An Abstract-Based Approach for Text Classification

Quoc Dinh Truong[1(✉)], Hiep Xuan Huynh[1], and Cuong Ngoc Nguyen[2]

[1] College of Information and Communication Technology, Can Tho University,
Campus 2, 3/2 Street, Ninh Kieu District, Can Tho City, Vietnam
{tqdinh,hxhiep}@ctu.edu.vn

[2] Department of Computer and Mathematical, The People's Security University,
Km9 Nguyen Trai Street, Ha Dong District, Ha Noi, Vietnam
cuongnn@hvannd.edu.vn

Abstract. Text classification is a supervised learning task for assigning text document to one or more predefined classes/topics. These topics are determined by a set of training documents. In order to construct a classification model, a machine learning algorithm was used. Training data is often a set of full-text documents. The training model is used to predict a class for new coming document. In this paper, we propose a text classification approach based on automatic text summarization. The proposed approach is tested with 2000 Vietnamese text documents downloaded from vnexpress.net and vietnamnet.vn. The experimental results confirm the feasibility of proposed model.

Keywords: Text classification · Automatic text summarization · Machine learning

1 Introduction

Text classification is one of the basic problems of text mining. Text classification is the process of assigning text documents to predefined categories based on their content. Text classification has been used in a number of application fields such as information retrieval, text filtering, electronic library and automatic web news extraction. Text classification can be achieved manually or can be automated successfully using machine learning techniques: support vector machines – SVM [1], tolerance rough set model approach [2] and association rules approach [3]. Whatever approach used, the challenge is that full-text content of documents is always taken into account so classification process often deals with large number of features.

Traditionally, before deciding to read or buy a document (book, scientific article) we often read the abstract of this document to catch the main idea. This proves that abstract reflects the main content of document and it can be used for classifying text documents.

Nowadays, research in the field of automatic text summarization has achieved some initial successes. We can list some of the most significant works: automatic text summarization based on word-clusters and ranking algorithms [4], multi-document

P.C. Vinh et al. (Eds.): ICTCC 2016, LNICST 168, pp. 237–245, 2016.
DOI: 10.1007/978-3-319-46909-6_22

summarization by sentence extraction [5], inferring strategies for sentence ordering in multi-document news summarization [6]. These methods have been applied for solving the problem of Vietnamese text summarization.

Based on the success of research in the field of automatic text summarization and the assumption that topic of text document can be simply identified through its abstract, we propose an abstract-based approach for text classification. The objective of our research is to answer the question of whether abstract should be used for classifying text documents.

In the field of data mining, the problem of text classification is often solved by using machine learning techniques: support vector machine-SVM [1], decision tree [7], k-nearest neighbor [8] and neural network [9] in which the most used techniques are SVM and decision tree. So in the context of our research, we use SVM and decision tree as classifier for both classification models: our proposed model uses abstract of text documents as input samples and the baseline model uses full-text content of documents as input samples. To compare the performance of models, 2000 Vietnamese text documents are collected from two websites vnexpress.net and vietnamnet.vn. The abstracts of these text documents are generated by module which is programed according the model we proposed in [10]. Experimental results show that the proposed approach is effective and promising.

2 Proposed Model

The overall model architecture is depicted in Fig. 1. First a classification is trained and then this model is used for classifying input data in which abstract of full-text document is automatically generated.

