ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.237.8: 81’322.2
DOI: 10.17223/19988605/33/1

M.F. Ashurov, V.V. Poddubny

TEXT CLASSIFICATION STREAM-BASED *R*-MEASURE APPROACH USING FREQUENCY OF SUBSTRING REPETITION

The publication is prepared within the framework of the scientific project No. 14-14-70010 supported by the Russian Humanitarian Scientific Fund. This work is performed under the state order No. 1.511.2014/K of the Ministry of Education and Science of the Russian Federation.

Stream-based approach of *R*-measure using frequency of substring repetition in text classification is offered. Comparative quality analysis of classificators based on the truncated *R*-measure using frequencies of test text substring repetition and without one is performed on a text set of Russian fiction of the 19th century and the 90th of 20th century. An accuracy of text classification is estimated by Van Rijsbergen’s effectiveness measure known as *F*-measure. The fact that in case of genre mixing free into author’s text classes accounting frequency of test text substring repetition in supertexts increases the classification accuracy is confirmed.

Keywords: stream-based classification approach; *R*-measure; frequency of substring repetition; classification accuracy; *F*-measure.

Issues of automatic text classification arise in the creation and updating of search systems, scientific research linked with recognition style features of fiction or published texts, author text matching etc. The formal description of the text classification issue is described, as example, in [1]. As mentioned in [1] all text classification approaches can be divided on two groups: feature-based and stream-based. Feature-based approaches do not deal with text directly, but these use text-value representation as an array of feature values. It is necessary to determine a sufficient feature set by which the classificatory calibrated before using any feature-based approach of text classification. In case of feature set selection is poor; the classification result becomes extremely unsatisfactory. Stream-based approaches unlike feature-based do not require selection of any features for text classification and deal with text directly. Text X is considered as a sequence (stream) $x_1 x_2 \ldots x_n$ of n elements from an alphabet A.

Among the stream-based approaches are two main groups described in [1–4]. The first group is based on substring repetition counting (*R*-, *C*- and other measures). The second group uses data compression (for example, the off-the-shelf algorithm using PPM (Prediction by Partial Matching) compression method).

Khmelev [5] did a comparison of the methods based on substring repetition counting with the methods based on data compression. The comparison was held on the news articles (journalistic articles) in which an author style is involved much and a genre specifics influence is decreased. To estimate an accuracy of text classification better the additional comparison should be done on different types of texts, particularly, on fictions.

We offer a modification of *R*-measure which can use frequencies of substring repetition. This article describes the comparative quality analysis of the classification approaches based on the truncated *R*-measure and its modification in case of Russian fiction texts of different ages.

1. Stream-based approaches and truncated *R*-measure

Approaches proposed by Khmelev and Teahan in [6] are based on both of the above groups, but there is the approach which can be especially highlighted. This approach named *R*-measure (from repetition) counts all substrings of a test text that are present in a class supertext. The supertext is formed by concatenating all texts of the particular class. So *R*-measure has the following equation:

$$R(X \mid S) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{n-k} c\left(x_i \ldots x_{i+k} \mid S\right)}{\frac{1}{2}n(n+1)}, \tag{1}$$

where X is a test text; S – a supertext of current class; n – length of the test text; k – length of pattern substring used in string search and $c\left(x_i \ldots x_{i+k} \mid S\right)$ – indicator function, indicating presence of a substring into the supertext:

$$c\left(x_i \ldots x_{i+k} \mid S\right) = \begin{cases} 1, x_i \ldots x_{i+k} \in S, \\ 0, x_i \ldots x_{i+k} \notin S. \end{cases} \tag{2}$$

As the volumes of fiction, texts are so huge R-measure calculation should be done quickly. In that case, the suffix arrays [1, 3, 4] are used in the R-measure calculation to search a substring quickly in the huge text.

The indicator function based on suffix array approach can be implemented in an algorithm using dichotomous search in the suffix array. So the indicator function should return 1 value if a pattern substring is matched with a suffix in the array at first time (other matching is not used here), or return 0 value, if no matching with the suffixes.

