Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model*

SU Gui-yang (苏贵洋)†, LI Jian-hua (李建华), MA Ying-hua (马颖华), LI Sheng-hong (李生红)

(*Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200030, China*)

†E-mail: sugy@sjtu.edu.cn

Received June 18, 2003; revision accepted Nov. 2, 2003

Abstract: With the flooding of pornographic information on the Internet, how to keep people away from that offensive information is becoming one of the most important research areas in network information security. Some applications which can block or filter such information are used. Approaches in those systems can be roughly classified into two kinds: metadata based and content based. With the development of distributed technologies, content based filtering technologies will play a more and more important role in filtering systems. Keyword matching is a content based method used widely in harmful text filtering. Experiments to evaluate the recall and precision of the method showed that the precision of the method is not satisfactory, though the recall of the method is rather high. According to the results, a new pornographic text filtering model based on reconfirming is put forward. Experiments showed that the model is practical, has less loss of recall than the single keyword matching method, and has higher precision.

Key words: Pornographic text filtering, Content based filtering, Information filtering, Network content security
doi:10.1631/jzus.2004.1106 **Document code:** A **CLC number:** TN915.08

INTRODUCTION

There are many filtering systems (Honorguard; SurfControl; S4F) used in schools, families, libraries, etc. to keep people, especially children, away from offensive information in the Internet. Methods to filter such information can be classified into two kinds: metadata based and content based.

Metadata are exterior information data, such as author, publishing date, links between documents, origin of information, and so on. The most popular information metadata in Internet used by filtering systems is origin of information, which includes URL and IP address. Nowadays, the most popular filtering method is based on URL or IP blocking. How to update the blocking list quickly enough is the key problem of URL or IP blocking method, as there are more than 500 000 pornographic websites (S4F), and more and more newer such sites are connected into the Internet each day. There are researches hammering at resolving the problem. For example, a research of the National University of Singapore (Ding *et al*., 1999) showed that applying heuristic rules to analyze links between documents and compound words in WebPages can accelerate the updating of URL blocking list; and that the efficiency of filtering can increase from 60% to 90%. But there are servers in Internet offering services such as proxy, which can help users access blocked information.

Development of distributed systems worsen the problem. There are systems designed to resist information censorship (Infranet; Amos and Jared,

*Project (Nos. 2001AA142160, 2002AA145090) supported by the Hi-Tech Research and Development Program (863) of China

2002; Waldman *et al.*, 2000; Waldman and Mazieres, 2001), which can hide information metadata such as origin and author to make metadata based filtering method useless. Therefore, analyzing the information content is of great importance for the filtering task in such systems.

A large number of existing systems deal with the filtering of pornographic information. Other information being filtered also include information on violence, drug abuse, etc. Our research here focuses on filtering Chinese pornographic text, which is the most popular and the most easily produced type of information crammed in the Internet. In pornographic text filtering, keyword matching is the most commonly used method of content based filtering in many filtering systems (S4F), but few literature dealing with the evaluation of this method are available, especially on its accuracy. So, several experiments were conducted to roughly evaluate the performance of keyword matching. Based on the results of those experiments, a new content based system model for automatically filtering pornographic text is put forward. The validity of this system model had been confirmed.

The rest of this paper is organized as follows: In Section 2, results and analyses of experiments on keyword matching method are given. The new content based system model is introduced in Section 3. Experiments to confirm the validity of the new model and its analyses are presented in Section 4. The last section summarizes this paper and proposes future work.

KEYWORD MATCHING METHOD

General steps of keyword matching are: (1) finding particular string (keyword) in text; (2) using heuristic rule (for example, Boolean) to judge whether the information should be filtered. The most simple heuristic rule is “Boolean And”. If more than N (N>1) particular strings coexist in a text, the text will be considered as one which should be filtered. Particular strings can be words, phrases, sentences and even paragraphs. Particular strings and heuristic rules compose a profile of certain kind of information. The profile can be obtained manually or automatically, and there are various ways to complete the job.

