DOI:10.26104/IVK.2019.45.557

ФИЛОЛОГИЯ ИЛИМДЕРИ
ФИЛОЛОГИЧЕСКИЕ НАУКИ
PHILOLOGICAL SCIENCES

Танг овогт Уонг Хенг

ТЕКСТ БОЮНЧА КАРДАРЛАРДЫН ПИКИРЛЕРИН КЛАССИФИКАЦИЯЛООНУН НАТЫЙЖАЛАРЫ

Танг овогт Уонг Хенг

РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ ОТЗЫВОВ КЛИЕНТОВ ПО ТЕКСТУ

Tang Yong Heng

RESULTS OF CLASSIFYING CUSTOMER COMMENTS BY TEXT

УДК: 83.02

Бул макаланын максаты монгол кириллицасы менен комментарийлерди чыгарып, текстти классификациялоо ыкмасынын негизинде позитивдүү жана терс мазмуну боюнча классификациялоо мүмкүнчүлүгүн текшерүү болуп саналат. Иштин жыйынтыгында көптөгөн колдонуучулардын туура жана туура эмес жазылышын карадык. Изилдөө учурунда ар башка маанидеги комментарийлерин топтоп чыгып, эки сөздүктү түзүп, колдондук. Алар оң жана терс жактары салыштырылды. Статистикалык кол менен классификациялоо жүргүзүлүп, KNN жана Naive Bayes тарабынан автоматтык түрдө баалоо алгоритми иштелип чыгып, текшерилген. Маалыматтарды классификациялоо үчүн KNN методунун тактыгы 87,50 пайызды жана Naive Bayes ыкмасынын тактыгы 87,50 пайызды түзөт. Бул изилдөөдөгү маалыматтар жана алгоритмдер монгол кирилл текстин ар кандай түрлөргө классификациялоо боюнча андан аркы изилдөөлөргө негиз боло алат.

Негизги сөздөр: *текст, ирettөө методу, кириллица эмоция, классификациялоо, оң жагы, терс жагы, туура, туура эмес, жазуу.*

Целью данной статьи является проверка возможности классификации текста на положительное и отрицательное содержание на основе метода классификации текста путем публикации комментариев на монгольской кириллице. В результате мы посмотрели, сколько пользователей написали правильно и неправильно. В ходе исследования мы собрали комментарии разного значения, создали и использовали два словаря. Они сравнили плюсы и минусы. Была проведена статистическая ручная классификация, и алгоритм автоматической оценки был разработан и протестирован KNN и Naive Bayes. Точность метода KNN для классификации данных составляет 87,50 процента, а точность наивного метода Байеса - 87,50 процента. Данные и алгоритмы этого исследования могут послужить основой для дальнейших исследований по классификации монгольского кириллического текста на разные типы: сторона, правильная, неправильная, написание.

Ключевое слова: *текст, метод сортировки, кириллица эмоция, классификация, положительная сторона, отрицательная сторона, правильный, неправильный, написание.*

The purpose of this article tests whether it is possible to classify comments with Mongolian Cyrillic text into positive and negative comments based on the text classification method. As a result of the work, we collected comments from many users with correct and incorrect spellings as well as different meanings to created data of two parts: positive and negative. The statistical classification was also performed, and an automatic feedback grading algorithm was developed and tested by KNN and Naive Bayes. The accuracy of the KNN method for data classification is 87.50 percent and the accuracy of the Naive Bayes method is 87.50 percent. The data and algorithms created by this work will be the basis for further research to classify the Mongolian Cyrillic text into different types.

Key word: *text, sorting method, cyrillic emotion, classification, positive side, negative side, correct, incorrect, spelling.*

Introduction. With the rapid development of information technology, the number of electronic sources of information available on social networks and the Internet are increasing exponentially. it is the opportunity to extract knowledge from others. users can get accurate and valuable information from the vast amount of information online. The text classification method can help solve this problem..

Text classification is an important part of language processing, information retrieval, data mining, and document sorting. It is possible to automatically determine customer satisfaction on social networks based on information mining technology. It is possible to determine a person's feedback or product satisfaction in a short time and at low cost by automatically identifying and calculating feedback.

Nowaday severyone is free to express their impressions by text on social networks. As a result, this type of information has become an important topic of research. Research has been conducted to categorize user reviews on social media sites such as Facebook, Twitter, and which are popular around the world. In this study, user feedback was often categorized as positive, simple,

DOI:10.26104/IVK.2019.45.557

or negative. From the above research, it can be seen that the database-based, machine learning methods have the highest results in the automatic determination of impressions. In recent years, the method of deep learning has become more suitable for classifying the specific features of a language and the sentences that are written illegally in the online environment.

