Baltic J. Modern Computing, Vol. 3 (2015), No. 4, pp. 307–317

Concept Pattern Based Text Classification System Development for Georgian Text Based Information Retrieval

Manana KHACHIDZE, Magda TSINTSADZE, Maia ARCHUADZE,
Gela BESIASHVILI

Iv. Javakhishvili Tbilisi State University, University St.3, 0179 Tbilisi, Georgia

manana.khachidze@tsu.ge, magda.tsintsadze@tsu.ge, maia.archuadze@tsu.ge,
gela.besiashvili@tsu.ge

Abstract. Presented work outlines the text classification system developed with appropriate four main modules and the algorithm of the text classification for the Georgian Language. The Heuristic Analysis is used to develop the concept-pattern describing appropriate class in a document collection and a new algorithm was developed and applied to get the set of term classes for every detected stem. The TF-IDF (Term Frequency-Inverse Document Frequency) scheme for term weight calculation is performed. The novelty of the method is that one generalized concept is created on the basis of each category of the documents sourcing the whole database. The generalized concept contains all high weight terms along with the terms not to be presented in the definition of the concept, so the concept contains information what terms are efficient for the class and what terms "should not" be present (not to contain the defined term is a characteristics itself). The generalized concept pattern, based on the above mentioned scheme, is compared to the patterns (high weight term set) formed for each document.
The system was tested on Georgian text based on 900 documents collected from 6 categories (classes - 150 documents in each). Using the generalized concept formation method the accuracy was increased by 11% and recall by 10 %, compared to k-Nearest Neighbors algorithm (KNN). On the basis of the received results we may conclude that the patterns constructed with the help of Analytical Heuristics method used for retrieval is quite promising and future modifications for better results are possible.

Keywords: Text Categorization, Information Retrieval, Concept-Patterns, Analytical Heuristics Method , Text Evaluation, Classification

1 Introduction

Information retrieval (IR) is one of the important issues in Information Technologies. Information Retrieval is not a result of only one kind of operation. The evaluation of the performance and relevancy of information retrieval systems mainly depend on the retrieval cycle precision and recall. One of the most important aspects in this cycle might be considered the process of document (text) classification that in fact represents the beginning of the search process.

Term frequencyinverse document frequency (TF-IDF) is one of the most commonly used term weighing schemes in the modern retrieval systems. It represents an empirical method 2003 used to evaluate words importance within the text retrieved in a document collection or corpus. There are lots of variations of this scheme used by search engines or text classification platforms. It is considered to be a good filter used for stop words selection in text summarization/classification issues as TF-IDF is not a single method, but a variety of techniques using statistical measure. The following algorithms: Nave Bayes(NB), Support Vector Machine(SVM), N-Grams, K-Nearest Neighborhood(KNN), Back Propagation Network, Genetic Algorithm (GA) 2008, 1994, 2009 are best known and used in text classification system development. They differ according to the accuracy of the results and the request responding speed.

It is impossible to give a recipe which algorithm is better in general. Research argues that depending on the document collection, language, etc. the results of categorization vary even if the same algorithms were used 1999.

In 2008 four methods of spam filtering are evaluated: Nave Bayesian (NB), neural network (NN), support vector machine (SVM) and relevance vector machine (RVM). According to the authors results SVM and RVM behaves in the same manner: results are less relevant, but they are fast.

2006 argues about new modified KNN algorithm for Hierarchical document classification that leads to 13,86% improvement of the original KNN results.

In works 2010, 2007 for Chinese and Arabic language text classification, the Naive Bayesian classifier effectiveness was higher than KNN.

Along with SVM the NB was better performing compare to distance based classifiers in 1999.

The SVM and KNN combination was successful for Multi-class Text Classification 2008, some good results were provided in 2010 by joint scheme GA and KNN.

Compared with other languages, research performed on Georgian Language based text classification issues is in the range from none to very little. According to the results of leading scientists in IR, statistical methods even without any morphological analysis of the language 2001, 2011 are quite promising, but relying on the fact of Georgian language specific characteristics, we decided to use heuristic analysis method for concept-pattern formation.

