

Word Segmentation for Vietnamese Text Categorization: An online corpus approach

Thanh V. Nguyen, Hoang K. Tran, Thanh T.T. Nguyen and Hung Nguyen

Abstract—This paper extends a novel Vietnamese segmentation approach for text categorization. Instead of using annotated training corpus or lexicon which is still lack in Vietnam, we use statistic information extracted directly from a commercial search engine and genetic algorithm to find the most reasonable way of segmentation. The extracted information is document frequency of segmented words. We conduct many thorough experiments to find out the most appropriate mutual information formula in word segmentation step. Our experiment results on segmentation and categorization obtained from online news abstracts clearly show that our approach is very optimistic. It achieves results in nearly 80% human judgment on segmentation and over 90% micro-averaging F_1 in categorization. The processing time is less than one minute per document when enough statistic information was cached.

Index Terms—Genetic Algorithm, Text Categorization, Web Corpus, Word Segmentation.

I. INTRODUCTION

It has clearly known that word segmentation is a major barrier in text categorization tasks for Asian languages such as Chinese, Japanese, Korean and Vietnamese. Although Vietnamese is written in extended Latin characters, it shares some identical characteristics with the other phonographic southeast Asian languages. Asian languages are hard in determining word boundaries, as well as have different phonetic, grammatical and semantic features from Euro-Indian languages. Thus, it is difficult in trying to fit Vietnamese into wide- and well-investigated approaches on Euro-Indian languages without acceptable Vietnamese word segmentation.

Why is identifying word boundary in Vietnamese vital for Vietnamese text categorization? According to [18] and our survey, most of top-performing text categorization methods: the Support Vector Machine ([8]), k-Nearest Neighbor ([16]), Linear Least Squares Fit ([17]), Neural Network ([15]), Naïve

Bayes ([1]), Centroid-based ([13]) all require probabilistic or statistics or weight information of word¹. By examining and evaluating these methods, we realize that word segmentation is the very first and important step on Vietnamese text categorization.

And what Vietnamese characteristics make identifying word boundary be a difficult task? The element unit of Vietnamese is the syllable (“tiếng”), not the word (“từ”). Some unanimous points of the definition of Vietnamese words ([5]) are:

- They must be integral in several respects of form, meaning and be independent in respect of syntax.
- They are structured from “tiếng” (Vietnamese syllable)
- They consist of simple words (1-tiếng, monosyllable) and complex words (n-tiếng, $n < 5$, polysyllable), e.g. reduplicative and compound words.

On the other hand, in English, “Word is a group of letters having meaning separated by spaces in the sentence” (Webster Dictionary). Thus, we summarize some main different features between Vietnamese and English that make Vietnamese word segmentation be a difficult and challenging task.

Characteristic	Vietnamese	English
Basic Unit	Syllable	Word
Prefix or Suffix	No	Yes
Part of Speech	Not Unanimous	Well-Defined
Word Boundary	Context meaningful combination of syllable	Blank or Delimiters

Table 1. Summary of main differences between English and Vietnamese.

And what is the biggest obstacle for Vietnamese text categorization? Currently, there is not a standard lexicon or well balanced, large enough annotated Vietnamese text training corpus. Due to Vietnamese characteristics, building such lexicon and corpus requires much time and cost. We affirm that this is the most concerned problem for any works on Vietnamese text categorization, natural language processing or information retrieval.

In this paper, we focus on how to segment Vietnamese text in some *acceptable* ways *without* relying on any *lexicon* or *annotated training corpus* for text categorization tasks. Remarking to the problem of how to find the most satisfied way to segment words in a sentence, we apply Genetic Algorithm to evolve a population in which each individual is a particular way of segmenting. Statistics information for the fitness function is the document frequency of the segmented words extracted directly from Internet by a search engine.

The organization of this paper is as follows. After this

Manuscript received December 4, 2005. This work was supported in part by the University of Natural Sciences 2004-2005 Research Grant for young lecturer.