We used some machine learning methods to collect comments in Mongolian Cyrillic to create a marked vocabulary then classify comments into positive and negative by KNN.

RELATED WORDS.

With the rapid development of information technology, the number of information sources on social networks has increased dramatically. For users, it is important to obtain the accurate and valuable information people need from a large amount of information available online. In recent years, the method of text classification has become widely used. For example Companies' products, in order to improve their service, we are able to understand customer satisfaction by receiving customer feedback on our products and services. There are many methods of classifying text, but there is a lack of experiments on data in Mongolian Cyrillic.

In the late 20th century, International researchers have begun to study the method of text classification. H. P. Luhn first introduced frequency statistics in the text category [1]. Since the 1990s, machine learning and statistical methods have provided more opportunities for text classification. It improved the accuracy of automatic text classification [2].

The purpose of this research is to extract user feedback from Mongolian Cyrillic text and it can be classified as good or bad according to the text classification method. Since there are still many methods to classify text, this study used the highly classified KNN method to test the Mongolian Cyrillic alphabet, which is now widely used in other languages. Chinese researcher Huang Juan Juan improved the performance of the KNN method by studying specific word weights, classification methods, and classification performance ratings [3]. Cheng Bo, a Chinese researcher developed a multi-level classification system for website texts. The test results show that the classification performance Cheng Bo, a Chinese researcher, developed a multi-level classification system for website texts. The test results show that the classification performance in the system is good in the system [4].

Chinese researcher Chen Jin Jie proposed a method for recognizing handwritten numbers based on the KNN algorithm. Based on the similarity of the sample classification, the KNN classification model was studied to identify the handwritten numbers through training. [5]. The method of the KNN classification algorithm and two different decision rules are introduced. Experiments show that the KNN method with similar weights is better. [6].

Based on the theory of Naive Bayes, a method of classifying emotions is proposed in a new generation of Chinese texts. Researcher (Yang Ding) uses emotion vocabulary to create text classifications based on Naive Bayes theory [7]. The naive Bayes classification algorithm, is effective and effective by the algorithm for grouping news texts on social networks [8].

Chinese researchers, Qi Yuan and Qiao Yu have suggested some ways to improve the accuracy of classifying by machine learning models. It develops and implements a text classification model, The python programming language is used to compare the test evaluation results that calculate the potential weight interaction and improve the calculation stability. [9]. Chinese researcher Jia Yun Fan has studied two methods, mainly KNN and SVM. The combination of SVM and KNN algorithms is very effective than KNN (K Nearest Neighbor) and SVM (Support Vector Network) to analyze the advantages and disadvantages of these two Chinese text classification methods. [10].

SURVEY OF METHOD

A. KNN (K-Nearest Neighbors) method

The KNN algorithm is widely used in the text classification industry, and the KNN algorithm is the most suitable classification algorithm for many text classification algorithms.

The K neighbors of the test text calculate the important features of each class. The calculation formula is as follows: Compare the values of the classification weights and divide the test text into the categories with the highest weights.

The KNN algorithm is often used as a model for Euclidean space classification. The Euclidean distance is defined as follows: For example, X=(x_1,x_2,…,x_n)、Y=(y_1,y_2,…y_n) Two points in N-dimensional space, then the Euclidean distance between these two points are:

$$d\ (x, y) = \sqrt{\sum_{k=1}^{n} (x_k, y_k)^2} \qquad (2)$$

The n is the space,, xk and yk are the k sequence attribute values of x and y. The KNN algorithm categorizes search text to find the nearest k training text according to the Euclidean distance formula. The "majority Suggestions" method then determines the search text classification among the K training text by the number of text types. For example, 1 is positive, 0 is negative, and 2 is neutral.