Also note that R-measure equation (1), in fact, has the number of used substrings in the denominator. This number is used to normalize the value of the proximity between the texts so that it should be in the interval [0,1], where 1 value means that a class supertext contains full test text (the maximum similarity of a test text with a class supertext), and 0 value means no matching between the test text and the supertext.

In the first place, the main disadvantage of R-measure is the fact that the substrings with length in 1 or 3 symbols and the substring with length over 200 symbols are absolutely unsuitable, because these substrings cannot be used to recognize specific characteristics for each class. The second disadvantage of R-measure is the issue that the length n of test fiction texts is around 150 thousand symbols on average so R-measure calculation under that condition of test text length requires a lot of computing resources and time on the typical personal computers in the classification using real data. Even if using the best on-time performance algorithms of substring search and data structures, R-measure calculation requires about 6 days for each class at the length of test text around 200 thousand symbols. A lot of time required in the R-measure classification is completely unsuitable in the real world. To solve these issues the modification of R-measure, named truncated R-measure, which uses only particular substring lengths from the full range of available substrings has been proposed.

Truncated R-measure similarity described in [1] counts all repetitions of substrings of test text length n in the supertext. These substrings have lengths from k_1 to k_2 unlike classical R-measure equation (1) where $k_1 = 1$, $k_2 = n$:

$$R(X \mid S) = r(X \mid S) / N, \tag{3}$$

where X is a test text; S – a supertext of particular class; N – a number of used substrings. The function to calculate a measure of similarity is described as follows:

$$r\left(X \mid S\right) = \sum_{k=k_1}^{k_2}\sum_{i=0}^{n-k} c\left(x_i \ldots x_{i+k} \mid S\right). \tag{4}$$

Normalizing of the measure is based on the number of used substrings N, which is calculated by following equation:

$$N = \left(2(n+1) - (k_1 + k_2)\right)\left(k_2 - k_1 + 1\right) / 2. \tag{5}$$

The issue of the k_1 and k_2 parameters selection used in truncated R-measure calculation has been solved by natural language features. The minimum length k_1 is equal to 10 symbols because shorter substrings can match with many words in Russian language, which are common for all authors so authors specifics cannot be pick out. The maximum length k_2 is equal to 45 symbols that allows to process several words from the test text. Using the greater length of substrings seems not useful as different authors cannot repeat such long phrases many times or it could be unlikely. So a time calculation of the truncated R-measure is decreased compared with the classical R-measure because of the reduction of substrings amount. For example, the time to calculate the truncated R-measure on a test text which length is equal to 150 thousand symbols for each class is reduced to the acceptable level about one or two minutes.

2. Modification of *R*-measure using frequencies of substring repetition

Previous comparison of approaches based on *R*-measure and PPMD (PPMd – Prediction by Partial Matching (dynamic) data compression algorithm without loss) compression described in [7, 8] shows that the accuracy of classification by *R*-measure is less than accuracy by the PPMD approach in some case. These cases are explained by the fact that *R*-measure approach does not use frequencies of substring repetition unlike the PPMD compression approach. To estimate feasible performance of using frequencies in classification the approach based on *R*-measure modification that can use frequencies of substring repetition is offered. This *R*-measure modification is named *RF*-measure (repetition frequency). The *RF*-measure detailed description and its comparison with *R*-measure are described in this article.

RF-measure is based on the same arrangement of truncation used in equations (3) and (4):

$$RF(X \mid S) = rf(X \mid S) / N, \tag{6}$$

$$rf(X \mid S) = \sum_{k=k_1}^{k_2} \sum_{i=0}^{n-k} rf(x_i \ldots x_{i+k} \mid S). \tag{7}$$

In that case, indicator function $c(x_i \ldots x_{i+k} \mid S)$ used in *R*-measure equation (3) is removed and the function, which determines the degree of similarity of substring frequencies between a test text and a supertext, is used in the similarity measure:

$$rf(x_i \ldots x_{i+k} \mid S) = \begin{cases} rfc(x_i \ldots x_{i+k}),\ x_i \ldots x_{i+k} \in S, \\ 0,\ x_i \ldots x_{i+k} \notin S. \end{cases} \tag{8}$$

