Automatic Categorization of Japanese Patents based on Surrogate Texts

Yuen-Hsien Tseng, Da-Wei Juang*, and Chi-Jen Lin*
National Taiwan Normal University, Taipei, Taiwan, R.O.C., 106
samtseng@ntnu.edu.tw
*WebGenie Information LTD., Taipei, Taiwan, R.O.C., 106
*{david, dan}@webgenie.com.tw

Abstract

This paper describes our work at the fifth NTCIR workshop on the subtask of patent classification. We use KNN (K-Nearest Neighbors) as our classifier and the English PAJ (Patent Abstract Japan) as the patent surrogate for classification. Based on the knowledge and experience learned from our previous experiments with other document collections, we leverage on the parameters to achieve above-average performance in an efficient way.

Keywords: *patent surrogate, text categorization, KNN, test collections.*

1. Introduction

Patent documents contain important research results that are valuable to the industry, business, law, and policy-making communities. If carefully analyzed, they can show technology details and relations, reveal business trends, inspire novel industrial solutions, or help make investment policy [1-2]. In recent years, patent analysis had been recognized as an important task. Public institutions in China, Japan, Korean, Singapore, and Taiwan have invested various resources in the patent analysis task [3-5].

The Classification Subtask of the Patent Retrieval Task in the NTCIR Workshop 5 is one of the attempts to automate the patent analysis task. By labeling patent applications with proper predefined and uniform categories, structured analysis such as patent mapping or data mining can be easily done.

Specifically, the Classification Subtask demands each Japanese patent application be automatically classified based on the *F-term* classification system, which is a multi-dimensional classification structure used by Japan Patent Office (JPO). In this system, over 2,000 *themes* are in use. Each theme denotes a technical field and may contain several *viewpoints* such as “PURPOSE”, “MEANS”, or “OPERATION MODE“. The content and the number of the viewpoints vary from theme to theme. Each viewpoint, in turn, has its *elements* to describe the viewpoint in different aspects. Pairs of viewpoints and its elements (with their corresponding theme omitted because it is often processed separately) are called “F-terms”.

Because Japanese patent was classified by themes and F-terms before their publication, enough patent documents can be collected for the Classification Subtask to train and test the machine classifiers proposed by the participants.

We participate in this subtask based on our past experiences in automated text categorization. The techniques we used are explained in Section 2. The lessons we learned from previous experiments with other document collections are reported in Section 3. Section 4 describes our results in this subtask. Finally, Section 5 concludes this report.

2. Categorization Techniques

The techniques of text categorization (TC) have been extensively explored in recent years. Many machine classifiers have been proposed. To know which classifiers are better, Yang and Liu [6] had conducted a comparative experiment based on the Reuters-21578 test collection. Their results showed that

{SVM, KNN} > *LLSF* > *MLP* > > *MNB*

where SVM denotes the Support Vector Machine method, KNN denotes K-Nearest Neighbors, LLSF is Linear Least Square Fit, MLP is MultiLayered Perceptrons, MNB is Multinomial Naïve Bayes, ‘>’ denotes ‘better’, and ‘>>’ is ‘far better’. In another experiment conducted by Joachims [7], where only the largest 10 categories of the Reuters-21578 test collection were tested, the microF measure showed that

SVM (0.864) > KNN (0.823) > {Rocchio (0.799), C4.5(0.794)} > naïve Bayes (0.72)

As can be seen, SVM and KNN are among the best performing classifiers.

However, the effectiveness of automatic TC is affected by several factors; including term indexing, feature selection, document surrogate selection, classification methods (classifiers), and the number of training documents [8-9]. Using a best performing classifier alone without optimizing other factors does not guarantee that the best results can be obtained. Below we describe the methods implemented in our TC system for each factor.

The goal of indexing in text categorization is to choose a term formation strategy such that the indexed terms are both predictive and discriminative. By predictive, we mean that a term should not only occur in the training documents, but it should have high probability that it will occur in the testing documents to be classified. By discriminative, we mean that a term is highly indicative of the categories to which the documents containing the term belong. As such, exhaustive indexing is not necessary, as in text retrieval. On the contrary, frequent terms that are specific to some categories are preferred. To have such terms, the keyphrase extraction algorithm used in the SLIR subtask of NTCIR Workshop 3 for Chinese, Japanese, and Korean texts is used [10]. The algorithm extracts those maximally repeated strings in the texts followed by a stopword filtering process to get important and high TF (Term Frequency) words or phrases [11]. Low DF (document frequency) terms (DF<2) were further filtered from the index.