Here, we use a lexicon as a profile of pornographic texts. If more than N (N>1) words in the lexicon appear in a text, the text is filtered. Eight-hundred and fifteen words listed in the lexicon are automatically extracted from 1600 pieces in the corpus (which is collected manually includes 4000 pieces of Chinese pornographic text in pure text format) of pornographic texts, and finally confirmed manually. The remaining 2400 pieces in the corpus are used to roughly calculate the recall of keyword matching method.

According to the number of times of a keyword appeared in pieces of corpus, 4 more lexicons are constructed from the 815 lexicon (lexicon containing 815 entries). If 236 words in the 815 lexicon appear in more than 2.5% (40=1600×0.025) of the pieces in the corpus, those words form a 236 lexicon. In a similar way, a lexicon with 154 entries includes words that appeared in more than 5% of the pieces, a lexicon with 78 entries includes words that appeared in more than 10% of the pieces, and a lexicon with 34 entries includes words that appeared in more than 20% of the pieces.

Two measures are borrowed from Information Retrieval to evaluate the performance of the keyword matching method: Recall and Precision (Rocchio, 1971). Precision is the proportion of examples filtered out positive by the system that should truly be filtered out. Recall is the proportion of positive examples filtered out positive by the system. Table 1 shows how to compute precision and recall.

Table 1 How to compute precision and recall

Predicted classes	Actual classes	
	Pornographic	Others
Positive	a	b
Negative	c	d

$$\text{Precision } (P) = \frac{a}{a+b} \qquad \text{Recall } (R) = \frac{a}{a+c}$$

Sample recall varies with variation of the lexicon size and the filtering threshold (N). The result of the recall experiment is shown in Table 2.

Table 2 Examples of recall on pornographic texts varying with variation of the lexicon size and the filtering threshold (N)

Lexicon	815	236	154	78	34
N=3	99.8%	99.8%	99.8%	99.6%	98.0%
N=5	99.7%	99.4%	99.3%	98.7%	95.4%
N=8	98.9%	98.5%	98.1%	96.3%	91.0%
N=15	96.0%	94.5%	93.6%	89.7%	74.3%

Table 2 shows that:

1) As the number of entries in the lexicon increased, the Recall is increased too. But the rising trend of recall is much slower than that of the numbers. Lexicon with few entries is already capable of high Recall. In actual filtering system, addition of new items (keywords and heuristic rules) into profiles should be done very cautiously.

2) As N increased, the recall decreased rapidly, especially when the lexicon's size is small. The proportion of keywords in lexicons that appeared in each pornographic text is small, even though those keywords most likely have high correlation with pornographic texts. In fact, there is no word that appeared in all pornographic texts.

When N=3, using a small lexicon with only 78 entries, the keyword matching filtering method can reach a high Recall of 99.6%. High Recall is one of most important reasons for the wide use of this method. To estimate the precision of the method, the following experiment was done on the 78 lexicon.

We submitted the keywords in the 78 lexicon to the Google search engine. Each submission used three keywords. The search string was "S_i and S_j and S_k", where S is a word in the 78 lexicon, and $0<i, j, k\leq 78$, $i\neq j\neq k$. More than 42000 results were returned. The first 8000 results were analyzed manually. Only 22% of results returned were pornographic texts. Others were scientific texts about sex or medicine (30%), texts about sports or entertainment news (16%), texts about fashion, beauty and fitness (27%), and other kinds (5%). That is, on such experiment condition mentioned above, the precision of keyword matching was only about 22%. At the same time, most pornographic texts were filtered out; many other texts with similar glossary were filtered out too.

To find the percentage of scientific texts about sex filtered out, 2500 samples of sex knowledge texts were collected. Experiment to calculate the recall of the keyword matching method using lexicons mentioned before was conducted. The result of the experiment is shown in Table 3.

Table 3 Examples of recall on scientific texts varying with variation of the lexicon size and the filtering threshold (N)

Lexicon	815	236	154	78
N=3	73.7%	70.0%	63.1%	43.9%
N=5	57.1%	51.8%	41.3%	27.3%
N=8	34.4%	29.6%	22.5%	10.3%

A part (43.9%) of scientific texts about sex can be filtered out in our pornographic filtering system using single keyword matching method with the 78 lexicon. With increased size of lexicon, the recall increased rapidly. Words which frequently appear in pornographic texts also appear frequently in scientific texts about sex, in news about sports or entertainment and in texts about fashion, beauty and fitness. So, though keywords listed in lexicons are selected carefully by man and have high correlation with pornographic texts, they are also have high correlation with other kinds of texts. A keyword listed in lexicons is not necessarily 'pornographic' itself.