During the selection of similarity measure, it was decided to use the relative frequencies of substring repetition of the test text and the supertext in the measure calculation. The main condition of the measure selection is that calculated value should satisfy the normalization principles. It means that the similarity measure value should be in the interval [0,1], where 1 value means that compared frequencies are the same. This condition can be satisfied easily if similarity measure is based on the ratio of the minimal frequency to maximum one. Accordingly, the similarity measure using the ratio of frequencies of substrings repetition of test texts and supertexts is following:

$$rfc(x_i \ldots x_{i+k}) = \frac{\min(f_X(x_i \ldots x_{i+k}), f_S(x_i \ldots x_{i+k}))}{\max(f_X(x_i \ldots x_{i+k}), f_S(x_i \ldots x_{i+k}))}.$$

Thus, any difference in the relative frequencies results in a reduction of the similarity measure, whereby a test text belonging to the class is reduced too.

The relative frequencies of substrings repetition for test text *X* and supertext *S* is as follows:

$$f_X(x_i \ldots x_{i+k}) = \frac{cc_X(x_i \ldots x_{i+k})}{N_X} \text{ and } f_S(x_i \ldots x_{i+k}) = \frac{cc_S(x_i \ldots x_{i+k})}{N_S},$$

where a function $cc(x_i \ldots x_{i+k})$ unlike the indicator function (2) returns the amount of pattern substring matching in the test text and supertext (relevant marks *X* and *S* is in the subscripts), and *N* is the lengths of texts *X* and *S* respectively.

The function $cc(x_i \ldots x_{i+k})$ can be implemented in an algorithm using the dichotomous substring search in a suffix array. This algorithm after the completion of substring search process and the first success of full pattern matching selects next suffixes in series that satisfy the pattern matches and counts the total number of that matches. To decrease time of counting matches the special array containing the lengths of a common prefix of the next suffix for each particular suffix can be used. In the algorithm the procedure of full matching substring pattern and suffix can be skipped in case a length of the next suffix prefix less than a length of the substring pattern. Using the common prefix array calculation time of counting all substring matches for substring lengths more than 10 symbols could be decreased about 5–10%.

3. Accuracy analysis of classification

In this article the analysis of classification accuracy of *R*-measure and *RF*-measure have been done for two text samples. The first text sample consists of 126 prose fiction texts of nine authors of the 19th century representing the classes. This text sample is described in [9, 10]. More than 90 texts have been used as a training sample to build corresponding supertexts. The second text sample used in [8] work is the prose fiction texts of the 90th of 20th century. This text sample contains 138 texts by 21 authors. The volume of the supertexts is the same for each class necessarily and text lengths of test samples are equal to 100 thousand symbols.

To test the classification accuracy software module that allows classify a text by *R*-measure and *RF*-measure has been designed and implemented. The issue of parameters k_1 and k_2 selection has been solved by natural language features as described above.

Classification accuracy for each class is estimated by Van Rijsbergen's measure [1, 11] known as *F*-measure. *F*-measure is harmonic mean of precision *p* (the number of correct positive results divided by the number of all positive results) and recall *r* (the number of correct positive results divided by the number of positive results that should have been returned):

$$F = 2\frac{r \times p}{r + p}.$$

In general, the rating of the classification accuracy is calculated as an arithmetical mean of *F*-measures of all classes, named *F*-macro, or as *F*-measure is calculated on average of precisions and recalls of all classes, named *F*-micro.

4. Results

As for the first text sample represented by 9 classes the result characteristics have been calculated for truncated *R*-measure and *RF*-measure, shown in table 1.

Table 1

Precision, recall and *F*-measure for authors of the 19th century

Author	Precision		Recall		*F*-measure	
	***R*-measure**	***RF*-measure**	***R*-measure**	***RF*-measure**	***R*-measure**	***RF*-measure**
Chehov	0,60	0,56	0,75	0,75	0,67	0,64
Dostoevskii	1,00	0,56	0,42	0,42	0,59	0,48
Gogol	0,92	1,00	1,00	1,00	0,96	1,00
Goncharov	0,67	0,63	1,00	1,00	0,80	0,77
Kuprin	1,00	0,50	0,50	0,42	0,67	0,45
Leskov	0,71	1,00	0,83	0,75	0,77	0,86
Saltikov-Shedrin	0,40	0,40	0,17	0,17	0,24	0,24
Tolstoi	1,00	1,00	0,92	0,92	0,96	0,96
Turgeniev	0,57	0,71	1,00	1,00	0,73	0,83

For clarity, the total ratings of classification accuracy on the first text sample are shown in table 2.

