Feature Selection for Natural Language Call Routing Based on Self-Adaptive Genetic Algorithm

To cite this article: A Koromyslova et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 173 012008

View the article online for updates and enhancements.

Related content
- Discriminative feature selection for visual tracking
  Junkai Ma, Haibo Luo, Wei Zhou et al.
- Sentiment analysis of feature ranking methods for classification accuracy
  Shashank Joseph, Calvin Mugauri and S Sumathy
- A new Self-Adaptive dispatching system for local clusters
  Bowen Kan, Jingyan Shi and Xiaofeng Lei
Feature Selection for Natural Language Call Routing Based on Self-Adaptive Genetic Algorithm

A Koromyslova 1, M Semenkina 1, R Sergienko 2

1 Reshetnev Siberian State Aerospace University, Krasnoyarsky rabochy avenue 31, Krasnoyarsk, Russia, 660037
2 Ulm University, Lise-Meitner-Str. 9, Ulm, Germany, 89081

E-mail: akoromyslova@mail.ru, semenkina88@mail.ru, roman.sergienko@uni-ulm.de

Abstract: The text classification problem for natural language call routing was considered in the paper. Seven different term weighting methods were applied. As dimensionality reduction methods, the feature selection based on self-adaptive GA is considered. k-NN, linear SVM and ANN were used as classification algorithms. The tasks of the research are the following: perform research of text classification for natural language call routing with different term weighting methods and classification algorithms and investigate the feature selection method based on self-adaptive GA. The numerical results showed that the most effective term weighting is TRR. The most effective classification algorithm is ANN. Feature selection with self-adaptive GA provides improvement of classification effectiveness and significant dimensionality reduction with all term weighting methods and with all classification algorithms.

1. Introduction

Natural language call routing is an important problem in the design of modern automatic call services and the solving of this problem could lead to improvement of the call service [21]. Generally natural language call routing can be considered as two different problems. The first one is speech recognition of calls and the second one is topic categorization of users utterances for further routing. Topic categorization of users utterances can be also useful for multidomain spoken dialogue system design [12]. In this work we treat call routing as an example of a text classification application.

In the vector space model [16] text classification is considered as a machine learning problem. The complexity of text categorization with a vector space model is compounded by the need to extract the numerical data from text information before applying machine learning algorithms. Therefore, text classification consists of two parts: text preprocessing and classification algorithm application using the obtained numerical data. Text preprocessing comprises three stages:

- Textual feature extraction.
- Term weighting
- Dimensionality reduction.

The first one is the textual feature extraction based on raw preprocessing of the documents. This process includes deleting punctuation, transforming capital letters to lowercase, and additional procedures such as stop-words filtering [4] and stemming [14]. Stop-words list contains pronouns, prepositions, articles and other words that usually have no importance for the classification. Using stemming it is possible to join different forms of the same word into one textual feature.

The second stage is the numerical feature extraction based on term weighting. For term weighting we use “bag-of-words” model, in which the word order is ignored. There exist different unsupervised and supervised term weighting methods. The most well-known unsupervised term weighting method is TFIDF [15]. The following supervised term weighting methods are also considered in the paper:
Gain Ratio (GR) [3], Confident Weights (CW) [10], Term Second Moment (TM2) [22], Relevance Frequency (RF) [11], Term Relevance Ratio (TRR) [9], and Novel Term Weighting (NTW) [18]; these methods involve information about the classes of the documents.

As a rule, the dimensionality for text classification problems is high even after stop-words filtering and stemming. Due to the high dimensionality, the classification may be inappropriate time-consuming, especially for real-time systems such as natural language call routing. Therefore, the next stage of preprocessing is the dimensionality reduction based on numerical features; it is possible with feature selection or feature transformation. In our research we use a feature selection method based on genetic algorithm or GA-based wrapper.

One of the most complicated problems with GA applications is setting algorithm parameters. A conventional genetic algorithm has at least three methods of selection (proportional, tournament, and rank), three methods of recombination (one-point, two-point, and uniform). Mutation probability requires tuning as well. The amount of various combinations can be estimated at tens. Exhaustive search of combinations requires a lot of time and computational power, especially for such time-consuming problems as GA applications for machine learning. Parameters combination selection at random can be also insufficient as algorithm efficiency on same problem can differ very much for various parameters setting. This problem can be solved with self-configuring GA [17] or co-evolutionary GA [19]. Therefore, we propose a use of self-adaptive GA for the feature selection in the field of natural language call routing.