Generally speaking the aim of text classification is to refer a document to an appropriate pre-defined class of notions (concepts). The problem might be considered as a particular problem in pattern recognition and appropriate methods can be used for solution. We are considering Georgian language based texts retrieval problem and use the method based on imitation of human way concept formation process - the heuristic analysis method presented in 1974 year by Georgian Academician V. Chavchanidze 1974.

One can judge the complexity of Georgian verbal system according to the term screeve(the set of six verb forms inflected for person and number) used by linguists more than separate terms ”tense”, ”aspect”, ”mood”, etc. 2009. Verbs are mainly divided into four types: transitive verbs, intransitive verbs, verbs with no transitive counterparts and indirect verbs. Each type uses different strategies to build the verb complex. There are also many irregular verbs in Georgian, requiring different formations. Because of its extremely complex structure we decided to avoid dealing with verbs and perform stemming only for nouns. The form-processing of noun in Georgian is easier, it has only one root, but declension is applied. The noun declension depends on the ending of the root: If it ends with a vowel, the declension can be either truncating (roots ending with -e or -a) or non-truncating (roots ending with -o or -u). In the truncating declensions, the last vowel of the word stem is lost in the genitive and the instrumental cases. The Georgian nominal has a series of morpheme slots that must be filled in a specific order: noun root + plural suffix + case suffix (+ postposition). The task of document classification is based on the text indexation performed using stemming - commonly used in Information Retrieval but also beneficial in machine learning based models applications to improve the system quality. Keeping in mind the structure or language and using database of 30000 nouns, the new stemming algorithm was developed and used in text-processing initial module of our presented system.

The paper is organized as follows: Section 2 introduces the main modules of the classification system, while section 3 is dedicated to the System testing results and the conclusion is given in section 4. Finally paper is supported by the bibliography of appropriate references.

2 The system performance algorithm

The presented Classification System has four main modules: 1. Text initial processing module; 2. Knowledgebase module; 3. Concept Class establishing module; 4. Knowledgebase future updating module;(Fig. 1).

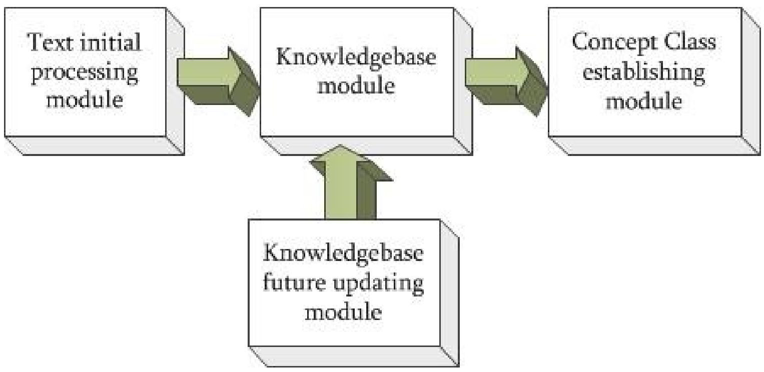

Fig. 1: The Classification System

Text initial processing module is designed for the initial classification processing of the document. The result of this process is the construction of term-vector defining and appropriate of this text. The length of term-vector primary depends on text size: for text up to 600 words it consists of 10 components (terms), from 600 to 3000 words 20 components and from 3000 to 10 000 words we have 30 terms. We have to mention that the amount of terms was gained from fulfilled several practical experiments. For the text defining term-vector construction we used the TF-IDF weights.

2.1 The text initial processing

Let us describe the text initial processing module performance, which in fact represents the document normalization process: the sequence of the following operations:

1. Creating the list of so called stop words (for Georgian language these are conjunctions, interjections, pronouns, etc.) and their replacement with wildcard characters;

2. The suffix stripping or stemming operation 1980: the standard approach of removing longest one from all available suffixes is not working for Georgian Language 2009. In order to calculate the frequency of words in text correctly (not to count different forms of the same word as different words), we have to process the words: leave only the constant part (root) of the word while word-building. Root definition issue for Georgian language primary associates with its complex grammar, remarkably different from European languages, with its split ergativity and a poly-personal verb agreement system 2009. As the Georgian verbal system is extremely complex, our target group represents nouns and adjectives. Sure there are other parts of speech in Georgian, but they are quite little and their influence on text classification might be neglected.