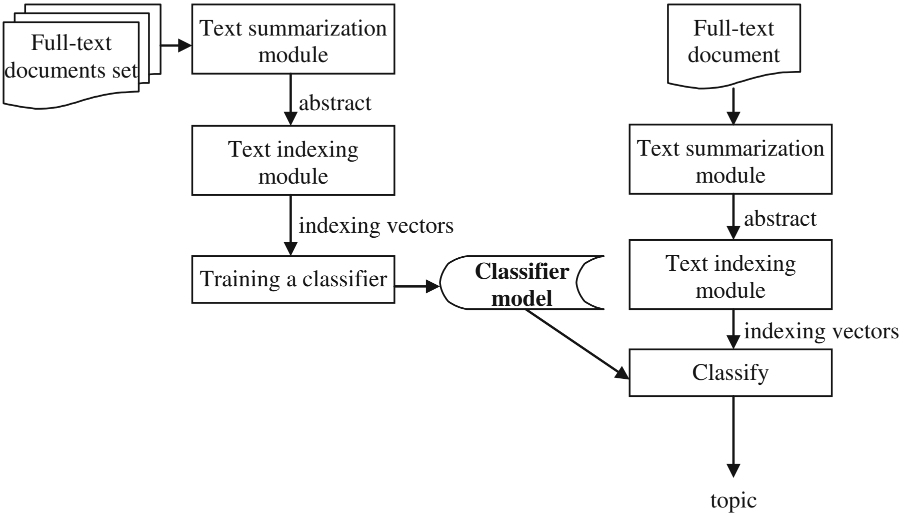

Fig. 1. Proposed model architecture

2.1 Text Document Representation

We use bag of words (BoW) model to represent text document. In order to do this, first the text document is segmented into word tokens using a tokenizer and then a term-weighting scheme was applied to compute the weights of word tokens. There exist some tools to separate Vietnamese words, by example vnTokenizer [11] which is used in this study. That tool was built based on the maximum matching method with the database used is the Vietnamese syllabary and Vietnamese dictionary with Java open source. Thus, it can be easily integrated into other systems. The following is an example of Vietnamese text segmentation:

- Original text: "Để có thể thực hiện rút trích tự động tóm tắt cũng như phân lớp văn bản với máy học vectơ hỗ trợ thì văn bản cần được biểu diễn dưới dạng thích hợp".
- Text after segmentation: "Để có_thể thực_hiện rút trích tự_động tóm_tắt cũng_như phân lớp văn_bản với máy học vectơ hỗ_trợ thì văn_bản cần được biểu_diễn dưới dạng thích_hợp".

To compute the weights of word tokens, we use TF-IDF term-weighting scheme, originated from information retrieval field, which is widely used in natural language processing field. The TF-IDF scheme combines two aspects of a word: the importance of a word within a document (TF: Term Frequency) and its discriminative power within the whole collection (IDF: Inverse Document Frequency). There are many variants of the TF-IDF scheme and the most used is:

- TF: number of occurrences of a word token in a given text document
- IDF: $\log(1 + \frac{N}{n})$, where N is number of documents in the collection and n is the number of documents which contain given word token.

2.2 Automatic Text Summarization

Our proposed model [10] for Vietnamese text summarization provides good result. The proposed model is based on the notion of similarity between sentences. The ranking scores of sentences are computed by using the advanced PageRank algorithm and then sentences with highest scores are extracted as abstract (summary).

The main steps for text summarization are the following:

- Split text document into sentences represented as vector in the space of indexing terms.
- Construct graph for representing document in which nodes are sentences and arcs represent the similarity between sentences.
- PageRank algorithm [12] has been modified to better suit the new context – undirected weighted graph. Sentence is selected to put into the summary based on its PageRank score.

Our proposed model belongs to the class of "extraction" approaches. The main advantages of this approaches are: since they are unsupervised approaches so no training set is needed; we can identify exactly how many sentences have been extracted.

First, after sentences are splitted by using separation characters such as '.', '?', '…' etc., we use bag of words model and TF-IDF weighting scheme to model sentences. In the next step, an undirected weighted graph will be constructed in which nodes represent sentences and each arc represents the similarity between two sentences. The arc's weight is computed by using the Jaccard coefficient [13]. We proved in [10] that by using Jaccard coefficient we can get a good result for the problem of automatic text summarization. In the last step, the PageRank algorithm that have been adjusted to suit the context of undirected weighted graph is used to compute the "importance" of nodes. The "importance" of nodes will be fixed after a number of iterations. The "importance" of nodes are updated every iteration by using following equation:

$$PR(A) = \frac{1-d}{N} + d(W_{AB}\frac{PR(B)}{L(B)} + W_{AC}\frac{PR(C)}{L(C)} + \ldots)$$

Where

d is often chosen equal to 0.85
W_{ij} is the weight of arc connected two nodes i and j
L(i) is N−1, N is the number of sentences.

