Effect of Vietnamese Text Retrieval System based on Topic Modeling

Cuong Nguyen Ngoc
Department of Computer and Mathematical The people's Security University

Abstract
Text retrieval is an important field when the data is growing day by day. So the big problem in a text retrieval field is the large of features number. Researchers always find the ways that can reduce features in texts. In this paper, we proposed a Vietnamese text retrieval method based on text classification and topic modeling. The experimental with 220 documents shown that, our method is really effective, higher accuracy and can reduce the complexity of computing and easy to build an application system for Vietnamese text.

Keywords:
Text retrieval, Vietnamese text, Topic modeling, feature reduction, text classification.

1. Introduction

Studies in the text mining field getting attractive recently because of growing very fast day by day from many sources of electronic documents on the Internet. Not mention to structure documents, numbers of unstructured documents are also very large. Motivation of text mining allows users to extract information that they need from one or many sources of documents through some other tools: text classification, text summarization, text extraction, text clustering, text retrieval…

Information retrieval is a branch of computer science that aims to store and allows quick access a large amount of information. And text retrieval is a sub field of information retrieval. 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 now, how to reduce complexity and still effectively on the data processing system is needed. If we don’t care about features dimensional, these text retrieval systems can encounter some disadvantage: need much time for processing and accuracy is not high.

If we need much time to process, it means, the retrieval system has a low speed, can’t satisfy user’s what they need. And if we aren't processing features dimensional well, Data has noise much, so that, it can reduce the accuracy of the system. Information retrieval based on data clustering was mentioned in some previous study, but the accuracy of this approach will be less than information retrieval based on data classification [30].

For Vietnamese text, there is not any text retrieval system has been built before. Because of, Vietnamese is a single syllable language, it’s very hard to identify words, if only based on white spaces, and it needs some word processing tool for identifying words in a text. So, time need much for processing and accuracy is not high (accuracy of Vietnamese text segmentation tool about 85%).

In our research, we find a way to improve effectiveness of the Vietnamese text retrieval system by using a topic model, documents in the corpus were classified, and then, we only represent queries and documents in the corpus through set of words that much less than the number of words in the original text. When using the topic model, we don’t need to use any text segmentation tool. In this way, we can:

- Improving speed of text retrieval system.
- Improving accuracy of retrieving

We will be demonstrated effective of method in the experiment is really significant. Structures of this paper as follows: Section 2 presents some studies in information retrieval, the methodology of Vietnamese text retrieval will be presented in section 3, and section 4 is the experimental and evaluate of our method and finally is conclusion.

2. Related Work

The earliest works of text retrieval have started with a task of finding by keyword (keyword search), this is the simple text retrieval perform by input some words and find out all documents that includes these words in it. After then, for enhance of system’s speed, some studies index these documents and sorting by relative with question. Called vector space model [18].

In general, text retrieval methods focus to two directions. The first approach based on machine learning: clustering, keyword search, and feedback from user… Another direction based on optimization of search engine by reducing of features. In the machine learning approach, used Hidden Markov Model for information retrieval system [5] or probabilistic model of information retrieval [8], [14], [15]. Some studies use feedback by users and after update scores for training. For this approach,

Manuscript received November 5, 2015
Manuscript revised November 20, 2015

dimensionality hasn't reduced, so that, features are so high and time for processing need much [17], [22], [31].

For enhancing speed of retrieval system, some approaches were introduced. Other directions that are being investigated [3] and relevance feedback [31]. Buscher et al in 2008 introduced a new approach for enhancing the speed of the system by using the concept of the physiology of eye, eye movement point to point on the documents and they record track of these points. So that, it can be reduced time for processing when browse sequentially from head to bottom of documents. However, we need much of documents for training and cost for building training sets are expensive, and this training set only useful for text retrieval systems, hard to apply to other fields [1]. Beside of this approach, many researchers also proposed methods that can reduce complex of time by using topic model [19], [25], [26].

