ICIC Express Letters
Volume **11**, Number **5**, May **2017**

 ISSN 1881-803X
pp. **955–959**

AN EFFECTIVE DATA ORGANIZING METHOD FOR VIETNAMESE DOCUMENT RETRIEVAL

Ha Nguyen Thi Thu[1], Linh Bui Khanh[1], Tinh Dao Thanh[2] and Vinh Ho Ngoc[3]

[1]Faculty of Information Technology
Vietnam Electric Power University
235 Hoang Quoc Viet, Tu Liem, Hanoi 100000, Vietnam
{ hantt; linhbk }@epu.edu.vn

[2]Faculty of Information Technology
Le Quy Don Technical University
236 Hoang Quoc Viet, Tu Liem, Hanoi 100000, Vietnam
tinhdt@mta.edu.vn

[3]Faculty of Information Technology
Vinh University of Technology Education
Hung Dung ward, Vinh City, Nghe An 460000, Vietnam
hongocvinh@gmail.com

Received November 2016; accepted February 2017

Abstract. *The development of amount of the Internet information is increasing. The current search engine needs to be improved of high precision and speech as quickly as possible. In addition to the development of the algorithms, upgrade computer systems, we should organize the data in the database; it can help this engine search better and more exactly. In this paper we present a method to search the related documents based on clustering. In this, the features are assigned weight by supporting. Experimental results show that the proposed method is really effective, high accurate and the response results are quick.*

Keywords: Support, Data mining, Text retrieval, Information retrieval, Clustering, Document retrieval

1. **Introduction.** The development of Internet brings an explosive amount of information on the web. Sometimes, it makes users feel quite hard to read and search information that they need. Therefore, data mining is hot and related field such as information retrieval, information extraction, and data clustering are concerned [1,2].

Information retrieval is a sub field of data mining that aims to store and allow quick access to a large amount of information. The simplest studies are described by matching the words that are entered as a search query and the documents in the data warehouse [3,10]. After that, to increase the effectiveness of search engines, there are several studies suggesting data organizing task, index documents in the warehouse or ranking data [1,4]. Some others concerned how to select features and reduce it to speed up search engines [5,6].

Most of data organizing methods use machine learning to cluster (classify) data like HAC, SVM, neural network or decision tree. After clustered (classified), documents are organized in clusters with similar kinds of semantic or content [7,8].

Feature reduction is a solution to speed up the search engine. Some studies showed that, the full features often make system slower. Therefore, to speed up effectively, feature vectors are needed to reduce. However, the selection of useful features and removing unneeded features is a difficult problem [5,6]. In this paper, we present a method that uses support and feature reduction to improve speed of the search engine. Data is also

organized into clusters. The documents in each cluster are similar of content. National words in these documents are considered as features. They include nouns, adjectives and verbs.

The rest of the paper is organized as follows. Section 2 is the presentation of our method for data organizing. Methodology of Vietnamese document retrieval will be presented in Section 3. Experiments and results will be shown in Section 4. And finally, Section 5 is a conclusion and future work.

2. Document Organizing Based on Clustering.

2.1. Feature selection.

Feature selection is one of the key topics in machine learning and other related fields. Real-life datasets are often characterized by a large number of irrelevant or redundant features that may significantly hamper model accuracy and learning speed if they are not properly excluded. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features.

To overcome the disadvantages of large feature vectors we selected by using a word segmentation tool for separating word and selecting only national words. A national word set defined is a set of words that include verb, noun and adjective.

Example 2.1.
"Singapore vừa triển khai thử nghiệm dịch vụ truyền hình video gắn trực tiếp vào máy ATM, cho phép khách hàng có thể "chát" trực tiếp với nhân viên ngân hàng khi gặp vấn đề."
National word set= {Singapore, triển_ khai, thử_ nghiệm, dịch_ vụ, truyền_ hình, video, gắn, máy, ATM, cho_ phép, khách_ hàng, chat, nhân_ viên, ngân_ hàng}.

2.2. Document organizing based on clustering.

Clustering algorithms group a set of documents into subsets or clusters. The algorithms' goal is to create clusters that are coherent internally, but clearly different from each other. In other words, documents within a cluster should be as similar as possible; and documents in one cluster should be as dissimilar as possible from documents in other clusters. We use HAC algorithm and the similarity between two documents based on Euclidean distances is as Equation (1).

$$d(i,j) = \sqrt{\left(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2\right)} \tag{1}$$

where $d(i,j)$ is the distance of document i and document j. x_{ik} and x_{jk} are the features that are extracted from document i and document j (in this, they are national words).

