Verifying a Chinese Collection for Text Categorization

Yuen-Hsien Tseng
Dept. of Library & Information Science
Fu Jen Catholic University
Taipei, Taiwan, R.O.C. 242
tseng@lins.fju.edu.tw

William John Teahan
School of Informatics,
University of Wales, Bangor
United Kingdom
wjt@informatics.bangor.ac.uk

ABSTRACT
This article describes the development of a free test collection for Chinese text categorization. A novel retrieval-based approach was developed to detect duplicates and label inconsistency in this corpus and in Reuters-21578 for comparison. The method was able to detect certain types of similar and/or duplicated documents that were overlooked by an alternative repetition-based method [1]. Experiments showed that effectiveness was not affected by the confusing documents.

Categories and Subject Descriptors
[**INFORMATION STORAGE AND RETRIEVAL**]: Information Storage – *Record Classification.*

General Terms
Algorithms, Documentation, Performance, Experimentation, Reliability, Standardization, Languages, Verification.

Keywords
Chinese collection, duplicate detection, consistency verification.

1. Introduction
Test collections such as OHSUMED, Reuters-21578, 20NG, and RCV1 have been important resources for the evaluation of automatic text categorization (TC) technologies. However, they are not perfect; duplicates and label inconsistencies that may bias classifiers' behavior occur in these corpora. In the development of a new collection for Chinese TC, these anomalies were detected before its release to the public to show its possible impact on the effectiveness of the classifiers.

The Chinese test collection, called FJU CTC, originated 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.

To develop this corpus into a test collection for TC, we follow these guidelines: (1) As many documents are included to better utilize the label information that already exists. (2) Each category has 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.

SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Accordingly, a total of 28011 documents were identified. They were divided into a training set of 19901 documents and a test set of 8110 documents. Each document was labeled with one to four categories. The average number of labels per documents is 1.0286. In total there are 82 categories. 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 http://www.lins.fju.edu.tw/~tseng/Collections/Chinese_TC.html. Since the documents spread over 17 years, consistency of the label assignment may be a problem. This was verified below.

2. Verification of the Test Collection
In [1], R-measure was proposed as a means to detect duplicates and label inconsistency in test collections. The R-measure of a document is a normalized sum of lengths of all substrings of the document that are repeated in other documents of the collection. But direct use of the R-measure as in [1] is problematic for some information retrieval tasks as texts are often stemmed, stopword-stripped, and noise- or punctuation-removed. Hence the number of duplicates can be different under different indexing criteria. Although R-measure can be applied after text cleaning and word normalization, words may be in slightly different order among plagiarisms, a case that may reduce the effectiveness of R-measure since it relies on exact order of substrings for similarity detection. Thus we additionally used a novel retrieval-based approach to verify this Chinese collection.

In this approach, the whole documents are indexed by a retrieval system. Then each document is submitted as a query to match against all documents by using a retrieval model that reports similarity scores scalable between 0 and 1. A report is generated which produces a list of similar documents with similarity scores for each document. From the list, statistics about the duplicates and label inconsistency can be obtained by giving a score threshold and the label information of each document.

This approach has been applied to the Chinese collection and the widely used ModApte split of the Reuters-21578 corpus. A commercial retrieval system [2] was used for indexing and retrieval. In Reuters-21578, English words were stemmed, but no stopwords were used. Thus every English word was submitted for retrieval. This amounts to 133 words per document on average. In the Chinese collection, texts were segmented using existing index terms based on a non-backtracking and longest-match principle. Furthermore, functional words and non-semantic bearing words were removed to reduce submitted terms and save retrieval time. The average submitted words were estimated to be 38.42% or 88 words per document. The system uses a vector-space model for document retrieval, where bytesize instead of cosine is used to normalize the document length [3]. Since bytesize normalization

is an approximation of cosine normalization, similarity scores sometimes were larger than 1. In such cases, they were set to 1.

Results of the retrieval-based approach are shown in Table 1. Compared to the R-measure in [1], which reported 177 duplicates in Reuters-21578, this approach (S-measure for simplicity) detected 269 duplicates. An additional fuzzy string matching technique based on dynamic programming (D-measure) was applied to double-check the 269 documents. All were confirmed to be real duplicates. The most dissimilar pair was 2782 and 2880 (with R=0.17, D=0.83, and S=1.0, all are the larger one between the pair) as shown in Table 2. As can be seen, these two documents are similar in fragments, a case that R-measure overlooks. At low S level, similar documents can still be found. Examples are document 519 and 7204 which have S=0.53. Document 519 contains only the title: "FED SETS 1.5 BILLION DLR CUSTOMER REPURCHASE, FED SAYS", and 7204 is also a title: "FED SAYS IT SETS 1.5 BILLION DLRS OF CUSTOMER REPURCHASE AGREEMENTS". In contrast, the R- and D-measure for these two documents are 0.15 and 0.37, respectively.