Topic modeling demonstrates the semantic relations among words, which should be helpful for information retrieval tasks. Xing Yi and James Allan proposed a text retrieval method based on topic model, and they evaluated the effectiveness of different types of topic models within those retrieval approaches. They show that topic models are effective for document smoothing, should utilize topics discovered in the top feedback documents instead of coarse-grained topics from the whole corpus [27]. XingWei in his PhD thesis also demonstrated topic modeling really effective for text retrieval and reducing dimensional [28]. Ivan Vulic et al., using a topic modeling in some his researches: text classification, text retrieval and evaluating in these methods is the evident effectiveness of the topic modeling [25].

3. Features reduction based on topic modeling

Topic models are corpus for discovering the main topic from a large number of unstructured documents. The first concept of topic model was launched in 2002 by Griffiths and Steyvers. And then, Some researchers proposed methods for building topic model: Blei, Hofmann, Hanna..., Almost of them built a corpus based on probability theory combine with LDA, Bayes or HMM [24].

Vietnamese hasn't got any studying for the topic model before and building VietwordNet now, but it hasn't completed yet. Currently, Studies focus on word processing tools, grammar analysis tools and other studies focus on text mining tools but these tools haven't applied yet. If we want to start building topic model, we need many times, cost and human.

So that, we find a new method that can reduce time for building, cost and human based on the core terms and conditional probability.

Supposed, topics defined by humans through the task of classifying sets of documents, we find a word that has the most likelihood in each subset of the document that was called core term. Signs of the core terms is represent all the words of each topic, and remain of words in each topic always relative with it through distance. If the distances between its approximates 1, indicate it is in the different topics.

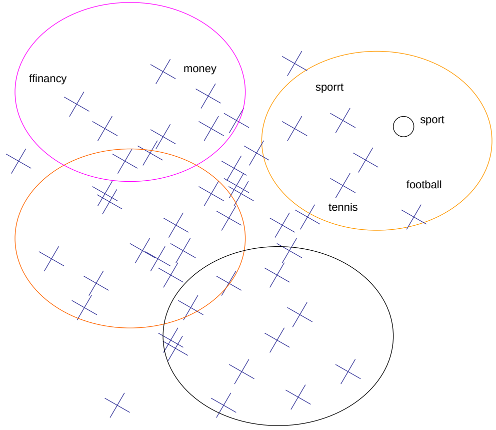


Fig 1. Topic model based on Core terms and conditional probabilistic

In the figure 1 illustrates topic modeling, these circles is topics. In each topic, "x" sign is the words that belong to topic and "O" sign indicates a core term. Somehow, topics can have a part of overlapping and some words can belong to or there another topics [10].

4. Text Retrieval Based on Text Classification and topic modeling

A. Text classification

Text classification based on topic modeling can help to reduce complex calculating process, because we don't need to process with some stop words, redundant words or not relative words. For enhancing effectiveness of retrieval system based on contents, we used a classified corpus by topic (that inherit from building topic model task). In our work, we used Naive Bayes for text classification.

Maximize the posterior (Maximum a posteriori-MAP).

$$c_{map} = \arg\max_{c \in C}(P(c \mid d)) = \arg\max_{c \in C}\left(P(c)\prod_{1 \le k \le n_d} P(t_k \mid c)\right) \quad (1)$$

In which:

- T_k is the word of the text.
- C is the subject
- $P\ (c \mid d)$ is the conditional probability of class c and given text d

- *P (c)* is the prior probability of class *c*
- *P (t_k | c)* is the conditional probability of class *c* from t_k to have.

Use the Laplace formula (1) is converted to:

$$P(t|c)=\frac{T_{ct}+1}{\sum_{t\in V}(T_{ct'}+1)}=\frac{T_{ct}+1}{\sum_{t'\in V}T_{ct'}+B'} \qquad (2)$$

In which *B'* is the number of topic words.

B. Text retrieval based on text classification

i. Topic feature representation:

Information retrieval in the document sets that have been classified before. With each topic, we have:

Training set corresponding contain m document $D_k=\{d_{k1},d_{k2},...,d_{km}\}$. In this dki is the document in training set with topic k.

Term sets corresponding with topic $T_{D_k}=\{t_{Dk1},t_{D_{k2}},...,t_{Dkj}\}$. TDki is the term in topic k.