Table 2

Values of *F*-macro and *F*-micro for each approach

Approach	***F*-measure (macro)**	***F*-measure (micro)**
R-measure	0,71	0,75
RF-measure	0,69	0,71

For the first text sample the *F*-measure of the classificator based on *R*-measure is 2-4% better than the *F*-measure of the classificator based on *RF*-measure. Further research has illustrated the fact that stylistic characteristics of the test texts and supertexts of some classes are different visibly. In case using *RF*-measure that difference between texts can significantly affects the numerical value of similarity measure and results of text classification.

To minimize the effect of the above issue, depending on the quality of training and test samples, it has been decided to make these samples randomly from all texts of the class. The process of sample generation must be satisfy two conditions – test texts should not be included in the training sample, which is used to build a supertext, and vice versa, and the volumes of all supertext should be the same for each class. After repeated classifying all values of *F*-measure calculated for each class on the randomly generated samples are average for each class. Average values of the class *F*-measure for the truncated *R*-measure и *RF*-measure, and total *F*-macro of the classificators are presented in table 3.

Table 3

Average *F*-measure for authors of the 19th century on randomly generated samples

Author	*F*-measure	
	***R*-measure**	***RF*-measure**
Chehov	0,95	0,90
Dostoevskii	0,41	0,57
Gogol	0,88	0,92
Goncharov	0,91	0,89
Kuprin	0,89	0,69
Leskov	0,89	0,90
Saltikov-Shedrin	0,61	0,73
Tolstoi	0,94	0,94
Turgeniev	0,73	0,85
Total F-mesuare (macro)	0,80	0,82

As illustrated by results in table 3 total (average for the classes) the *F*-measure value of *RF*-measure is higher than the *F*-measure value of *R*-measure. Note that only two classes have *F*-measure results using *RF*-measure worse than results using *R*-measure on 2–5%. It means using *RF*-measure improves the classification accuracy in general. This fact also can be confirmed by the classification results on the second text sample. The text sample is specific because it contains modern text of little-known authors, texts are in most cases represented by one genre and the used set of words is not wide as for prose authors of the 19th century.

As for the second text sample based on 21 classes the results for truncated *R*-measure and *RF*-measure are shown in table 4.

Table 4

Precision, recall and *F*-measure for authors of the 20th century

Author	Precision		Recall		*F*-measure	
	***R*-measure**	***RF*-measure**	***R*-measure**	***RF*-measure**	***R*-measure**	***RF*-measure**
Agafonov	0,88	0,88	1,00	1,00	0,93	0,93
Aristov	1,00	1,00	1,00	1,00	1,00	1,00
Azarov	0,96	1,00	1,00	1,00	0,98	1,00
Baganov	0,79	0,84	1,00	1,00	0,88	0,91
Belkin	1,00	1,00	1,00	1,00	1,00	1,00
BelobrovPopov	1,00	1,00	1,00	1,00	1,00	1,00
Belomlinskaya	1,00	1,00	1,00	0,80	1,00	0,89
Belov	1,00	1,00	0,78	0,78	0,88	0,88
Bonch	1,00	1,00	0,27	0,27	0,43	0,43
Bronin	1,00	1,00	1,00	1,00	1,00	1,00
Burmistrov	0,50	0,50	1,00	1,00	0,67	0,67
Galkin	1,00	1,00	0,29	0,29	0,45	0,45
Gergenreder	1,00	1,00	1,00	1,00	1,00	1,00
Glushkin	1,00	1,00	1,00	1,00	1,00	1,00
Svetlana	1,00	1,00	1,00	1,00	1,00	1,00
Velboi	0,10	0,26	0,20	0,90	0,13	0,40
Vershovskii	1,00	1,00	0,19	0,19	0,32	0,32
Veter	0,57	0,80	1,00	1,00	0,73	0,89
Vitkovskii	1,00	1,00	0,33	0,33	0,50	0,50
Voronov	1,00	1,00	1,00	1,00	1,00	1,00
Vulf	1,00	1,00	1,00	1,00	1,00	1,00

For clarity, the total ratings of classification accuracy on the second text sample are shown in table 5.