Thanh V. Nguyen is a lecturer of the Faculty of Information Technology, University of Natural Sciences, HoChiMinh, Vietnam. He is now a graduate student at the Department of Computer Science & Engineering, University of Minnesota at Twin Cities, MN 55455 USA (phone: (1) 651-399-9557; e-mail: thnguyen@cs.umn.edu).

Hoang K. Tran and Thanh T.T. Nguyen are seniors at the Faculty of Information Technology, University of Natural Sciences, HoChiMinh, Vietnam. (e-mail: {azury_thanh, trankhaihoang}@yahoo.com).

Hung Nguyen is a lecturer of Vietnam National University, HoChiMinh, Vietnam (e-mail: hung64vn@yahoo.com).

¹ Sometimes called “word stem” or “term”

introduction, we will look back to state of the art of Chinese and Vietnamese word segmentation. Section 3 expresses our principle of internet-based statistic. In the next section, we describe in detail our genetic algorithm approach. Section 5 shows some experimental results and discussions. Finally, we conclude and provide directions for future research.

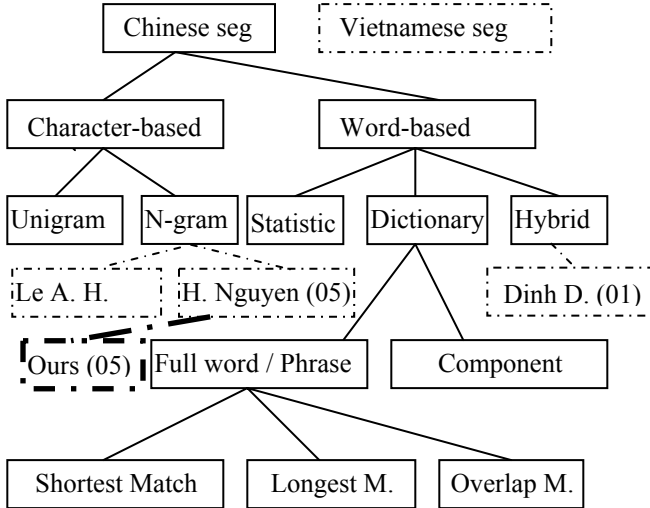


Figure 1. Basic Approaches of Chinese Segmentation and Current Works of Vietnamese Segmentation.

II. RELATED WORKS

In this section, we examine some significant prior work on Vietnamese segmentation and also try to categorize them based on the art of Chinese segmentation ([7]).

Word-based approaches, with three main categories: statistics-based, dictionary-based and hybrid, try to extract complete words from sentences. *Statistics-based* approaches must rely on statistical information such as term, word, character frequencies or co-occurrences in a set of preliminary data. Therefore, its effectiveness is significantly depended on a particular training corpus. In *dictionary-based* approaches, segmented texts must be matched with the ones in dictionary. Unfortunately, it is unfeasible to build a complete Vietnamese dictionary or a well-balanced, large enough training corpus as we stated above. *Hybrid* approaches try to apply different ways to take their advantages. Dinh et al ([6]) have built their own training corpus (about 10MB) based on Internet resources, news and e-books. Of course, they are small and not well-balanced corpus. To sum up, we argue that word-based approaches are not suitable for Vietnamese text categorization until we have a good lexicon and/or a large and trusted training corpus.

Character-based approaches (syllable-based in Vietnamese case) purely extract certain number of characters (syllable). It can further be classified into single-based (unigram) or multi-based (n-gram) approaches. Although they are simple and straightforward, many significant results in Chinese are reported ([7]). Some recent publications for Vietnamese segmentation also follow this one. Le ([9]) has built a 10 MB raw corpus and used dynamic programming to maximize the sum of the probability of chunks (phrases separated by delimiters). In a recent publication of H. Nguyen

et al ([12]), instead of using any raw corpus, he extracted the statistic information directly from Internet and used genetic algorithm to find most optimal ways of segmenting the text. Although his work is still preliminary and lack of thorough experiments, we believe that this novel approach is promising. Our work will extend this idea, give significant changes and make some considerate experimental evaluations to find the best mutual information formula, the key point of this approach.