For the Chinese collection, 113 documents were found with R=1.0 and 1965 documents were found with S=1.0. The same fuzzy string matching detects 22 documents among the 1965 with fuzzy similarity lower than 0.5. A number of human inspections revealed that even among these dissimilar documents, most are plagiaries of the others in parts or in whole. The Chinese collection not only has a high percentage of similar documents, but also has an extreme distribution of label inconsistency among these similar documents, compared to the English corpus.

3. Effectiveness Experiments

The label inconsistency between similar documents may confuse any machine learning classifiers such that effectiveness becomes unpredictable and unreliable. To have an idea of what this effect may have on these collections under the inconsistency level we found, we removed these similar documents from the collections to see how the effectiveness changes.

We used KNN classifiers for both collections. Since KNN's performance highly depends on the similarity measure, we used a bytesize normalized vector-space model (VSM) [3] and a probabilistic model called BM11 [4] for comparison. The K in the KNN method was set to 20 in all our experiments. Each document in the Chinese collection was assigned exactly one category from those that have the maximum category score of the classifier. Each document in the English collection was assigned at most two categories, one with the largest category score, the other with the second largest. The second ranked category was not assigned to the document if its score is less than the half of the largest. Feature selection was based on averaged chi-square values. Terms whose document frequency is less than 2 were not used for both collections. Both microF and macroF were reported as the effectiveness measures.

Results are shown in Table 3. The BM11 retrieval model performs better than the VSM in all cases. In the Chinese corpus, although inconsistent similar documents are far more than the consistent ones, the effectiveness was not getting obviously better as more confusing documents were removed. In the English corpus, the effectiveness was not getting obviously worse as more consistent similar documents were removed. This suggests that given these levels of duplication and given these levels of consistency imbalance, the classifiers were not affected.

4. Conclusions

Our experiments showed: (1) Better classifiers perform better independent of duplicates and label inconsistency. (2) The high percentage of the confusing documents in the Chinese corpus does not prevent it from being a test collection candidate for TC. Actually, the 17 years of the collection time span provides another facet of TC that may be worthy of further exploration.

Table 1: Percentage of similar documents and label inconsistency at different S levels, where S is similarity score.

Collections	label	S>=1.0		S>=0.8		S>=0.6	
		# Doc	%	# Doc	%	# Doc	%
FJU CTC	Incon.	1922	6.86	7894	28.18	9618	34.34
	Con.	43	0.15	351	1.25	2438	8.70
	Total	1965	7.02	8245	29.43	12056	43.04
Reuters-21578	Incon.	62	0.57	121	1.12	246	2.28
	Con.	207	1.92	628	5.82	1165	10.80
	Total	269	2.49	749	6.94	1411	13.08

Table 2: The most dissimilar pair with S=1.0. Only the first paragraph is shown. The 2nd paragraph of both is the same.

2782: TECHNIGEN PLATINUM CORP> IN METALS FIND. Technigen Platinum corp said it initial results of a 13-hole drilling program on its R.M. Nicel platinum property in Rouyn-Noranda, Quebec, indicate "extensive" near-surface zones "highly" enriched in gold, platinum and palladium were found in rocks on the periphery of a sulphide deposit.
2880: TECHNIGEN PLATINUM CORP IN METALS FIND. Technigen Platinum corp said initial results of a 13-hole drilling program on its R.M. Nicel platinum property in Rouyn-Noranda, Quebec, indicate extensive near-surface zones highly enriched in gold, platinum and palladium. The metals were found in rocks on the periphery of a sulphide deposit.

Table 3: MicroF (upper half) and MacroF (lower half).

Collection	Model	All docs.	S<1.0	S<0.8	S<0.6
FJU CTC	VSM	0.4383	0.4412	0.4494	0.4504
	BM11	0.4605	0.4679	0.4787	0.4788
Reuters-21578	VSM	0.8192	0.8170	0.8161	0.8156
	BM11	0.8240	0.8253	0.8247	0.8232
FJU CTC	VSM	0.2881	0.2916	0.2844	0.2844
	BM11	0.3177	0.3236	0.3182	0.3080
Reuters-21578	VSM	0.3681	0.3676	0.3603	0.3428
	BM11	0.4211	0.4206	0.4072	0.3823

5. Acknowledgments

This work is partly supported by NSC 92-2213-E-030-017-.

6. References

[1] Dmitry V. Khmelev and William J. Teahan, "A Repetition Based Measure for Verification of Text Collections and for Text Categorization," ACM SIGIR, 2003, pp.104-110.

[2] WebGenie 3.23, http://www.webgenie.com.tw

[3] Amit Singhal, Gerard Salton and Chris Buckley, "Length Normalization in Degraded Text Collections" Symp. on Document Analysis and Info. Retr., 1996, pp. 149-162.

[4] S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," Proc. of ACM SIGIR, 1992, pp.42-49.