Table 5

Values of *F*-macro and *F*-micro for each approach

Approach	***F*-measure (macro)**	***F*-measure (micro)**
R-measure	0,80	0,85
RF-measure	0,82	0,88

The classificator based on *RF*-measure is a little better than one based on *R*-measure on the text sample. And only one class has the *F*-measure value of *RF*-measure less than one of *R*-measure.

Moreover, note the fact that there are classes, classification accuracy which is extremely low. Additional researches show that this aspect occurs because the similarity measure between the supertext of that classes and the supetext of another classes is more than 20%. The aspect is illustrated in figure 1.

Identify author's stylistic features of such text requires using additional instruments. It can be an analyzer, which takes class specific substrings from the text, or a filter, which excludes substrings that are absolutely common for the classes (for example, it can be a common phrase or idioms of Russian language or a series of function words without semantic). These instruments can become a dynamic dictionary, which can be extended by adding new texts in the class supertext. Moreover, the proposal instruments are heuristic tools based on the language features and its performance is questionable in general if a stream contains other entities instead of text characters (stream codes of the human genome for example).

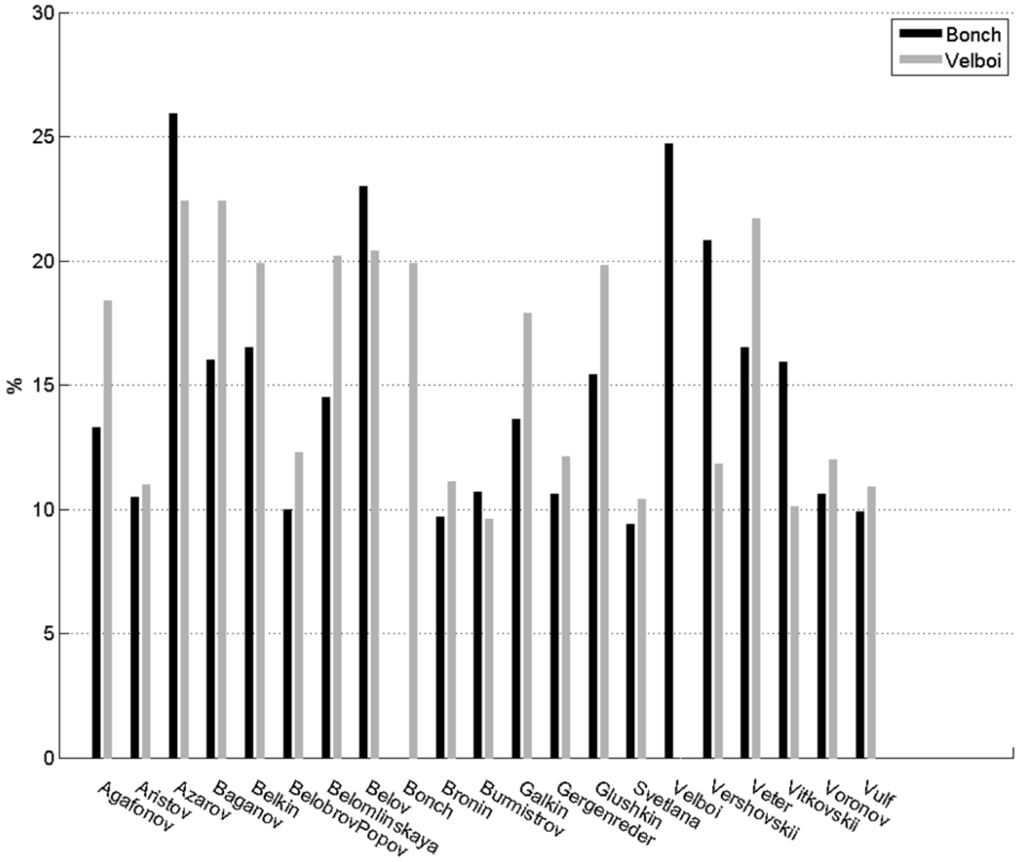

Fig. 1. Similarity measure of Bonch and Velboi supetexts by other classes

As for the second text sample, note that *RF*-measure has a better classification accuracy for that class types allowing highlight the stylistic features of author style well. So the accuracy of the classification based on *RF*-measure is better on text samples, class supetext of which can be very similar. In that case, the precisions of two classes have increased more than 50%.

