2015 Fifth International Conference on Communication Systems and Network Technologies

Building Vietnamese Topic Modeling Based on Core Terms and Applying in Text Classification

Ha Nguyen Thi Thu
Department of E-commerce
Electric Power University
Hanoi, Vietnam
hantt@epu.edu.vn

Tinh Dao Thanh
Information Technology Faculty
Le Quy Don Technical University
Hanoi, Vietnam
tinhdt@mta.edu.vn

Thanh Nguyen Hai
Vietnam Ministry of
Education and Training
Hanoi, Vietnam
thanhk37@yahoo.com

Vinh Ho Ngoc
Vinh University of
Technology Education
NgheAn, Vietnam
hnvinh.skv@moet.edu.vn

***Abstract*—in the languages, the occur of words are indicated about meaning of contents in text. Generative models for text, such as the topic model, have the potential to make important contributions to the statistical analysis of large document collections, and the development of a deeper understanding of human language learning and processing. In this paper, we proposed a novel method for building Vietnamese topic model based on core terms and conditional probability. With this approach, we reduced cost of time for building corpus. After that, we perform with Vietnamese text classification and the experimental show that, this corpus will help text classification system really effectively than traditional methods, higher accuracy and reduced complex data processing.**

Keywords—Vietnamese text, Text mining, Topic modeling, text classification, word processing.

I. INTRODUCTION

The research on text mining are attracted researcher and scholar when the number of electronic documents on the Internet increased rapidly from difference sources. Both of structured document and unstructured document are also increased too huge. Text mining can help user some task through effective tools like: retrieval, summary, classification, clustering ...

Vietnamese is a single syllable language, so that very difficult for identifying word through white spaces. When processing Vietnamese text, people often use word segmentation tool for separating words. So that, it is very hard to developing Vietnamese text mining tools on the Internet. Because it spends a lot time to process and accuracy is not high.

Because word segmentation tools sometimes is not the most powerfull for treatment with Vietnamese text (accuracy achieved about 80%). So that, it's very difficult to building Vietnamese text mining tools, because of:

- Quality of text mining system is not high
- Complex of computational problem

Figure 1. is the illustration of Vietnamese text mining system, when we building any Vietnamese text mining, we need to used at least one word processing tool like: word segmentation, Pos tagging,... Both of training phase and Test phase need word processing tool.So that, it always need a lot of time to process.

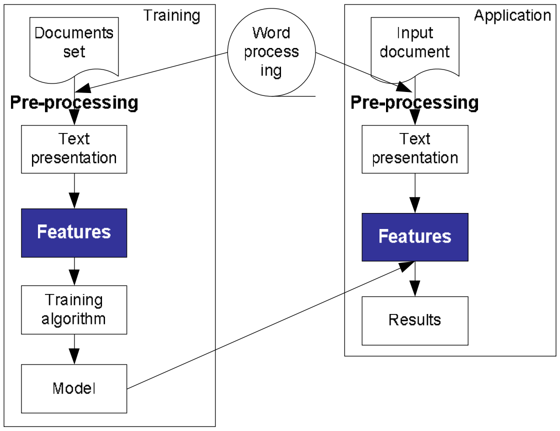


Fig. 1. Vietnamese text mining system.

Therefore, to boost speed and performance of these systems, need to optimize two main steps in Fig 1.

- Improve the quality of word processing tools
- Reduce dimension of features

Through the study of topic model for English, we found that, topic model can be the best solutions for building Vietnamese text mining tools. However, Vietnamese text now hasn't topic model, so that we must to build a new corpus. But, to build a new corpus need a lot of time and cost and experts. Therefore, in this paper, we present a new method for building topic model by identifying core terms based on document training set that were labeled by topic, after that, topic words will be calculated through conditional probability and calculated distances from its to core terms.