Many text features can be used in profiles. Word is one of those features. Other features are genre, structure, paragraph, tone of text, and so on. Aside from the co-occurrence aspect of words we used in our experiments, other aspect of words, such as distribution, density and collocation between words, can be used. Features and various word aspects used in profiles are all positive characters extracted and refined from positive samples. So, even if positive features listed in profiles are elaborated, they can only assure high correlation with positive topic (pornographic texts). It is very difficult to assure those positive features of low correlation with negative topics (non-pornographic texts), as the negative topics are not one or several certain kinds of text but all kinds of text except the positive one.

The keyword matching method uses positive features to distinguish certain texts from others. It is understandable that the recall is affected by the correlation between the positive features and the positive topic, but the prediction is essentially affected by the correlation between the positive features and the positive topic, and by the interlinking correlation between the positive features and the negative topic. That is why the keyword matching method yields high recall but low precision.

To sum up, though the keyword matching method recall is high, the precision is rather low. That keeps the filtering system based on single keyword matching method from being practical. What we should pay close attention to is that, texts which are filtered incorrectly by keyword matching are converged in three categories, with texts in these three categories occupying 93% of all. Therefore, we bring forward an idea: texts filtered out by keyword matching method are called suspicious texts, and intelligent methods of classification will be used to distinguish truly pornographic texts from those suspicious texts. According to the idea, we design a new model of pornographic text filtering system.

A new model of pornographic text filtering

The new model of pornographic text filtering is shown in Fig.1. On the basis of the present method of keyword matching shown on the left of Fig.1, a new module named Classification for Reconfirming (Right part of Fig.1) is added. In this module, intelligent technologies of automatic classification are used to distinguish true pornographic texts from suspicious texts, which are filtered out by the keyword matching module. The high recall and high speed of the filtering system can be preserved in the keyword matching module, and high precision can be ensured by the classification for reconfirming module.

Texts for filtering are first processed by the keyword matching module. Texts filtered out by keyword matching module are given to the new classification module for reconfirming. In our system, if more than N (N=3) keywords (entries of the 78 lexicon) are found coexisting in the text, the text is considered as a suspicious one. Suspicious texts will be classified by a new module. If a suspicious text is identified as a pornographic one by the classifier of the new module, the text will be finally filtered out. If not, the suspicious text will be

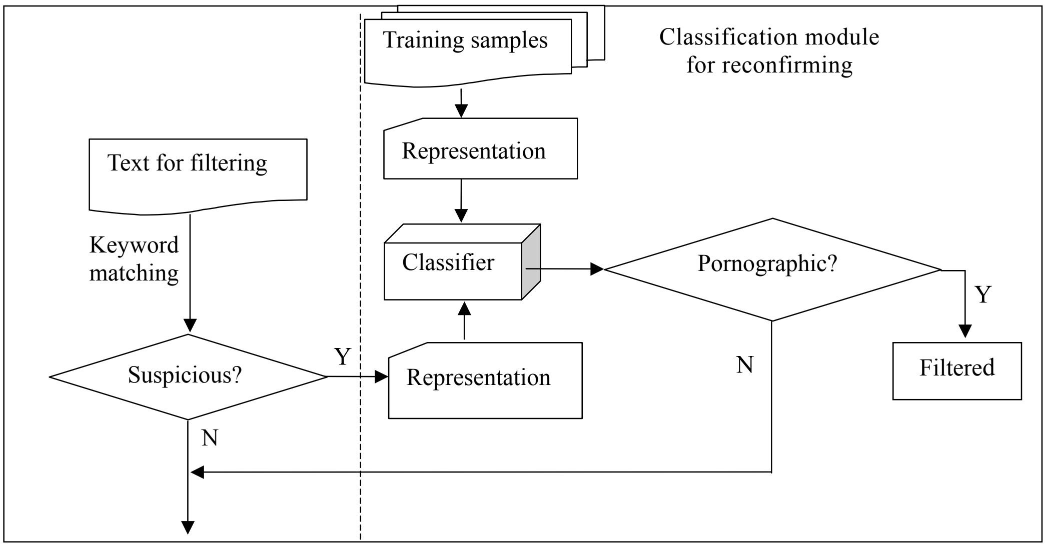

Fig.1 A new content based pornographic text filtering model

passed as a non-pornographic text.

