***Aidan Finn's algorithm***
Dantri.com.vn	99.02%	47.12%
Vietnamnet.vn	99.67%	65.71%
VnExpress.net	99.00%	48. 87%

3.2 News web content classification

We use a learning machine method called support vector machine (SVM) to train a 5 types classifier for news classification on webs. This is a multi-class classification problem. The idea of solving a multi-class classification problem is to convert it into two-class problems by constructing multiple classifiers. The common multi-class classification strategy are: One-Against_One (OAO), and One-Against- Rest (OAR).

With OAR strategy (Fig 1), we will use K-1 binary classifiers to build K-class. The K-class classification problem is converted into K-1 two-class classification problems. In particular, the i^{th} two-class classifier built on the i^{th} class and all the remaining classes. The i^{th} decision function for the i^{th} classifier and the remaining classes has the form

$$y_i(x) = w_i^T(x) + b_i \qquad (1)$$

Hyper-plane $y_i(x) = 0$ will form optimal division hyper-plane, the support vector of class (i) puts $y_i(x) = 1$ and the remaining support vector class to be satisfying $y_i(x) = -1$.

If the data vector x satisfying the conditions $y_i(x) > 0$ for only one i, x will be assigned to the i^{th} class.

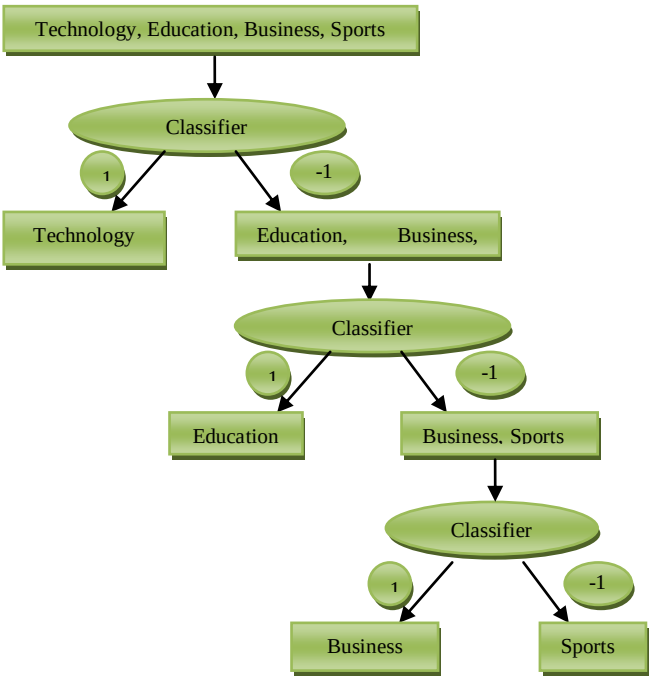

Fig 1: OAR Strategy

OAO strategy (Fig 2) uses K* (K-1)/2 binary classifiers constructed by pairing two classes so this strategy should also be referred to as the pair (pairwise) and used the following the method of combining multiple parts of this class to determine the final classification results. The number of classifiers never exceeds K*(K-1)/2

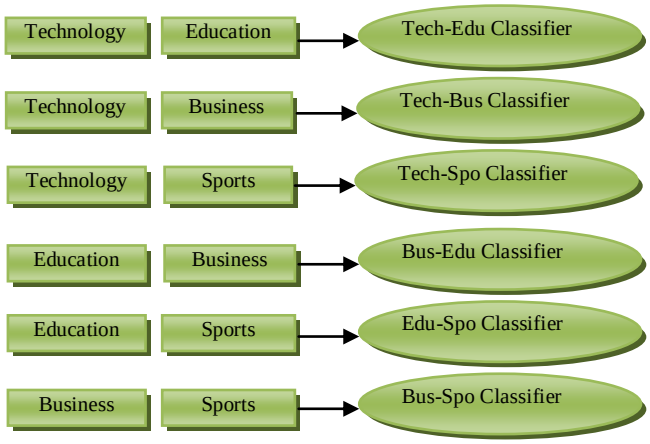

Fig 2: OAO Strategy

Compared with OAR strategy, the advantage of the strategy is not only to reduce the region which cannot be classified, but also to increase the accuracy of the classification. The strategy OAR just needs K-1 classifiers meanwhile the OAO strategy needs K*(K-1)/2 classifiers.