Sentences are arranged in descending order according to their importance scores. A certain percentage of sentences with highest scores is selected as summary. In this work, 15 % of sentences or at least 2 sentences will be included in summary. The following example shows the result of summarization module:

Original text: Windows XP ngừng hỗ trợ vào ngày 8/4 năm sau. Nhiều nhân viên bán hàng bảo hiểm tại Nhật Bản sẽ được chuyển từ máy tính cũ lên tablet chạy Windows 8 để tương tác tốt hơn với khách hàng. Microsoft tại Nhật Bản hôm nay thông báo đang giúp một công ty bảo hiểm lớn của Nhật Bản là Meiji Yasuda nhằm nâng cấp hàng loạt máy tính chạy hệ điều hành sắp tròn 12 tuổi. Các thiết bị mới sẽ chạy Windows 8 do Fujitsu sản xuất cùng nhiều phần mềm và tiện ích cài đặt sẵn. "Trước đây, đội ngũ bán hàng sẽ chuẩn bị các đề xuất trên máy tính chạy Windows XP và sau đó in ra để chia sẻ với các khách hàng. Tuy nhiên, hệ thống thiết bị mới sẽ giúp chấm dứt các bước làm phiền toái này", thông báo của Microsoft có đoạn. Ngoài trang bị phần cứng mới, hãng phần mềm Mỹ cũng sẽ tổ chức khóa đào tạo và hướng dẫn sử dụng thao tác trên phần mềm mới. Các khách hàng cũng sẽ thuận tiện hơn trong việc sử dụng như đăng ký thông tin, tìm hiểu trực tiếp các gói bảo hiểm mà không phải ngập trong đống giấy tờ, văn bản như trước đây. Meiji Yasuda cũng sẽ là công ty bảo hiểm nhân thọ đầu tiên của Nhật thông qua việc sử dụng hoàn toàn hệ điều hành Windows 8 Pro. Microsoft dự kiến sẽ chấm dứt hỗ trợ hệ điều hành Windows XP từ ngày 8/4/2014. Tuy nhiên, đây vẫn là hệ điều hành có số lượng người dùng khổng lồ, kém không nhiều so với vị trí dẫn đầu thuộc về Windows 7. Hỗ trợ công ty bảo hiểm Nhật Bản là một trong những động thái "mạnh tay" của Microsoft giúp Windows XP sớm "nghỉ hưu" và nhường sự phát triển cho các hệ điều hành mới hơn.

Abstract (summary): Nhiều nhân viên bán hàng bảo hiểm tại Nhật Bản sẽ được chuyển từ máy tính cũ lên tablet chạy Windows 8 để tương tác tốt hơn với khách hàng. Microsoft tại Nhật Bản hôm nay thông báo đang giúp một công ty bảo hiểm lớn của

Nhật Bản là Meiji Yasuda nhằm nâng cấp hàng loạt máy tính chạy hệ điều hành sắp tròn 12 tuổi. “Trước đây, đội ngũ bán hàng sẽ chuẩn bị các đề xuất trên máy tính chạy Windows XP và sau đó in ra để chia sẻ với các khách hàng. Tuy nhiên, hệ thống thiết bị mới sẽ giúp chấm dứt các bước làm phiền toái này”, thông báo của Microsoft có đoạn. Microsoft dự kiến sẽ chấm dứt hỗ trợ hệ điều hành Windows XP từ ngày 8/4/2014.

2.3 Text Classification

To do the classification, a classifier should be trained using training dataset in which the topic of each document is known in advance. The goal of our study was to verify the usability of the proposed model using abstract of text document instead of using the full text content so in this study, the abstracts of text documents are extracted automatically to perform the test. We use libSVM [14] and decision tree J48 which are integrated in WEKA [15] to verify the model.