Conclusions

Classification accuracy of both classificators on considered text data is more than 70%. Furthermore, the approach proposed in this article has a prospect of further improving. Individual features of the constructed training and test samples can explain the issue that *RF*-measure has the low values or it is less than *R*-measure for some classes. The impact of this issue can be decreased by frequentative classifications on the training and test text samples created randomly from all texts of the class. In the future, we see an opportunity to improve a classification accuracy of *RF*-measure by:

1) using a threshold filtering;

2) modification of counting relative frequencies of a supertext;

3) modification of the function using ratio of the minimum of relative frequencies to the maximum to the weighted average.

The first aspect is connected with the issue that natural language texts have some substrings that do not contains the semantic content (author's thoughts or natural logic), but the number of this substring repetition can reach 200 or 300 repetitions in the supertext. It affects relative frequencies vastly so these substrings, in the simplest case, can be skipped.

The second aspect occurs when a class supertext contains many different texts, so a length of the supertext is many times greater than a length of the test text. The current approach uses the relative frequency calculation as a ratio of the number of pattern substring repetitions to length of the supertext. So relative frequency of the supertext is less than the relative frequency of the test text if the numbers of repetitions are the same. Moreover, as for natural languages, if the supertext consists of different texts then the relative frequency of the supertext cannot recognize clearly author style used in particular text included in the supertext. Therefore, it requires finding a function which calculates a relative frequency of the supertext removing the effects of absolute value of the supertext length.

The third aspect is based on the above issue that the difference between relative frequencies of the supertext and test text results in little value of the frequencies ratio. This means that, in practice, calculated value of the ratio is less than 1 if there is a little absolute difference between the relative frequencies. The little difference of frequencies is a normal in fiction texts. Therefore, a function, which returns values around 1 in little difference of relative frequencies, should be found.

The most interesting to improve the *RF*-measure is using an additional markup of the text in classification. It allows creating a better similarity measure which can use more text features for each substring. This aspect is useful not only for the fiction text classification. In general, many streams (streams of characters, codes, markup entities) can be used in the similarity measure calculation. Also note that the calculation can be done in case of availability of additional information or if it is not.

REFERENCES

1. *Шевелёв О.Г.* Методы автоматической классификации текстов на естественном языке : учеб. пособие. Томск : ТМЛ-Пресс, 2007. 144 с.
2. *Humnisett D., Teahan W.J.* Context-based methods for text categorization // Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR). The University of Sheffield. UK. 2004.
3. *Ukkonen E.* Constructing Suffix-trees On-Line in Linear Time // Algorithms, Software, Architecture: Information Processing. 1992. № 1(92). 484 p.
4. *Kärkkäinen J., Sanders P.* Simple linear work suffix array construction // ICALP 2003, LNCS 2719 / eds. by J.C.M. Baeten et al. 2003. P. 943–955.
5. *Хмелёв Д.В.* Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение. 2003. URL: http://compression.graphicon.ru/download/articles/classif/intro.html
6. *Khmelev D.V., Teahan W.J.* Verification of text collections for text categorization and natural language processing // Technical Report AIIA 03.1. School of Informatics. University of Wales. Bangor. 2003.
7. *Ашуров М.Ф., Поддубный В.В.* Метод классификации текстов художественной литературы на основе *R*-меры // Новые информационные технологии в исследовании сложных структур : материалы Десятой рос. конф. с междунар. участием. Томск : Издательский Дом Томского государственного университета, 2014. С. 63–64.
8. *Ашуров М.Ф.* Сравнение потоковых методов классификации текстов художественной литературы на основе сжатия информации и подсчета подстрок // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2014. № 4(29). С. 16–22.

