An Effective of Data Organizing Method Combines with Naïve Bayes for Vietnamese Document Retrieval

Khanh Linh Bui[1], Thi Ngoc Tu Nguyen[1], Thi Thu Ha Nguyen[1](✉), and Thanh Tinh Dao[2]

[1] Vietnam Electric Power University, 235 Hoang Quoc Viet, Tuliem, Hanoi, Vietnam
{linhbk,hantt}@epu.edu.vn
[2] Le Quy Don Technical University, 100 Hoang Quoc Viet, Tuliem, Hanoi, Vietnam
tinhdt@mta.edu.vn

Abstract. Data is uploaded to Internet daily that make more and more difficult to mine it. Currently, the available of data mining tools still cannot discover knowledge from data that need semantic with difference dimensions. In this paper we present a method to search the related documents based on clustering that grouped by content. In this, the features are assigned weight by supporting. Experimental results show that the proposed method is really effective, high accuracy and the response results are quickly.

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 as information retrieval, information extraction, data clustering are concerned [1, 2].

Information retrieval is a sub field of data mining that aims to store and allows quick access a large amount of information. The text is often considered as documents, books, articles, etc. However, this is not an easy task, because the booklets in the information systems often have to deal with tens of thousands or tens of millions of documents. So, the search engine can not process more quickly if we don't use any technique to reduce time for processing and enhance accuracy of system [1, 4].

There are several proposed approaches previously mentioned organizing data and feature reduction that have been able to effective search engines [5, 6]. However, it is very difficult to determine feature and how to reduce it. In this paper, we present a method to search effectively by reducing the feature and enriching the semantics of features by using support measure that improve from association rule. It is really better than two – dimensional tf - idf before.

P. Cong Vinh et al. (Eds.): ICCASA 2017/ICTCC 2017, LNICST 217, pp. 205–213, 2018.
https://doi.org/10.1007/978-3-319-77818-1_20

The rest of the paper is organized as follows: In Sect. 2, we will introduce some related works. In Sect. 3 is the presentation of our method for data organizing, methodology of Vietnamese document retrieval will be presented in Sect. 4. Experiments and results will show in Sect. 5. And finally, Sect. 6 is a conclusion and future works.

2 Related Works

The earliest studies on the task of information retrieval are described through keywords. It is the simplest approach by matching the words that are entered as a search query and the documents in the data warehouse [3, 10]. To increase the effectiveness of search engines, there are several studies suggested to organizing data task, index documents in the warehouse or ranking data [1]. The other studies also added matching problem that can enhance accuracy between query and data [3]. They also concerned how to select features and reduce it to speed up search engines [5, 6].

The problem of organizing data, the number of studies often uses clustering or classification based on machine learning methods as HAC, SVM, neural network or decision tree. After clustered or classified, documents is organized in clusters with similar kinds of semantic or content [7, 8].

To enhance accuracy of the search engines, some researches focus on relevant feedback. They proved the effectiveness of the search engine when receive feedback from the users [9, 10].

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 remove unneeded features is a difficult problem [5, 6].

3 Organize Documents in the Warehouse

3.1 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 is define is a set of words that are include verb, noun and adjective.

3.2 Organize Document 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 score to cluster documents.

4 The Methodology of Effective Document Retrieval

4.1 Calculating Score of Features Based on Support

In the clustering process (Sect. 3.2), there are n clusters made. It is called *C* and presented as below

$$C = \left\{ C_1, C_2, \ldots, C_n \right\} \tag{1}$$

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

$$D = \left\{ d_1 \ldots d_m \right\} \tag{2}$$

Suppose that, in each cluster *C*, if we consider a document is a transaction, frequency of national word is considered an item, we have a table like this:

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

Table 1. Transactions and item set

TID	Term
d_1	t_{11}, t_{12},…
d_2	t_{21}, t_{22}…
……	
d_k	t_{k1}, t_{k2},…

$$\mathrm{supp}(t_i \to C_j) = \frac{n(t_i)}{N} \tag{3}$$

In which:

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

Finally, we built a relationship of national words and topics. In this, each national word has a score to topics. We can set a threshold to adjust amount of national words in each topic. It called feature reduction (Fig. 1).