As classification algorithms we use the k-NN algorithm, the linear SVM, and artificial neural networks (ANN). Some comparative studies of machine learning algorithms in the field of text classification showed high classification effectiveness of k-NN, SVM-based algorithms, and ANN [2, 7, 8, 10, 13].

The tasks of our research are the following:
- Perform research of text classification for natural language call routing with different term weighting methods and classification algorithms.
- Investigate the feature selection method based on self-adaptive GA.

The paper is organized as follows: In Section 2, we describe the considered corpus for natural language call routing. Section 3 describes the considered term weighting methods. The GA-based feature selection is described in Section 4. The self-adaptive GA are described in Section 5. Section 6 contains short description of classification algorithms. The results of numerical experiments are presented in Section 7. Finally, we provide concluding remarks in Section 8.

2. Corpus description

The data for testing and evaluation consists of 292,156 user utterances recorded in English language from caller interactions with commercial automated agents. Utterances are short and contain only one phrase for further routing. The database contains calls in textual format after speech recognition. The database is provided by the company Speech Cycle (New York, USA). Utterances from this database are manually labelled by experts and divided into 20 classes (such as appointments, operator, bill, internet, phone and technical support). One of them is a special class TE-NOMATCH which includes utterances that cannot be put into another class or can be put into more than one class.

The database contains 45 unclassified calls and they were removed. The database contains also 23,561 empty calls without any words. These calls were placed in the class TE-NOMATCH automatically and they were also removed from the database. As a rule, the calls are short in the database; many of them contain only one or two words. The average length of an utterance is 4.66 words, the maximal length is 19 words. There are a lot of identical utterances in the database; the corpus contains only 24,458 unique non-empty classified calls. The corpus is unbalanced. The largest class contains 27.05% and the smallest one contains 0.04% of the unique calls.

Due to the very high frequency of a small number of utterances in the corpus, we formulate two different problem definitions.
**Problem definition 1.** The whole database with 268,550 classified non-empty calls is used for training and test sets forming. Numbers of repetitions of the utterances in training and test sets are used as weights for classification. This problem definition is the closest to the real situation but frequently repeated utterances decrease difference between preprocessing and classification methods. Additionally, there are some identical utterances in training and test sets simultaneously. In this case the over-fitting problem of classification may be hidden. Therefore, this problem definition is not very appropriate for the comparative study.

**Problem definition 2.** Before training and test samples forming, all utterance duplicates were removed from the database. It means that there is no intersection between training and test sets and frequency of utterances is ignored. Therefore, the problem definition 1 is suitable for the quality estimation of the real natural language call routing system; the problem definition 2 is the most appropriate for the comparative study of different preprocessing and classification methods.

For statistical analysis we performed 20 different divisions of the database into training and test samples randomly. This procedure was performed for two problem definitions separately. The train samples contain 90% of the calls and the test samples contain 10% of the calls. For each training sample we have designed a dictionary of unique words which appear in the training sample after deleting punctuation and transforming capital letters to lowercase. The size of the dictionary varies from 3,275 to 3,329 words for problem definition 1 and from 3,277 to 3,311 for problem definition 2.

### 3. Term weighting methods

As a rule, term weighting is a multiplication of two parts: the part based on the term frequency in a document (TF) and the part based on the term frequency in the whole training database. The TF-part is fixed for all considered term weighting methods and is calculated as following:

$$TF_{ij} = \log(f_{ij} + 1) ; \quad tf_{ij} = \frac{n_{ij}}{N_j},$$

where $n_{ij}$ is the number of times the $i^{th}$ word occurs in the $j^{th}$ document, $N_j$ is the document size (number of words in the document). The second part of the term weighting is calculated once for each word from the dictionary and does not depend on an utterance for classification. We consider seven different methods for the calculation of the second part of term weighting.

#### 3.1. Inverse Document Frequency (IDF)

IDF is a well-known unsupervised term weighting method which was proposed in [15]. There are some modifications of IDF and we use the most popular one:

$$idf = \log\left(\frac{|D|}{n_i}\right),$$

where $|D|$ is the number of documents in the training set and $n_i$ is the number of documents that have the $i^{th}$ word.