3. The Methodology of Effective Document Retrieval.

3.1. Calculating score of features based on support.

In the clustering process (Section 2.2), there are n clusters made. It is denoted as C and presented as below

$$C = \{C_1, C_2, \ldots, C_n\} \tag{2}$$

In each cluster C, we have a set of documents D.

$$D = \{d_1, d_2, \ldots, d_m\} \tag{3}$$

Suppose that, in each cluster C, if we consider a document is a transaction, frequency of national word is considered as an item, and we have Table 1.

After that, we calculate score of term. We use the improving support (in the association rule) to assign value to terms. With each term in Table 1, support of it with each C is calculated as

$$\text{supp}\,(t_i \to C_j) = \frac{n(t_i)}{N_{C_j}} \tag{4}$$

In which:
- $n(t_i)$: number of document in cluster C_j that includes t_i
- N_{C_j}: number document in each cluster C_j.

TABLE 1. Transactions and item set

TID	Term
d_1	$t_{11}, t_{12}, \ldots$
d_2	$t_{21}, t_{22}, \ldots$
$\vdots$	$\vdots$
d_k	$t_{k1}, t_{k2}, \ldots$

3.2. **Document retrieval.** In the entered query Q, we perform to extract national words.

$$Q = \{w_1, w_2, \ldots, w_k\} \tag{5}$$

Then, we calculate total of national words in the query Q with each cluster C.

$$\text{total\_supp}\,(Q_{C_i}) = \sum_{j=1}^{k} \text{supp}(w_j) \tag{6}$$

In which:
- $\text{supp}(w_j)$ is the support of the term w_j with cluster C.

The highest of total support is the cluster that is the most similar with the query.

Here is the algorithm to document retrieval. It is called REBDO (Retrieval Effectively Based on Data Organizing).

REBDO Algorithm

Input: query Q, national word dictionary V
Output: d_i
Initialization
$L = \emptyset$; $M = \emptyset$; $i = 0$;
Begin
1. *Segment real words in query Q*
   For $i \leftarrow 1$ to ***length (Q)***
   If $Q[i] = Match(Q[i], V)$ ***then***
   $L \leftarrow Q[i]$;
2. *Calculate sum of supporting in each cluster*
   For each cluster C_k
   M←total_supp(L(i))
3. *Ranking clusters*
   For $i := 1$ to ***count(M) do***
   $K \leftarrow M[i]$;
4. *Output results*
   Sorting (K)';
   Output d_i in K;

End;

FIGURE 1. REBDO algorithm

4. Experiments.

4.1. **Corpus.** There is no standard corpus for Vietnamese text retrieval now. Therefore, we built corpus manually. Documents in corpus are downloaded from websites as: http://vietnamnet.vn, http://tin247.com. There are 1,230 documents in the corpus. Table 2 presented some documents in corpus and number sentences in each document.

Table 2. Corpus

Document	Source	Sentences	File name
Ứng dụng Twitter trong lớp học	thongtincongnghe.com	28	18-10.txt
Hacker "sờ tới" website chính phủ Malaysia	Vietnamnet.vn	15	11-5.txt
Yahoo ra mắt công cụ tìm kiếm app cho Android	Ngoisao.net	12	12-9.txt
TQ phủ nhận điều tra chống độc quyền Microsoft	Tin247.com	21	13-8.txt
Cấu hình tối thiểu để nâng cấp lên Mac OS X Lion	Sohoa.vnexpress	18	16-3.txt
Chọn hệ điều hành của bạn	pcworld.com	69	21-10.txt
Linux ở khắp mọi nơi	Vietbao.vn	71	22-1.txt
Màn hình cảm ứng: Đằng sau những cú chạm	Pcworld	86	25-4.txt
Phanh phui bí mật thế giới ngầm hacker Việt Nam	Echip.com	137	33-4.txt
Người dùng di động quan tâm giá cả hơn sáng tạo công nghệ	baomoi.com	39	33-7.txt

All file downloaded from website will be saved in corpus by *.txt and preprocessed.

4.2. **Word segmentation.** We build a dictionary of national words and use VnTagger tool that is downloaded from vlsp website to segment words. VnTagger is published on Internet via address: http://vlsp.hpda.vn:8080/demo/?page=home [11].

4.3. **Evaluation.** At present, Vietnamese does not have any standard assessment method; we use recall measure for evaluation. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$recall = \frac{\#(\text{relevant items retrieved})}{\#\ (\text{relevant items})} \tag{7}$$

For the evaluation, we build a retrieval system and use it to retrieve automatic documents. It is developed based on C # on Microsoft Visual Studio 2015.

