УДК 004.93

TEXT CATEGORIZATION BY COEVOLUTIONARY GENETIC ALGORITHM

T. O. Gasanova[1], R. B. Sergienko[1], W. Minker[1], E. S. Semenkin[2]

[1]Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany
E-mail: taniagasanova@yandex.ru, romaserg@list.ru, wolfgang.minker@uni-ulm.de
[2]Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarskiy Rabochiy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: eugenesemenkin@yandex.ru

In this paper we propose a model of text categorization, which combines a word clustering algorithm with coevolutionary genetic algorithm. Clustering is used in order to reduce the feature space dimension and genetic algorithm optimizes parameters of the model. The proposed method can be applied in large-scale information retrieval and data mining problems and it can be easily transportable to different domains and different languages since our approach does not require any domain-related or linguistic information. The performance on the data from the text-mining campaign DEFT'08 shows that the proposed method can compete with existing information retrieval models.

Keywords: text categorization, term relevance estimation, clustering, coevolutionary algorithm.

КАТЕГОРИЗАЦИЯ ДОКУМЕНТОВ КОЭВОЛЮЦИОННЫМ ГЕНЕТИЧЕСКИМ АЛГОРИТМОМ

Т. О. Гасанова[1], Р. Б. Сергиенко[1], В. Минкер[1], Е. С. Семенкин[2]

[1]Ульмский Университет
Германия, 89081, Ульм, Аллея Альберта Эйнштейна, 43
E-mail: taniagasanova@yandex.ru, romaserg@list.ru, wolfgang.minker@uni-ulm.de
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: eugenesemenkin@yandex.ru

Предлагается модель категоризации документов, которая комбинирует алгоритм кластеризации слов и коэволюционный генетический алгоритм. Кластеризация используется для сокращения признакового пространства, а генетический алгоритм для оптимизации параметров модели. Предложенный метод был применен для задач извлечения информации и анализа данных высокой размерности и может быть без затруднений адаптирован для решения задач из различных предметных областей на различных языках, так как наш подход не требует дополнительной предметной или лингвистической информации. Показана эффективность на задачах DEFT'08 в сравнении с существующими моделями извлечения информации.

Ключевые слова: категоризация документов, оценка релевантности термов, кластеризация, коэволюционный алгоритм.

Nowadays, Internet and the World Wide Web generate a huge amount of textual information. It is increasingly important to develop methods of text processing such as text categorization. Text categorization can be considered to be a part of natural language understanding, where there is a set of predefined categories and the task is to automatically assign new documents to one of these categories. There are many approaches to the analysis and categorization of text, but they could be roughly divided into statistical approaches, rule-based approaches and their combinations. Furthermore, the method of text preprocessing and text representation influences the results that we obtained even with the same methods.

Related work was done by the participants of the fourth edition DEFT text mining campaign [1–8], which have worked with the same data using some classic approaches for text categorization and combinations of standard algorithms with the original ideas.

The challenge of 2008 year involves the articles classification by genre and category. This paper reports on the results obtained using only the second task of the campaign and focuses on detecting the category, however articles are still from two sources: Le Monde (a French daily newspaper), and Wikipedia. Articles from this task are divided into five categories: France, International, Literature, Science, Society and are labeled by experts.

The proposed approach consists of preprocessing step, when we extract all words from the train set regardless of the case of the letters and excluding the punctuation. Then using our formula for word relevance estimation and applying hierarchical clustering algorithm we obtain a set of clusters and assign a common value to the whole cluster. However these common values do not provide the maxi-

mum classification quality, and therefore we suggest using hybrid genetic algorithm to improve the values corresponding to a single category and coevolutionary genetic algorithm with cooperation scheme [9] to improve all values in parallel.

Genetic algorithms are well-known and widely used methods of global optimization since they make no assumptions about the problem being optimized; they can work with algorithmically defined criterions on the binary, real, nominal, mixed etc. data types, they search in parallel in different regions of the feature space. To provide local convergence genetic algorithms are often hybridized with methods of local optimization. Due to these features in our approach genetic algorithms combined with local search have been chosen as an optimization tool. In order to improve the accuracy rate using all words we propose a coevolutionary genetic algorithm where individual algorithms do not compete with each other, but exchange the information between subpopulations.

This paper is organized as follows: Section 2 describes the task and the DEFT'08 corpora. Section 3 explains our approach. In Section 4 we discuss our experimental results in comparison with results observed by DEFT participants. Finally, we draw some conclusions and discuss future research directions in Section 5.