#### 3.2. Gain Ratio (GR)

Gain Ratio (GR) is mainly used in term selection [24], but in [3] it was shown that it could also be used for weighting terms. The definition of GR is as follows:
\[
GR(t_i, c_j) = \frac{\sum_{c \in \mathcal{C}} \sum_{t \in \mathcal{I}} M(t, c)}{-\sum_{c \in \mathcal{C}} P(c) \cdot \log P(c)},
\]

\[
M(t, c) = P(t, c) \cdot \log \frac{P(t, c)}{P(t) \cdot P(c)},
\]

where \( P(t, c) \) is the relative frequency that a document contains the term \( t \) and belongs to the category \( c \); \( P(t) \) is the relative frequency that a document contains the term \( t \) and \( P(c) \) is the relative frequency that a document belongs to category \( c \). Then, the weight of the term \( t \), is the max value between all categories as follows:

\[
GR(t_i) = \max_{c_j \in \mathcal{C}} GR(t_j, c_j),
\]

where \( \mathcal{C} \) is a set of all classes.

### 3.3. Confident Weights (CW)

This supervised term weighting approach has been proposed in [20]. Firstly, the proportion of documents containing term \( t \) is defined as the Wilson proportion estimate \( p(x, n) \) by the following equation:

\[
p(x, n) = \frac{x + 0.5Z_{\alpha/2}^2}{n + Z_{\alpha/2}^2},
\]

where \( x \) is the number of documents containing the term \( t \) in the given corpus, \( n \) is the number of documents in the corpus and \( \Phi \left( Z_{\alpha/2} \right) = \alpha/2 \), where \( \Phi \) the \( t \)-distribution (Students law) when \( n < 30 \) and the normal distribution when \( n \geq 30 \).

In this work \( \alpha = 0.95 \) and \( 0.5Z_{\alpha/2}^2 = 1.96 \) (as recommended by the authors of the method). For each term \( t \) and each class \( c \) two functions \( p_{\text{pos}}(x, n) \) and \( p_{\text{neg}}(x, n) \) are calculated. For \( p_{\text{pos}}(x, n) \) \( x \) is the number of documents which belong to the class \( c \) and have term \( t \); \( n \) is the number of documents which belong to the class \( c \). For \( p_{\text{neg}}(x, n) \) \( x \) is the number of documents which have the term \( t \) but do not belong to the class \( c \); \( n \) is the number of documents which do not belong to the class \( c \). The confidence interval \((p^-, p^+)\) at 0.95 is calculated using the following equation:

\[
M = 0.5Z_{\alpha/2}^2 \sqrt{\frac{p(1-p)}{n + Z_{\alpha/2}^2}},
\]

\[
p^- = p - M; \quad p^+ = p + M.
\]

The strength of the term \( t \) in the category \( c \) is defined as the follows:

\[
\text{str}(t, c) = \begin{cases} 
\log_2 \frac{2p_{\text{pos}}^-}{p_{\text{pos}}^+ + p_{\text{neg}}^-}, & \text{if } p_{\text{pos}}^+ > p_{\text{neg}}^-; \\
0, & \text{otherwise.}
\end{cases}
\]
The maximum strength \((\text{Maxstr})\) of the term \(t_i\) is calculated as follows:

\[
\text{Maxstr}(t_i) = \max_{c_j \in \mathcal{C}} \text{str}(t_i, c_j)^2.
\]

### 3.4. Term Second Moment (TM2)

This supervised term weighting method was proposed in [22]. Let \(P(c_j|t)\) be the empirical estimation of the probability that a document belongs to the category \(c_j\) with the condition that the document contains the term \(t\); \(P(c_j)\) is the empirical estimation of the probability that a document belongs to the category \(c_j\) without any conditions. The idea is the following: the more \(P(c_j|t)\) is different from \(P(c_j)\), the more important the term \(t_i\) is. Therefore, we can calculate the term weight as the following:

\[
\text{TM2}(t_i) = \sum_{j=1}^{C} (P(c_j|t) - P(c_j))^2,
\]

where \(\mathcal{C}\) is a set of all classes.

### 3.5. Relevance Frequency (RF)

The RF term weighting method was proposed in [11] and is calculated as the following:

\[
\text{rf}(t_i) = \max_{c_j \in \mathcal{C}} \text{rf}(t_i, c_j),
\]

\[
\text{rf}(t_i, c_j) = \log_2 \left( 2 + \frac{a_j}{\max(1, a_j)} \right),
\]

where \(a_j\) is the number of documents of the category \(c_j\) which contain the term \(t_i\) and \(\overline{a_j}\) is the number of documents of all the other categories which also contain this term.