III. PRINCIPLE OF INTERNET-BASED STATISTIC

We agree with H. Nguyen et al ([12]) that through commercial search engines, we can extract useful statistic information from Internet. This is the *document frequency* (df), the number of indexed documents containing this word. To approximate the probability of a word randomly occurred on the Internet, we normalize the df value by dividing it by a MAX value, which is the number of indexed Vietnamese documents.

$$p(w) = \frac{df(w)}{MAX}$$

As we do not know exactly how many Vietnamese documents have been indexed, by testing some common words, we choose MAX to be $1 * 10^9$.

Vietnamese	English	df
có	has / have	$21.3 * 10^6$
của	of	$20.4 * 10^6$
một	one	$14.4 * 10^6$

Table 2. Document frequencies of some common Vietnamese words.

Vietnamese word contains consecutive syllables, thus, we need a statistic measure of syllable associations. *Mutual information* (MI), an important concept of information theory, has been used in natural language processing to capture the relationship between two specific words x, y ([3]):

$$MI(x, y) = \log \frac{p(x, y)}{p(x) \times p(y)}$$

However, we not only look at pairs of syllables, bigrams, but also consider n-grams as well. Many formulas were introduced to measure the relationship of n consecutive syllables. However, it is difficult to find the most appropriate formula for our task. So, we will experiment on three approaches given by [2], [12] and us.

Chien et al ([12]) suggests calculate the mutual information for Chinese n-gram as follow:

$$MI(cw) = \frac{p(cw)}{p(lw) + p(rw) - p(cw)} = MI1$$

where cw is composed of n single syllables ($cw = s_1 s_2 \dots s_n$), lw and rw are the two longest composed substrings of cw with the length $n-1$, i.e., $lw = s_1 s_2 \dots s_{n-1}$ and $rw = s_2 s_3 \dots s_n$. Basically, if $MI(cw)$ is large, lw and rw seem to occur together on Internet, i.e., cw is likely a compound word.

Another formula introduced by [12] specifies for Vietnamese:

$$MI(cw) = \frac{p(cw)}{\sum_{i=1}^n p(s_i) - p(cw)} = MI2$$

However, we argue that both above formulas have some drawbacks. Most of Vietnamese 4-grams are actually the combination of two 2-syllable words, for example “hội nghị khoa học”. So, instead of comparing a 4-syllable word with two sub 3-grams, we should consider it with two sub distinct 2-grams. Meanwhile, the latter favors words having one or two syllables since the more syllables a word contains, the larger denominator it gets. Consequently, it has a low MI. With above intuition, we represent a new way to calculate the mutual information of Vietnamese n-grams:

$$MI(cw) = \frac{p(cw)}{p(lw)+p(rw)-p(cw)} = MI3$$

where lw and rw are the two composed substrings of cw with the length $\lceil n/2 \rceil$. We can easily find that our formula is similar to the one given by Chien et al ([2]) for 2-grams and 3-grams but different from words having four, or more, syllables.

In the next section, we will introduce genetic algorithm approach to find the global optimal MI for a given text, i.e. the most acceptable segmentation for this text.

IV. GENETIC ALGORITHM APPROACH

The search space of word segmentation is very large since there are many ways to combine syllables into words. Base on the principle of evolution and heredity, Genetic Algorithm (GA) has long been known for its ability to traverse very large search space efficiently and find approximate global optimal solutions instead of local optimal solutions ([10]). GA will evolve a number of generations. For each generation, we will select top N best quality individuals after performing cross-over, mutation and reproduction. The quality of an individual will be calculated by a fitness function.