But the number of training record for each classifier in OAO is less than OAR and the classification is also simpler. So the OAO strategy has higher accuracy, but the cost to build is equivalent to OAR strategy. The decision function to subclass of class i to class j in the OAO strategy is:

$$y_{ij}(x) = w_{ij}^T(x) + b_{ij} \quad (2)$$

However, both strategies lead to a vague region in the classification (Fig 3). We can avoid this problem by building K-Class based on K-linear functions of the form:

$$y_k(x) = w_k^T x + b_{k0} \quad (3)$$

And a point s is assigned to the class C_k if: $y_k(x) > y_j(x)$ with every $j \neq k$.

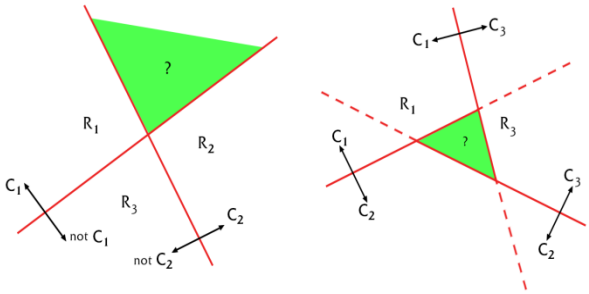

Fig 3: The vague region in subclass

Table 2. Category label of each topic

Topic	*Label*
Technology	1
Education	2
Business	3
Laws	4
Sports	5

The specific steps in the training phase as the following

Step 1: Download the HTML page corresponding to the news page links to filter and retrieve content saved in plain text format (file.txt), remove the documents which are over the size allowed (1KB) and have duplicate contents.

Step 2: Separate words (integrate VnTokenize) according to [20] and remove the stop words, select features [4] (the selection of features will be presented in detail in section 4)

Step 3: Represent a news article as a vector as follows: <*class*$_i$><*label*$_1$>:<*value*$_1$><*label*$_2$>:<*value*$_2$> ... <*label*$_n$>:<*value*$_n$>

Where:

- *Class*$_i$ is the category label of each topic with i = 1 ÷ 5 (Table 2).

- $Label_j$ is the index of the j[th] feature word in the feature space which may appear in the training news with j = 1÷ n.
- $Value_j$ is the weight of the $index_j$ which is calculated by the TF.IDF formula, if the $value_j$ = 0, then do not write that feature itself. This format complies to the input data format of the program SVM[Multiclass] [21].

Step 4: Train classification model based on multi-class SVM algorithm applying OAO strategy with the optimal model parameters (by experiment and using a number of methods such as GrisSeach, Genetics)

The specific steps in the classifying phase as the following:

Step 1: Allow the user to select seed words then generate queries randomly.

Step 2: Perform a search for each query through Google and store links of the found news site after filtering out invalid links such as: music, video, and forums…

Step 3: Use the method of [1] to extract text from the download link, check and remove the text files which do not meet requirement (size<1KB) and text files having duplicate content.

Step 4: Perform to separate words (integrated VnTokenize) and remove stop words from the text. Represent the text as feature vector which is input format of the SVM algorithm.

Step 5: Perform classification (under 5 labels), and save the results in a database.

4. Experiments and Evaluation

4.1. Pre-processing data and experiments on classification model training

Training and testing data were built semi-automatically by the authors. We have developed software to automatically extract content of the text by updating the RSS links of two news electronic sites that are VietNamnet.net and Vnexpress.net by date for each topic (Technology, Education, Business, Laws, Sport). The data obtained is the links of news sites after removing duplicate links, invalid links. The content of news sites are extracted based on the method described in [1]. Then, the preprocessing steps of separating and removing the stop words made as step 2 in the training phase presented in section 3.2. After separating of words, the words will be re-weighted to carry out the selection of features vector for articles. Each article in the data set will be represented by an n-dimensional vectors, each dimension corresponds to a feature word. We choose n dimensions corresponding to n words with the highest weights. Representation of all news articles described in step 3, Section 3.2 in which:

$Value_i$ of the i[th] word in vector representing j[th] news is the weight of the word which is calculated by the formula (4)

$$weight(i,j) = \begin{cases} \left(1 + \log\left(tf_{ij}\right)\right) log\left[\frac{N}{df_i}\right] & if\ tf_{ij} \geq 1 \\ 0 & if\ tf_{ij} = 0 \end{cases} \quad (4)$$

Where:

$df_i < cf_i\ and\ \sum_j tf_{ij} = cf_i$

tf_{ij} (Term frequency): The number of occurrences of the word w_i in document d_j

df_i (Document frequency): The number of documents that contain the word w_i

cf_i (Collection frequency): The number of occurrences of word w_i in the corpus.

