2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.128

Performance Analysis of Ad Hoc Classifiers for Categorization of Uyghur texts

Palidan Tuerxun[1, 2], Fang Dingyi[1], and Askar Hamdulla[2]

[1] School of information and technology, Northwestern University, Xi'an, China

[2] School of Software, Xinjiang University, Urumqi, China

Keywords: Uyghur; text classification; stemming; classifier.

Abstract. This paper starts from the characteristics and writing rules of Uyghur, have established a relatively large text corpus which include 20 categories, 300 documents for each category. And studied the KNN, Naive Bayes (NB), and SVM classification algorithms more thoroughly, which have widely been used in domestic and foreign academic research fields, then classified the Uyghur text by using these algorithms, and analyzed the performance of each algorithm separately. Finally, some research directions on Uyghur text classification are also given in this paper.

Introduction

With the rapid development of computer and network technology, the internet has received a wide popularity from all corners of society. The rapid growth of Web information brought a challenge to information retrieval, and made our work more difficult to find the needed information from it. Text classification plays a key and effective role for dealing with clutter information, also has important applications in areas such as information retrieval, search engines, digital library management.

With the rapid development of information construction in Xinjiang Uyghur autonomous region, the west part of China, a lot of text information in Uyghur language and other minority languages are being presented in digital form and the text information is growing continually. A vast amount of paper-based text information which had been accumulated in the past begins to be stored in digital form. Many application areas are requiring a computer automatic classification method to integrate and use the mass text messages effectively. So, how to automatically classify a large number of text data of minority languages has become an important research topic of natural language processing of minority languages in Xinjiang, including the Uyghur language.

Uyghur language belongs to Turkic family of Altaic language system. It has some features listed as follows: (1) the writing direction of Uyghur is from right to left, from top to bottom. (2) Uyghur has all of 32 letters in which some characters are borrowed from Arabic and Persian. (3) Uyghur text is completely different from the Chinese and English due to its agglunatative nature. In this type of language, the word (word) is the smallest independent unit. Word is constituted of one or more letters which are connected to the front and rear. (4)Uyghur word is constituted by a root or stem and additional word formation which usually follow after the formation of structure. The stem contains the lexical meaning of word, and it can be retrieved after removing all additional ingredients of word.

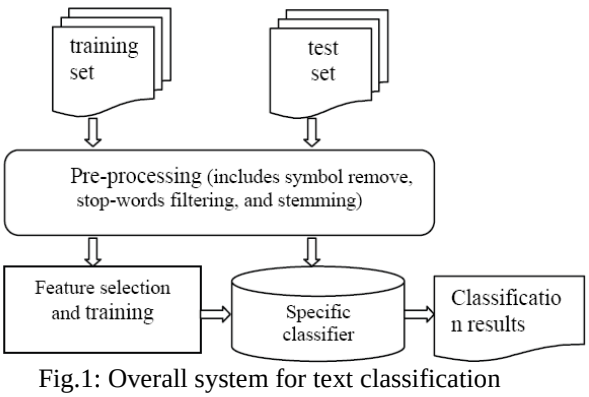

Fig.1: Overall system for text classification

Due to Uyghur text classification research started relatively late, the recall rate of those researchers such as Prof. Wenira team[1] and Dr. Alimujiang [2] and others in the Uyghur text classification are all about 70% (based on the experimental results of a small amount of expected).

In this paper, based on large-scale text corpus, with c # development platform, we designed and implemented the Uyghur text classification system based on the ad-hoc classifier such as KNN, NB and SVM shown in Figure 1, and the test and evaluation results will be given below.

Experiments setup

Corpus Design

As Uyghur text classification is a late-started research area, there is not publicly available text corpus for intensive study. Therefore, this work collected a large amount of internet based Uyghur text documents, and established a large scale text corpus library by classifying the documents into various categories using manual work.

The text corpus library consists some 20 categories(or class) of real-estate, health, education, military, tourist attractions, star (culture), people, mobile phone (common sense, market, transaction), literature, political issues, computer, traffic, economy, history, language habits and customs, the car (common sense, market, trade), social and legal system, sports, recruitment and employment, religion.etc,300 text documents for each text category, the text corpus contaied altogether 6000 Uyghur text documents. The configuration of text corpus participated in classification research in this paper are given in table 1.

Table 1: Experimental data

Corpus Name	Number of contained categories	Total amount of documents	The amount of trained documents	The amount of tested documents
D5	5	1500	5×220=1100	5×80=400
D10	10	3000	10×220=2200	10×80=800
D15	15	4500	15×220=3300	15×80=1200
D20	20	6000	20×220=4400	20×80=1600

The categories contained in Corpus named D10 are given in Table 2.

Table 2: Experimental text corpus in D10

Category Name	Number of text documents	Number of trained text documents	Number of tested text documents
Traffic	300	220	80
People	300	220	80
Sports	300	220	80
Health	300	220	80
Military	300	220	80
Real estate	300	220	80
Education	300	220	80
Scenic spots	300	220	80
Economy	300	220	80
Computer	300	220	80
Total number of text documents	3000	2200	800

The categories contained in Corpus named D20 are given in Table 3.

