The Impact of OCR Accuracy on Automatic Text Classification

Guowei Zu[1,2], Mayo Murata[1], Wataru Ohyama[1],
Tetsushi Wakabayashi[1], and Fumitaka Kimura[1]

[1] Mie University, Faculty of Engineering,
1515 Kamihama-cho, Tsu-shi, Mie, 5148507, Japan
http://www.hi.info.mie-u.ac.jp/
[2] Toshiba Solutions Corporation, Systems Integration Technology Center
Toshiba Building, 1-1, Shibaura 1-chome, Minato-ku, Tokyo 105-6691, Japan

Abstract. Current general digitization approach of paper media is converting them into the digital images by a scanner, and then reading them by an OCR to generate ASCII text for full-text retrieval. However, it is impossible to recognize all characters with 100% accuracy by the present OCR technology. Therefore, it is important to know the impact of OCR accuracy on automatic text classification to reveal its technical feasibility. In this research we perform automatic text classification experiments for English newswire articles to study on the relationships between the accuracies of OCR and the text classification employing the statistical classification techniques.

1 Introduction

With the development of internet and the information processing technology in these ten years, the main ways and means of information exchange has been shifted from the traditional paper to the digital data. Because the digital data, e.g. text, image, audio and so on, is transferred and retrieved much more quickly and easily, the digital publishing and the digital library will become the main resources of information in the twenty first century. Therefore, the traditional library should consider converting a great deal of paper media to the digital data in order to provide them on the internet. Current general digitization approach is converting paper media into the digital images by a scanner, and then reading them by an OCR to generate ASCII text for full-text retrieval. However, it is impossible to recognize all characters with 100% accuracy by the present OCR technology. Especially, the recognition accuracy can be quite low for classic books with special character font and handwritings. Therefore, it is important to know the impact of OCR accuracy on automatic text classification to reveal its technical feasibility.

In this research we perform automatic text classification experiments for English newswire articles to study on the relationships between the accuracies of OCR and the text classification employing the statistical classification. While the impact of OCR accuracy on information retrieval has been studied and reported in [1], [2], the impact on text classification has not been reported, to the best knowledge of the authors.

2 The basic classification technology

In this research we employed the statistical classification technique for classifying a feature vector composed of frequencies of lexicon words that appear in a text. The approach is learning a classification scheme from labeled training examples then using it to classify unseen textual documents [3]. Several classification techniques based on the Euclidean distance, Fisher's linear discrimination function, projection distances [4], and the support vector machine (SVM) are employed in the classification test for the English text collection (the reuters-21578).

A drawback of the statistical classification technique is that the dimensionality of the feature vector can increase together with the lexicon size. For example, the lexicon size and the feature dimensionality grow to 34,868 for Reuters-21578 articles, which requires enormous computational time and storage for the text classification. To solve this problem we need to employ a statistical feature extraction technique which extracts small number of features with high separability to reduce the feature dimensionality without sacrificing the classification accuracy. In this research the dimension reduction based on the principal component analysis (PCA) was employed.[5]

3 The classification experiment

3.1 Used Data

To study the impact of OCR accuracy on the automatic text classification, a set of texts that are pre-classified to their true category is required. The Reuters-21578 test collection is frequently used by many researchers as a typical test collection for English text classification. The Reuters-21578 is composed of 21578 articles manually classified to 135 categories. The Reuters-21578 data is a set of ASCII texts in which the predefined marks are embedded according to the SGML format to indicate the structural elements of the text. In this experiment total of 750 articles, 150 articles/category randomly selected from five categories (acq, crude, earn, grain, trade), were used. Since the sample size is not enough large, the sample is divided into three subsets each of which includes 50 articles/category. When a subset is tested, the rest of the two subsets are used as learning sample in order to keep the learning sample size as large as possible while keeping the independency between the samples for learning and test. Classification tests are repeated for three subsets and the correct classification rates are averaged to evaluate the classification accuracy.

3.2 The procedure of the experiment

The procedure of the experiment consists of three general steps for (1) text image generation, (2) ASCII text generation by OCR and (3) the automatic text classification.