The problem was to develop algorithm defining whether two arbitrary elements represent the same lexical unit or not. It is clear that the set of all forms of a noun is transit and creates the equivalent classes, where each represents the set of all forms of one lemma. We wanted to put an identifier in accordance with each equivalency class and thus used the words grammar root for it.

Again come from the complexity of the language, we used the database of words represented in a nominative case to form the root - as taking off words case suffix is not working for example ded-is (mothers)- we will get ded as rood which is not correct- the root is deda, but for megobar- is (friends) root is megobar and now it is correct, the issue is getting even more complex when words are compressed or truncate.

For the beginning all possible forms of each word is generated in the database using all forms of declension (suffixes for each noun case) and one root will be assigned to each word. If the form refers to more than one root we choose one that contains less morpheme (case suffix and particle).

Note: The generalization (word form generating process) should be performed after each database update. Con of it is that database is becoming bigger, but we are using indexation for database search optimization.

3. Text representation in form of term/weight couple. From the received sets (stage 2), the TF weights 2012 are gained for every different term. TF weight defining rule is also based on text size: if the amount of words is less than 600, then each t_i term weight is calculated with formula:

$$tf(t_i, d) = \frac{n_i}{\sum_l n_l}, \tag{1}$$

where n_i is the number of t_i term in d document, and - $\sum_l n_l$ is the number of all words in this document. For middle sized texts (600-3000 words) we have:

$$tf(t_i, d) = 1 + \log f_{t_i,d} \tag{2}$$

where $f_{t_i,d}$ represents the raw frequency and for large ones (3000-10000 words) respectively:

$$tf(t_i, d) = \frac{1}{2} + \frac{\frac{1}{2} \times f_{t_i,d}}{\max\{f_{t_i,d} : t_i \in d\}}, \tag{3}$$

Above mentioned parsing leads us to respectively first 10, 20 or 30 high-weight word selection further used in text-defining term-vector construction (Fig. 2)

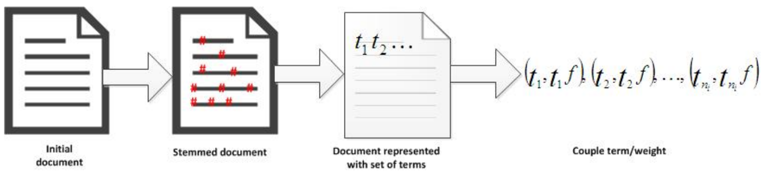

Fig. 2: Text initial processing module scheme

2.2 The Knowledgebase module

The Knowledgebase module represents the key part of the system. The proposed systems logical performance is mainly based on this modules accuracy and recall. The knowledge base represents the union of concrete class defining concept-patterns 2014a. Each class is described in Disjunctive Normal Form, where terms are in role of implicants. Let us describe the technics of knowledgebase construction.

Let us have $D = \{d_1, d_2, ..., d_R\}$ set of R documents (collection), distributed in $C = \{c_1, c_2, ..., c_K\}$ classes (K is the number of classes). Each $c_i, i = 1, ..., K$, class corresponds the documents subset D_i from D, $D_i = \bigcup_{j=1,...,J} d_{i_j}$, J is the number of c_i classs corresponding documents.

The method of heuristic analysis might be used to develop the concept-pattern describing appropriate c_i class from each D_i 2014b. Using above mentioned method let us construct term-vector for each d_{i_j} document :

$$\begin{array}{r} d_{i_1} \rightarrow t_{i_1}^1, t_{i_1}^2, ..., t_{i_1}^M \\ d_{i_2} \rightarrow t_{i_2}^1, t_{i_2}^2, ..., t_{i_2}^M \\ ... \\ d_{i_j} \rightarrow t_{i_j}^1, t_{i_j}^2, ..., t_{i_j}^M \end{array} \tag{4}$$

d_{i_j} is the corresponding document (text) of c_i class;

Note: As we are using different size patterns for different size (volume) texts, the size of the vector will vary.