Two key parts of the new module are the classifier and the text representation. According to Section 2, suspicious texts can be roughly sorted in three kinds: scientific texts about sex or medicine, text about sports or entertainment news, and texts about fashion, beauty and fitness. As each kind has its specialties, in order to gain high accuracy of classification, classifiers and representations of text should be designed respectively. Our classification experiment confirmed the validity of the new model. Experiments showed that this model of keyword matching method reinforced by a new module of classification for reconfirming is practical; as it gains much higher precision with little loss of recall compared to the single keyword matching method.

EXPERIMENTS AND RESULTS

Experiments on classification

Many automatic classification algorithms, such as k-Nearest Neighbor (KNN) (Creecy *et al.*, 1992; Yang, 1994), Bayes (Cheeseman *et al.*, 1988; John *et al.*, 1995), Decision trees (Quinlan, 1986; Lewis and Ringuette, 1994; Apte *et al.*, 1998), Neural Network (NN) (Wiener *et al.*, 1995; Ng *et al.*, 1997), Support Vector Machines (SVM) (Thorsten, 1998). According to Yang and Liu (1999), KNN and SVM are the most outstanding algorithms in English texts classification. After comparing 14 text classification algorithms (Yang, 1997), a conclusion was drawn that KNN is the only one which is suitable for various domains. In Chinese text processing (He *et al.*, 2000a; 2000b; 2003), compared to SVM and ARAM, KNN is a better choice. Therefore, in our experiments, we chose the KNN algorithm to do the classification.

Among many different kinds of text representation methods, “bag of words” representation and POS (part of speech) are the most popular. Others are punctuation, special symbol, text structure and so on. Chinese text can also be represented by characters. How to represent a text is much more important for accuracy in text classification than how to choose a good algorithm (Mladenic, 1999). So, in our experiments, four text representation methods: character, bag of words, POS and punctuation were compared.

In text representation, there are many methods, such as Boolean, relative frequency, square root, logarithm, TFIDF, TFC, LTC and so on, for calculating the weight of features (Aas and Eikvil, 1999). For convenient comparison, the easiest weighting system of Boolean was chosen for “bag of words” representation and character representation, and relative frequency weighting were used for other representations.

The corpus for classification was that used in experiments described in Section 2, and included two categories of texts: pornographic text and scientific text of sex and medicine. Three-thousand and two-hundred samples (1600 for each category) of pornographic and scientific texts for training were randomly chosen, and 1420 (710 for each category) were chosen for testing.

Considering the great number of Chinese characters, only characters listed in GB5007-85 are processed in character representation. There are 6763 characters in it, and all of them are in common use. In punctuation representation, 11 punctuations are used. In POS representation, an open POS labeling software provided by Northeastern University (http://www.nlplab.com) was used to extract POS data from text. Twelve POS labels were used. As for bag of words representation, in order to reduce computing time, only 815 words in the feature lexicon of pornographic texts mentioned before were processed. The result of experiments is shown in Table 4.

In KNN algorithm, the fewer features used in the text representation, the faster is the algorithm. The number of features of POS and Punctuation are

Table 4 Result of reconfirming classification experiment

Precision	Word	Character	POS	Punctuation and others
Scientific texts	96%	90%	66%	91%
Pornographic texts	100%	100%	79%	95%
Average	98%	95%	72.5%	93%

the least. But because extracting features of POS from Chinese texts is a rather complicated process, the fastest classifier is the one using features of punctuation.