The rest of the paper is structured as follows: in section 2 paper will present the related works, section 3 presents methods to build models for Vietnamese topic, the 4 presents the experimental results and applications to Vietnamese text classification and evaluate the accuracy of the method that used topic model with traditional methods that only used word segmentation tool and finally is conclusion.

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.22

II. SOME RELATED WORKS

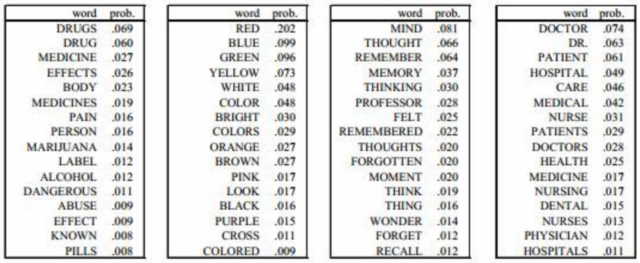
Topic models are corpus for discovering the main topic from large number of unstructured documents. The first concept of topic model was lauched in 2002 by Griffiths and Steyvers. And then, Some researchers proposed methods for building topic model, almost of these methods built corpus based on probability theory combine with latent Dirichlet allocation (LDA), Bayesian network or Hidden Markov Model (HMM).

LDA is extended by LSI concept, it is a generative models, that can give us a part of content or semantics in the documents by a subset of presentation word in its. LDA was first presented as topic model by David Blei, Andrew Ng, and Michael Jordan [3], [4], [7], [14], [25].

After then, David Blei et al described a latent Dirichlet allocation developed with WordNet by using an unsupervised probabilistic topic model includes word sense as a hidden variable. In experimentals, they two data sets: SEMCOR with many nouns labeled with the correct Wordnet and the British National Corpus [4].

Hanna M. Wallach launched the idea of model-based topic models from the bag of word, in their study, they used a probabilistic model with the combination of n-grams and latent topic variables by extending a unigram topic model to include properties of a hierarchical Dirichlet Bigram language model. They also performed experiments on two data sets, the first set was built by 150 abstracts from the Psychological Review Abstract data provided by Griffiths and Steyvers. The first set included 100 documents and used to infer, and remain of 50 documents for models evaluating. An other set included 150 documents, drawn from the 20 Newsgroups data. 100 documents in this were used for inference, while 50 were retained for evaluating predictive accuracy [24]

Mark Steyvers was introduced a probabilistic topic models and he think the probabilistic of topic model can represent semantic ambiguity through uncertainty over topics better than topic model without probabilistic [21].



word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Fig. 2. Some topic in TASA corpus with probabilistic

Mark Andrews, Gabriella Vigliocco in 2010 have described a model of semantic representations that learns from statistical language. It is based on ideas from the bag of word model and perform inference by inheritance chain of natural language data and be deduced from the hidden Markov model, developed from a bag of word from the Bayesian model [1].

TABLE I. SOME TOPICS DESCRIBED BY ANDREWS.

Theatre	Music	League	Prison	Rate	Pub	Market
Stage	Band	Cup	Years	Cent	Guinness	Stock
Arts	Rock	Season	Sentence	Inflation	Beer	Exchange
Play	Song	Team	Jail	Recession	Drink	Demand
Dance	Record	Game	Home	Recovery	Bar	Share
Opera	Pop	Match	Prisoner	Economy	Dringking	Group
Cast	Dance	Division	Serving	Cut	Alcohol	news

Ivan Vulic et al., introduced some good applications when using topic models for text mining in multilingual like: Cross lingual clustering, Cross-lingual document classification, Cross lingual semantic word similarity and Cross lingual information retrieval. And in their experimental, they selected randomly examples from cross lingual topics for retrieval tasks. Topics are discovered by BiLDA trained on Wikipedia for various language pairs: French-English (FR-EN), Dutch-English (NL-EN), Italian-English (IT-EN), and Spanish-English (ES-EN). And they observd strong intra semantic coherence as well as strong inter semantic coherence [25-29].

