Word Segmentation for Vietnamese Text Categorization

An Internet-based Statistic and Genetic Algorithm Approach

Hung Nguyen Thanh[1], Khanh Bui Doan[2]

[1] High School for the Gifted, VNU-HCM
hung64@yahoo.com

[2] University of Paris 6
vhquan@yahoo.com

Résumé

Ce papier présente une nouvelle approche de la segmentation du vietnamien pour la catégorisation de texte. Au lieu d'utiliser des corpus d'entraînement annotés ou des lexiques (qui font défaut pour le vietnamien) nous utilisons des informations statistiques extraites directement d'un moteur de recherche commercial et des algorithmes génétiques pour trouver les segmentations les plus probables. Les informations extraites incluent la fréquence des documents et l'information mutuelle des n-grams. Nos résultats expérimentaux obtenus sur la segmentation et la catégorisation de résumés de nouvelles montrent que notre approche est très prometteuse. Elle offre des résultats semblables à 80 % avec le jugement humain sur la segmentation et à 90 % en catégorisation. Le temps de traitement est inférieur à une seconde par document quand l'information statistique est maintenue en cache.

Mots-clés : catégorisation de texte, segmentation de texte, algorithmes génétiques.

Abstract

This paper suggests a novel Vietnamese segmentation approach for text categorization. Instead of using an annotated training corpus or a lexicon which are still lacking in Vietnamese, we use both statistical information extracted directly from a commercial search engine and a genetic algorithm to find the optimal routes to segmentation. The extracted information includes document frequency and n-gram mutual information. Our experiment results obtained on the segmentation and categorization of online news abstracts are very promising. It matches near 80 % human judgment on segmentation and over 90 % micro-averaging F_1 in categorization. The processing time is less than one second per document when statistical information is cached.

Keywords: text categorization, text segmentation, genetics algorithms.

1. 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. It is hard to determine word boundaries in these languages which have different phonetic, grammatical and semantic features from Euro-Indian languages. Thus, it is difficult to make Vietnamese fit into wide-

and well-investigated approaches on Euro-Indian languages without acceptable Vietnamese word segmentation.

So why identifying word boundary in Vietnamese is vital for Vietnamese text categorization? According to Yang and Xiu (1999) and our survey, most of top-performing text categorization methods require probabilistic or statistics or weight information on the words: the Support Vector Machine (Joachims, 1998), Linear Least Squares Fit (Yang and Chute, 1994), Naïve Bayes (Baker and Mccallum, 1998), Centroid-based (Shankar and Karypis, 2000), etc. To examine and evaluate these methods on Vietnamese text categorization, we realize that word segmentation will be the first and important step.

And what Vietnamese characteristics make identifying word boundary is so difficult? We will learn about Vietnamese linguistics through some general instructions. The element unit of Vietnamese is the syllable ("tiếng"), not the word ("từ"). Following is some unanimous points of the definition of Vietnamese words (Dinh Dien, 2000):

- They must be integral in respects of form, meaning and be independent in respect of syntax.
- They are structured from "tiếng" (Vietnamese syllable).
- They consists of simple words (1-tiếng, mono-syllable) and complex words (n-tiếng, n<5, poly-syllable), *e.g.* reduplicative and compound words.

Comparing with English word definition, "Word is a group of letters having meaning separated by spaces in the sentence" (Webster Dictionary), we summarize some main different characteristics between Vietnamese and English. All of these features 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? Recently, many encouraged results in text categorization for Chinese and other southeast Asian languages have been published. Trying to apply these corpus-based approaches, however, into Vietnamese is sometimes impossible. 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 various ways to segment words in a sentence and the problem of how to find most satisfied ways, we apply Genetic Algorithm to evolve a population in which each individual is a particular way of segmenting. The fitness function will be the statistics information extracted directly from Internet using a commercial search engine. The extracted information includes document frequency and n-gram mutual information.

The organization of this paper is as follows. After this 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. In the final section, we conclude and provide directions for future research.

2. Related works

Following is a review by Foo and Li (2004) of Chinese segmentation and my statistic on Vietnamese segmentation (figure 1).