B. NBM (Naive Bayes Model) method

The Naive Bayes classifier is a linear classifier constructed using the Bayes theorem. This algorithm is

DOI:10.26104/IVK.2019.45.557

widely used in many fields and received well by industry professionals. The algorithm can be used to predict relationships between class members. Despite the shortcomings of the text classification algorithm, many experiments have shown that the algorithm has shown good classification performance. The calculation formula is as follows：

Set the text d, this text depends on a specific category C = { C_1,C_2,...... C_n} Middle Cj class. According to Base's law:

$$P(\boldsymbol{c_j}|\mathrm{d}) = \frac{P(c_j)P(d\,|c_j)}{\mathrm{p(d)}} \qquad (2)$$

Equal:

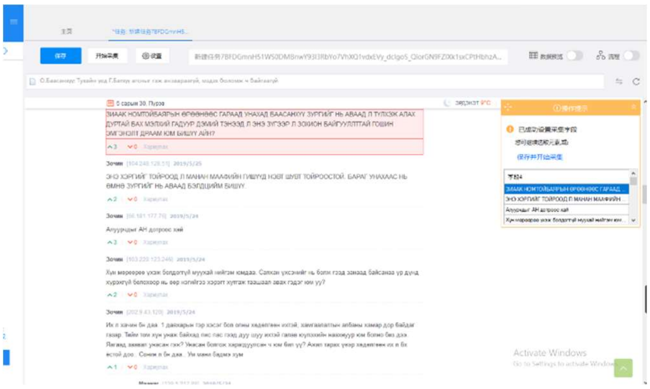
$$\mathrm{P(d)} = \textstyle\sum_{j=1}^{k} P(c_j)P(d|c_j) \qquad (3)$$

The document given by the above formula shows that the d belongs to the category Cj probability, calculates the value of P (Cj | d), ie the text d belongs to the category that calculates P (Cj | d) to obtain the maximum value, then:

$$\mathrm{P(Cj|d\,)} = MAX_{j=1}^{k}\ \{\mathrm{P}(c_j|\mathrm{d})\ \} \qquad (4)$$

C. Data collection

We collected user’s comments from the Mongolian news website IKON.MN by the Chinese software “Ba zhua yu”. The operating process of the program is as follows. As follows:



О.Баасанхүү Тухайн үед Г.Батхүү агсныг хэн ч анзааралгүй, ...

6 сарын 30. Пүрэв

ЗНААК НОМТОЙБАЯРЫН ӨРӨӨНӨӨС ГАРААД УНАХАД БААСАНХУУ ЗҮРГИЙГ НЬ АВААД Л ТҮЛХЭЖ АЛАХ ДУРТАЙ БАХ МЭЛХИЙ ГАДУУР ДЭМИЙ ТЭНЭЭД Л ЭНЭ ЗҮГЭЭР Л ЗОХИОН БАЙГУУЛАЛТТАЙ ГОШИН ЭМГЭНЭЛТ ДРААМ ЮМ БИШҮҮ АЙН?

^3 v0 Хариулах

Зочин [104.248.128.51] 2019/5/25

ЭНЭ ХЭРГИЙГ ТОЙРООД Л МАНАН МААФИЙН ГИШҮҮД НЭВТ ШҮЕТ ТОЙРООСТОЙ. БАРАГ УНАХААС НЬ ӨМНӨ ЗҮРГИЙГ НЬ АВААД БЭЛДЦИЙМ БИШҮҮ.

^2 v0 Хариулах

Зочин [66.181.177.76] 2019/5/24

Алуурчдыг АН дотроос хай

^3 v0 Хариулах

Зочин [103.222.123.246] 2019/5/24

Хүн мөрөөрөө ухаж болдоггүй муухай нийгэм юмдаа. [illegible]

^2 v0 Хариулах

Зочин [202.9.43.120] 2019/5/24

[illegible]

^1 v0 Хариулах

Figure 1. Ba zhua yu software review

Using this program, out of 684 comments on 16 types of information, only 368 comments in Cyrillic were collected and tested. 316 Galician or English comments were excluded.

D. Data cleaning

There were a number of issues with the collection of comments. For example, there may be many misspelled words, non-standard entries in Latin or English. Therefore, words other than the Cyrillic text were deleted, and then the misspelled word was corrected. In addition, the following actions were performed. As follow:

- Separate words, characters, and punctuation
- Links images, other news, and the web are considered insignificant because they are based on text-only impressions, and it is a separate study to identify impressions from external sources.
- Delete special characters: Some unnecessary characters have been removed and emotional characters (emoji) have been left.

E. Evaluation index

We use the accuracy level to evaluate the classification algorithm mainly. It compares the weights of the categories and divides the test text into the categories with the highest weights.

- Accuracy is our most common evaluation indicator. accuracy = (TP + TN) / (P + N). It divides the number of samples by the number of samples. Generally, the higher correct data is the better classification result [11].
- Sensitive is sensitive = TP/P. Represents a pair of total positive cases. The ability of the classifier should be measured to determine the positive case;Precision is a measure of accuracy, for example, a positive example ratio divided by positive examples or precision = TP / (TP + FP) [12].
- Precision positive example ratio divided by positive examples, or precision = TP / (TP + FP).
- Recall is a measure of coverage. Recall = TP/(TP+FN)=TP/P=sensitive.it's similar to the recall feeling.We mainly used the accuracy level to evaluate the results of the classification algorithm [13].