Goal. Let the given text t be composed of n syllables: $t=s_1s_2\dots s_n$. The goal of this GA process is to find most acceptable ways to segment t to m segments: $t=w_1w_2\dots w_m$ which $w_k=s_{i_1}\dots s_{i_j}$ ($1 \leq k \leq m$, $1 \leq i,j \leq n$) can be either a simple or complex word.

Representation. The population (pop) is represented as a set of individuals (id) which are strings of 0s and 1s bit. Each bit is corresponding to a syllable. So a word will be a meaningful consecutive string of bits. For example:

0	0	1	0	0
Học sinh # học # sinh học (pupil study biology)				
w_1		w_2		w_3

Initialization. In this step, we must set several parameters for the GA such as number of generations, population size, cross-over fraction, mutation fraction and reproduction fraction. We also have to randomly build an initial population, randomizing a 0s and 1s string. However, we make some restrictions on the random string for optimization. Table 3 a statistic derived from an online usual dictionary² containing 72994 words and phrases.

Through this statistic, we see that there are over 67% of the words containing two syllables and about 30% consisting of one, three or four syllables. Longer words, many of which are idiomatic expressions, are about 3%. These lead us to define

some restrictions on the initial random string. First, we limit the length of each segment w_k with four. Second, when randomizing, we set a bias ratio to generate more segments having the length 2 than the others. Besides, we also apply the simple form of the Left Right Maximum Matching algorithm ([14]) to build two specific individuals, forward / backward ones. Consequently, the initial population will have some local optimal individuals.

Word length	Frequency	Percentage
1	8933	12.2
2	48995	67.1
3	5727	7.9
4	7040	9.7
≥ 5	2301	3.1
Total	72994	100

Table 3. Statistics of word lengths in a dictionary.

Cross-over. We apply the standard one-point cross operation on bit strings. For a couple of two individuals id_1 id_2 , the two new offsprings are combined the beginning of id_1 with the ending of id_2 and vice-versa. However, if a child individual breaks the above restriction, each segment w_k cannot have the length greater than four(4), we will normalize them by flipping over all exceeding bits at the end of this segment.

Mutation. Instead of using random inversion mutation, we invert only boundary bits of a segment. Like the cross-over, we apply the normalization to ensure the mutative individual satisfying the restriction.

Reproduction. After performing cross-over and mutation, we will mingle a proportion of the parent individuals into child individuals for the selection step of next generation.

Selection. For each generation, we only select $top N$ individuals from child and reproduction parent candidates for the next generation. The selection is based on the following fitness function

$$fit(id) = fit(w_1w_2\dots w_m) = \sum_{k=1}^m MI(w_k)$$

$$fit(pop) = \sum_{i=1}^N fit(id_i)$$

where $id=w_1w_2\dots w_m$ is a particular individual of the population, $pop = \{id_1, \dots, id_N\}$

Convergence. The GA process tries to improve the fitness of the individual i.e. the quality of word segmentation. Thus, we will stop the GA process when the fitness value of the next generation is convergent or the number of generations reaches a pre-defined maximum.

Example: “Nhà nước # xây # cao ốc # thương mại.”

(The government builds commercial buildings.)

Initialization: $id_1 = 0110101$ $fit(id_1) = 0.020$

$id_2 = 0011011$ $fit(id_2) = 0.699$

Cross-over:

$$id_1 = 011 \times 0101 \rightarrow id_1 = 011\mathbf{1011} \quad fit(id_1) = 0.464$$

$$id_2 = \mathbf{001} \times \mathbf{1011} \rightarrow id_2 = \mathbf{0010101} \quad fit(id_2) = 0.255$$

Mutation:

$$id_2 = 0010\mathbf{101} \rightarrow id_2 = 0010\mathbf{011} \quad fit(id_2) = 0.704$$

(Nhà nước # xây # cao ốc # thương mại.) (convergence)