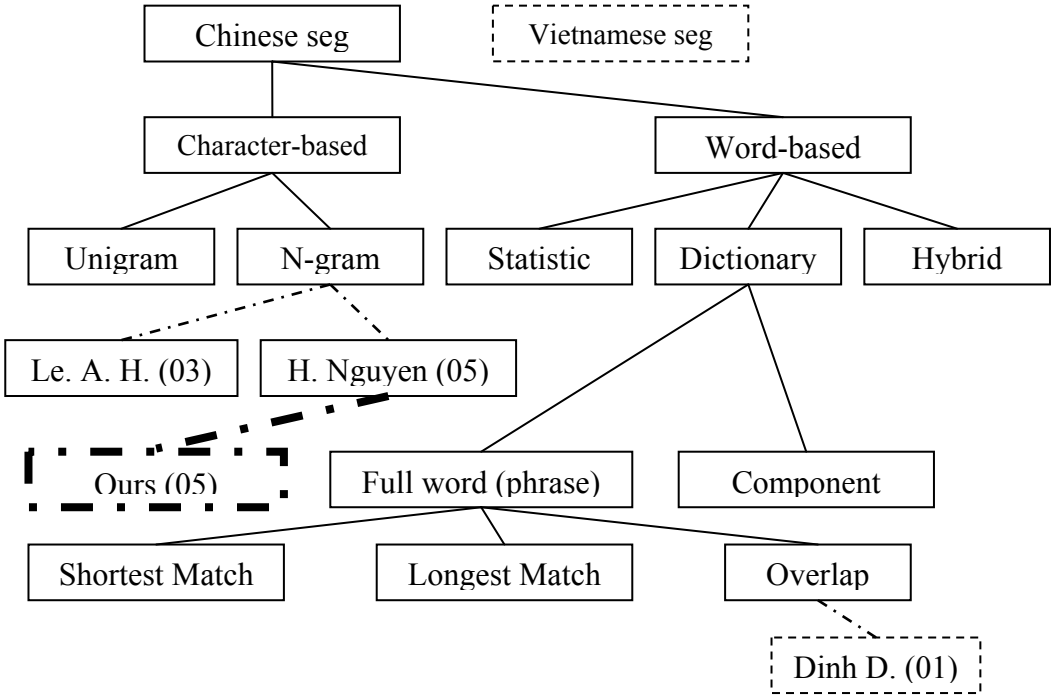

Figure1. Basic Approaches of Chinese Segmentation and Current Approaches of Vietnamese Segmentation

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 or character frequencies and co-occurrences in a set of preliminary data. So its effectiveness is significantly dependent on a particular training corpus. Unfortunately, it is a big problem for Vietnamese word segmentation as we state above. Dinh *et al.* (2001) build their own training corpus (about 10 MB) based on Internet resources, news and e-books. Of course, it is a small and not well-balanced corpus. In *dictionary-based* approaches, segmented texts are matched against a dictionary. Again, it is unfeasible to build a complete Vietnamese dictionary that contains all Vietnamese words and phrases. *Hybrid* approaches try to apply different ways to take their advantages. 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 validated 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 (uni-gram) or multi-based (n-gram) approaches. Although they are simple and straightforward, many

significant results in Chinese are reported (Foo and Li, 2004). Some recent publications for Vietnamese segmentation also follow this one. Le (2003) builds a 10 MB raw corpus and uses dynamic programming to maximize the sum of the probability of chunks (phrases separated by delimiters). In a recent publication of H. Nguyen *et al.* (2005), instead of using any raw corpus, he extracted the statistic information directly from Internet and use 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 experimental evaluations.

3. Principle of Internet-based Statistics

We agree with H. Nguyen *et al.* (2005) that through commercial search engines, we can extract useful statistic information from Internet. This is the *document frequency* (*df*), the number of indexed documents which contain this word. We normalize the d*f* value by dividing it to a *MAX* value, the number of indexed Vietnamese document, to approximate the probability of a word occurrence on the Internet.

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

Because we can not know exactly how many Vietnamese documents have been indexed, by testing some common word *df*, 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

Because Vietnamese word contains consecutive syllables, 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* (Church et al 1991). However, we not only look at pairs of syllables, bigrams, but consider n-grams as well. Like Chien et al (1997), we extend mutual information for bigram to n-gram with some changes:

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

where *cw* is composed of *n* single syllables ($cw=s_1s_2\ldots s_n$), *lw* and *rw* are the two longest composed substrings of *cw* with the length *n*-1, *i.e.*, $lw=s_1s_2\ldots s_{n-1}$ and $rw=s_2s_3\ldots s_n$. Basically, if *MI*(*cw*) is large, *lw* and *rw* seem to occur together on Internet, *i.e.*, *cw* is likely a compound word. We will clarify this with the following illustration.

Given a Vietnamese phrase, “đại học khoa học tự nhiên” (“đại học # khoa học tự nhiên”, “university of natural sciences”), we will compare whether “khoa học tự nhiên” or “học khoa học tự” is more likely a compound word. We found that “khoa học tự nhiên” (natural sciences) has MI exceed significantly “học khoa học tự” (no meaning).