Table 3: Experimental text corpus D20

Category Name	Number of text documents	Number of trained text documents	Number of tested text documents
Traffic	300	220	80
People	300	220	80
Sports	300	220	80
Health	300	220	80

Military	300	220	80
History	300	220	80
Religion	300	220	80
Real estate	300	220	80
Mobile Phone	300	220	80
Recruitment & Employment	300	220	80
Politics	300	220	80
Education	300	220	80
Art	300	220	80
Scenic spots	300	220	80
Star	300	220	80
Ethnic& Customs	300	220	80
Vehicle	300	220	80
Society & Law	300	220	80
Economy	300	220	80
Computer	300	220	80
Total number of text documents	6000	4400	1600

Experimental environmnet configuration

The experiments in this paper are conducted on WIN7 operating system istalled PC，that with the hardware configuration of E7300 dual core, 2.66GH CPU, 2G RAM. and used the Microsoft C#2010 Express programming software.

Experiments and results

Experiment 1: Comparing the Precision of KNN，NB and SVM classifiers

Followings are the information of data, algorithms and parameters applied in classification Experiment 1 in this paper.

1. Text corpuses: D5, D10, D15, D20
2. Classification algorithms：KNN, NB, SVM
3. Feature selection algorithm：CHI
4. K value is 20
5. Preprocessing: symbol removing, stop-words filtering and stemming

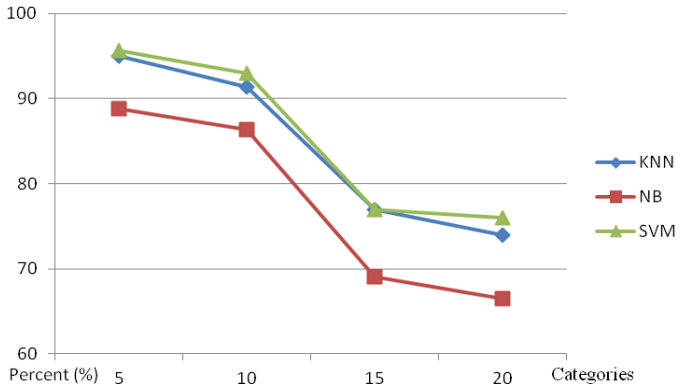

Fig.2: Comparing the Precision of KNN，NB and SVM classifiers

The results from experiment1 gives the new classification records of KNN，NB and SVM classifiers on Uyghur text categorization, with cotegory number of 5, 10, 15, 20 respectively. Fig.1 shows that SVM performs with higher precision than KNN and NB.

Experiment 2: Comparing the time consuming of KNN，NB and SVM classifiers

The information of data, algorithms and parameters applied in classification Experiment 2 are as follows:

1. Text corpuses: D5, D10, D15, D20
2. Classification algorithms：KNN, NB, SVM
3. Feature selection algorithm：CHI
4. K value is 20

5. Preprocessing: symbol remove, stop-words filtering and stemming

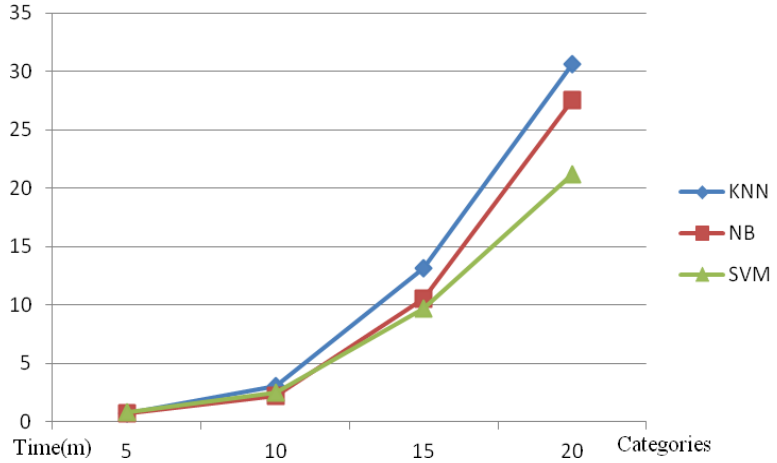

Fig.3: Comparing the time efficency of KNN，NB and SVM classifiers

Experiment 2 offers the time consuming of KNN，NB and SVM classifiers on Uyghur text categorization, with cotegory number of 5, 10, 15, 20 respectively. Figure 1 shows that SVM performs with less time consuming than KNN and NB.

Conclusions

This paper studied the widely used classification algorithms such as KNN, NB and SVM more thoroughly on Uyghur Text classification, and analyzed their performance. The experimental results show that SVM performs much better than KNN and NB both in time consuming and classification precision. Still some must do works such as phrases and concept analysis, N-gram introducing, improving the accuracy of stem extracting and spelling correction need special study.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China under grant of 61163033.

References

[1] Wangzhen. Uyghur, Kazak, Kirgiz automatic classification technology study results of search engine. Xinjiang university master's thesis, 2010.

[2] Alimjan et al. The Uyghur text classification based on machine learning research. *Computer Engineering and Applications*, (5), pp.110-112, 2012.

[3] Li tengfei, Li jun. *Data mining and knowledge discovery*. Beijing: Higher Education Press, 2003.

[4] Zhang pengzhao. Based X2 statistics of Chinese text classification feature selection methods research. Chongqing University master's thesis, 2008.

[5] Ahmatjan ablat, Turdi tuhti, Askar hamdulla. Naive Bayes based Uyghur text classification algorithms and performance analysis. *Computer Applications and Software*, 29(12), pp. 27-29, 2012,

[6] Wang xiaoqin. Chinese text classification feature selection methods research. Southwest University master's thesis, 2010.