Term set TDk with each topic k is considered the feature vector that representation each topic. Terms is inherited when we build the topic modeling, but the value of each term in this haven't been determined yet. To determind value of feature we calculate based on average score of terms has been contained in topic. Sign $\overline{x}_{Dki}$, and $\overline{x}_{Dki}$ is determined as follows:

$$\overline{x}_{Dki}=\frac{\sum_{l=1}^{m}N_l(t_{Dki})}{Z} \qquad (3)$$

In which:

- $\overline{x}_{Dki}$: is the average score of term ith in the document set D of topic k
- $N_l(t_{Dki})$: is the number of occuracy of term ith in document l in document set D.
- Z: Number of occuracy of term tDki in document set D.
- Let PDk is the representation vector of document set of topic k, PDk can be rewritten:

$$T_{D_k}=\{<t_{Dk1},\overline{x}_{Dk1}>,<t_{D_{k2}},\overline{x}_{Dk2}>,...,<t_{Dkj},\overline{x}_{Dki}>\}$$

Finally, term set built can be representationed by:

$$V=\{T_{Dk1}\cup T_{Dk2},...,\cup T_{Dkn}\}$$

ii. Queries representation

Each query s is represented by extract terms from it and then calculate score of each term.

$$s=\{<t_1,x_1>,<t_2,x_2>,...,<t_p,x_p>\}$$

In which:

- t_i : is the terms have been extracted from s based on dictionary V.
- x_j: score of term t_j in query s.

Figure 2. below illustrate feature vector of query *s*.

REPRESENTATION OF QUERY (ROQ)

```
Input           s: query;
                V: vocabulary of terms;
Output          W: list of weight of term in s;
Initialization
                N←∅; O←∅; d←0; W←∅;
1.          For i=1 to length(s)
1.1             if (s[i] ∈ V) then
1.1.1               if (s[i] ∉ N) then
1.1.1.1                 N ←s[i];
1.1.1.2                 d++;
1.1.1.3                 O[i] ←1;
1.1.2               if (s[i] ∈ N) then O[i] ←O[i]+1;
2.          For i=1 to count (N)
2.1             W[i] ←O[i]/d;
```

Fig 1.Representation of query.

To calculate similarity score between of query s and topics, we used Euclidean distance method

$$d(s,D_k)=\sqrt{\sum_{i=1}^{g}(t_{si}-t_{Dj})^2} \qquad (4)$$

After extract classes that have least similarity with query s, we will be ranking document in its.

Experimentals

A. Corpus

We build Vietnamese topic model based on set of documents which was labeled. We used a part of Vietnamese text classification from previous studies [11] and other documents that were downloaded from the website http://vnexpress.net, http://vietnamnet.vn, after that, these documents were labeled by human.

B. Text classification result

Using Naive Bayes combination with topic model that was built before, Vietnamese text classification result is is shown as Table 1 below.

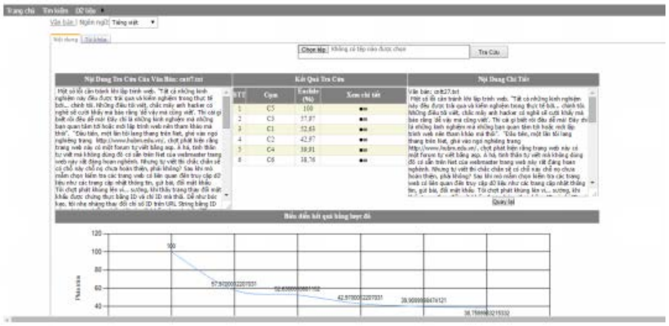
Table 1. Results of text classification

Topic	Number of documents	Traditional Method		Method applied dimensional reduction with topic model	
		Average of features	*Precision*	*Average of features*	*Precision*
Art	50	1120	86%	435	91.6%
Sport	30	835	88%	251	96%
Technology	40	456	85.4%	216	97%
Market	25	727	78%	304	93%
Finance	30	883	80.33%	378	94.8%
Land	45	954	82%	452	92%

Based on the assessment used to measure the accuracy and comparing with traditional methods show that, our methods can be reduced features dimensional effectively, the number of features when use topic model was decrease 40.9% than the tradition method on the 220 documents (6 different topic). The accuracy average of 6 subjects increased from 83% to 94.07%.