In the representation method of "bag of words", we used word selection to avoid data sparseness, decrease the text vector dimension and speedup the KNN classifier. We selected a small number of words called subject words which appeared much more frequently in the text to represent it. The result of experiment using 3, 5, 8, 15 subject words is given in Table 5.

Table 5 Relationship between precision of classification and number of subject words extracted from text

	3 words	5 words	8 words	15 words
Pornographic texts	88.0%	92.5%	97.1%	100%
Scientific texts	91.5%	96.4%	96.4%	95.0%
Average	89.7%	94.4%	96.7%	97.5%

The result of experiment showed that small number of subject words could accelerate the classifier but decreased accuracy slightly. Using only 8 subject words to represent the text, the KNN classifier can process 2.7 pieces every second, as quickly as the KNN classifier using punctuation representation.

Two similar text classification experiments using bag of words representation were conducted. One was classified pornographic texts from sports or entertainment news and the other classified pornographic vs texts about fashion, beauty and fitness. Sample texts of these two kinds were retrieved from the http://www.sohu.com.cn, one of the most popular websites in China. There are 3500 (1750 for training and 1750 for testing) texts about fashion, beauty and fitness and 2400 (1200 for training and 1200 for testing) about sports or entertainment news. Using 8 subject words to represent the text, the KNN classifier can distinguish 93.4% pornographic texts from texts about fashion, beauty and fitness, and 98.5% pornographic texts from text about news of sports or entertainment.

The experiments above showed that though suspicious texts have glossary similar to that of pornographic texts, intelligent classification can distinguish them very well.

Experiment on the new filtering model

According to results of classification experiments, KNN algorithm and 8 subject words representation method were used to construct the classifier in our filtering system. We continued the performance experiment mentioned in Section 2, using the new classification module of the new model to distinguish pornographic texts from suspicious texts filtered by the keyword matching module. Suspicious texts were the 8000 texts which were labeled manually. There were 1758 pornographic texts, 2411 scientific texts about sex or medicine, 1205 texts of news about sports or entertainment, 2157 texts about fashion, beauty and fitness and 469 other texts.

The KNN classifier filtered out 1977 'pornographic' texts, of which 1662 were pornographic texts. The precision of the new model was 84.0%, much higher than the precision of the single keyword matching method. As 96 true pornographic texts are mis-classified, the recall of the new model was lowered to 94.5×99.6%=94.1%.

Under the condition of the experiment mentioned above, the precision of the whole filtering system rose from 22% to 84%, and the recall decreased from 99.6% to 94.1%. As non-suspicious texts only were processed by the keyword matching module, the overall speed of the system was not slowed down. The new model improved greatly the precision of the filtering system with little loss of recall and speed.

CONCLUSION

The keyword matching method is widely used in harmful text filtering systems. The speed is crucial for filtering systems. Outstanding advantages of this method are high speed and simplicity to realize. Profiles which describe certain kinds of texts used in the method can be inspected, adjusted and modified directly by man via man-machine interface. Description of certain kinds of text is different in different systems; and items in the profiles are different too. The quality and quantity of items in profiles have great influence on the filtering performance. It is difficult to accurately evaluate the performance of filtering systems, as any change in the profile will influence the performance more or less.

Our survey of several pornographic texts filtering systems revealed that most systems use merely words to construct the profile; and that their effects on the systems are not good enough to be practical. Then, we design a simple experiment to evaluate roughly the precision and recall of the keyword matching method and intend to find the key problems. After analyzing the results of experiments using the present keyword matching method, the main problem of low precision was found. The problem cannot be solved by improving the present keyword matching method or the method of constructing the profile, to say nothing about modifying the profile. The reason is that though all items in profile are positive features of pornographic texts; all those features are more or less correlated with other texts. So, there will doubtless be other texts filtered out by the keyword matching method.

In order to distinguish those other texts from true pornographic texts, a new model based on reconfirmation is put forward. A new module named Classification Module for Reconfirming is added to the model. The new module using intelligent technologies of classification distinguishes true pornographic texts from texts filtered out by the keyword matching method, which include pornographic texts and not pornographic.