For training a classifier using WEKA, training dataset should be in ARFF format. The sparse ARFF format is structured as follow:

@RELATION <relation-name>
@ATTRIBUTE <Attribute-name> <datatype>
…
@ATTRIBUTE class {class-label$_1$, class-label$_2$ …}
@DATA
{<index$_1$> <value$_1$> … <index$_n$> “class-label”}

3 Results and Discussion

The goal of our research was to verify the feasibility of the text classification model based on automatic text summarization so we will compare the results of using abstracts of text documents with the results of using full text documents. We use a PC with a CORE i3 CPU and 4 GB RAM to perform the test.

3.1 Experimental Dataset

For experiments, we use 2000 articles, under ten different topics, which are collected from online newspapers (vnexpress.net and vietnamnet.vn). The distribution of ten topics is shown on Table 1.

After collecting 2000 articles, text summarization method (described in Sect. 2.2) is applied on this set of articles to produce 2000 abstracts/summaries. The execution time is about 1 s per article (see Table 2).

Table 1. Distribution of 10 topics

Topic	Number of articles
IT	200
Business	200
Law	200
Education	200
Health	200
Sports	200
Science	200
Travel	200
Society	200
Culinary	200

Table 2. Execution time for creating the summaries

Topic	Dataset size (MB)	Execution time (second)
IT	5.91	201
Business	6.84	280
Law	6.01	229
Education	6.59	273
Health	6.21	230
Sports	6.28	229
Science	6.94	229
Travel	6.46	186
Society	6.89	202
Culinary	6.28	242

3.2 Exeperimental Results

We use two most frequently used machine learning methods for the case of text classification: support vector machine (libSVM) and decision tree (J48) for verifying the feasibility of the proposed model. In both case, we use 10-fold cross validation to test the model.

3.2.1 Using libSVM

We show the classification result in the case of using full text documents in Table 3.

In the case of using summaries dataset, classification result is shown in Table 4.

With this experiment, we can conclude that by using libSVM the classification result on the summaries dataset is better than the one on the full text dataset for all metrics. Especially, we can improve more than 7 % of the TP Rate on average.

3.2.2 Using J48

We show the classification result in the case of using full text documents in Table 5.

In the case of using summaries dataset, classification result is shown in Table 6.

Table 3. Classification result using libSVM on full text documents dataset

Topic	TP rate	FP rate	Precision	Recall	F-measure
IT	0.975	0.112	0.491	0.975	0.653
Business	0.735	0.004	0.948	0.735	0.828
Law	0.930	0.024	0.812	0.930	0.867
Education	0.795	0.007	0.924	0.795	0.855
Health	0.815	0.009	0.906	0.815	0.858
Sports	0.850	0.003	0.971	0.850	0.907
Science	0.910	0.002	0.978	0.910	0.943
Travel	0.700	0.011	0.881	0.700	0.780
Society	0.768	0.007	0.921	0.768	0.837
Culinary	0.855	0.005	0.950	0.855	0.900
Average	**0.833**	**0.019**	**0.878**	**0.833**	**0.843**

Table 4. Classification result using libSVM on summaries dataset

Topic	TP rate	FP rate	Precision	Recall	F-measure
IT	0.975	0.048	0.691	0.975	0.809
Business	0.799	0.002	0.975	0.799	0.878
Law	0.965	0.024	0.818	0.965	0.885
Education	0.859	0.004	0.961	0.859	0.907
Health	0.885	0.006	0.941	0.885	0.912
Sports	0.950	0.002	0.979	0.950	0.964
Science	0.960	0.006	0.946	0.960	0.953
Travel	0.805	0.009	0.904	0.805	0.852
Society	0.840	0.007	0.928	0.840	0.882
Culinary	0.945	0.003	0.969	0.945	0.957
Average	**0.898**	**0.011**	**0.911**	**0.898**	**0.900**

Table 5. Classification result using J48 on full text documents dataset

Topic	TP rate	FP rate	Precision	Recall	F-measure
IT	0.790	0.068	0.564	0.790	0.658
Business	0.665	0.037	0.668	0.665	0.667
Law	0.650	0.027	0.726	0.650	0.686
Education	0.865	0.017	0.848	0.865	0.856
Health	0.635	0.024	0.747	0.635	0.686
Sports	0.835	0.017	0.843	0.835	0.839
Science	0.709	0.020	0.797	0.709	0.750
Travel	0.620	0.030	0.697	0.620	0.656
Society	0.747	0.031	0.725	0.747	0.736
Culinary	0.840	0.022	0.808	0.840	0.824
Average	**0.736**	**0.029**	**0.742**	**0.736**	**0.736**