$t_{i_j}^m, (m = 1, ..., M)$ is respectively the first M high-weight words in d_{i_j} document. $M = \{10, 20, 30\}$ according to the document text size. Each different $t_{i_j}^m$ term, appropriate to c_i class from the document collection denote as $w_{i_j}^n, n = 1, ..., N$. Then in frames of heuristic analysis, the d_{i_j} "term-vector" corresponding to c_i class will have the following view:

$$d_{i_j} \rightarrow \breve{w}_{i_j}^1 \& \breve{w}_{i_j}^2 \& ... \& \breve{w}_{i_j}^N \tag{5}$$

Where N is the number of different high-weight terms in whole document collection;

$$\breve{w}_{i_j}^n = \begin{cases} w_{i_j}^n & \text{if this word is contained in appropriate text high-weight word collection} \\ \overline{w_{i_j}^n} & \text{if not} \end{cases} \tag{6}$$

Let us represent unity of term-vectors in Disjunctive Normal Form:

$$c_i = \bigvee_{d_{i_j}} \breve{w}_{i_j}^1 \& \breve{w}_{i_j}^2 \& ... \& \breve{w}_{i_j}^N \tag{7}$$

Finally at the knowledgebase we will have 1 level k(amount) patterns. The set of patterns are the union of implicants and are representing the main elements of the knowledgebase.

define P^i patterns of appropriate c_i classes in form of the following implicants:

$$P^i = I_1^i \vee I_2^i \vee ... \vee I_{j_i}^i \tag{8}$$

Where j_i is the number of implicants derived from minimization of (7), apropriate to c_i class.

In case classification process requires more focus on details like subareas, subthemes and etc., same procedure will be fulfilled for each c_i class. Let us have $D_i = \{d_{i_1}, d_{i_2}, ..., d_{i_N}\}$ set of documents assigned to c_i class (i_N is the number of documents assigned to this class).

c_i is split to c_{i_j}, $i_j = i_1, ..., i_k$ subclasses, thus D_i documents are in appropriate (D_{i_1},...,D_{i_k}) subsets. On the basis of each D_{i_j} the second level pattern will be defined and it will be included in Knowledgebase formation. If additional preciseness is required, the above described procedure will be repeated for each c_{i_j} subclass.

Thus, processor gives the following information: the list of all classes in the knowledgebase along with similarity measures for appropriate documents to classify. In case of insufficient results, document will be sent for updating module for future processing.

The term vector derived from documents initial processing is compared with every $I_1^i, I_2^i, ..., I_{j_i}^i$ rule of P^i pattern. The highest similarity coefficient is calculated and if it is less then 0.7 , other patterns for different size texts are processing as well. If the procedure gives no results (similarity is less then 0.7 still) system will refer this document to the highest coefficient appropriate class and such documents along with their evaluation data will be sent to the knowledge base future updating module. As for the knowledgebase future updating module- it represents the set of ill-classified documents used for knowledge base periodic update.

Let us define the classification process based on concept-patterns for the following example:

Suppose we have $k = 3$ different classes c_1, c_2, c_3.

Five different $d_{i_j}, j = 1, ..., 5$ documents (text size: less than 600 words) are appropriate to each c_i class. For each text 10 high-weight different $t_{i_j}^m, m = 1, ..., 10$. terms(words) were selected. Thus for each c_i class we will have five term-sets and 15 term-sets for the whole document collection: for c_1 class well have:

$$
\begin{array}{r}
d_{1_1} \rightarrow t_{1_1}^1, t_{1_1}^2, ..., t_{1_1}^{10} \\
d_{1_2} \rightarrow t_{1_2}^1, t_{1_2}^2, ..., t_{1_2}^{10} \\
... \\
d_{1_5} \rightarrow t_{1_5}^1, t_{1_5}^2, ..., t_{1_5}^{10}
\end{array}
\tag{9}
$$