Syllables	*Wf*	*MI*
khoa học tự nhiên	39,200	0.92
khoa học tự	41,800	
học tự nhiên	39,900	
học khoa học tự	14,900	0.27
học khoa học	28,600	
khoa học tự	41,800	

Table 3. Examples of n-gram mutual information

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

4. Genetic Algorithm Approach for Word Segmentation

For each text, we want to find the most reasonable ways of segmentation. However, the search space will be 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 spaces efficiently and find approximate global optimal solutions instead of local optimal solutions. GA will evolve a number of generations for a population including many individuals toward global optimization through the use of selection, cross-over, mutation and reproduction operators. The quality of an individual will be calculated by a fitness function and for each generation, we will select top N best quality individuals after performing cross-over, mutation and reproduction. A sketch of our GA approach for Vietnamese word segmentation is presented below:

Goal. Let the given text t be composed of n syllables: $t=s_1s_2...s_n$. The goal of this GA process is to find most acceptable ways to segment t to m segments: $t=w_1w_2...w_m$ which $w_k=s_i...s_j$ ($1\leq k\leq m$, $1\leq i,j\leq n$) 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:

học sinh học sinh học 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. Here is a statistic derived from an online usual dictionary containing 72,994 words and phrases.

Word length	Frequency	Percentage
1	8,933	12.2
2	48,995	67.1
3	5,727	7.9
4	7,040	9.7
≥5	2,301	3.1
Total	72,994	100.0

Table. 4. Statistics of word lengths in a dictionary

We want to remind that at this time there is no standard general dictionary or specific dictionary using for language processing, so we decide to choose a common dictionary for our statistics. However, through this statistic, we see that there are over 67 % words containing two syllables and about 30 % of the words consist of single syllable or three or four syllables. Longer words constitute about 3 % of the dictionary. A further subjective analysis showed that many of longer words are idiomatic expressions. These lead us to define some restrictions for the random initial string. First, we limit each segment wk can not have the length greater than four. Second, when randomizing, we set a bias ratio to generate more segment having the length 2 than the others. Moreover, we also apply the simple form of the Left Right Maximum Matching algorithm (Tsai, 2000) to build two specific individuals, forward / backward ones. So, the initial population will have some local optimal individuals.

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 offspring are the combining the beginning of id_1 with the ending of id_2 and vice-versa. However, if the child individuals break the above restriction, each segment w_k can not have the length greater than four, we will normalize them by flipping 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 for 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:

$$fit(id) = fit(w_1 w_2 .. w_m) = \sum_{k=1}^{m} MI(w_k) \text{ and } fit(pop) = \sum_{i=1}^{N} fit(id_i)$$

where $id=w_1 w_2 ... w_m$ is a particular individual of the population, $pop = \{id_1, ..., id_N\}$.

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.

5. Experimental Results and Discussions

Evaluating the accuracy of Vietnamese word segmentation is very problematic, especially without a manual segmentation test corpus. Moreover, as we state above, word segmentation is only the first step of text categorization, further steps should be made before we can evaluate a word segmentation approach in an equitable way. 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.

We build a corpus for testing purpose. Because our approach use internet-based statistic, we harvest news abstracts from many online newspapers, thus somehow balanced in styles and genres. Moreover, for the text categorization experiment, we classify these abstracts into different topics: society, world, business, science, culture, health and sport. Finally, we collect a 0.5 MB testing corpus containing 700 abstracts and 41,219 syllables, 100 documents for each 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

5.1. Word Segmentation Experiment

In this experiment, we have involved two natives, 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 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 5 represents the human judgment for our word segmentation approach.

Judgment	Prefect	Acceptable
Linguist Professor	368	538
	52.27 %	76.86 %
Graduate Student	431	554
	61.57%	79.14%

Table 5. Human judgment for word segmentation experiment

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 judges. We believe 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 segmentations outcome do not make the readers misunderstand the meaning. This is exactly what we expect. 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.

5.2. Text Categorization Experiment

The testing corpus consists of a set of documents, $D = \{d_1, d_2, ..., d_n\}$, where each document will be labeled with a unique category from a set of classes $C = \{c_1, c_2, ..., c_m\}$. Each category will contain a list of representative keyword $K = \{k_1, k_2, ..., k_u\}$. 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. Finally, d will be represented as follows: $d = g_1 g_2 ... g_r$ where g_i is a group of syllables after pre-processing.