If $Value_i = 0$, do not need to keep this feature.

The data set for the training and testing phases includes 2114 articles in total in which 1000 articles belong to the training data set and all remaining 1114 the testing belong to data set. Table 3 lists the number of the training data set and testing on each topic.

Table 3. Number of training data set and testing

Topic	*Training data set*	*Test data set*
Technology	200	123
Education	200	128
Business	200	236
Sports	200	62
Laws	200	565

To choose the value for the dimension of feature vectors, we perform experiments with different values of n listed in Table 4. Evaluation results of SVM and Bayes classifier with different dimensions of feature vectors on the same set of training and testing data set are shown in Table 3 with the accuracy suitable evaluated by the formula (5) and (6). This is the basis for selecting the dimension of the feature space for classifiers: The given criteria evaluating on is the high classification results and narrow fluctuations in a certain data region. Based on Table 4, authors choose dimension n for SVM method is 2000 and Bayes method is 1500.

$$Pre = \frac{TD}{SD} \quad (5)$$

Where:

- Pre: Classification accuracy for a topic.
- TD: The number of correctly classified documents.
- SD: Total number of documents to be classified.

$$TPre = \frac{\Sigma \frac{TDC_i}{SDC_i}}{ST} \quad (6)$$

Where:

- Tpre: Total classification accuracy for topics.
- TDC_i: The number of correctly clssified documents belonging to the topic C_i.
- SDC_i: Total number of classified documents belonging to the topic C_i.
- ST: Total topics.

4.2 Classification experiment and evaluation

In order to automatically classify information on the web, the authors build applications which automatically classify 5 topics: Technology, Education, Laws, Business and Sports. This classification models are trained with SVM and Bayes algorithms with the dimension of the feature vectors selected in Section 4.1. The application is built following 5 specific steps in the classification phase presented in Section 3.2. To evaluate the classification model obtained after conducting the training, we tested classified documents of 1114 in different categories.

Table 4. Results of evaluation with the different lengths of vectors

Number of dimensions of feature vectors	*Accuracy of SVM algorithm*	*Accuracy of Naïve Bayes algorithm*
500	91.56%	89.77%
800	91.92%	90.04%
1000	92.01%	90.39%
1200	92.37%	90.57%
1500	93.00%	91.92%
1800	93.63%	90.48%
2000	93.81%	90.84%
2500	93.72%	90.79%

The results showed that the SVM method give results with an accuracy of approximately 94%, Naïve Bayes (NB) method with an accuracy of approximately 91%. The accuracies are calculated by the formula (5) and (6). For each topic, testing data and evaluation results are described in Table 5 according to the formula (5) and (6). Results of classification for the accuracy of each topic are different. Technology topics have the lowest accuracy and Sports have the highest accuracy.

Table 5. Results of evaluations are categorized by topic

Topic	*Number of news sites*	*NB Method*	*SVM Method*
Technology	123	77.23%	87%
Education	128	92.96%	96.88%
Business	236	94.91%	83.47%
Laws	62	83.87%	96.77%
Sports	565	94.51%	98.58%

5. Conclusion

This paper describes automatically classifying framework for Vietnamese news from news sites on the Internet. In the proposed method, the contents of Vietnamese news are extracted automatically by applying the improved performance extraction of the author group [1]. The news is classified by using two machine learning methods Naïve Bayes and SVM. Experiments on news sites (vietnamnet.vn and vnexpress.net) shows that using SVM method gives a higher (94%) accuracy while the method naïve Bayesian network for lower results. However, we find that classification accuracy is different with various topics and Sports news has the highest accuracy. In future, we will aim to improve automatic classification methods to increase classification accuracy for different news topics and extend widen the types of sites with different and more complicated content than news.