RESEARCH RESULTS

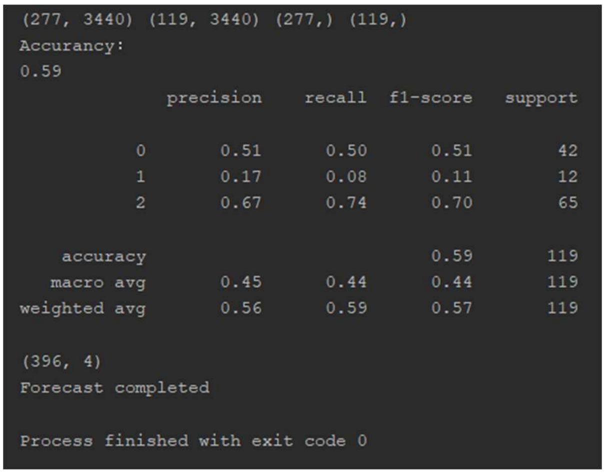
In this study, 684 comments on 16 types of information, only 368 comments were tested in Cyrillic. All Cyrillic comments were classified by KNN and Naive Bayes algorithms and statistically. Examples of statistically categorized reviews are shown in Table 1 and the results of using KNN to classify reviews are shown in Table 2. The results for classifying comments using the Naive Bayes method are shown in Table 3.

Table 1.

Results of statistical classification of total comments

№	The meaning of the topic	Positive comments			Negative comments			Statistical method
		Sentences	Total words	Positive words	Sentences	Total words	Negative words	
1.	topic 1	67	1346	12	54	1525	284	Positive
2.	topic 2	0	0	0	17	206	55	Negative
3.	topic 3	7	343	50	7	118	18	Neutral
4.	topic 4	7	113	5	10	282	16	Negative
5.	topic 5	13	371	11	22	717	53	Negative
6.	topic 6	0	0	0	6	143	15	Negative
7.	topic 7	11	155	4	6	103	21	Neutral
8.	topic 8	11	255	7	2	41	9	Neutral
9.	topic 9	37	223	18	16	462	35	Negative
10.	topic10	7	213	3	1	39	8	Positive
11.	topic 11	15	302	5	3	50	20	Positive
12.	topic 12	7	172	0	3	155	9	Negative
13.	topic 13	13	164	5	15	370	20	Negative
14.	topic 14	8	156	0	21	421	28	Neutral
15.	topic 15	9	265	2	1	47	15	Positive
16.	topic 16	21	492	3	13	230	19	Positive

The statistical results shown in Table 1 show that 6 out of 16 groups of comments were **positive**. Conversely, 6 sets of comments are **negative**. We classified the set of comments as **neutral** because they expressed values that were neither **positive** nor **negative** during the experiment. A total of 4 sets of comments were included in the **neutral** classification in this group.



```
(277, 3440) (119, 3440) (277,) (119,)
Accurancy:
0.59
              precision    recall  f1-score   support

           0       0.51      0.50      0.51        42
           1       0.17      0.08      0.11        12
           2       0.67      0.74      0.70        65

    accuracy                           0.59       119
   macro avg       0.45      0.44      0.44       119
weighted avg       0.56      0.59      0.57       119

(396, 4)
Forecast completed

Process finished with exit code 0
```

Figure 2. Results of an algorithm for classifying comments using the KNN method

Table 2

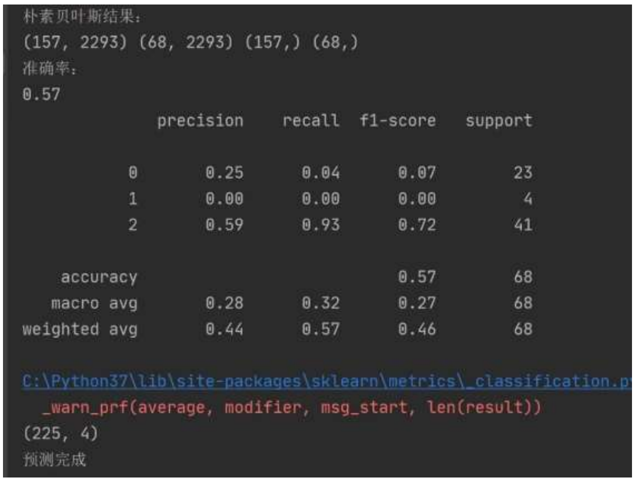
Results of KNN classification of total comments