C. Building Vietnamese text Retrieval System

So based on proposed method, we built a system for testing. We used C# programming laguage and SQL database management for building.



STT	Cụm	Euclide (%)	Xem chi tiết
1	C5	100	●■
2	C3	57.97	●■
3	C1	52.63	●■
4	C2	42.97	●■
5	C4	39.91	●■
6	C6	38.76	●■

Fig 2.Vietnamese text Retrieval system – retrieval page .

The figure 3. illustrated the retrieval page and the figure 4. below is the similarity between documents and query.

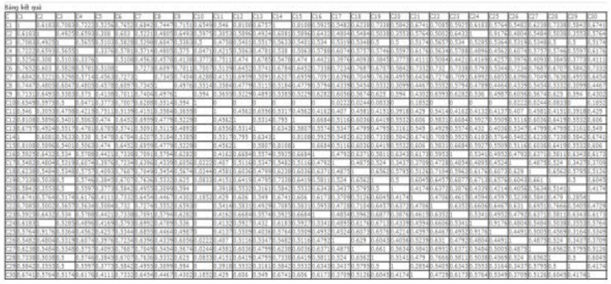

Bảng kết quả

Fig 3. Vietnamese text Retrieval system – similarity.

D. Vietnamese text Retrieval Result

After testing by the retrieval system that we built. We used Recall in evaluation Vietnamese text retrieval

$$recall = \frac{|\,relevant\ document \cap retrieved\ documents\,|}{|\,relevant\ documents\,|}$$

The table II is the result of retrieval with six topics.

Table 2. some TOPICS FOR RETRIEVING

Topics	Number of relevant documents	Recall
Market	52	0.173333
Land	54	0.18
Sport	46	0.153333
Travel	42	0.14
Technology	78	0.26
Art	36	0.11

Conclusion

Topic modeling helps text retrieval problem become clearer and easier, especially when we process in Vietnamese text or other single syllable languages like: Chinese, Japanese and Korean...

We proposed information retrieval method based on data that has been classified, so accuracy better than some previous methods. Because we optimized in some process like: data preprocessing, data classify and information retrieval.

Acknowledgment

We would like to thank people who help us in the lead up to this direction, support training sets, and their previous works. For Dr. Nguyen Le Minh, Dr. Nguyen Van Vinh, Dr. Nguyen Phuong Thai in Vietnam National University and some experts in Vietnamese language processing Community.

References

[1] Buscher, G., Dengel, A., and 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, Singapore, 387–394, 2008.

[2] Christos Faloutsos, Douglas W. Oard, "A survey of information retrieval and filtering methods" Technical Report, University of Maryland at College Park, 1995.

[3] Chirita, P.-A., Firan, C. S., and Nejdl, W., "Personalized query expansion for the web". Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Amsterdam, The Netherlands, 7–14, 2007.

[4] Dakka, W., Gravano, L., and Ipeirotis, P. G. "Answering General Time Sensitive Queries. IEEE Transactions on Knowledge and Data Engineering", 24(2): IEEE Computer Society Press. 220-235, 2012.

[5] David R.H.Miller, Tim Leek, Richard M.Schwartz, "A hidden Markov model information retrieval system", Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 214-221, 1999.

[6] Dung, T.D. and Phuc, "Develop a Search Engine supports for Information Retrieval in the field of Information Technology", HCMC University of Science, 2006.

[7] Efron, M. "Query-Specific Recency Ranking: Survival Analysis for Improved Microblog Retrieval", In Proceedings of the TAIA'12 Workshop associated to SIGIR'12. Portland, USA. August 16, 2012.

[8] N. Fuhr, 'Optimum polynomial retrieval functions based on the probability ranking principle', ACM Transactions on Information Systems, vol. 7, no. 3, pp. 183-204, 1989.