9. *Ашуров М.Ф., Поддубный В.В.* Потоковый метод классификации текстов художественной литературы на основе *С*-меры // Информационные технологии и математическое моделирование (ИТММ-2013) : материалы XII Всерос. науч.-практ. конф. с международ. участием им. А.Ф. Терпугова (29–30 ноября 2013 г.). Томск : Изд-во Том. ун-та, 2013. Ч. 2. С. 85–89.
10. *Shevelyov O.G., Poddubnyj V.V.* Complex investigation of texts with the system «StyleAnalyzer» // Text and Lanquage / ed. by P. Grzyber, E. Kelih, J. Macutek. Wien : Praesens Verlag, 2010. P. 207–212.
11. *Van Rijsbergen C.J.* Information Retrieval. London : Butterworths, 1979.

Ashurov Mikhail Faritovich. E-mail: ashurov.mf@gmail.com
Poddubny Vasiliy Vasilievich, doctor of science, professor. E-mail: vvpoddubny@gmail.com
Tomsk State University, Russian Federation.

Поступила в редакцию 24 сентября 2015 г.

Ашуров М.Ф., Поддубный В.В. (Томский государственный университет, Российская Федерация).

Основанный на *R*-мере подход к потоковой классификации текстов, использующий частоты повторения подстрок.

Ключевые слова: потоковые методы классификации; *R*-мера; частота повторения подстрок; качество классификации; *F*-мера.

DOI: 10.17223/19988605/33/1

Предлагается потоковый метод классификации текстов на основе *R*-меры с использованием частот повторения подстрок. На материале текстов русской художественной прозы XIX в. и 90-х гг. XX в. проводится сравнительный анализ качества работы алгоритмов распознавания авторских классов классификаторами, построенными на основе усечённой *R*-меры как с использованием, так и без использования частот повторения подстрок испытуемого текста. Качество классификации оценивается известной *F*-мерой Ван Ризбергена. Показано, что в отсутствие жанрового смешения текстов внутри авторских классов учёт частот повторения подстрок исследуемого текста в текстовых суперклассах приводит к повышению качества классификации.

Усечённая *R*-мера близости текстов учитывает все возможные повторения всех подстрок длин от k_1 до k_2 (для неусечённой *R*-меры $k_1 = 1$, $k_2 = n$) испытуемого текста длины n в супертексте:

$$R(X\,|\,S) = r(X\,|\,S)/N,\quad r\left(X\,|\,S\right) = \sum_{k=k_1}^{k_2} c_k\left(X\,|\,S\right),\quad N = \left(2(n+1) - (k_1 + k_2)\right)\left(k_2 - k_1 + 1\right)/2,$$

$$c_k\left(X\,|\,S\right) = \sum_{i=k}^{n} c\left(x_{i-k+1}\ldots x_n\,|\,S\right),\quad c\left(x_{i-k+1}\ldots x_n\,|\,S\right) = \begin{cases} 1,\ x_{i-k+1}\ldots x_n \subset S, \\ 0,\ x_{i-k+1}\ldots x_n \not\subset S, \end{cases}$$

где X – испытуемый текст, S – супертекст (объединённый текст) исследуемого класса, N – число подстрок.

В работе предлагается метод потоковой классификации, основанный на модификации *R*-меры с учётом частот повторений подстрок. Эта мера (обозначим её как *RF*-мера) использует тот же механизм усечённости:

$$RF(X\,|\,S) = rf(X\,|\,S)/N,\quad rf\left(X\,|\,S\right) = \sum_{k=k_1}^{k_2}\sum_{i=0}^{n-k} rf\left(x_i\ldots x_{i+k}\,|\,S\right),$$

$$rf\left(x_i\ldots x_{i+k}\,|\,S\right) = \begin{cases} rfc(x_i\ldots x_{i+k}),\ x_i\ldots x_{i+k} \in S, \\ \quad 0,\ x_i\ldots x_{i+k} \notin S, \end{cases}\quad rfc(x_i\ldots x_{i+k}) = \frac{\min(f_X(x_i\ldots x_{i+k}), f_S(x_i\ldots x_{i+k}))}{\max(f_X(x_i\ldots x_{i+k}), f_S(x_i\ldots x_{i+k}))}.$$