Table 6. Classification result using J48 on summaries dataset

Topic	TP rate	FP rate	Precision	Recall	F-measure
IT	0.845	0.023	0.805	0.845	0.824
Business	0.729	0.028	0.744	0.729	0.736
Law	0.835	0.021	0.819	0.835	0.827
Education	0.859	0.011	0.895	0.859	0.877
Health	0.775	0.032	0.731	0.775	0.752
Sports	0.920	0.007	0.934	0.920	0.927
Science	0.845	0.022	0.813	0.845	0.828
Travel	0.830	0.015	0.860	0.830	0.845
Society	0.755	0.021	0.799	0.755	0.776
Culinary	0.850	0.016	0.854	0.850	0.852
Average	**0.824**	**0.020**	**0.825**	**0.824**	**0.825**

Again, the obtained results indicate that the classification result on the summaries dataset is better than the one on the full text dataset for all metrics. We can improve more than 10 % of the TP Rate on average.

4 Conclusion

In this paper, we propose a text classification model based on automatic text summarization. This is a relatively new approach and there are not many studies in the literature. The initial results that we obtained are satisfactory. Indeed, the model we propose gives the better results than the traditional one on all evaluation metrics.

These positive results can be explained by several reasons: (1) abstract/summary of a text covers the main ideas of the whole text so it can be used to identify the topic; (2) by using the right words segmentation method (vnTokenizer library in this work) we do not lose too much semantic information; (3) the proposed text summarization method is effective. Indeed, as we mentioned in [10] our method can produce the summary of the text with the accuracy of 52 % (an acceptable accuracy for automatic text summarization).

Although the experimental results show the feasibility of the proposed model, we have also remaining issues: (1) the volume of experimental data is not large enough; (2) only Vietnamese texts are collected. In future, we will continue updating this work, e.g., increasing the volume of experimental dataset as well as improving the text summarization model accuracy.

References

1. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, pp. 137–142 (1998)

2. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering by a tolerance rough set model. Intl. J. Fuzzy Logic Intell. Syst. **17**(2), 199–212 (2012)
3. Zaïane, O.R., Antonie, M.-L.: Classifying text documents by associating terms with text categories. In: Proceedings of the 13th Australasian Database Conference, pp. 215–222, Melbourne, Victoria, Australia (2002)
4. Amini, M.R., Usunier, N., Gallinari, P.: Automatic text summarization based on word-clusters and ranking algorithms. In: Proceedings of the 27th European Conference on Advances in Information Retrieval Research, Santiago de Compostela, Spain (2005). doi:10.1007/978-3-540-31865-1_11
5. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, pp. 40–48, Seattle, Washington (2000). doi:10.3115/1117575.1117580
6. Barzilay, R., Elhadad, N., McKeown, K.: Inferring strategies for sentence ordering in multidocument news summarization. J. Artif. Intell. Res. **17**, 35–55 (2002)
7. Johnson, D., Oles, F., Zhang, T., Goetz, T.: A decision tree-based symbolic rule induction system for text categorization. IBM Syst. J. **41**(3), 428–437 (2002)
8. Han, E.H., Karypis, G., Kumar, V.: Text categorization using weighted-adjusted k-nearest neighbor classification. In: PAKDD Conference (2001)
9. Ruiz, M., Srinivasan, P.: Hierarchical neural networks for text categorization. In: ACM SIGIR Conference (1999)
10. Truong, Q-D., Nguyen, Q-D.: Automatic Vietnamese text summarization (in Vietnamese). In: Proceeding of The Fifteenth National Conference, pp. 233–238, Hanoi, Vietnam (2012)
11. 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). doi:10.1007/978-3-540-88282-4_23
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web (1999)
13. Jaccard P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547–579
14. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27 (2011)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)