[9] Jeff Dean, Challenges in Building Large-Scale Information Retrieval Systems, Stanford, , 2014

[10] Ha Nguyen Thi Thu, Tinh Dao Thanh, Thanh Nguyen Hai, Vinh Ho Ngoc, Building Vietnamese Topic Modeling Based on Core Terms and Applying in Text Classification, Proc. Of The Fifth IEEE International Conference on Communication Systems and Network Technologies, pp: 1284 -1288, 2015, DOI 10.1109/CSNT.2015.22.

[11] M.J.Gardner, J.Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, K. Seppi, "The topic browser: An interactive tool for browsing topic models", NIPS Workshop on Challenges of Data Visualization. MIT Press, 2010.

[12] J.H. Lau, D. Newman, S. Karimi, T. Baldwin. "Best topic word selection for topic labelling", Coling 2010: Posters, pages 605–613, 2010.

[13] T.-Y. Liu, "Learning to rank for information retrieval", SIGIR Tutorials, pp. 904, 2010.

[14] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA", Information Retrieval, pp. 1–26, 2010.

[15] M.F.Moens and I. Vulic, "Monolingual and cross lingual probabilistic topic models and their application in information retrieval", Pro. Of the 35th European Conference on Information Retrieval, pp 875-878, Moscow, Russian Federation, 2013.

[16] L. A. Park, K. Ramamohanarao, "The sensitivity of latent Dirichlet allocation for information retrieval", ECML PKDD, pp. 176–188. Springer-Verlag, 2009.

[17] Salton, G. and Buckley, C. 1990. Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Sciences 41, 4, 288–297

[18] Salton, G. and McGill, M., "Introduction to Modern Information Retrieval". McGraw Hill, New York, N. Y., USA, 1983.

[19] A. Smola, S. Narayanamurthy, "An architecture for parallel topic models", Proc. VLDB Endow., 3:pp. 703–710, 2010.

[20] Shokouhi, M., Azzopardi, L., and Thomas, P., "Effective query expansion for federated search", In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Boston, MA, USA, 427–434, 2009.

[21] Ricardo Campos, Gael Dias, Alipio M. Jorge, Adam Jatowt, "Survey of Temporal Information Retrieval and Related Applications"ACM Computing Surveys, Vol. 47, No. 2, 2014.

[22] Rocchio, J. J.,"Relevance feedback in information retrieval", In The SMART Retrieval System, G. Salton, Ed. Prentice-Hall, Englewood Cliffs, N. J., USA, 313–323, 1971.

[23] Ruthven, I. , "Re-examining the potential effectiveness of interactive query expansion", In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and De-velopment in Information Retrieval. ACM Press, Toronto, Canada, 213–220, 2003

[24] Y. W. Teh, D. Newman, M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation", In NIPS. MIT Press, 2007.

[25] I.Vulic, W. De Smet, M.F.Moens, "Cross language information retrieval models based on latent topic models traned with document aligned comparable corpora Ử, Information Retrieval, vol 16, no.3, pp.331-368, Springer, 2013.

[26] X. Yi, J. Allan, "A comparative study of utilizing topic models for information retrieval", In ECIR, pp. 29–41. Springer-Verlag, 2009.

[27] Xing Yi, James Allan, "A Comparative Study of Utilizing Topic Models for Information Retrieval", Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, pp. 29 – 41, Springer-Verlag Berlin, 2009.

[28] Xing Wei, "Topic models in information Retrieval", PhD thesis, University of Massachusetts Amherst, 2007.

[29] L. Zhang, Y. Zhang, "Interactive retrieval based on faceted feedback", In SIGIR, pp 363–370. ACM, 2010.

[30] P.P.T.M. van Mun, "Text Classification in Information Retrieval using Winnow", http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.1279

[31] Wang, X., Fang, H., and Zhai, C.,"A study of methods for negative relevance feedback", In Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Singapore, Singapore, 219–226, 2008.

Cuong Nguyen Ngoc is a Lecture in Department of Computer and Mathematical, The people's Security University. He received Ph.D degree in Mathematics and Physic in 1994. He interested in Natural Language Processing (NLP), computational techniques, data analyzing and processing. Since 2006, I have been developing the new science of network security and high technology for crime investigation.