Problem Description. The data for testing of DEFT 2008 edition is related to the text classification by categories and genres. The data consists of two corpora containing articles of two genres: articles extracted from French daily newspaper Le Monde and encyclopedic articles from Wikipedia in French language.

In this paper we use the second database, focusing on detecting the category, but nevertheless with the two types of documents. The given database is divided into five common categories: France (FRA), International (INT), Literature (LIV), Science (SCI), and Society (SOC). The data is divided into train (60 % of the whole number of articles) and test set (40 %). The main difficulty of the given database is that the articles in every category are written in two different genres. To apply our algorithms we extracted all words which appear in train set regardless of the letter case and we also excluded dots, commas and other punctual signs. At the end we obtained 262400 words which we enumerated and each article is represented as the list of word numbers. We have not used any additional filtering as excluding the stop or ignore words.

The initial database has been preprocessed to be a binary matrix with rows representing utterances and columns representing the words from the vocabulary. An element from this binary matrix, a_{ij}, equals to 1 if in utterance i the word j appears and equals to 0 if it does not appear.

Utterance duplicates were removed. The preprocessed database consisting of 24458 utterances was divided into train (22 020 utterances, 90,032 %) and test set (2438 utterances, 9,968 %) such that the percentage of classes remained the same in both sets. The size of the dictionary of the whole database is 3464 words, 3294 words appear in training set, 1124 words appear in test set, 170 words which appear only in test set and do not appear in training set (unknown words), 33 utterances consisted of only unknown words, and 160 utterances included at least one unknown word.

The Proposed Approach. The aim of this work is to develop algorithms which are able to solve the given task without extra knowledge related to domain or language and nevertheless the method should provide good classification quality in comparison with the participants of DEFT'08.

The idea is that every word that appears in the article has to contribute some value to the certain class and the class with the biggest value we define as a winner for this article. For each term we assign a real number term relevance that depends on the relative frequency of the word occurrence. Term relevance is calculated using a modified formula of fuzzy rules relevance estimation for fuzzy classifier [10]. Membership function has been replaced by word frequency in the current class.

Let L be the number of classes; n_i is the number of articles of the i^{th} class; N_{ji} is the number of j^{th} word occurrence in all articles of the i^{th} class; $T_{ji} = \frac{N_{ji}}{n_i}$ is the relative frequency of j^{th} word occurrence in the i^{th} class.

$R_j = \max_i T_{ji}, S_j = \arg(\max_i T_{ji})$, where S_j is the number of class which we assign to jth word. The term relevance, C_j, is given by

$$C_j = \frac{1}{\sum_{i=1}^{L} T_{ji}} (R_j - \frac{1}{L-1} \sum_{\substack{i=1 \\ i \neq S_j}}^{L} T_{ji}).$$

C_j is higher if the word occurs often in few classes than if it appears in many classes. In our approach as a decision rule: for the given article we calculate the values A_i (for each i^{th} class): $A_i = \sum_{j:S_j=i} C_j$.

Then we find the number of class which achieves maximum of A_i. The results are shown in table 1.

The average F-score obtained by DEFT'08 participants on the test data was 81.1 and ours reached only 80.26. In the next Sections we describe some ways to improve our results.

Word Clustering. Over 250 000 words have been extracted from the train data. For each word we assigned the C value and the number of class where it contributes. Due to the size of the dictionary it is time consuming and nontrivial to directly apply any optimization method. Therefore we suggest to preprocess our dictionary in the way that words that have equal or similar C values be in the same cluster and one common C value will be assigned to all words from this cluster. It should be mentioned that our preprocessing stage does not use any specific linguistic information, expert knowledge or domain related information. Therefore it can be easily transportable to the data from another domain or even in another language.

In the table 2 we show how dramatically the size of dictionary decreases if we combine only words which have identical C values which mean that there is no difference between them in terms of our approach.

Table 1

Precision, Recall and F-score obtained by using C-values to estimate the word relevance

	Precision	Recall	F-score	Accuracy
Train set	91,4912	89,5783	90,5246	90,9045
Test set	82,5306	78,1136	80,2614	80,6347

Table 2

Total number of words and number of words with unique C values per category

Category	Number of words	Number of words with unique C values
France	29145	5192
International	46102	7238
Literature	65919	8068
Science	80412	5760
Society	40822	4793
Total	*262400*	*31051*

However from the Table 2 one can see that the number of clusters is still too large and in order to reduce the dictionary size we take hierarchical agglomerative clustering

As a common C value of the cluster we calculate the arithmetic mean of all word C values from this cluster. To choose which clusters are joint on the current step we calculate all distances between clusters: $dist(X,Y) = \frac{1}{N}\frac{1}{M}\sum_i \sum_j \left\| X_i - Y_j \right\|$, where N is the number of words in cluster X and M is the number of words in cluster Y; and we unite the closest clusters.