However, term filtering by TF and DF may not be enough to get high discriminative terms. Some feature selection methods may be needed. Yang and Pedersen [12] had compared five different methods. They found that Chi-square is among the best that lead to highest performance. The Chi-square method computes the relatedness of term T with respect to category C in the manner:

$$\chi^2(T,C) = \frac{(\text{TP} \times \text{TN - FN} \times \text{FP})^2}{(\text{TP + FN})(\text{FP + TN})(\text{TP + FP})(\text{FN + TN})}$$

where TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) denote the number of documents that belong or not belong to C and that contain or not contain T, respectively. The Chi-square is exactly the square of the correlation coefficient (CC) method, which is:

$$Co(T,C) = \frac{(\text{TP} \times \text{TN - FN} \times \text{FP})}{\sqrt{(\text{TP + FN})(\text{FP + TN})(\text{TP + FP})(\text{FN + TN})}}$$

Ng et al [13] had pointed out that CC selects exactly those words that are highly indicative of membership in a category, whereas the Chi-square method will not only pick out this set of terms but also those terms that are indicative of nonmembership in that category. This is especially true when the selected terms are in small number. Thus we use CC as our feature selection method when necessary.

Although feature selection may reduce noisy terms for effective text categorization, excessive term reduction may hurt effectiveness in some real-world test collections. For example, Bekkerman et al [14] showed that the more terms are used, the better the categorization results in the 20NG test collection. Yang and Pedersen also showed the similar results for the OHSUMED test collection [12]. Thus whether feature selection should be used depends on the collections to be classified.

Document surrogate selection may be useful for long documents. The goal of document surrogate selection is similar to that of feature selection. They are both to reduce noisy texts and text volume as well such that better performance in efficiency and effectiveness is possible. Thus for short documents, feature selection may be suitable, while for long documents, document surrogate and/or feature selection may be profitable. For example, only titles and abstracts are used in many scientific papers classification cases. As another example, some sort of summaries in patent documents may lead to better classification results than their full texts. Fall et al had confirmed this observation in an American patent classification experiment using Naïve Bayes, KNN, and SVM as classifiers. They show that even using only the first 300 words from the abstract, claims, and description sections, the performance is better than those using the full texts regardless of which classifiers are used [15].

For the classifiers, SVM and KNN are among the best as is mentioned. We chose SVMlight [16] as our SVM classifier. We accepted all the default values of the system and used the sign of the class output to determine the predicted class, as suggested by the system manual. But it was found that many documents have no class output. We then forced them to have one or two, depending on the average number of categories a document has, by selecting the categories with the maximum output. This strategy always improves the effectiveness of the SVM classifier in our experiments.

For the KNN classifier, we implemented one by ourselves. The document similarity is calculated by the inner product of the corresponding document vectors, whose elements are weighted by taking into account the term frequency and inverse document frequency. Byte size (BS) normalization [17] (denoted as KNN-BS) instead of cosine normalization is used due to its simplicity and better performance observed in previous IR studies. Since KNN's performance highly depends on the similarity measure, we also implemented a probabilistic model called BM11 [18] (denoted as KNN-BM11) for comparison. The K in the KNN method was set to 20 in most of our experiments. For some categorization tasks, we also limit the maximum number of terms used from a document when the document was classified by the KNN method.

As to the number of training documents, a rule of thumb is that the more the training examples used, the better the effectiveness. This can be seen in the experimental results in the next section.

3. Experiments on Other Collections

Before we participated in the Japanese patent classification task, we experimented with the above techniques on five test collections to learn experience. Among them, one is an English corpus. The other four are all Chinese. Some statistics about these collections are shown in Table 1, where the first column is the name of the collection, the second is the number of documents for training, the third is the number of documents for testing, the fourth is the number of total categories in the collection, the fifth is the average number of categories assigned to a document, and the last column shows the average number of characters in each document.

Table 1. Some statistics about the 5 test collections.