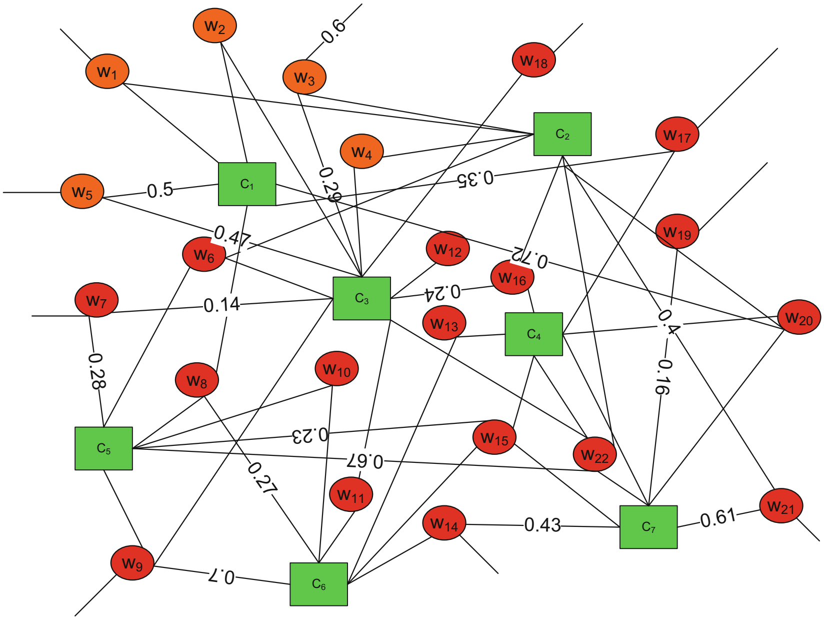

Fig. 1. Relationship between national words and topics

4.2 Calculating Similarity Between Query and Clusters

In the entered query Q, we perform to extract national words (Fig. 2).

$$Q = \{q_1, q_2, \ldots, q_k\} \tag{4}$$

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

$$\text{total\_supp}(Q_{C_i}) = \sum_{j=1}^{k} \text{supp}(\text{w}_j) \tag{5}$$

In which:

– *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 (Fig. 3). The algorithm likes below:

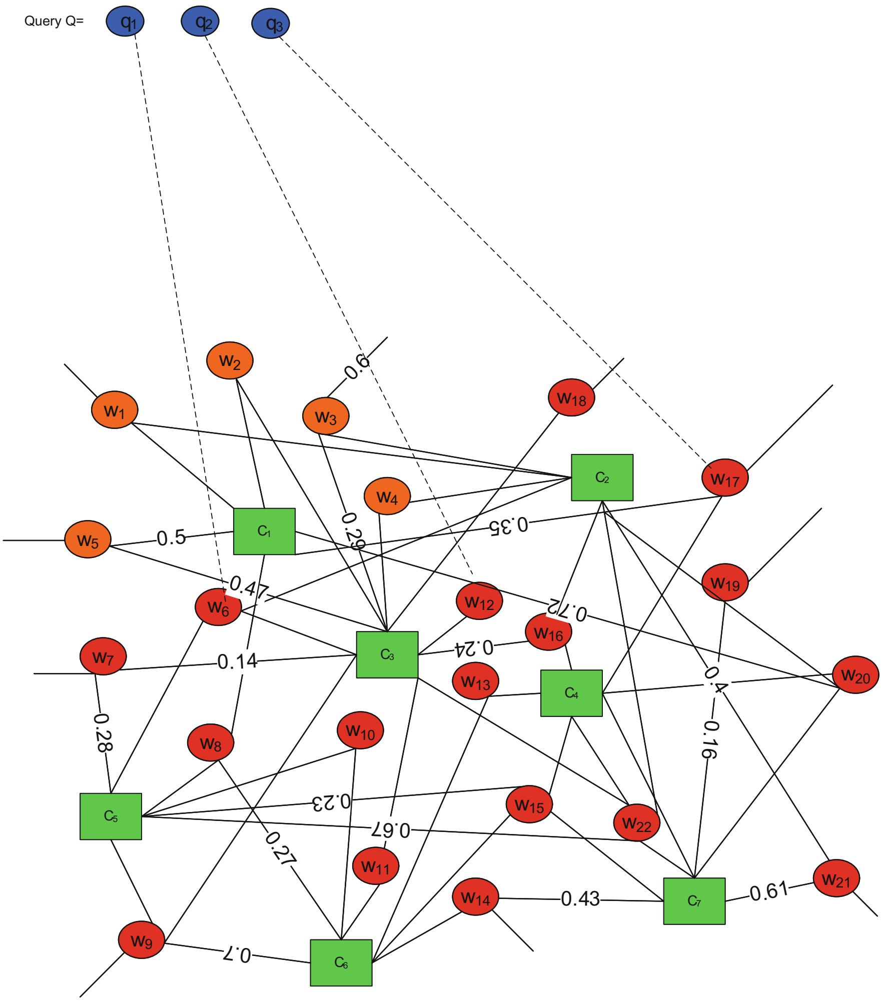

Fig. 2. Similarity between query Q and topics

```
REBDO Algorithm

Input: query $Q$, national word dictionary $V$
Output: $d_i$
Initialization
        $L=Ø; M= Ø; i=0;$
Begin
  1. Segment real words in query Q
       For i←1 to length(Q)
       If Q[i] ЄV then
       L←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←M[i];
  4. Output results
       Sorting (K)'
       Output $d_i$ Є cluster in $K$;
End;
```