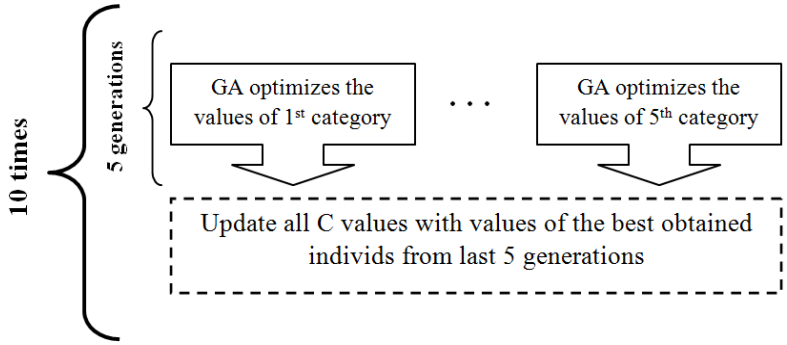
Genetic Algorithm. After we clustered words in the dictionary there is a hierarchical tree for each category and assigned values to all clusters. The question if these values are global optimum remains open. There is no evidence that the current values are even a local maximum of classification quality function.

To optimize C values when there is a predefined set of clusters for the certain category we suggest to apply genetic algorithm hybridized with local search due to its relative simplicity and global convergence, and it does not require any a priori information about behavior of the classification quality function.

In this work we apply a local search algorithm only to the best obtained individual to make sure that it reaches at least a local maximum. The C values of other categories are fixed and only the C values of the current class are being optimized. Each individual represents C values for the current category encoded to a binary string. As a fitness function we use the F-score on train set calculated with the fixed C values and C values of the individual.

Coevolutionary Genetic Algorithm. In order to take advantage of all C values improvement we propose to apply cooperative coevolutionary genetic algorithm with local search. The main stages of applied method are shown in figure.

On the first phase all individual genetic algorithms work separately (for each of them other C values are fixed and the task is to optimize C values which belong to the corresponding class), the length of this phase defines how many generations individual algorithms can work without information exchange. Then we stop all separate algorithms and update all C values which have been obtained by the best individuals of all algorithms. We repeat these two stages until we reach the maximum number of generations.



Coevolutionary genetic algorithm for C values optimization

This variant of coevolutionary algorithm uses cooperative scheme in order to achieve higher performance than each individual algorithm, in this case subpopulations do not exchange individuals, only information that influences the fitness function calculation.

Experimental Results. After the The DEFT ("Défi Fouille de Texte") Evaluation Package has been used for algorithms application and results comparison. In order to evaluate obtained results with the campaign participants we have to use the same measure of classification quality: precision, recall and F-score.

Precision for each class *i* is calculated as the number of correctly classified articles for class *i* divided by the number of all articles which algorithm assigned for this class. Recall is the number of correctly classified articles for class *i* divided by the number of articles that should have been in this class. Overall precision and recall are calculated as the arithmetic mean of the precisions and recalls for all classes (macro-average). F-score is calculated as the harmonic mean of precision and recall.

First, we present our results obtained by hybrid genetic algorithm with 200 individuals and 200 generations, tournament selection (tournament size: 3), weak mutation, uniform crossover (table 3).

From the table 3 one can see that words from the category SOC influence more on the classification quality than words which belong to other classes and with the results of category INT and SOC we achieve better performance than the average result of participants.

Finally, Table 4 provides results obtained by coevolutionary algorithm with different number of clusters.

From the Table 4 one can conclude that with clusters number from 40 to 70 we achieve F-scores which outperform the average results of the DEFT'08 participants. The best obtained F-score on the test data achieved with 60 clusters and then if we increase the clusters number we will see decreasing of the classification quality.

Conclusion and Future Directions. In this paper we present how hybrid genetic algorithms can improve the classification quality in term of precision, recall, F-score and accuracy. To reduce the feature space dimension we applied the hierarchical agglomerative clustering using the alternative formula for word relevance estimation which gives better results on the given data than the simple relative frequency. This approach has been applied for the corpora from DEFT'08 and it can compete with the results of campaign participants despite that the method does not use any linguistic or domain related information.

The obtained results are promising, however the methods require more investigation, e. g. there were no experiments with different number of clusters for each category. Our next step will be to design an automatic method that optimizes the number of clusters for each class. Future work also involves applying our method to detect genre and category in the first task of the database.