Experiments showed that the model achieved high precision with little loss of recall. In order to get better performance of the new model based on reconfirming, our further research will focus on:

1) Research methods to construct and update the profile of every kind of offensive or undesirable text automatically, and find general methods to describe characteristics of offensive or undesirable text.

2) Investigate how to select key features in classification and design and test more algorithms to construct faster classifier.

3) Carry out long-term testing, and apply the reconfirming based model to other kinds of filtering systems for offensive or undesirable information.

References

Aas, K., Eikvil, L., 1999. Text Categorisation: A Survey. http://citeseer.nj.nec.com/aas99text.html.

Amos, F., Jared, S., 2002. Censorship Resistant Peer-to-Peer Content Addressable Networks. Proceedings of Symposium on Discrete Algorithms.

Apte, C., Damerau, F., Weiss, S., 1998. Text Mining with Decision Rules and Decision Trees. Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web.

Ding, C., Chi, C.H., Deng, J., Dong, C.L., 1999. Centralized Content-Based Web Filtering and Blocking: How Far Can It Go? IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics, **2**:115-119.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., Freeman, D., 1988. Autoclass: A Bayesian Classification System. Proc. Fifth Int. Conf. on Machine Learning. San Mateo, California, p.54-64.

Creecy, R.H., Masand, B.M., Smith, S.J., Waltz, D.L., 1992. Trading mips and memory for knowledge engineering: Classifying census returns on the connection machine. *Comm. ACM*, **35**:48-63.

Honorguard. Christian Filtered Internet Service on the Internet. http://www.honorguard.net/.

Infranet. A System that Enables Clients to Surreptitiously Retrieve Sensitive Content via Cooperating Web Servers Distributed across the Global Internet. http://nms.lcs.mit.edu/projects/infranet.

He, J., Tan, A.H., Tan, C.L., 2000a. Machine Learning Methods for Chinese Web Page Categorization. In proceedings, ACL'2000 International Workshop on Chinese Language Processing, Hong Kong.

He, J., Tan, A.H., Tan, C.L., 2000b. A Comparative Study on Chinese Text Categorization Methods. Proceedings, PRICAI'2000 International Workshop on Text and Web Mining, Melbourne, p.24-35.

He, J., Tan, A.H., Tan, C.L., 2003. On Machine Learning Methods for Chinese Document Classification. *Applied Intelligence*, **18**:311-322.

John, G.H., Langley, P., 1995. Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, San Mateo, p. 338-345.

Lewis, D.D., Ringuette, M., 1994. Comparison of Two

Learning Algorithms for Text Categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR’94).

Mladenic, D., 1999. Text-learning and related intelligent agents: a survey. *IEEE Intelligent Systems*, **14**(4): 44-54.

Ng, H.T., Goh, W.B., Low, K.L., 1997. Feature Selection, Perception Learning, and A Usability Case Study for Text Categorization. 20th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’97), p.67-73.

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning*, **1**:81-106.

Rocchio, J., 1971. Relevance Feedback in Information Retrieval. *In*: G. Salton, ed., The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, p.313-323.

SurfControl. Internet Filtering Solutions Stop Unwanted Content. http://www1.surfwatch.com/.

S4F. Provides Internet Content Filtering Solutions and Internet Blocking of Pornography, Adult Material, Criminal Activity, Chat Blocking, and Many More Categories. http://www.s4f.com/.

Thorsten, J., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning (ECML).

Waldman, M., Rubin, A., Cranor, L., 2000. Publius: A Robust, Tamper-Evident, Censorship-Resistant Web Publishing System. Proc. of the 9th USENIX Security Symposium.

Waldman, M., Mazieres, D., 2001. Tangler: A Censorshipresistant Publishing System Based on Document Entanglements. Proc. 8th ACM Conf. on Computer and Communications Security.

Wiener, E., Pedersen, J.O., Weigend, A.S., 1995. A Neural Network Approaches to Topic Spotting. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR’95).

Yang, Y., 1994. Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. 17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’94), p.13-22.

Yang, Y., Liu, X., 1999. A Re-examination of Text Categorization Methods. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99), p.42-49.

Yang, Y., 1997. An Evaluation of Statistical Approach to text Categorization. Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University.