№	The meaning of the topics	KNN (accurancy)
1.	Comment 1	0.77
2.	Comment 2	0.17
3.	Comment 3	0.25
4.	Comment 4	0.17
5.	Comment 5	0.73
6.	Comment 6	0.00
7.	Comment 7	0.83
8.	Comment 8	0.75
9.	Comment 9	0.60
10.	Comment 10	1.00
11.	Comment 11	0.67
12.	Comment 12	0.33
13.	Comment 13	0.44
14.	Comment 14	0.56
15.	Comment 15	1.00
16.	Comment 16	0.64

The results of the KNN method shown in Table 2 show that 10 sets of 16 groups of comments, or comments 1, 5, 7-11, and 14-16, were **positive**, with an accuracy of 0.56% -1.0%. Conversely, 6 sets of 2-4, 6, 12-13 comments were **negative**, with an accuracy of 0.00% -0.44%.

DOI:10.26104/IVK.2019.45.557

Because this method is calculated by a computer program, it is not classified as neutral. It means that the set of comments are neither positive nor negative. because we assume a positive value if the accuracy of the calculated result is 0.51-1.00%. However, if the accuracy of the calculated results is 0-50%, we consider it negative.



```
朴素贝叶斯结果:
(157, 2293) (68, 2293) (157,) (68,)
准确率:
0.57
              precision    recall  f1-score   support

           0       0.25      0.04      0.07        23
           1       0.00      0.00      0.00         4
           2       0.59      0.93      0.72        41

    accuracy                           0.57        68
   macro avg       0.28      0.32      0.27        68
weighted avg       0.44      0.57      0.46        68

C:\Python37\lib\site-packages\sklearn\metrics\_classification.py
  _warn_prf(average, modifier, msg_start, len(result))
(225, 4)
预测完成
```

Figure 3. Results of an algorithm for classifying comments using the Naive Bayes method

Table 3

Results of Naive Bayes classification of total comments

№	The meaning of the topics	Naive Bayes (accurancy)
1.	Comment 1	0.80
2.	Comment 2	0.33
3.	Comment 3	0.25
4.	Comment 4	0.33
5.	Comment 5	0.45
6.	Comment 6	0.50
7.	Comment 7	0.83
8.	Comment 8	0.75
9.	Comment 9	0.47
10.	Comment 10	1.00
11.	Comment 11	0.67
12.	Comment 12	0.67
13.	Comment 13	0.56
14.	Comment 14	0.67
15.	Comment 15	1.00
16.	Comment 16	0.64

The results of the Naive Bayes method shown in Table 3 show that 10 sets of 16 groups of impressions, or 1, 7-8, and 10-16 comments, were **positive**, with an accuracy of 0.56% -1.0%. Conversely, 6 is a set of 2-6, and the 9th comments are **negative**, with an accuracy of 0.25% - 0.50%. Because this method is calculated by a computer program, it is not classified as neutral, meaning that the set of comments are neither positive or negative.

Table 4.

Statistics, results classified by KNN and Naive Bayes methods

№	The meaning of the topics	Statistical method	KNN		Naive Bayes (accurancy)	
			accurancy	point	accurancy	point
1.	Comment 1	Positive	0.77	1	0.80	1
2.	Comment 2	Negative	0.17	1	0.33	1
3.	Comment 3	Neutral (-)	0.25	1	0.25	1
4.	Comment 4	Negative	0.17	1	0.33	1
5.	Comment 5	Negative	0.73	0	0.45	1
6.	Comment 6	Negative	0.00	1	0.50	1
7.	Comment 7	Neutral (+)	0.83	1	0.83	1
8.	Comment 8	Neutral (+)	0.75	1	0.75	1
9.	Comment 9	Negative	0.60	0	0.47	1
10.	Comment 10	Positive	1.00	1	1.00	1
11.	Comment 11	Positive	0.67	1	0.67	1
12.	Comment 12	Negative	0.33	1	0.67	0
13.	Comment 13	Negative	0.44	1	0.56	0
14.	Comment 14	Neutral (+)	0.56	1	0.67	1
15.	Comment 15	Positive	1.00	1	1.00	1
16.	Comment 16	Positive	0.64	1	0.64	1
Estimated percentage				87.50		87.50

DOI:10.26104/IVK.2019.45.557

Note: Accuracy is the accuracy of conformity, 1 point is consistent with the results of the statistical method, and 0 point is inconsistent. In order to compare the results of these two methods, the set of comments with the meaning of neutral was reconsidered and added as positive (+) and negative (-).