Table 3

Results of genetic algorithm application for each class (averaged by 50 runs)

Class	FRA	INT	LIV	SCI	SOC
Test set	80,4053	82,0047	80,7685	80,6156	84,4262
Train set	91,4055	93,1635	91,7792	91,3461	94,0594

Table 4

Results of coevolutionary genetic algorithm application with different clusters number for each category

Number of clusters for each category	Train set	Test set
10	93,5851	84,063
20	93,5244	84,3051
30	93,4268	84,3688
40	93,6178	84,4644
50	93,7986	84,591
60	94,0533	84,925
70	93,5512	84,4979

References

1. Bechet F., Beze M. E., Torres-Moreno J.-M. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, p. 27–36.

2. Charnois T., Doucet A., Mathet Y., Rioult F. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8–13, 2008). DEFT '08. TALN, Avignon, France, p. 37–46.

3. Charton E., Camelin N., Acuna-Agost R., Gotab P., Lavalley R., Kessler R., Fernandez S. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8–13, 2008). DEFT '08. TALN, Avignon, France, p. 47–56.

4. Cleuziou G., Poudat C. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, p. 57–64.

5. DEFT (DÉfi Fouille de Textes) http://deft.limsi.fr/.

6. European Language Recourses Association. DEFT'08 Evaluation Package http://catalog.elra.info/product_info.php?cPath=42_43&products_id=1165.

7. Plantie M., Roche M. and Dray G. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8–13, 2008). DEFT '08. TALN, Avignon, France, 65–74.

8. Trinh A.-P., Buffoni D., Gallinari P. 2008. Proceedings of the 4th DEFT Workshop (Avignon, France, June 8–13, 2008). DEFT '08. TALN, Avignon, France, p. 75–86.

9. Potter M. A., De Jong K. A. 2000. Cooperative coevolution: an architecture for evolving coadapted subcomponents. Trans. Evolutionary Computation, 8 (Jan. 2000), p. 1–29.

10. Ishibuchi H., Nakashima T. and Murata T. 1999. Trans. on Systems, Man, and Cybernetics, vol. 29, p. 601–618.

УДК 004.93

A NEW METHOD FOR NATURAL LANGUAGE CALL ROUTING PROBLEM SOLVING

T. O. Gasanova[1], R. B. Sergienko[1], W. Minker[1], E. A. Zhukov[2]

[1]Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany
E-mail: taniagasanova@yandex.ru, romaserg@list.ru, wolfgang.minker@uni-ulm.de
[2]Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: zhukov.krsk@gmail.com

Natural Language call routing remains a complex and challenging research area in machine intelligence and language understanding. This paper is in the area of classifying user utterances into different categories. The focus is on design of algorithm that combines supervised and unsupervised learning models in order to improve classification quality. We have shown that the proposed approach is able to outperform existing methods on a large dataset and do not require morphological and stop-word filtering. In this paper we present a new formula for term relevance estimation, which is a modification of fuzzy rules relevance estimation for fuzzy classifier. We propose to split the classification task into two steps: 1) "garbage" class identification; 2) further classification into meaningful classes. The performance of the proposed algorithm is compared to several standard classification algorithms on the database without the "garbage" class and found to outperform them with the accuracy rate of 85,55 %. Combination of our approach with 9-NN algorithm for two-stage classification problem definition provides the accuracy rate of 77,11 % for test sample at whole.

Keywords: call classification, term relevance estimation, natural language processing.

НОВЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ МАРШРУТИЗАЦИИ ВЫЗОВОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Т. О. Гасанова[1], Р. Б. Сергиенко[1], В. Минкер[1], Е. А. Жуков[2]

[1]Ульмский Университет
Германия, 89081, Ульм, Аллея Альберта Эйнштейна, 43.
E-mail: taniagasanova@yandex.ru, romaserg@list.ru, wolfgang.minker@uni-ulm.de
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: zhukov.krsk@gmail.com

Маршрутизация вызовов, основанная на обработке естественного языка, представляет собой сложную и перспективную область исследований в интеллектуальных машинных методах и интерпретации языка. Рассмотрена категоризации пользовательских заявок. Сделан акцент на комбинировании технологий машинного обучения с учителем и без учителя в целях повышения точности классификации. Показано, что разработанный подход способен превзойти существующие алгоритмы на больших базах данных и не требующих морфологического анализа или фильтра в виде «стоп-слова». В предлагаемом подходе осуществляется декомпозиция задачи классификации, к которой сводится маршрутизация вызовов, на две стадии: обнаружение «мусорного» класса и отнесение объектов к значимым классам. Предлагается новая формула оценки релевантности термов при определении значимых классов, которая является модификацией оценки релевантности нечетких