for c_2 class well have:

$$
\begin{array}{r}
d_{2_1} \rightarrow t_{2_1}^1, t_{2_1}^2, ..., t_{2_1}^{10} \\
d_{2_2} \rightarrow t_{2_2}^1, t_{2_2}^2, ..., t_{2_2}^{10} \\
... \\
d_{2_5} \rightarrow t_{2_5}^1, t_{2_5}^2, ..., t_{2_5}^{10}
\end{array}
\tag{10}
$$

for c_3 class well have:

$$
\begin{array}{r}
d_{3_1} \rightarrow t_{3_1}^1, t_{3_1}^2, ..., t_{3_1}^{10} \\
d_{3_2} \rightarrow t_{3_2}^1, t_{3_2}^2, ..., t_{3_2}^{10} \\
... \\
d_{3_5} \rightarrow t_{3_5}^1, t_{3_5}^2, ..., t_{3_5}^{10}
\end{array}
\tag{11}
$$

From this term collection let us define each different $t_{i_j}^m$ term by $w_{i_j}^n, n = 1, ..., N$ where N is the number of different terms in the whole term collection ($10 \leq N \leq 150$). Suppose $N = 20$, thus well receive the following Disjunctive Normal Form for each class:

$$
\begin{array}{l}
c_1 \rightarrow \breve{w}_{1_1}^1 \& \breve{w}_{1_1}^2 \& ... \& \breve{w}_{1_1}^{20} \vee \breve{w}_{1_2}^1 \& \breve{w}_{1_2}^2 \& ... \& \breve{w}_{1_2}^{20} \vee ... \vee \breve{w}_{1_5}^1 \& \breve{w}_{1_5}^2 \& ... \& \breve{w}_{1_5}^{20} \\
c_2 \rightarrow \breve{w}_{2_1}^1 \& \breve{w}_{2_1}^2 \& ... \& \breve{w}_{2_1}^{20} \vee \breve{w}_{2_2}^1 \& \breve{w}_{2_2}^2 \& ... \& \breve{w}_{2_2}^{20} \vee ... \vee \breve{w}_{2_5}^1 \& \breve{w}_{2_5}^2 \& ... \& \breve{w}_{2_5}^{20} \\
c_3 \rightarrow \breve{w}_{3_1}^1 \& \breve{w}_{3_1}^2 \& ... \& \breve{w}_{3_1}^{20} \vee \breve{w}_{3_2}^1 \& \breve{w}_{3_2}^2 \& ... \& \breve{w}_{3_2}^{20} \vee ... \vee \breve{w}_{3_5}^1 \& \breve{w}_{3_5}^2 \& ... \& \breve{w}_{3_5}^{20}
\end{array}
\tag{12}
$$

The minimization of this disjunctions will lead us to c_i class appropriate pattern ”binary” (is word/is not word) form. (Fig. 3)

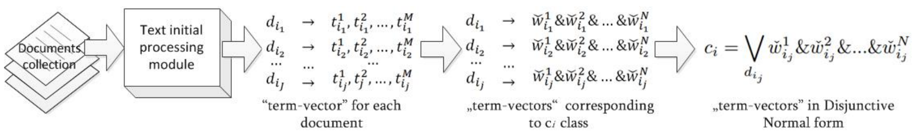

Fig. 3: Concept pattern formation process

Thus finally our patterns will have the following view:

$$\begin{array}{l} P^1 = I_1^1 \vee I_2^1 \vee ... \vee I_{j_1}^1 \\ P^2 = I_1^2 \vee I_2^2 \vee ... \vee I_{j_2}^2 \\ P^3 = I_1^3 \vee I_2^3 \vee ... \vee I_{j_3}^3 \end{array} \quad (13)$$

where $I_{j_1}, I_{j_2}, I_{j_3}$ have value $1, ..., 5$.

Note: The formation of Knowledge base is a different procedure and is not contained in functional environment of the system.