### 3.6. Term Relevance Ratio (TRR)

The TRR method [9] uses \(tf\) weights and it is calculated as the following:

\[
\text{TRR}(t_i, c_j) = \log_2 \left( 2 + \frac{\overline{P}(t_i|c_j)}{P(t_i|c_j)} \right)
\]

\[
P(t_i|c_j) = \frac{\sum_{k=1}^{V} t_{i,k}}{\sum_{l=1,k}^{V} \sum_{k=1}^{V} t_{i,k}},
\]

\[
\overline{P}(t_i|c_j) = \frac{\sum_{k=1}^{V} t_{i,k}}{\sum_{l=1}^{\overline{c}_j} \sum_{k=1}^{V} t_{i,k}},
\]

\[
\text{TRR}(t_i) = \max_{c_j \in \mathcal{C}} \text{TRR}(t_i, c_j),
\]

where \(c_j\) is a class of the document, \(\overline{c}_j\) is all of the other classes of \(c_j\), \(V\) is the vocabulary of the training data and \(T_c\) is the document set of the class \(c\).
3.7 Novel Term Weighting (NTW)
This method was proposed in [1, 18]. The details of the procedure are the following. Let $L$ be the number of classes; $n_i$ is the number of documents which belong to the $i^{th}$ class; $N_{ij}$ is the number of occurrences of the $j^{th}$ word in all documents from the $i^{th}$ class. $T_{ij} = N_{ij}/n_i$ is the relative frequency of occurrences of the $j^{th}$ word in the $i^{th}$ class; $R_j = \max_i T_{ij}$; $S_j = \arg \max_i T_{ij}$ is the class which we assign to the $j^{th}$ word. The term relevance $C_j$ is calculated by the following:

$$C_j = \frac{1}{\sum_{i=1}^{L} T_{ij}} \left( R_j - \frac{1}{L-1} \cdot \sum_{i=1, i \neq S_j}^{L} T_{ij} \right).$$

4. GA-based feature selection
The popular kind of wrappers is feature selection based on genetic algorithms (GA). GA is used as an optimization algorithm for finding the optimal or the sub-optimal subset of the original feature set with the predefined classification algorithm [23]. The details of the GA-based wrapper are the following:

1. Randomly initialize the population of the binary strings with the length equals to dimensionality of the considered classification problem. Set generation counter $g = 0$.
2. Apply the classification algorithm on the validation set for all individuals form the population (with including only features with 1 value in the chromosomes)
3. Set the classification effectiveness measure (i.e. $F$-score) on the validation set as a fitness function value for all individuals.
4. Check if the generation counter is greater than some predefined value $G$ (maximal number of generations): if $g > G$ than go to the step 7; otherwise go to step 5.
5. Using GA operators: selection, crossover and mutation, form the next population.
6. Increment the generation counter $g = g + 1$. Go to step 2.
7. Put the best individual as a solution of the feature selection procedure; apply the chosen feature subset for the test set.

5. Self-adaptive GA
For solving the problem of GA setting parameters we use the self-configuring algorithm that was proposed in [17]. The scheme of the self-configuring GA is presented in Figure 1.

In the self-configuring GA (SCGA), different types of selection, recombination, and different levels of mutation are performed simultaneously. In the beginning of the algorithm, all types of GA operators have the same probability to be use for a new off-spring generation. After that, the dynamic adaptation of probabilities is performed according “usefulness” of GA operator types in terms of fitness function.
The details of the self-adjusting GA are the following:

1. Put the probabilities of all used types of GA operators: \( p_{ij} = 1/N_i \), where \( j = 1, \ldots, N_i \), with \( N_i \) being the number of types of the \( i \)th operator, \( i = 1, \ldots, N \), where \( N \) is the number of GA operators. In our case \( N = 3 \): selection, recombination, and mutation.

2. Set threshold probabilities for all used types of GA operators: \( \widetilde{p}_{ij} = 3/(10 \cdot N_i) \).

3. Generate new population. For each offspring, we randomly choose types of selection, recombination, and mutation according to probabilities \( p_{ij} \).