² <http://dict.vietfun.com>

Word	df	p	MI3
nhà	2180000	2.18E-03	2.18E-03
nước	1840000	1.84E-03	1.84E-03
nhà nước	771000	7.71E-04	2.37E-01
nước xây	9360	9.36E-06	2.38E-03
xây	2100000	2.10E-03	2.10E-03
xây cao	287	2.87E-07	2.13E-05
cao	11400000	1.14E-02	1.14E-02
cao ốc	35300	3.53E-05	3.04E-03
ốc	239000	2.39E-04	2.39E-04
ốc thương	277	2.77E-07	1.12E-04
thương	2230000	2.23E-03	2.23E-03
thương mại	1260000	1.26E-03	4.62E-01
mại	1760000	1.76E-03	1.76E-03
nước xây cao	0	0.0E+00	0.0E+00

Table 4. Statistics of n-grams in “Nhà nước xây cao ốc thương mại”.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Evaluating the accuracy of Vietnamese word segmentation is very problematic, especially without a manual segmentation test corpus. Therefore, we perform two experiments, one is done by human judgment for word segmentation result, and the other is a text categorization evaluation based on our word segmentation approach.

Due to the fact that our approach use internet-based statistic, we harvest news abstracts from many online newspapers³ to build a corpus for testing purpose. Thus, somehow our data is balanced in styles and genres. Moreover, for the text categorization experiment, we automatically classify these abstracts into two levels of topics based on the current categorization of news websites (Table 5).

Level 1	Level 2
Science	
Society	Education, Study abroad , Life style, Travel
Business	Estate, Stock, Foreign trade
Sport	Football ,Tennis
Culture	Fashion, Cinema
Health	Making up, Sex

Table 5. Two levels of topics of testing corpus.

Since each online newspaper has its own topic categorization, we choose the most common topics from these websites. In summary, we collect a 10MB testing corpus containing 1400 abstracts and 82,219 syllables, 100 documents for each sub topic.

For our experiments, we set genetic parameters as follows:

- Generation limit = 100
- Population size = 100
- Cross-over fraction = 0.8
- Mutation fraction = 0.1
- Reproduction fraction = 1
- Top N selection = 100

A. Word Segmentation Experiment

In this experiment, we ask two native, one is a linguistic professor and the other is a computer science graduate student, who usually reads online news. These people will examine our segmentation results and answer two questions:

- Whether or not he absolutely agrees with the segmentation result. (This question is used for calculating *perfect* segmentation).
- Whether or not the segmentation result makes the reader understand the meaning correctly. (This question is used for calculating *acceptable* segmentation).

We argue that, for text categorization task, we just need *acceptable* ways of segmentation, i.e. the *important* words are segmented correctly while *less important* words may be segmented incorrectly. Table 6 represents the human judgment for our word segmentation approach.

Judgment	MI	Perfect	Acceptable
Linguist Professor	MI1	703 (50.18%)	1076 (76.86%)
	MI2	736 (52.57%)	1074 (76.72%)
	MI3	787 (56.24%)	1132 (80.86%)
Graduate Student	MI1	759 (54.19%)	1088 (77.71%)
	MI2	747 (53.36%)	1063 (75.94%)
	MI3	862 (61.57%)	1175 (83.91%)

Table 6. Human judgment for word segmentation experiment.

Our experiment shows that there is no significant difference in word segmentation result among three MI formulas. However, our proposed MI gets the best performance. This is not a surprising result. We believe that the our MI3 formula overcomes the drawbacks of the other MI formula in evaluating words having four or more syllables while reserving the existing evaluation for 2,3-grams.

Overall, the perfect segmentation percentage seems to be low as we expect. Moreover, there is considerable difference in the agreement of how a sentence is segmented perfectly among judge. The reason is that part-of-speech system of Vietnamese is not well-defined. This causes the inhomogeneous phenomenon in judgment word segmentation.