We perform with the queries that relevant several topics as finance, health, sport, travel, tablet and electricity. Table 3 shows the result of travel topic is not better than others, because it includes some sub topics as culture, culinary, weather and geography.

5. Conclusion.

The task of information retrieval based on content has been concerned by researchers and scholars when the current systems still search by keyword or phrase. In this paper, we propose an effective method for information retrieval based on content and added objectives are fast and accurate. With the results of experimental it shows that, our method is really effectively to reduce complex computing and time for processing when performing with Vietnamese text. In the near future, we will develop a system that is based on our proposed method and test with real data on the Internet.

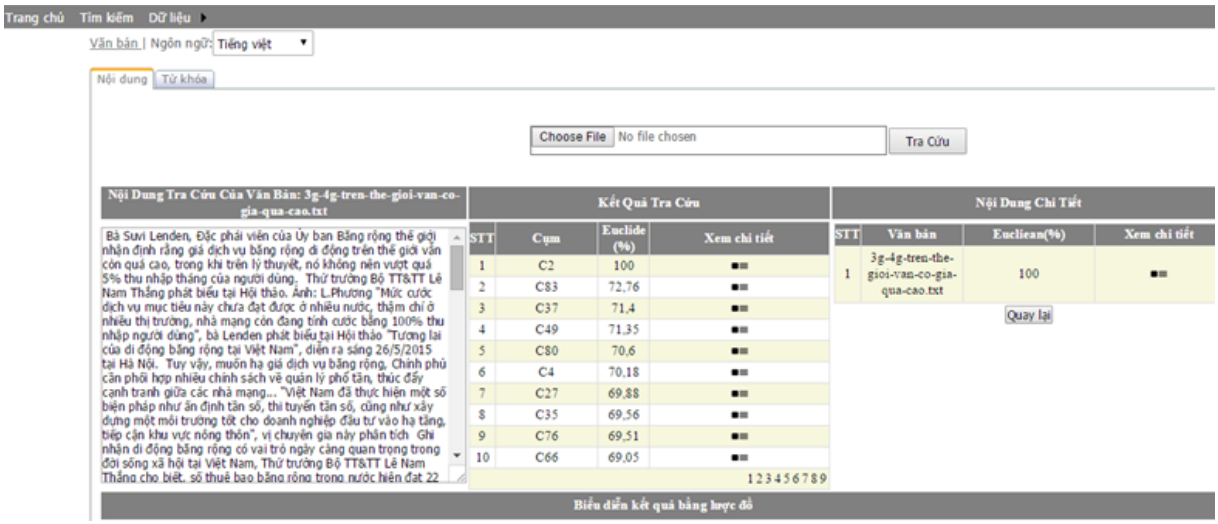

FIGURE 2. The Vietnamese document retrieval system

TABLE 3. Some topics for retrieving

Topic	Number of relevant documents	Recall
Finance	26	0.23
Health	20	0.167
Sport	63	0.155
Travel	42	0.514
Tablet	30	0.113
Electricity	78	0.12

REFERENCES

[1] P. Bhattacharyya and J. Datta, *Ranking in Information Retrieval*, 2010.
[2] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali and S. Quarteroni, *Web Information Retrieval*, 2013.
[3] G. Buscher, A. Dengel and L. van Elst, Query expansion using gaze-based feedback on the subdocument level, *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, pp.387-394, 2008.
[4] M. V. Zadeh, *Improving the Performance of Text Information Retrieval (IR) System*, Ph.D. Thesis, Porto University, 2012.
[5] J. H. Lau, D. Newman, S. Karimi and T. Baldwin, Best topic word selection for topic labelling, *Proc. of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, China, pp.605-613, 2010.
[6] M. F. Moens and I. Vulic, Monolingual and cross lingual probabilistic topic models and their application in information retrieval, *Proc. of the 35th European Conference on Information Retrieval*, Moscow, Russian Federation, pp.875-878, 2013.
[7] H. Park et. al, Agglomerative hierarchical clustering for information retrieval using latent semantic index, *IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp.19-21, 2015.
[8] C. Kalyanasundaram, S. Ahire, G. Jain and S. Jain, Text clustering for information retrieval system using supplementary information, *International Journal of Computer Science and Information Technologies*, vol.6, no.2, pp.1613-1615, 2015.
[9] L. Kuhn and C. Eickhoff, Implicit negative feedback in clinical information retrieval, *Medical Information Retrieval Workshop (MedIR)*, Pisa, Italy, 2016.
[10] J. J. Rocchio, *Relevance Feedback in Information Retrieval*, 1971.
[11] *http://vlsp.hpda.vn:8080/demo/?page=home*.