4. Recalculate the probabilities with the following:

   4.1. For each \( i \)th operator do:
      4.1.1. Set \( S_i = 0 \).
      4.1.2. For each \( j \)th type of the \( i \)th operator do:
         4.1.2.1. If \( p_{ij} < \widetilde{p}_{ij} + 1/(T \cdot N_i) \) AND \( p_{ij} > \widetilde{p}_{ij} \), where \( T \) is the number of generations, then:
            \[ S_i = S_i + (p_{ij} - \widetilde{p}_{ij}) ; \quad p_{ij} = \widetilde{p}_{ij}. \]
         4.1.2.2. If \( p_{ij} > \widetilde{p}_{ij} + 1/(T \cdot N_i) \) then:
            \[ S_i = S_i + 1/(T \cdot N_i) ; \quad p_{ij} = p_{ij} - 1/(T \cdot N_i). \]
      4.1.2.3. Calculate average fitness function \( F_{ij} \) of all off-springs of the current generation that were generated with the \( j \)th type of the \( i \)th operator.
      4.1.3. Find the best type \( d \) of the \( i \)th operator with the maximal fitness function on the current generation and recalculate its probability:
            \[ p_{id} = p_{id} + S_i. \]
   5. Check stop criterion. If TRUE then: END; else: go to the step 3.

Moreover two co-evolutionary algorithms (CEA) [19] have been realized. Both use island model with four islands and cooperative-competitive strategy for resources allocating. First of them (CEA) used four best genetic algorithms for different types of problems, second one use four independent (on the start and on the probabilities recalculation stage) SCGA (CESCGA).

6. Classification Algorithms

   As classification algorithms we use the k-NN algorithm, the linear SVM algorithm, and ANN. The classification criterion is the macro \( F \)-score [6] which is appropriate for classification problems with
unbalanced classes. For k-NN we performed validation of $k$ from 1 to 15 on the validation sample. We used 80% of the train sample for the first level of learning and 20% for the validation. The same scheme of validation is used for feature selection based on self-adjusting GA.

6.1. Genetic Algorithms for Automated Artificial Neural Network Design

Artificial neural network (ANN) is a set of interconnected neurons. Typically, the transfer functions of all the neurons in the neural network are fixed, and the weights are the parameters of the neural network.

In case of GA using for ANN weights tuning: weights are recorded sequentially in the chromosome as a binary code. The number of bits that are used to encode a single weighting coefficient depends on the accuracy and the spread of possible values of weights.

In case of GA using for ANN structure design (GA-ANN):
1. Hidden layers are coded sequentially;
2. Each neuron is encoded in four bits;
3. Each neuron will exist with probability equal to 0.3.
4. If in the network a neuron is not presented, its place in the chromosome is marked with zeros. Otherwise, it is randomly selected as one of the fifteen activation functions, whose number is written in binary code.

Since the efficiency of the genetic algorithm for ANN design depends on the dimension of the problem in hand, it is reasonable to avoid the use of uninformative features. The modification of the genetic algorithm for the choice of the most informative features during the automated design of ANN assumes the use of additional bits in the GA chromosomes. These bits determine whether an input is included in the input layer or not.

7. Results of numerical experiments

Tables 1-10 show the results of the numerical experiments for problem definitions 1 and 2 without dimensionality reduction and after feature selection with the self-adaptive GA-based wrapper. The best results in terms of $F$-score are bold (for each case independently). The column “Features” contains number of features after GA-based wrapper performing.

The numerical results showed that the most effective term weighting is TRR. The most effective classification algorithm is ANN. The most effective optimization technique is SCGA. Feature selection with self-configuring GA provides improvement of classification effectiveness and significant dimensionality reduction with all term weighting methods and with all classification algorithms.

**Table 1. Results for problem definition 1 with k-NN**

| Term weighting method | F-score | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection |
| IDF                   | 0.778    | 0.800 | 2781 |
| GR                    | 0.759    | 0.795 | 2851 |
| CW                    | 0.763    | 0.803 | 2537 |
| RF                    | 0.822    | 0.863 | 2509 |
| TM2                   | 0.742    | 0.778 | 3054 |
| TRR                   | **0.851** | **0.887** | 2346 |
| NTW                   | 0.847    | 0.881 | 2379 |