Acknowledgments

We would like thank to Phuong le Hong PhD providing useful tools for word processing on the Web Vietnamese.

References

[1] Phan Thi Ha and Ha Hai Nam, "Automatic main text extraction from web pages", Journal of Science and Technology, Vietnam, Vol. 51, No.1, 2013.

[2] Yang and Xin Liu, "A re-examination of text categorization methods", Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999.

[3] Rong Hu, "*Active Learning for Text Classification*", Ph.D Thesis, Dublin Institute of Technology, Dublin, Ireland, 2011.

[4] Joachims T., "*Text categorization with Support Vector Machines: Learning with many relevant features"*, in Proc. of the European Conference on Machine Learning (ECML), 1998, pages 137–142.

[5] Fabrizio Sebastiani, "Text categorization**",** In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, 2005, pp. 109-129.

[6] Tong Zhang and Frank J. Oles. "Text categorization based on regularized linear classification methods", *Information Retrieval*, Vol. 4:5-31, 2001

[7] Tran Vu Pham, Le Nguyen Thach, "*Social-Aware Document Similarity Computation for Recommender Systems*", in Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, 2011, Pages 872-878

[8] Tran Vu Pham, **"***Dynamic Profile Representation and Matching in Distributed Scientific Networks"*, Journal of Science and Technology Development, Vol. 14, No. K2, 2011

[9] David Gibson, Kunal Punera, Andrew Tomkins, "The Volume and Evolution of Web Page Templates". In WWW\'05: *Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005.

[10] Aidan Finn, Nicholas Kushmerick, Barry Smyth, "*Fact or Fiction: Content Classification for Digital Libraries",* Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin City University, Ireland, 2001.

[11] Aidan. Fin, R. Rahman, H. Alam and R. Hartono, "Content Extraction from HTML Documents", in WDA: *Workshop on Web Document Analysis*, Seattle, USA, 2001.

[12] C.N. Ziegler and M. Skubacz, "Content extraction from news pages using particle swarm optimization on linguistic and structural features," in WI '07: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 242–249.

[13] Ion Muslea, Steve Minton, and Craig Knoblock, "A hierarchical Approach to Wrapper Induction", in Proceedings of the third annual conference on Autonomous Agents, 1999, Pages 190-197.

[14] Tim Weninger, William H. Hsu, Jiawei Han, "CETR-Content Extraction via Tag Ratios". In *Proceedings of the 19th international conference on World wide web*, 2010, Pages 971-980

[15] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, C. Lee Giles, "Automatic Identification of Informative Sections of Web-pages*"*, Journal IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 9, 2005, pages 1233-1246

[16] http://nhuthuan.blogspot.com/2006/11/s-lc-v-k-thut-trong-vietspider-3.html

[17] www.majestic12.co.uk/projects/html_parser.php

[18] Vu Thanh Nguyen, Trang Nhat Quang, " *Ứng dụng thuật toán phân lớp rút trích thông tin văn bản FVM trên Internet*", Journal of Science and Technology Development,Vol. 12, No. 05, 2009.

[19] Ngo Quoc Hung, "*Tìm kiếm tự động văn bản song ngữ Anh-Việt từ Internet*", MS thesis, Ho Chi Minh City University of Science, Vietnam, 2008

[20] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

[21] http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

[22] http://mason.gmu.edu/~montecin/htmltags.htm#htmlformat

[23] http://www.w3schools.com/tags/

First Author: Dr. Phan Thi Ha is currently a lecturer of the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam. She received a B.Sc.in Math & Informatics, a M.Sc. in Mathematic Guarantee for Computer Systems and a PhD. in Information Systems in 1994, 2000 and 2013, respectively. Her research interests include machine learning, natural language processing and mathematics applications.

Second Author: M.Sc Nguyen Quynh Chi is currently a lecturer of the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam. She received a B.Sc.in Information Technology in Hanoi University of Technology in Vietnam, a M.Sc. in Computer Science in University of California, Davis, USA (UCD) and became PH.D Candidate at UCD in 1999, 2004 and 2006, respectively. Her research interests include machine learning, data mining.