The results in Table 4 show that for Comment 1, KNN is 0.77%, and Naive Bayes 0.80%. In particular, the Naive Bayes method, with an accuracy of 0.80%, seemed to be the best classification. Comment 2, on the other hand, agrees with KNN 0.17%, Naive Bayes 0.33%, and statistically accurate.

In order to select the most effective method from the above methods, we compared the statistical or manual classification results with KNN 87.50% and Naive Bayes 87.50%. Therefore, we assume that the KNN and Naive Bayes methods are classified by similar results.

DISCUSSION

As for the Mongolian language, D.Zolboo and others conducted the first study to classify electronic texts. This was the first experiment, and those 1,000 texts were divided into relatively many categories, indicating a lack of training data. On the other hand, this article does not mention the quantity and nature of the experimental data so the validity of the experiment did not explain clearly.

Munkhjargal's Zoljargal, Dambasuren's Nanzadragchaa, Chagnaa's Altangerel, and Jaimain Purev have developed a basic model of a machine-generated machine learning tool that records the impressions of Mongolian text. For this work, a basic model of machine learning was created with a bunch of Mongolian-language impressions of the text. Experiments were also conducted to classify comments using depth training. However, the results of the experiment did not specify what to focus on and what was lack part.

In my research, we used KNN and Naive Bayes to collect and analyze user comments in Cyrillic. Our results have a KNN of 0.59 percent and a Naive Bayes of 0.57 percent for comments rating accuracy. However, we have tried only two methods. In the future, we will try the additional classification method again and compare it with the results.

CONCLUSION. In our study, KNN and Naive Bayes used 368 Cyrillic text comments. We also collected user feedback, cleaned up the data, created a vocabulary of positive and negative words in the comments. The algorithm was developed and analyzed by KNN and Naive Bayes methods. The test results were compared with the manual classification results. In the future, another text classification method will be used to confirm this result.

References:

1. Luhn H.P. “The Automatic Creation of Literature Abstracts.”, IBM Journal, April 1958.
2. Sebastiani F. Machine learning in automated text classification [J]. Computer Science, 2015, 34(l): l-47.
3. Huang Juan Yun ,Research and Improvement on Feature Selection and Classification Algorithms for Text Classification Based on KNN, 2014.
4. Cheng Bo, Research and System for Classification of Web Text, 2010.05.
5. Tian Shao xing, Chen Jin Jie, Handwritten Numeral Recognition Based on KNN, October 2017.
6. LU Zheng Yu , ZHAO Shuang , LIN Yong MI, Research of kNN in Text Classification, 2008.
7. Yang Ding, Classification approach of Chinese texts sentiment based on semantic lexicon and naive Bayesian,YANG Ai Min, 2010.
8. Wang Jun Qiang , Liu Jian Ping, Research on Social Network Information Text Algorithm Based on Native Bayes, 2015
9. Text Classification Based on Naive Bayesian, Yaun Qi, Yu Qiao, 2017.05
10. Jia Yun Fan, Classification Method Research Method, 2017.6
11. Bat-Erdene Nyandag, Li Ru, G. Indruska, “Performance Analysis of Optimized Content Extraction for Cyrillic Mongolian Learning Text Materials in the Database”, Journal of Computer and Communications, Volume 4, No 10, August 31, 2016, China, Online, 79-89. doi:10.4236/jcc.2016.410009.
12. Bat-Erdene Nyandag, Li Ru, Orgil Demberel., “Keyword Extraction Based on Statistical Information for Cyrillic Mongolian script”, The 29th Chinese Control and Decision Conference (2017 CCDC), 2250-2255, May 28–30, 2017, DOI: 10.1109/CCDC.2017.7978889
13. Bat-Erdene Nyandag, Li Ru, Khishigdalai Ulaankhuu., “Improving Determine Lexical Meanings for Mongolian Cyrillic Script and Its Used to Keyword Extraction”, 7th IEEE International Conference on Logistics, Informatics and Service Sciences (LISS’2017), Kyoto,Japan,24-27 July, 2017.
14. Ташбаев А.М. Цифровая трансформация и состояние применения информационно-коммуникационных технологий в сфере образования. Наука, новые технологии и инновации Кыргызстана. 2019. №. 10. С. 109-115.