**Table 2. Results for problem definition 2 with k-NN**

| Term weighting method | F-score | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection |
| IDF                   | 0.643    | 0.667 | 2678 |
| GR                    | 0.629    | 0.671 | 2745 |
Table 3. Results for problem definition 1 with SVM

| Term weighting method | F-score   | Features |
|-----------------------|----------|----------|
|                        | All terms | Feature selection |
| IDF                   | 0.749    | 0.762    | 2617 |
| GR                    | 0.724    | 0.784    | 2789 |
| CW                    | 0.776    | 0.803    | 2537 |
| RF                    | 0.798    | 0.851    | 2386 |
| TM2                   | 0.712    | 0.765    | 3012 |
| TRR                   | 0.812    | 0.874    | 2239 |
| NTW                   | 0.803    | 0.869    | 2314 |

Table 4. Results for problem definition 2 with SVM

| Term weighting method | F-score   | Features |
|-----------------------|----------|----------|
|                        | All terms | Feature selection |
| IDF                   | 0.627    | 0.668    | 2775 |
| GR                    | 0.619    | 0.659    | 2948 |
| CW                    | 0.632    | 0.672    | 2896 |
| RF                    | 0.639    | 0.681    | 2583 |
| TM2                   | 0.611    | 0.647    | 3007 |
| TRR                   | 0.654    | 0.693    | 2469 |
| NTW                   | 0.648    | 0.685    | 2541 |

Table 5. Results for problem definition 1 with SCGA-ANN

| Term weighting method | F-score   | Features |
|-----------------------|----------|----------|
|                        | All terms | Feature selection |
| IDF                   | 0.699    | 0.796    | 2679 |
| GR                    | 0.731    | 0.784    | 2526 |
| CW                    | 0.766    | 0.812    | 2475 |
| RF                    | 0.837    | 0.850    | 2337 |
| TM2                   | 0.715    | 0.771    | 3145 |
| TRR                   | 0.844    | 0.891    | 2369 |
| NTW                   | 0.832    | 0.876    | 2215 |

Table 6. Results for problem definition 2 with SCGA-ANN

| Term weighting method | F-score   | Features |
|-----------------------|----------|----------|
|                        | All terms | Feature selection |
| IDF                   | 0.679    | 0.684    | 2598 |
| GR                    | 0.662    | 0.675    | 2657 |
| CW                    | 0.654    | 0.691    | 2431 |
| RF                    | 0.682    | 0.713    | 2231 |
| TM2                   | 0.621    | 0.663    | 2856 |
| TRR                   | 0.703    | 0.727    | 2258 |
| NTW                   | 0.699    | 0.717    | 2231 |
Table 7. Results for problem definition 1 with CEA-ANN

| Term weighting method | F-score  | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection | |
| IDF                   | 0.673    | 0.78     | 2563   |
| GR                    | 0.711    | 0.758    | 2408   |
| CW                    | 0.739    | 0.795    | 2353   |
| RF                    | 0.82     | 0.828    | 2212   |
| TM2                   | 0.69     | 0.754    | 3023   |
| TRR                   | 0.825    | 0.866    | 2241   |
| NTW                   | 0.813    | 0.852    | 2075   |

Table 8. Results for problem definition 2 with CEA-ANN

| Term weighting method | F-score  | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection | |
| IDF                   | 0.655    | 0.663    | 2456   |
| GR                    | 0.647    | 0.657    | 2529   |
| CW                    | 0.634    | 0.67     | 2291   |
| RF                    | 0.655    | 0.698    | 2085   |
| TM2                   | 0.599    | 0.641    | 2738   |
| TRR                   | 0.682    | 0.703    | 2143   |
| NTW                   | 0.675    | 0.696    | 2104   |

Table 9. Results for problem definition 1 with CESCGA-ANN

| Term weighting method | F-score  | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection | |
| IDF                   | 0.644    | 0.653    | 2384   |
| GR                    | 0.639    | 0.644    | 2461   |
| CW                    | 0.621    | 0.662    | 2219   |
| RF                    | 0.642    | 0.686    | 2021   |
| TM2                   | 0.586    | 0.629    | 2681   |
| TRR                   | 0.674    | 0.691    | 2084   |
| NTW                   | 0.667    | 0.687    | 2040   |

Table 10. Results for problem definition 2 with CESCGA-ANN

| Term weighting method | F-score  | Features |
|-----------------------|---------|----------|
|                       | All terms | Feature selection | |
| IDF                   | 0.67     | 0.773    | 2786   |
| GR                    | 0.697    | 0.76     | 2625   |
| CW                    | 0.738    | 0.785    | 2560   |
| RF                    | 0.817    | 0.821    | 2436   |
| TM2                   | 0.685    | 0.739    | 3238   |
| TRR                   | 0.822    | 0.87     | 2452   |
| NTW                   | 0.804    | 0.853    | 2314   |

8. Conclusions

The text classification problem for natural language call routing was considered in the paper. Seven different term weighting methods were applied. As dimensionality reduction methods, the feature
selection based on self-configuring GA is considered. k-NN, linear SVM and ANN were used as classification algorithms.