Text image generation

Each ASCII text of the Reuters collection is printed out by a LBP with the character size of times11 point. The text on paper is converted to a digitized image of 300 dpi by a scanner. In order to obtain the digitized text images of different OCR accuracies, Photoshop 6.0 software was used to intentionally reduce the dpi of the images to 240, 200, 150, 145, and 140. Figure 1(a)-(c) show the example of the text images of 300dpi, 150dpi and 140dpi respectively.

ASCII text generation by OCR

The text images generated in above are converted to the ASCII texts by OCR software "OKREADER2000" (Figure 2(a)-(c)).

The obtained ASCII text is compared with the original ASCII text in the Reuters collection to calculate the average character recognition rate and the average word recognition rate for each dpi. The average character recognition rate is defined by

$$c = \frac{(s-t)}{s} \times 100 \tag{1}$$

,where s and t is the number of total characters and the number of mis-recognized characters, respectively. The average word recognition rate is defined by

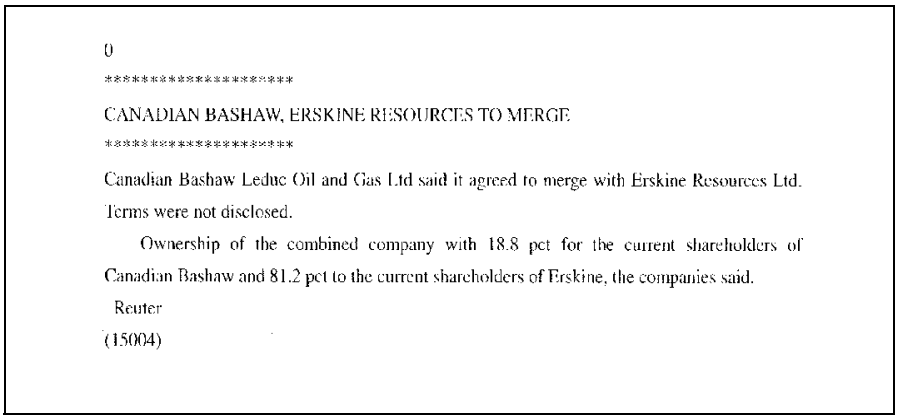
$$v = \frac{(w-u)}{w} \times 100 \tag{2}$$

, where w and u is the number of total words and the number of mis-recognized words, respectively.