III. VIETNAMESE TOPIC MODEL BASED ON CORE TERMS AND CONDITIONAL PROBABILISTIC

Vietnamese hasn't got any styding for topic model before and building VietwordNet now, but it hasn't completed yet. Currently, Studies focus on word processing tools, grammar analysic 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 hunman based on the core terms and conditional probability.

Supposed, topics defined by human through task of classify sets of documents, we find a word that has the most likelihood in each subset of document that was called core term. Significant 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 distances between its approximates 1, indicate its in the difference topics.

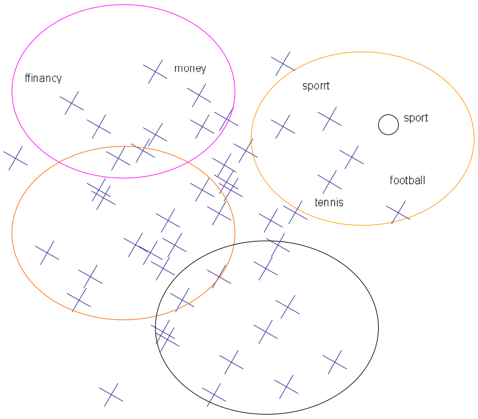


Fig. 3. Topic model based on Core terms and conditional probabilistic

In the figure 3 illustrate our method, these circles is topics. In each topic, x sign is the words that belong with topic and o sign indicates a core terms. Somehow, topics can have a part of overlapping and some words can belong two or there another topics.

Supposed that, We have A is k *–dimensional* space. Let $A = \{A_1, A_2 \ldots, A_k\}$. Ai is the topics. Each A_i include a set of words that belong its, we denote $P(w_{ij})=0$ when word w_j not appear in topic A_i and $P(w_{ij}) \neq 0$ when otherwise. If both $P(w_{ij}) \neq$ and $P(w_{mj}) \neq 0$, w_j can be in space A_i and can be also in space A_m. (A_i and A_m is intersect).

For example:

Supposed that we have 2 topics: Financy and stock and some words: market, stock, exchange, bank, money, share. Through Probability of its we can know its belong or not belong to a topic. Below is the table of the probability of words with financy topic and stock topic. And we call two topics is intersect because of they have some common words.

TABLE II. AN EXAMPLES FOR TWO INTERSECT TOPICS

	finance	Stock
Market	0.0239	0.34
Stock	0.28	0.67
Exchange	0.171	0.42
Bank	0.472	0.013
money	0.64	0.482
share	0.024	0.46

The core term in space A_i is the highest score in A_i. So, we need calculating distance from remain of words to the core terms through a degree, and we used conditional probablilistic to calculate its.

In fact, we built topic model based on core terms and conditional probability by these steps:

- Have a training set consist n documents, $D = \{d_1, d_2, \ldots, d_n\}$, and m topics $C = \{c_1, c_2, \ldots, c_m\}$.
- For each document is assigned to each topic. This step was done by human.
- Use VnTagger to separate words from D and select the nouns in it.
- Calculate the frequency of all the nouns in each topic an select the most likelihood in its and assign the core term. Finish this step, we have m core terms relative m topics in our corpus.
- Calculate the conditional probability of remain words with the core terms, and assign words to any topic that has conditional probabilities not equals to 0.

These steps is described by an algorithm in Figure 4. method to build Vietnamese topic model.

```
VIETNAMESE TOPIC MODEL
Input:
    - D: The documents set, assigned to topics
    - VnTagger: word identifying tools
    - C: Set of topics
Output:
    - T: Set of words assigned in to C
Initialization:
    V=$\phi$ ;
    n=count(S); n'=count(S');
    G=$\phi$ ;G'=$\phi$ ;
1. For each $d_i$ in $C_k$ do
1.1 $V_k \leftarrow$ Vntagger($d_i$);
2. For each $C_k$ do
2.1    If $w(j) \in V_k$ then
2.1.1    n(j)$\leftarrow$n(j) +1; // frequency of w(j)in each $C_k$
2.1.2    $N_k$=argmax$(n(j))$; // select the most likelihood of wj in each Ck
3. For each $C_k$ do
3.1    For all $w$ in $V$
3.1.1    if $Pr(w(i)|N_k)$ <>0 then $V_k \leftarrow w(i)$; // assign w(i) to $V_k$ of $C_k$
```

Fig. 4. Algorithm for building Vietnamese Topic Model.