Fig. 3. REBDO algorithm

4.3 Document Retrieval

In machine learning, Naïve Bayes belongs to classifier methods that use probability with features is independence (conditional independence). Naïve Bayes is applied widely in data classification. Assume that, there are two classes: C = {R, $\overline{R}$). In which, R is set of document that relate with query $\overline{R}$ is collected by unrelated document to query. So that, the information retrieval problem becomes to determine which documents in cluster are related to the query. Similarity of document d_j and query q is denoted by Bayes rule as

$$sim(\overrightarrow{d_j}|q) = \frac{\Pr(d_j|R) \times \Pr(R)}{P(d_j|\overline{R}) \times \Pr(\overline{R})} \tag{6}$$

In which:

- Pr(R): Probability of documents set that related to query.

- $\Pr(\overline{R})$: Probability of documents set that unrelated to query.
- $\Pr(d_j|R)$: Probability d_j with set of documents R that related to query.
- $P(d_j|\overline{R})$: Probability d_j with set of documents $\overline{R}$ that related to query.

Smoothing (6) by logarithm, obtain (7)

$$sim(\vec{d_j}|q) = \log \frac{\Pr(d_j|R)}{P(d_j|\overline{R})} + \log \frac{\Pr(R)}{\Pr(\overline{R})} \tag{7}$$

Assume that, term in query q is independence

$$q = \{t_1, t_2, \ldots, t_k\} \tag{8}$$

Then

$$\Pr(\vec{d_j}, R) = \prod \Pr(t_i|R) \tag{9}$$

$$\Pr(\vec{d_j}, \overline{R}) = \prod \Pr(t_i|\overline{R}) \tag{10}$$

Assume that set of training data is enough larger $R << \overline{R}$

$$sim(\vec{d_j}|q) \approx \log \frac{\prod \Pr(t_i|R)}{\prod \Pr(t_i|\overline{R})} \tag{11}$$

5 Experimental

5.1 Corpus

There is no standard corpus for Vietnamese text summarization now. Therefore, we built corpus by manual. Documents in corpus are downloaded from websites' news as: http://thongtincongnghe.com, http://echip.com, http://vnexpress.net, http://vietnamnet.vn, http://tin247.com. There are over 300 documents in it. 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.

5.2 Word Segmentation

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

5.3 Evaluation

At the 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 (Table 3)

Table 3. Some topics for retrieving

Topics	Number of relevant documents	Recall
Business	52	0.573333
Education	54	0.68
Football	46	0.453
Travel	42	0.514
Information	78	0.526
Technical	36	0.511

$$recall = \frac{|\{relevantdocument\} \cap \{retrieveddocuments\}|}{|\{relevantdocuments\}|} \tag{12}$$

6 Conclusion

The task of information retrieval based on content 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 show that, our method really effectively to reduce complex computing and time for processing when performing with Vietnamese text.

References

1. Bhattacharyya, P., Datta, J.: Ranking in information retrieval, 16 April 2010
2. Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni, S.: Web Information Retrieval. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39314-3. ISBN 978-3-642-39314-3
3. Buscher, G., Dengel, A., van Elst, L.: Query expansion using gaze-based feedback on the subdocument level. In: Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Singapore, pp. 387–394 (2008)
4. Zadeh, M.V.: Improving the performance of text Information Retrieval (IR) System, Ph.D thesis. Porto University (2012)
5. Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best topic word selection for topic labelling, Coling 2010, Posters, pp. 605–613 (2010)
6. Moens, M.-F., Vulić, I.: Monolingual and cross-lingual probabilistic topic models and their applications in information retrieval. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 874–877. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36973-5_106
7. Park, H., et al.: Agglomerative hierarchical clustering for information retrieval using latent semantic index. In: 2015 IEEE International Conference Smart City/Socialcom/Sustaincom (SmartCity), 19–21 December 2015
8. Kalyanasundaram, C., Ahire, S., Jain, G., Jain, S.: Text clustering for information retrieval system using supplementary information. Int. J. Comput. Sci. Inf. Technol. **6**(2), 1613–1615 (2015)
9. Kuhn, L., Eickhoff, C.: Implicit negative feedback in clinical information retrieval. In: Medical Information Retrieval Workshop (MedIR), Pisa, Italy, 21 July 2016
10. Rocchio, J.J.: Relevance feedback in information retrieval (1971)
11. http://vlsp.hpda.vn:8080/demo/?page=home