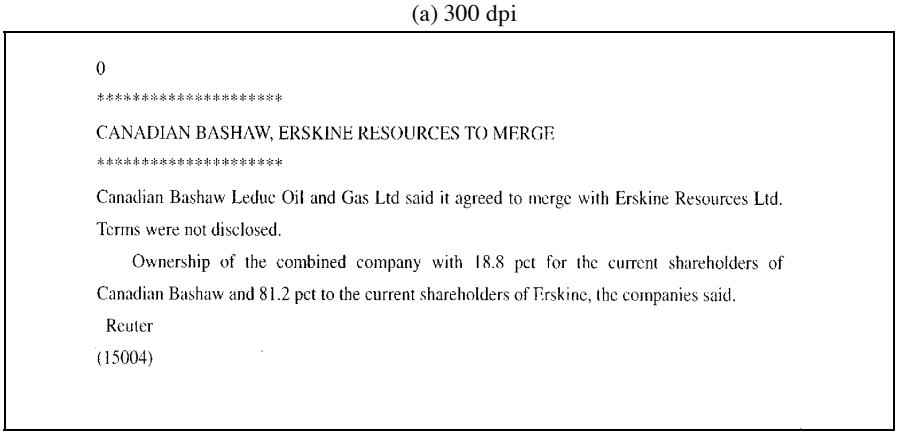
Text Classification

Feature vector generation

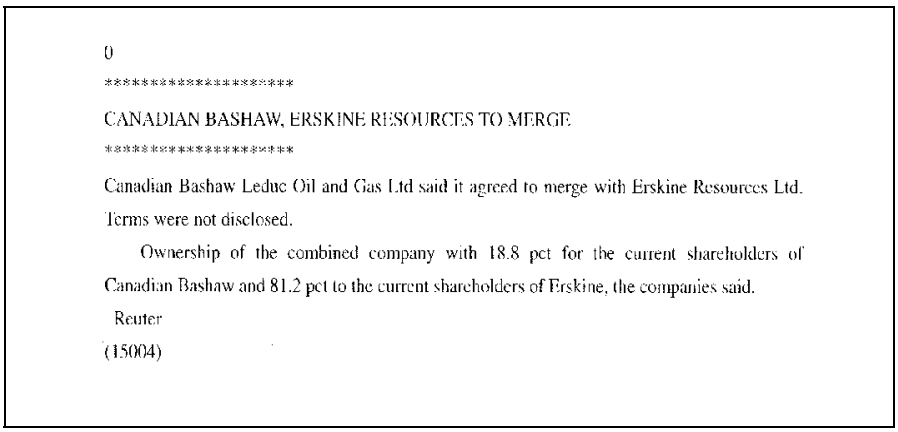
A lexicon is generated from the learning sample by picking up all words in the sample. Then the feature vector for each text is composed of the frequencies of the lexicon words in the text. The dimensionality of the feature vector is equal to the lexicon size and is denoted by n.


0
****
CANADIAN BASHAW, ERSKINE RESOURCES TO MERGE
****
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed to merge with Erskine Resources Ltd. Terms were not disclosed.
Ownership of the combined company with 18.8 pct for the current shareholders of Canadian Bashaw and 81.2 pct to the current shareholders of Erskine, the companies said.
Reuter
(15004)

(a) 300 dpi


0
****
CANADIAN BASHAW, ERSKINE RESOURCES TO MERGE
****
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed to merge with Erskine Resources Ltd. Terms were not disclosed.
Ownership of the combined company with 18.8 pct for the current shareholders of Canadian Bashaw and 81.2 pct to the current shareholders of Erskine, the companies said.
Reuter
(15004)

(b) 200 dpi


0
****
CANADIAN BASHAW, ERSKINE RESOURCES TO MERGE
****
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed to merge with Erskine Resources Ltd. Terms were not disclosed.
Ownership of the combined company with 18.8 pct for the current shareholders of Canadian Bashaw and 81.2 pct to the current shareholders of Erskine, the companies said.
Reuter
(15004)

(c) 140 dpi

Fig. 1. Examples of the text images

CANADIAN BASHAW, ERSKINE RESOURCES TO MERGE
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed to merge with Erskine Resources Ltd.
Terms were not disclosed.
Ownership of the combined company with 18.8 pet for the current shareholders of Canadian Bashaw and 81.2 pet to the current shareholders of Erskine, the companies said.
Reuter

(a) The ASCII text of 300dpi

CANADIAN BASHAW, ERSKINR RESOURCES TO MERGE
Canadian Bashaw Leduc Oil and Gas Ltd said it agreed lu merge with Rrskme Resources Lid.
Terms were not disclosed.
Ownership of the combined company with 18.8 pet tor the current shareholders of
Canadian Bashaw and 81.2 pel to the current. shareholders ofF.rskinc, I he companies said.
Keuler

(b) The ASCII text of 200dpi

CANADIAN yASHAW. HR^KJNh kI';SOlJRCn,S TO V1HRGR
Canadian Bashaw Leduc Oil and Gas I.Id s;ilii IT agreed ro merge wirii Hnkine RL^OUI-CCS Ltd.
Forms were not disclosed.
Ownership of ihe combined company with 1^.8 pet for the cuncnt shareholders of
C,m;idi,in B^sbaw and 81.2 pel to ihc eiirrL'nt sbaicholLlcrs cifF.rskrne, die companies said.
Rt-'uter

(c) The ASCII text of 140dpi

Fig. 2. Examples of the ASCII texts converted by OCR software

Dimension reduction

At first the total covariance matrix of the learning sample is calculated to find the eigenvalues and eigenvectors. Each feature vector is transformed to the principal components in terms of the orthonormal transformation with the eigenvectors as the basis vectors. To reduce the dimensionality of the feature vector the principal components which correspond to the m largest eigenvalues are selected to compose the feature vector of dimensionality m ($< n$).