Collection	Train	Test	All Cat.	Avg Cat.	Char.
Reuters-21578	7770	3990	90	1.3	133*
FJU CTC	19901	8110	82	1.0286	619
News	644	270	12	1.0	455
WebDes	1190	496	26	1.0	65
LawCase	8196	3512	7	1.0	735

*The 133 is in English word unit, not in Chinese character unit.

The Reuters-21578 is a widely used test collection in automatic TC community. There are 3 ways to split the collection into a training set and a testing set. To compare our results with the previous, the ModApte split of the Reuters-21578 corpus was used. Furthermore, we followed Yang's limitation on the categories [6] such that only those categories having at least one document in the training set and in the testing set are used.

The FJU CTC Chinese test collection originates from a special corpus of news manuscripts held by SCRC (Socio-Cultural Research Center) at Fu Jen Catholic University (FJU). These manuscripts are manually labeled and transcribed news broadcasts of Mainland China's radio stations between 1966 and 1982. In year 2000-2001, under a digitization project, SCRC had 42371 manuscripts keyed-in manually for the preservation and better use of this material. Among them, 30710 manuscripts have category labels and dates. This corpus was later developed by Yuen-Hsien Tseng [19] into a test collection for TC based on the following guidelines: (1) As many documents are included to better utilize the label information that already exists. (2) Each category should have documents in the training set and in the test set so that an effective training and testing of a machine classifier is possible. (3) The training documents predate all the test documents to reflect the ordinary use of an operational classifier. (4) Duplicates and highly similar documents with inconsistent labels should be isolated to reduce the unreliability of the evaluation results. Accordingly, a total of 28011 documents were identified. Since the corpus comes from the manual transcription of on-site news broadcasts, missing words or even missing snippets are not un-common in the documents. Statistics of the categories and the collection can be downloaded at [19]. Since the documents spread over 17 years, inconsistency of the label assignment may confuse any machine learning classifiers such that effectiveness becomes unpredictable and unreliable. This issue was explored by Tseng and Teahan [20]. Their result showed that even though there is as high as 34% inconsistent documents, better classifiers trained with these documents still perform better. Thus the inconsistency problem does not prevent this corpus as a good TC collection.

The News collection contains the news articles from PChome Online (www.pchome.com.tw), while the WebDes collection contains web page descriptions complied by the same website. All of WebDes categories are under the "Internet and Computer" category in PChome's Web directory. Collection News has a total of 914 documents and the average document length is 9.87 sentences, while WebDes has 1686 documents and the average length is 2.10 sentences. Category size in collection News ranges from 10 to 331 documents, while in collection WebDes it ranges from 4 to 374. Both collections exhibit skewed category distribution, like that in Reuters-21578 and the FJU CTC collections.

The last collection contains some lawsuit judgment documents. It has the longest documents compared to the other 4 collections. The category sizes are quite similar among its 7 categories.

With these 5 real-world collections, we experimented on them with the techniques discussed above. Table 2 shows the results, which are the best that we can obtain by varying on each factor that has been mentioned.

Specifically, all the training documents are used in the five collections. If not, performance can be affected severely. Table 3 shows the degradation in effectiveness as only a ratio of training documents is used for the News and WebDes collections. Table 3 also shows that a lower-performing classifier can perform better if more training documents were used.

Since documents in these collections are not very long, all the document texts instead of any document surrogates are used. However, feature selection varies from collection to collection. Generally, longer documents require feature selection more than those shorter ones. Table 4 shows the change of effectiveness with respect to the change of filtered terms based on the above correlation coefficient method. Over- and under-filtering do not lead to the best results.

Table 2. Effectiveness of the five test collections.

Collection	Method	MicroF	MacroF
Reuters-21578	KNN-BS	0.8192	0.3681
	KNN-BM11	0.8242	0.4211
	SVM	0.8459	0.4931
FJU CTC	KNN-BS	0.4383	0.2881
	KNN-BM11	0.4605	0.3177
	SVM	0.4993	0.3603
News	KNN-BS	0.79	0.73
	SVM	0.71	0.66
WebDes	KNN-BS	0.78	0.58
	SVM	0.78	0.67
LawCase	KNN-BS	0.7762	0.7664
	KNN-BS*	0.8243	0.8122

* See Table 4 for details.