IV. 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. Topic model

In the training process, words in documents were identified and separeted by Postagging tool (Vntagger can be downloaded at http://vlsp.vietlp.org:8080) and selected nouns and then stored in the database.

Table 2 below describes some of topic models: art, sport, technology, market, financy and land.

TABLE III. SOME TOPICS

Topics					
Art	***Sport***	***Technology***	***Market***	***Financy***	***Land***
Dân ca	Bóng đá	Lõi tứ	Giá	Cán cân	Bất động sản
Nghệ sĩ	Bóng chày	Tablet	Thực phẩm	Ngân hàng	Nhà đất
Showbiz	Cầu thủ	Điện thoại	Chứng khoán	Lãi suất	Lãi suất
Người mẫu	Thủ môn	Smartphone	Chỉ số	Tỉ lệ	Biệt thự
Ảnh	Cup	Iphone	Lương	Cắt giảm	Chung cư
Sân khấu	Tỉ số	Samsung	Người mua	Tài chính	Chủ thầu
Ca nhạc	Chelsea	Transformer	Hàng hóa	Chứng khoán	Bất động sản

C. Evaluate text categorization

To evaluate the effective of topic model in some applycations, we used topic model for Vietnamese text classification. We use Naive Bayes classifier and topic model for classification task.

We used the maximize the posterior (Maximum a posteriori-MAP)

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

In which:

- T_k is the word of the text.
- *C* is the subject
- *P (c | 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 operator formula (1) is converted to:

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

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

To evaluate text classification, we use precision and check number of features in each class of documents to compare the complexity of computing between traditional method (with word segmentation tool) and our method (use topic model).

Precision is represented by the following formula:

$$\pi = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} TP_i + FP_i} \qquad (3)$$

In which:

- TP: document that is true categories
- FP: document that is false categories

And the Average number of features in each class of documents can be calculate by

$$AF = \frac{\sum_{i=1}^{m} f_i}{N} \qquad (4)$$

In which:
f_i: number of features in the test set.
N: number of documents in the test set.

TABLE IV. RESULTS

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%.

V. CONCLUSION

Topic model applied in many of natural language processing field. Based on topic model, some text mining tools can be built and ensure that: it stabilize, effective, high accuracy and it also reduce cost of time for processing raw data. Therefore, if we use HMM or Bayesian to build Vietnamese topic model, it need a lot of time, cost and human.

In this paper, we use a different approach to building topic models, reducing the time, cost and human. It is really agreeably for Vietnamese, is one of the solutions to help solve the problem of building Vietnamese text mining tools.

With topic model, we have conducted experiments with text classification task, the experimental results showed the effectiveness of this approach, can reduce feature dimensionality more than 50% when compare with baseline method.

References

[1] Andrews, M. & Vigliocco, G. "The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation". Topics in Cognitive Science, Vol. 2, pp. 101-113, 2010.

[2] Bi Y., Bell D., Wang H., Guo G., Greer K., "Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization", MDAI, pp 127-138, 2004

[3] Blei, D et al, "Dynamic topic models". In Proceedings of the 23rd International Conference on Machine Learning, pp 113–120, 2006.

[4] Blei, D., Ng, A., and Jordan, M., "Latent Dirichlet allocation". Journal of Machine Learning Research, pp993–1022, 2003.