Learning

Parameters of each classification technique are determined in the training process using the learning sample. The Euclidean distance classifier employs the mean vector of each class. The linear discriminant function employs the weight vector determined by the mean vector of each class and the pooled within covariance matrix of entire classes. The projection distance and the modified projection distance employ the eigenvectors (and the eigenvalues) of the individual covariance matrix. As a support vector machine (SVM), C-support vector classification method (C-SVC) of linear type and of RBF type (with radial basis function) were employed for the classification tests. We used the SVM library (LIBSVM Version 2.33) developed by Chang and Lin (2002) [6].

Classification

The feature vector of reduced dimensionality is classified to the class the distance (or the discriminant function) of which is minimized. Referring to the subject field manually given to each article in Reuters-21578, the classification rate R is calculated by

$$R = \frac{x}{(x+y)} \times 100 \tag{3}$$

, where x and y is the number of articles correctly classified and incorrectly classified, respectively.

4 The experiment results

Table1 shows the character recognition rates and the word recognition rates for different dpi's. Table 2 shows the text classification rates of each classification technique for different character recognition rates, and for different word recognition rates. Figure3 shows the relationship between the text classification rate and the word recognition rate, and Figure 4 shows the relationship between the text classification rate and the character recognition rate.

Table 1. the character recognition rates vs. the word recognition rates

Resolution (dpi)	300	240	200	150	145	140
Word recognition rate (%)	97.54	94.84	92.69	82.77	66.40	63.33
character recognition rate (%)	99.28	98.72	98.14	95.35	91.14	90.04

Table 2. the text classification rates vs. character recognition rates and word recognition rates

Word recognition rate	100	97.54	94.84	92.69	82.77	66.40	63.33
character recognition rate	100	99.28	98.72	98.14	95.35	91.14	90.04
Euclidean distance	83.6	83.2	82.4	81.2	78.4	71.2	66.0
Linear discrimination function	95.2	95.2	93.6	94.0	93.2	86.8	86.4
projection distance	93.2	93.2	92.4	92.8	90.4	89.2	88.6
modified projection distance	95.2	95.2	95.2	94.0	93.2	92.8	91.2
SVM-Linear	95.6	95.6	95.6	95.2	95.2	93.6	93.2
SVM- RBF	93.6	93.6	93.2	92.4	92.4	92.0	89.6