Table 3. Effectiveness under different ratios of training documents.

Col.	News				WebDes			
	KNN-BS		SVM		KNN-BS		SVM	
Ratio	Mi.	Ma.	Mi.	Ma.	Mi.	Ma.	Mi.	Ma.
5%	0.47	0.30	0.40	0.19	0.64	0.32	0.67	0.35
10%	0.58	0.32	0.57	0.31	0.69	0.38	0.71	0.42
20%	0.70	0.50	0.63	0.45	0.67	0.45	0.65	0.46
40%	0.72	0.62	0.63	0.49	0.75	0.55	0.78	0.61
100%	0.79	0.73	0.71	0.64	0.78	0.58	0.78	0.67

Table 4. Effectiveness under different levels of feature selection for the LawCase collection.

CC>=x	No. of Terms Remained for TC	MicroF	MacroF
0.000	31329	0.7762	0.7664
0.010	4995	0.8015	0.7926
0.050	4129	0.8041	0.7958
0.075	2402	0.8141	0.8063
0.125	1068	0.8218	0.8120
0.100	1551	0.8172	0.8111
0.150	**756**	**0.8243**	**0.8122**
0.200	395	0.8186	0.8058
0.300	159	0.7091	0.6978

In summary, lessons learned from the above experiments include:

(1) The more the training documents were used, the better the performance.

(2) Improve the quality of the features, improve the performance. This can be done by three approaches: document surrogate selection (such as summary extraction), feature selection, and term indexing. Among them, feature selection may be the most important factor.

(3) SVM and KNN are the two high-performing classifiers. SVM often performs better than KNN. But its high training cost in computation may prohibit it from being used for a large amount of training documents.

4. Experiments on NTCIR's Collection

The collection of the Classification Subtask contains unexamined Japanese patent applications (JA) published from 1993 to 1999 and the corresponding English PAJ (Patent Abstract Japan) abstracts within the same year range. The patents published from 1993 to 1997 (five years) go to the training set, and the patents published from 1998 to 1999 (two years) go to the test set. A total of 2008 documents are selected from the test set for the Classification Subtask. There are two further subtasks: "theme categorization" and "F-term categorization". The former determines one or more themes for each patent application, while the later, given a specific theme, assigns one or more F-terms (i.e., pairs of viewpoints and its element) to each patent application. Both are evaluated independently.

As mentioned, the number of themes in use is about 2000. However, the number of F-terms is more than 200,000. For our operational system originally designed to deal with ordinary text collections, this vast number of F-terms exceeds the limit of the number of categories allowed in our system, which is 32,600. Therefore, we only participated in the theme categorization subtask.

We used as many training documents as possible, since it seems to be the most important factor in affecting the effectiveness. However, due to this decision, feature selection based on the CC method becomes an extremely time-consuming problem such that our PC-level hardware cannot support this calculation. Also, the often better-performing classifier SVM can not be used due its slow training process for such vast volume of training documents. Instead, the rote-learning method KNN is used. The similarity measure of which is the byte size normalization rather than the BM11 probabilistic model, again due to the inefficiency of the BM11 model.

Table 5: Dry Run Results.

RunID	Parameters	A-Precision	Recall at 5	R-Precision	F-measure	Retrieve
dt007	PAJ, 50, 20	0.5820	0.2507	0.5055	**0.1981**	21249
dt008	PAJ, 100, 50	**0.6249**	**0.2687**	**0.5376**	0.1217	42619
dt009	BAC*, 50, 20	0.5149	0.2254	0.4525	0.1797	22103
dt010	BAC, 100, 50	0.5397	0.2339	0.4625	0.1088	44750
dt011	BAC+PAJ, 50, 20	0.5503	0.2404	0.4885	**0.1962**	20977
dt012	BAC+PAJ, 100, 50	0.5717	0.2482	0.4946	0.1187	42382
Max		0.6320	0.2721	0.5376	0.4408	1179000
Min		0.1876	0.0775	0.1522	0.0040	1175
Avg		0.4670	0.2008	0.4005	0.1668	219992

Note: BAC= BIJ+ABJ+CLJ.

Table 6: Formal Run Results.