Относительные частоты повторения подстрок для исследуемого текста X и супертекста S вычисляются следующим образом:

$$f_X(x_i\ldots x_{i+k}) = \frac{cc_X(x_i\ldots x_{i+k})}{N_X}\ \text{и}\ f_S(x_i\ldots x_{i+k}) = \frac{cc_S(x_i\ldots x_{i+k})}{N_S},$$

где функция $cc\left(x_i\ldots x_{i+k}\right)$, в отличие от функции-индикатора $c\left(x_i\ldots x_{i+k}\,|\,S\right)$, возвращает количество совпадений подстроки поиска в исследуемом тексте или супертексте (обозначенные соответствующими индексами X и S), а N_X и N_S – количество символов в текстах X и S соответственно.

Качество классификации по каждому классу оценивалось по текстам контрольной выборки *F*-мерой Ван Ризбергена – средним гармоническим между полнотой r (долей текстов, приписываемых классу из всех текстов этого класса) и точностью p (долей текстов, правильно приписываемых этому классу из всех текстов, приписываемых этому классу).

Оба классификатора на рассматриваемых данных (текстах русской художественной литературы XIX и XX вв.) дают качество классификации более 70%, причём в условиях отсутствия жанрового смешения текстов внутри авторских классов учёт частот повторения подстрок исследуемого текста в супертекстах приводит к повышению качества классификации.

REFERENCES

1. Shevelev, O.G. (2007) *Metody avtomaticheskoy klassifikatsii tekstov na estestvennom yazyke* [Approaches of automatic natural language text classification]. Tomsk : TML-Press.
2. Humnisett, D. & Teahan, W.J. (2004) Context-based methods for text categorization. *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR)*. The University of Sheffield. UK. 2004. DOI: 10.1145/1008992.1009129
3. Ukkonen, E. (1992) Constructing Suffix-trees On-Line in Linear Time. *Algorithms, Software, Architecture: Information Processing*. 1(92).

4. Kärkkäinen, J. & Sanders, P. (2003) Simple linear work suffix array construction. *ICALP 2003, LNCS 2719*. pp. 943-955. DOI: 10.1007/3-540-45061-0_73

5. Khmelev, D.V. (2003) *Klassifikatsiya i razmetka tekstov s ispol'zovaniem metodov szhatiya dannykh* [Text classification and markup using data compression approaches]. [Online] Available from: http://compression.graphicon.ru/download/articles/classif/intro.html

6. Khmelev, D.V. & Teahan, W.J. (2003) Verification of text collections for text categorization and natural language processing. *Technical Report AIIA 03.1. School of Informatics*. University of Wales. Bangor.

7. Ashurov, M.F. & Poddubnyy, V.V. (2014) [Approach based on R-measure of fiction text classification]. *Novye informatsionnye tekhnologii v issledovanii slozhnykh struktur* [New information technologies in complex structure research]. Proc. Of the 10th Conference with International Participation. Tomsk: Tomsk State University. pp. 63-64.

8. Ashurov, M.F. (2014) Comparison of stream-based fiction text classification methods based on data compression and counting substrings. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 4(29). pp. 16-22. (In Russian).

9. Ashurov, M.F. & Poddubnyy, V.V. (2013) [Approach based on C-measure of fiction text classification]. *Informatsionnye tekhnologii i matematicheskoe modelirovanie (ITMM-2013)* [Information technology and mathematical modeling (ITMM-2013)]. Proceedings of the 12th Russian Scientific-Practical Conference with International Participation named after A.F. Terpugov. Tomsk. 29th – 30th November 2013. Tomsk: tomsk State University. pp. 85-89. (In Russian).

10. Shevelyov, O.G. & Poddubnyy, V.V. (2010) Complex investigation of texts with the system "StyleAnalyzer". In: Grzyber, P., Kelih, E. & Macutek, J. (eds) *Text and Lanquage*. Vienna: Praesens Verlag. pp. 207-212.

11. Van Rijsbergen, C.J. (1979) *Information Retrieval*. London: Butterworths.