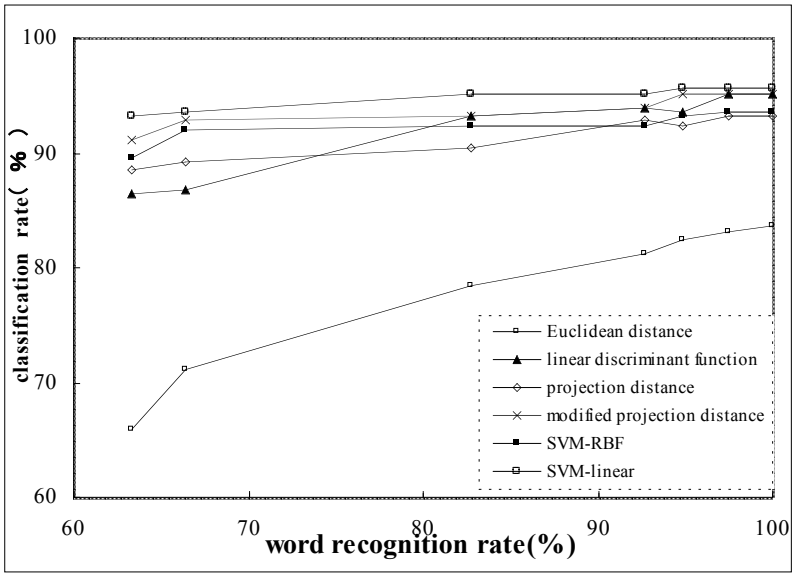


Fig. 3. the text classification rate vs. the word recognition rate

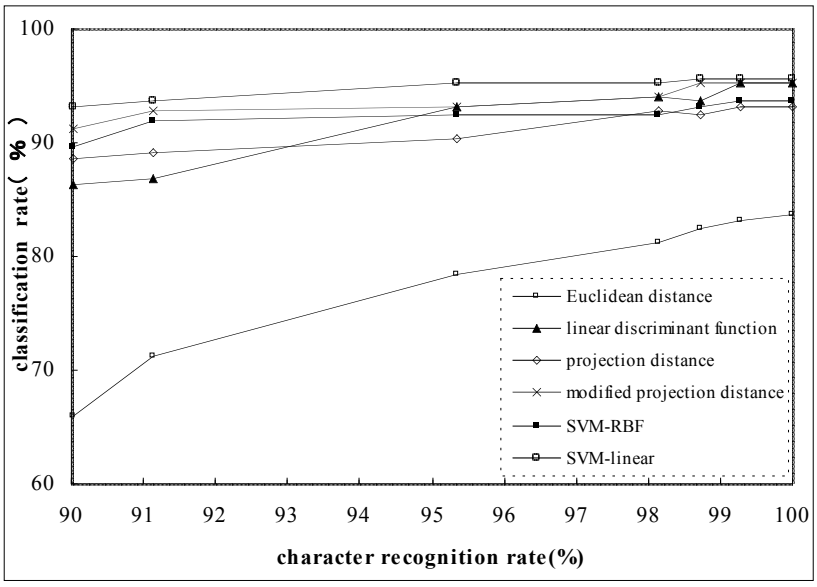


Fig. 4. the text classification rate vs. the character recognition rate

The results of experiment are summarized as follows:

1. The text classification rates of all classification techniques were not deteriorated significantly until the character recognition rate or the word recognition rate was deteriorated to 95% or 80%, respectively.
2. The text classification rate for the modified projection distance and the SVM-linear was kept over 90% even when the character recognition rate or the word recognition rate was further deteriorated to 90% or 60%, respectively.
3. The text classification rates for the linear discriminant function and the Euclidian distance were more rapidly deteriorated than other techniques.
4. The SVM-linear outperformed the others in the accuracy and the robustness of the text classification in this experiment.

5. The future study

In the experiment we dealt with five category case and obtained encouraging result, however, we need to deal with more categories in real world application of text classification. We will perform similar experiment with more categories to reveal the feasibility of the OCR input text classification.

Error correction of words by spelling check is also remaining as a future study to improve the text classification accuracy.

References

[1] Ohta,M., Takasu,A., Adachi,J.: "Retrieval Methods for English-Text with Missrecognized OCR Characters", *Proceedings* of the Fourth International Conference on Document Analysis and Recognition (ICDAR), pp.950-956, August 18-20, 1997,Ulm, Germany.

[2] Myka, A., Guntzer. U.: "Measuring the Effects of OCR Errors on Similarity Linking", *Proceedings* of the Fourth International Conference on Document Analysis and Recognition (ICDAR), pp.968-973, August 18-20, 1997, Ulm, Germany.

[3] Sebastiani, F.: "Machine Learning in Automated Text Categorization", *ACM Computing Surveys,* Vol. 34, No. 1, 1-47, March 2002.

[4] Fukumoto,T., Wakabayashi,T. Kimura,F. and Miyake,Y.: "Accuracy Improvement of Handwritten Character Recognition By GLVQ", *Proceedings* of the Seventh International Workshop on Frontiers in Handwriting Recognition Proceedings(IWFHR VII), 271-280 September 2000.

[5]Guowei Zu, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura,: "Accuracy improvement of automatic text classification based on feature transformation" *DocEng'03* (ACM Symposium on Document Engineering 2003), pp.118-120, November 20–22, 2003, Grenoble, France

[6] C.C. Chang, and C.J. Lin : “LIBSVM -- A Library for Support Vector Machines（Version 2.33)”, *http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html,* (2002.4)