RunID	Parameters (T, K, C, S)	A-Precision	Recall at 5	R-Precision	F-measure	Retrieve
ft020	50, 20, x	0.5455	0.2317	0.4782	0.1695	43234
ft021	100, 20, x	0.5688	0.2413	0.5000	0.1785	41803
ft022	100, 20, 3	0.5072	0.2033	0.4891	0.4017	5976
ft023	100, 50, x	**0.6004**	**0.2546**	**0.5198**	0.1047	88082
ft024	100, 50, 3	0.5303	0.2142	0.5086	**0.4201**	5996
ft025	200, 50, x	**0.6122**	**0.2574**	**0.5294**	0.1067	86576
ft026	200, 50, 3	0.5426	0.2181	0.5180	0.4287	5996
ft027	200, 50, x, 0.5	0.4683	0.1717	0.4580	**0.4959**	3085
ft028	200, 100, x	**0.6192**	**0.2616**	**0.5305**	0.0682	139540
ft029	*200, 100, 3*	*0.5432*	*0.2188*	*0.5182*	*0.4298*	6002
ft030	200, 100, 5	0.5819	**0.2613**	**0.5293**	0.3610	10005
ft031	200, 100, x, 0.5	0.4576	0.1654	0.4521	**0.4924**	2859
Max		0.6872	0.2905	0.5943	0.5269	139540
Min		0.2783	0.1249	0.1961	0.0682	2859
Avg		0.5288	0.2238	0.4540	0.3119	21057

Before this formal subtask, there is a dry run for participants to get acquainted with the whole process. The collection for the dry run is similar to that for the formal run, except that the year range is from 2000 to 2002.

Since patents are long documents and each segment has a specific function in describing the invention, we experiment on the use of different segments for TC in the dry run stage. Specifically, the following segments are used:

PAJ: Patent Abstract in English (from PAJ)
BIJ: Bibliography in JA
ABJ: Patent Abstract in JA
CLJ: Patent Claims in JA

Table 5 shows our results and the max, min, and average results from all the participants. There are three parameters separated by comma in the second column. The first parameter denotes the selected document surrogate, where BAC is the combination of the segment BIJ, ABJ, and CLJ. The second is the number of terms selected from the document to be classified for KNN calculation. Document terms are ranked by term frequency and then T top-ranked terms are selected. The third is the number K in the K-Nearest Neighbor method. Table 5 shows that the larger the T and K, the better the performance in terms of A-precision (average precision), Recall at top 5 suggested categories, and R-Precision (precision at top R suggested categories, where R is the number of relevant categories for the patent to be classified). However, in terms of F-measure, the less the T and K, the better the performance because less false drops result, as can be seen from the last column which indicates the total number of categories suggested. As to the difference in the document surrogates, the use of the combination of segment BIJ, ABJ, and CLJ does not perform better than the use of PAJ alone, regardless of which T and K are chosen. This may be due to the imperfect term indexing for Japanese based on the keyword extraction algorithm described above. Or it may be due to the reason that the conciseness of the English abstract does lead to better theme-based categorization.

Based on the above result, we only use PAJ as the patent surrogate in the formal run, because far more documents exist in the formal-run training set. Two additional parameters C and S are used, where C denotes the maximum number of suggested categories allowed and S denotes the threshold of the category confidence from the KNN calculation. In the second column of Table 6, the value x of C means no limitation on the suggested number of categories, while the omission of S means no threshold is applied. The value 0.5 of S means that the calculated categories whose confidence is half less than that of the top-ranked category are removed from being suggested.

Again, the higher the T and K without any limitation on C and S, the better the result in terms of the first three performance metrics. In contrast, the best F-measure is obtained when more conservative strategies are used in setting the parameters. To obtain high performance in terms of all metrics, the parameter setting (T, K, C) = (200, 100, 3) in italic in Table 6 seems to be the best.

5. Conclusions

We have used an operational TC system to participate in the Classification Subtask. Both efficiency and effectiveness are our concern in designing this system. By leveraging the factors that affect the performance based on the knowledge and experience learned from our previous experiments on other test collections, we show that an above-average performance can be obtained for the patent classification task. Our approach uses KNN as the classifier and the PAJ as the patent surrogate, such that all the training documents can be indexed on a low-cost machine in a short period of time. Although the simple vector space similarity model (byte size normalization) used in the KNN kernel performs slightly worse than the probabilistic model (BM11), it shortens the classification time by two to ten folds, depending on the number of training documents. Depends on which effectiveness metrics are used, we can optimize the performance by tuning the parameters as shown in Table 5 and 6. However, an ideal and robust way is to devise an efficient method that is high-performing in all effectiveness metrics. This would be our goal in the future workshops.