However, the acceptable segmentation percentage is satisfactory. Nearly eighty percent of word segmentation outcome does not make the readers misunderstand the meaning. This is exactly what we expected. Without training corpus, our approach achieves a considerable Vietnamese segmentation result. Therefore, we continually make a preliminary text categorization experiment to examine further our approach. We only use MI3 formula in word segmentation step for the next experiment.

B. Text Categorization Experiment

As we stated above, there are many approaches performing text categorization task. Nevertheless, the best performance approach for English may not be the best one for Vietnamese. To find the most appropriate text categorization approach for Vietnamese with our suggested word segmentation, we need many aggressive experiments with large data. We leave this task for future works.

In this part, we perform a simple text categorization experiment for testing our segmentation approach based on

³ <http://www.vnexpress.net>, <http://www.vnn.vn>, <http://www.tuoiitre.com.vn>, <http://www.thanhnien.com.vn>

the idea of Naïve Bayes approach ([1]). The testing corpus consists of a set of documents, $D=\{d_1, d_2, \dots, d_n\}$, where each document will be labeled with a unique category from a set of classes $C=\{c_1, c_2, \dots, c_m\}$. For each document d , we apply some pre-processing steps to speed up. First, we split d into many groups of syllables based on the delimiters and numbers. Second, using a stop word list, we remove common and less informative words based on a stop word list. Performing word segmentation task on d we get a segmented document. Finally, d will be represented as follows: $d = g_1g_2\dots g_r$ where g_i is a group of syllables, a word, after segmentation.

Naïve Bayes approach makes an assumption that the words $g_1:g_r$ are all conditionally independent of one another, given document d . We cite the following formula from [11]:

$$P(Y = c_k | g_1g_2\dots g_n) = \frac{P(Y = c_k) \prod_i P(g_i | Y = c_k)}{\sum_j P(Y = c_j) \prod_i P(g_i | Y = c_j)}$$

where c_k is the k^{th} topic and $d=g_1g_2\dots g_r$ is the document we want to categorize.

Nevertheless, given a topic c_k , we can not calculate the conditional probability $P(g_i | Y=c_k)$ that a word g_i belongs to that category c_k since we do not have a training corpus. So we have to utilize it approximately using the information from the search engine as follows:

$$\begin{aligned} P(X = g_i | Y = c_j) &= \frac{\#D\{X = g_i \wedge Y = c_j\}}{\#D\{Y = c_j\}} \\ &\approx \frac{p(g_i \& c_j) + 1}{\sum_k p(g_i \& c_k) + \|Y\|} \end{aligned}$$

where $\#D\{x\}$ operator returns the number of elements in the set D that satisfy property x while $p(g_i \& c_k)$ is calculated as described in Section 3. Moreover, we have to smooth the probability to avoid zero estimation by adding some additional number to the numerator and denominator.

$$P(Y = c_i) = \frac{\#D\{Y = c_i\}}{\|Y\|} \approx \frac{p(c_i)}{\sum_j p(c_j)}$$

With these modified formula, we now can calculate the probability $P(Y=c_k | g_1g_2\dots g_r)$ that a given document $d = g_1g_2\dots g_r$ belongs to a category c_k using the *document frequency* information returned from a commercial search engine. Since we are only interested in the most probable category for a document, we use the Naïve Bayes classification rule:

$$Y \leftarrow \arg \max_{c_k} \frac{P(Y = c_k) \prod_i P(g_i | Y = c_k)}{\sum_j P(Y = c_j) \prod_i P(g_i | Y = c_j)}$$

Our experiment assumption is that each document has and only has one category. We will use F_1 and *micro-averaging* F_1 measure described in [16] to evaluate performance.

Table 7 shows the results on our testing corpus for all level-1 topics and their micro-averaging. We compare our approach result with IGATEC introduced by [12].