[5] Bao Y. and Ishii N., "Combining Multiple kNN Classifiers for Text Categorization by Reducts", LNCS 2534, pp. 340-347, 2002.

[6] Brank J., Grobelnik M., Milic-Frayling N., Mladenic D., "Interaction of Feature Selection Methods and Linear Classification Models", Proc. of the 19th International Conference on Machine Learning, Australia, 2002.

[7] David M. Blei, "Probabilistic topic model", ACM, Volume 55 Issue 4, pp 77-84, 2012.

[8] Forman, G., "An Experimental Study of Feature Selection Metrics for Text Categorization". Journal of Machine Learning Research, pp. 1289-1305, 2003.

[9] Fragoudis D., Meretakis D., Likothanassis S., "Integrating Feature and Instance Selection for Text Classification", SIGKDD '02, July 23-26, 2002.

[10] Guan J., Zhou S., "Pruning Training Corpus to Speedup Text Classification", pp. 831-840, 2002.

[11] Ha Nguyen Thi Thu ; Quynh Nguyen Huu ; Khanh Nguyen Thi Hong ; Hung Le Manh, "Optimization for Vietnamese text classification problem by reducing features set", 6th IEEE International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM) , Page(s): 209 – 212, 2012.

[12] Han X., Zu G., Ohyama W., Wakabayashi T., Kimura F., ""Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination", LNCS, Volume 3309, pp. 463-468, 2004.

[13] Kim S. B., Rim H. C., Yook D. S. and Lim H. S., "Effective Methods for Improving Naive Bayes Text Classifiers", LNAI 2417, pp. 414-423, 2002.

[14] Mark Steyvers, Tom Griffiths, "Probabilistic Topic Models", In: In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds),Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum

[15] Novovicova J., Malik A., and Pudil P., "Feature Selection Using Improved Mutual Information for Text Classification", SSPR&SPR 2004, LNCS 3138, pp. 1010–1017, 2004

[16] Purver, M., K"ording, K., Griffiths, T., and Tenenbaum, J.. Unsupervised topic modelling for multi-party spoken discourse. In ACL, 2006

[17] Qiang W., XiaoLong W., Yi G., "A Study of Semi-discrete Matrix Decomposition for LSI in Automated Text Categorization", LNCS, Volume 3248, pp. 606-615, 2005

[18] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith, P. The author-opic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pages 487–494. AUAI Press, 2004.

[19] Soucy P. and Mineau G., "Feature Selection Strategies for Text Categorization", AI 2003, LNAI 2671, pp. 505-509, 2003 .

[20] Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F., "Feature Selection Algorithms to Improve Documents Classification Performance", LNAI 2663, pp. 288-296, 2003.

[21] Steyvers and Griffiths, "Probabilistic topic models". Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006.

[22] S. Gopal, Y. Yang. Multilabel classification withmeta-level features. ACM SIGIR Conference, 2010.

[23] Teh, Y., Jordan, M., Beal, M., and Blei, D. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581, 2007.

[24] Wallach, H. Topic modeling: Beyond bag-of-words. In W. W. Cohen & A. Moore (Eds.),Proceedings of the 23rd international conference on machine learning(Vol. 148, pp. 977–984). ACM, 2006.

[25] Wang, D. Blei, and D. Heckerman. "Continuous time dynamic topic models". In UAI, 2008.

[26] Vulic, I., De Smet, W., Moens, M.-F.: "Identifying word translations from comparable corpora using latent topic models". In: Proceedings of ACL, pp. 479–484, 2011.

[27] Vulic, I., De Smet, W., Moens, M.-F.: "Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus". LNCS, vol. 7097, pp. 37–48. Springer, 2011.

[28] Vulic, I., De Smet, W., Moens, M.-F.: "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora". In: Information Retrieval, 2012

[29] Vulic, I., Moens, M.-F.: "Detecting highly confident word translations from comparable corpora without any prior knowledge". In: Proceedings of EACL, pp. 449–459, 2012.