Because of the vast number of F-term categories, we did not participate in the F-term categorization subtask. After modifying the system’s limitation, we hope next time we can investigate our approaches and strategies to see if they work for the F-term classification.

Acknowledge

This work is supported in part by NSC under the grant number NSC 94-2213-E-003 -012 – and NSC 94-2524-S-003-014-.

References

[1] Sungil Jung, “Importance of Using Patent Information”, in the WIPO-Most Intermediate Training Course on Practical Intellectual Property Issues in Bussiness, organized by the World Intellectual Property Organization (WIPO), Geneva, Nov. 10-14, 2003.

[2] Campbell, R. S., “Patent Trends as a Technological Forecasting Tool”, World Patent Information , Vol. 5, No. 3, 1983, pp. 137-143.

[3] Shang-Jyh Liu, “Patent Map - A Route to a Strategic Intelligence of Industrial Competitive-ness,” The first Asia-Pacific Conference on Patent Maps, Taipei, Oct. 29, 2003, pp. 2-13.

[4] Young-Moon Bay, “Development and Applications of Patent Map in Korean High-Tech Industry” The first Asia-Pacific Conference on Patent Maps, Taipei, Oct. 29, 2003, pp. 3-23.

[5] Yuen-Hsien Tseng, Dai-Wei Juang, Yeong-Ming Wang, and Chi-Jen Lin, "Text Mining for Patent Map Analysis," Proceedings of IACIS Pacific 2005 Conference, May 19-21, 2005, Taipei, Taiwan, pp.1109-1116.

[6] Yiming Yang and Xin Liu, “A Re-Examination of Text Categorization Methods,” Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 42 – 49.

[7] Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proceedings of the European Conference on Machine Learning, 1998, Berlin, pp. 137-142.

[8] Yuen-Hsien Tseng, “Effectiveness Issues in

Automatic Text Categorization," (in Chinese) Bulletin of the Library Association of China, Vol. 68, June, 2002, pp. 62-83.

[9] Yuen-Hsien Tseng and Da-Wei Juang, "Document-Self Expansion for Text Categorization," Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '03, July 28 - Aug. 1, Toronto, Canada, 2003, pp.399-400.

[10] Da-Wei Juang and Yuen-Hsien Tseng, "Uniform Indexing and Retrieval Scheme for Chinese, Japanese, and Korean," Proceedings of the Third NTCIR Workshop on Evaluation of Information Retrieval, Automatic Text Summarization and Question Answering, Oct. 8-10, 2002, Tokyo, Japan, pp.137-141.

[11] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[12] Yiming Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the International Conference on Machine Learning (ICML'97), 1997, pp. 412-420.

[13] Hwee Tou Ng, Wei Boon Goh and Kok Leong Low, "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization," Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1997, Pages 67 – 73.

[14] Ron Bekkerman, Ran El-Yaniv, Yoad Winter, Naftali Tishby, "On Feature Distributional Clustering for Text Categorization," Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001, pp.146-153.

[15] C. J. Fall, A. Torcsvari, K. Benzineb, G. Karetka, "Automated Categorization in the International Patent Classification," ACM SIGIR Forum, Vol. 37, No. 1, 2003, pp. 10 - 25.

[16] Thorsten Joachims, SVMlight: Support Vector Machine, version 5, http://svmlight.joachims.org/.

[17] Amit Singhal, Gerard Salton, and Chris Buckley, "Length Normalization in Degraded Text Collections," Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 149-162.

[18] S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, August 1994, pp. 232-241.

[19] Yuen-Hsien Tseng , *FJU CTC : Test Collection for Chinese Text Categorization,* http://www.lins.fju.edu.tw/~tseng/Collections/Chinese_TC.html

[20] Yuen-Hsien Tseng and William John Teahan, "Verifying a Chinese Collection for Text Categorization," Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '04, July 25 - 29 Sheffield, U.K., 2004, pp.556-557.