Category	Ours	IGATEC
Science	95.2	90.9
Society	82.9	78.1
Business	88.5	87.4
Sport	93.7	87.6
Culture	96.4	96.0
Health	89.5	90.3
Micro-avg	91.03	88.38

Table 7. F_1 and micro-averaging F_1 performance of our approach and IGATEC for level-1 topics.

The experiment shows that our approach slightly outperforms IGATEC. Moreover, we claim that applying above pre-processing steps can help GA process reduce number of generation significantly. In practice, we realize that our GA iteration mean is just 52.3, comparing with the 500 iterations of IGATEC GA Engine. This, together with our less computational MI , makes our text categorization time is less than one minute per document on a normal personal computer⁴ when statistic information was cached.

During our experiments, we find that many documents may be categorized into more than one topic. To visualize this phenomenon, instead of choosing the highest probability topic for each document, we use relative gray scale discrimination (Figure 2). The higher the probability is, the darker it will get. For many topics like science, tennis, football, music etc ..., we get a very good result. Meanwhile, for some topics like sex or life style, the accuracy is low. Investigating further, we realize that our segmentation is not appropriate for these topics since these topics have less representative words using on Internet. A focus experiment is currently carried out on these topics.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we suggest to use a less computational but meaningful mutual information and some efficient pre-processing steps to segment and categorize Vietnamese text. The novel of this approach is that instead of using annotated training corpus or lexicon which is lack in Vietnamese, it uses statistic information extracted directly from a commercial search engine and genetic algorithm to find most reasonable ways of segmentation.

Through experiments, we show that our approach can get considerable result both in text segmentation and categorization with the micro-averaging F_1 over 90 percent. To sum up, we believe this is a potential approach for such languages like Vietnamese, lack of standard lexicon or annotated corpus. Moreover, we believe that our segmentation approach can benefit for many other computer science problems like natural language processing and information retrieval of Vietnamese. We will aggressively investigate this approach in following tasks.

⁴ Pentium IV, 1.50GHz, 256 MB RDRAM

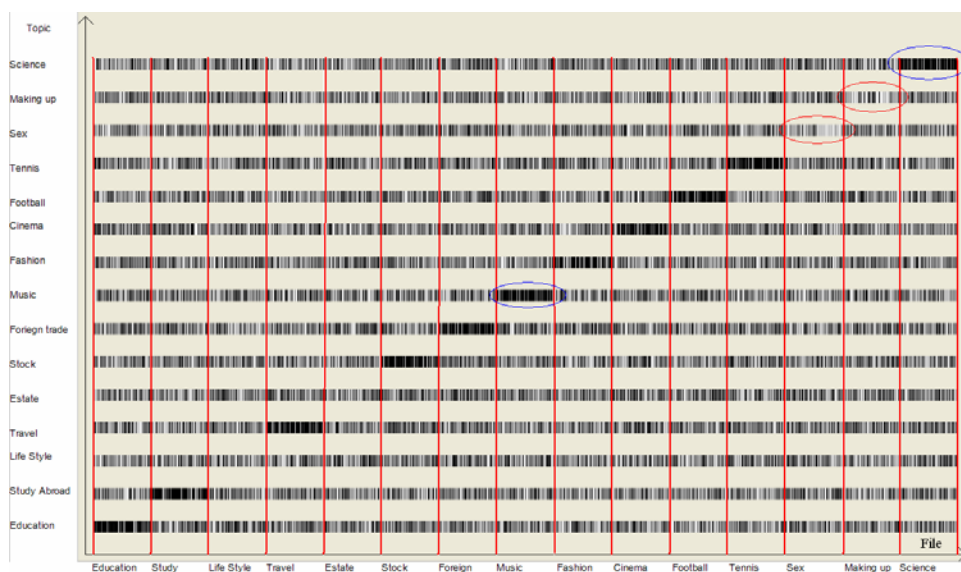


Figure 2. Gray scale discrimination for categorizing level-2 topics. The blue circle indicates the topic with high accuracy categorization while the red circle shows the topics with less performance

First, in a genetic algorithm, parameter tuning has an important role. In our approach, a text is segmented into groups of syllables with various lengths. We should build an auto parameter tuning scheme based on text length not a rigid one. This will speed up the processing time a lot.