The numerical results showed that the most effective term weighting is TRR. The most effective classification algorithm is ANN. Feature selection with self-configuring GA provides improvement of classification effectiveness and significant dimensionality reduction with all term weighting methods and with all classification algorithms.

Acknowledgements
The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Region Science and Technology Support Fund to the research project №6-41-240822.

References
[1] Akhmedova S, Semenkin E and Sergienko R 2014 Automatically generated classifiers for opinion mining with different term weighting schemes Informatics in Control, Automation and Robotics (ICINCO 2014) 2 pp 845–850
[2] Baharudin B, Lee L H and Khan K 2010 A review of machine learning algorithms for text-documents classification Journal of advances in information technology 1(1) pp 4–20
[3] Debole F and Sebastiani F 2004 Supervised term weighting for automated text categorization In Text mining and its applications pp 81–97
[4] Fox C 1989 A stop list for general text In ACM SIGIR Forum 24 pp 19–21
[5] Gasanova T, Sergienko R, Akhmedova S, Semenkin E and Minker W 2014 Opinion mining and topic categorization with novel term weighting Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL 2014 pp 84–89
[6] Goutte C and Gaussier E 2005 A probabilistic interpretation of precision, recall and f-score, with implication for evaluation Advances in information retrieval pp 345–359
[7] Han E-H S, Karypis G, and Kumar V 2001 Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification
[8] Joachims T 2002 Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms
[9] Ko Y 2012 A study of term weighting schemes using class information for text classification Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval pp 1029–1030
[10] Kwon O-W and Lee J-H 2003 Text categorization based on k-nearest neighbor approach for web site classification Information Processing & Management 39(1) pp 25–44
[11] Lan M, Tan C L, Su J and Lu Y 2009 Supervised and traditional term weighting methods for automatic text categorization IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4) pp 721–735
[12] Lee C, Jung S, Kim S and Lee G 2009 Example based dialog modeling for practical multi-domain dialog system Speech Communication 51(5) pp 466–484
[13] Morariu D I, Vintan L N and Tresp V 2005 Metaclassification using SVM classifiers for text documents Intnl. Jnl. of Applied Mathematics and Computer Sciences 1(1)
[14] Porter M F 2001 Snowball: A language for stemming algorithms
[15] Salton G and Buckley C 1988 Term-weighting approaches in automatic text retrieval Information processing & management 24(5) pp 513–523
[16] Sebastiani F 2002 Machine learning in automated text categorization ACM computing surveys (CSUR) 34(1) pp 1–47
[17] Semenkin E and Semenkina M 2012 Self-configuring genetic programming algorithm with modified uniform crossover *2012 IEEE Congress on Evolutionary Computation*

[18] Sergienko R, Gasanova T, Semenkin E and Minker W 2014 Text categorization methods application for natural language call routing *11th International Conference on Informatics in Control, Automation and Robotics (ICINCO ’2014)* 2 pp 827–831

[19] Sergienko R and Semenkin E 2010 Competitive cooperation for strategy adaptation in coevolutionary genetic algorithm for constrained optimization *IEEE Congress on Evolutionary Computation*

[20] Soucy P and Mineau G W 2005 Beyond tfidf weighting for text categorization in the vector space model *IJCAI* 5 pp 1130–1135

[21] Suhm B, Bers J, McCarthy D, Freeman B, Getty D, Godfrey K and Peterson P 2002 A comparative study of speech in the call center: Natural language call routing vs. touch-tone menus *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* pp 283–290

[22] Xu H and Li C 2007 A novel term weighting scheme for automated text categorization *Seventh International Conference on Intelligent Systems Design and Applications (ISDA ’2007)* pp 759–764

[23] Yang J and Honavar V 1998 Feature subset selection using a genetic algorithm *Feature extraction, construction and selection* pp 117–136

[24] Yang Y and Pedersen J O 1997 A comparative study on feature selection in text categorization *ICML 97* pp 412–420