Given a segmented text $t=w_1w_2\ldots w_m$, we calculate the relevance score of a topic c as follows:

$$p(k \mid w) = \frac{p(k \ \& w)}{p(w)}$$

$$rel(t,k) = \sum_{i=1}^{m} p(k \mid w_i)$$

$$rel(t,c) = \sum_{i=1}^{u} rel(t,k_i)$$

where $p(k \mid w)$ is the conditional probability of the keyword k given a particular word w. Following this equation, the higher the support degree is, the more possible the text belongs to that topic. We generalize it to calculate the support degree of a preprocessed document d for a topic c:

$$rel(d,c) = \sum_{i=1}^{r} rel(g_i,c)$$

In our experiment, we will classify the texts of the testing corpus into seven different topics: society, world, business, science, culture, health and sport, common topics on Vietnamese newspapers. The task of identifying list of keywords for a topic is not the subject of this paper and will be our future work. Thus, in this experiment, we only represent each topic by a keyword, the name of the topic.

Category	Ours	IGATEC
Society	87.2	83.9
World	90.5	91.4
Business	82.9	78.0
Science	88.5	87.4
Culture	85.7	83.6
Health	96.4	96.0
Sport	99.5	100.0
Micro-Avg	90.1	88.6

Table 6. F_1 and microaveraging F_1 performance of our approach and IGATEC

Our experiment assumption is that each document has a single category. We will use F_1 and *micro-averaging* F_1 measure described in Yang (1999) to evaluate performance. Table 6 shows the results on our testing corpus for all categories and their microaveraging. We compare our approach result with IGATEC introduced by H. Nguyen (2005).

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 our experiment, we realize that our GA iteration mean is only about 52.3, comparing with the 500 iterations of IGATEC GA Engine. This one, together with our less computational *MI*, makes our text categorization time is 0.5 seconds per document on a normal personal computer when statistic information was cached. One may question that our testing corpus is still small and the categories are not much confusable. We agree with this but we want to remind that there is neither standard training / testing corpus nor published Vietnamese text categorization algorithms. This initial experiment aims at testing our segmentation approach. An aggressive text categorization experiment is currently carried out.

6. 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 lacking 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 (Yang, 1999) 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 be 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.

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 algorithm with a self-adjusting parameter based on text length. 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 of Cilibrasi and Vitanyi (2005) 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 appropriate one for Vietnamese text categorization.

7. Acknowledgment

We would like to thank Mr. Nguyen Duc Hoang Ha at University of Natural Sciences, Vietnam National University for providing his IGATEC and valuable discussions. We would also like to thank Prof. Nguyen Duc Dan at University of Social Sciences and Humanities, Vietnam National University and Mr. Tran Doan Thanh at Kookmin University for their enthusiastic evaluation. This work is supported in part by University of Natural Sciences 2004 Research Grant.

References

BAKER L.D., MCCALLUM A.K. (1998). "Distributional clustering of words for text categorization". In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR'98)* : 96-103.

LEE-FENG CHIEN, HUANG, T.I., CHEN M.C. (1997). "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval ». In *Proceedings of 1997 ACM SIGIR Conference*. Philadelphia : 50-58.

CHURCH K, HANKS P., GALE W., HINDLE D. (1991). "Using Statistics in Lexical Analysis ». In Zernik U., *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

CILIBRASI R., VITANYI P. (2005). *Automatic meaning discovery of Google. A search for meaning, New Scientist*, 29 January 2005 : 21, by Duncan Graham-Rowe.

DINH DIEN. (2000). *Từ tiếng Việt* (Vietnamese words). Vietnam National University, HCMC.

DINH DIEN, HOANG KIEM, NGUYEN VAN TOAN (2001). "Vietnamese Word Segmentation". In *The*

Sixth Natural Language Processing Pacific Rim Symposium. Tokyo.

FOO S., LI H. (2004). "Chinese Word Segmentation and Its Effect on Information Retrieval". In *Information Processing & Management: An International Journal* 40 (1) : 161-190.

JOACHIMS T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In *European Conferences on Machine Learning (ECML'98)*.

LE AN HA (2003). "A method for word segmentation in Vietnamese". In *Proceedings of Corpus Linguistics*. Lancaster.

NGUYEN H., NGUYEN H., VU T., TRAN N., HOANG K. (2005). "Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese". In *Research, Innovation and Vision of the Future, the 3rd International Conference in Computer Science (RIVF 2005)*. Can Tho.

SHANKAR S., KARYPIS S. (2000). "Weight adjustment schemes for a centroid-based classifier". In *Text Mining Workshop on Knowledge Discovery in Data(KDD'00)*.

CHIH-HAO TSAI (2000). *MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm*.

YIMING YANG (1999). "An evaluation of Statistical Approaches to Text Categorization". In *Journal of Information Retrieval* 1 (1/2) : 67-88.

YIMING YANG, CHUTE C.G. (1994). "An example-based mapping method for text categorization and retrieval". In *ACM Transaction on Information System (TOIS'94)*: 252-277.

YIMING YANG, XIN LIU (1999). "A re-examination for text categorization methods". In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*.