Second, at this time, we only use the raw document frequency from the search engine. A recent publication ([4]) introduced many interesting distance measures and methods to extract meaning of words and phrases from internet using Google page counts. It may be helpful for our approach.

Finally, our long-term goal is applying and evaluating well and wide-studied text categorization approaches to find the most suitable one for Vietnamese text categorization.

ACKNOWLEDGMENT

We would like to thank Mr. Nguyen Duc Hoang Ha, lecturer at University of Natural Sciences, Vietnam National University for providing his IGATEC and valuable discussions. We would also like to thank Professor Nguyen Duc Dan at University of Social Sciences and Humanities, Vietnam National University and Mr. Tran Doan Thanh, graduate student at Kookmin University for their enthusiastic evaluation.

REFERENCES

- [1] L. D. Baker, A. K. McCallum. 1998. *Distributional clustering of words for text categorization*. Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR'98): 96-103.
- [2] Lee-Feng Chien, T. I. Huang, M. C. Chen. 1997. *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval*. Proceedings of 1997 ACM SIGIR Conference, Philadelphia, USA, 50-58.
- [3] K. Church, P. Hanks, W. Gale, and D. Hindle. 1991. *Using Statistics in Lexical Analysis*. U. Zernik Lexical Acquisition: Using On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates.

- [4] Rudi Cilibrasi, Paul Vitanyi, 2005. *Automatic meaning discovery of Google*. A search for meaning, New Scientist, 29 January 2005, p.21, by Duncan Graham-Rowe.
- [5] Dinh Dien. 2000. *Từ tiếng Việt (Vietnamese words)*. Vietnam National University, HCMC, Vietnam.
- [6] Dinh Dien, Hoang Kiem, Nguyen Van Toan. 2001. *Vietnamese Word Segmentation*. pp. 749-756. The Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.
- [7] Foo S., Li H. 2004. *Chinese Word Segmentation and Its Effect on Information Retrieval*. Information Processing & Management: An International Journal, 40(1):161-190.
- [8] T. Joachims, 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. European Conferences on Machine Learning (ECML'98)
- [9] Le An Ha. 2003. *A method for word segmentation in Vietnamese*. Proceedings of Corpus Linguistics 2003, Lancaster, UK.
- [10] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs*, 3rd edition, Springer-Verlag London, UK, 1996
- [11] Tom Mitchell. 2005. *Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression*, Machine Learning (draft version, 9/2005).
- [12] H. Nguyen, H. Nguyen, T. Vu, N. Tran, K. Hoang. 2005. *Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese*. Research, Innovation and Vision of the Future, the 3rd International Conference in Computer Science, (RIVF 2005), Can Tho, Vietnam.
- [13] S. Shankar, G. Karypis, 2000. *Weight adjustment schemes for a centroid-based classifier*, Text Mining Workshop on Knowledge Discovery in Data (KDD'00).
- [14] Chih-Hao Tsai, 2000. *MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm*. Web publication at <http://technology.chtsai.org/mmseg/>
- [15] E. Wiener, J.O. Pedersen, A.S. Weigend, *A neural network approach to topic spotting*. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95).
- [16] Yiming Yang, 1999. *An evaluation of Statistical Approaches to Text Categorization*. Journal of Information Retrieval, Vol 1, No. 1/2, pp 67-88.
- [17] Yiming Yang, C.G. Chute. 1994. *An example-based mapping method for text categorization and retrieval*. ACM Transaction on Information System (TOIS'94): 252-277. \
- [18] Yiming Yang, Xin Liu 1999. *A re-examination for text categorization methods*. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99).