3 System testing and evaluation

At the beginning of testing process six main classes (Politics, Law Science, Sociology, Computer Science, Sports, Economics) had been selected in order to fill out all functional modules. 200 documents for each different volume (up to 600 words, from 600 to 3000 words and from 3000 to 10000 words) had been selected for each class, thus totally the Document-Collection consisted of 1200 documents placed in the database. 50 documents per class were selected in a random manner, its processing led to construction of three level classification patterns of various classes. The number of such patterns were 15 and they formed the knowledgebase of the system. The documents used for knowledge base formation were withdrawn from the Document-Collection and rest used in the process of testing, thus the number of documents for retrieval process were 900. Following characteristics were defined for evaluation: Recall, Precision, Accuracy, F measure, ERR 1992, 1999, 2011. The results of testing is presented in the tables below: (Table. 1, Table. 2)

Categories	Not Retrieved	Retrieved	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
Politics	720	180	129	51	21	699
Law	757	143	114	29	36	721
Sociology	785	115	98	17	52	733
Comp.Science	715	185	132	53	18	697
Sports	765	135	113	22	37	728
Economics	685	215	137	78	13	672

Table 1: The Retrieval Results

Using the above mentioned well known evaluation tools we have the following picture(Fig. 4, Fig. 5):

Categories	Recall	Precision	F-Measure	Accuracy	ERR
Politics	0,86	0,717	0,782	0,92	0,08
Law	0,76	0,797	0,778	0,928	0,072
Sociology	0,653	0,852	0,74	0,923	0,077
Comp.Science	0,88	0,714	0,788	0,921	0,079
Sports	0,753	0,827	0,793	0,934	0,066
Economics	0,913	0,637	0,751	0,899	0,101

Table 2: The Evaluation of Results

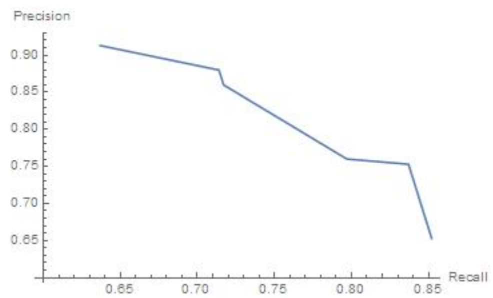

Fig. 4: The Curve: Precision-Recall

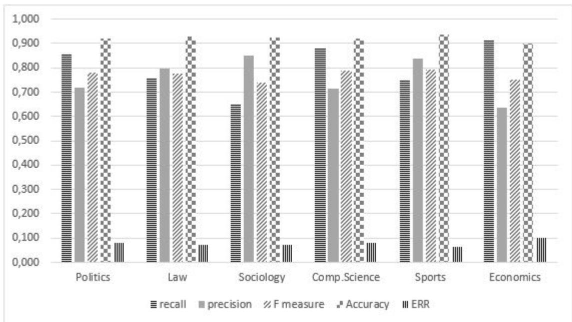

Fig. 5: The Result Summary

4 Conclusion

In the proposed article the heuristic analysis method application for Georgian language based texts classification system development is described. It was the first try ever made for Georgian language based text categorization using concept patterns. The gained results confirmed the effectiveness of the method, but in order to give it further conclusions and comparative analysis, authors plan to test the system for larger collection of texts and modify it using classical methods of IR.

References

Aizava Akiko (2003) *Information Processing and Management* Vol 39, 45–65.

Aronson, H (2009). *Reference grammar of Georgian* , SEELRC, Duke University

Bang, S. L,Yang, J. D., Yang, H. J.(2006). *Hierarchical document categorization with k-NN and concept-based thesauri.* Information Processing and Management . pp. 397406,

Cavnar, W.B, Trenkle, J.M. (1994). *N-Gram-Based Text Categorization* Collected papers. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval

Chavchanidze, V.V. (1974). *Towards The General Theory Of Conceptual Systems: (A New Point of View)* . Kybernetes. Vol. 3. Iss: 1, 17–25

Chinchor, N.(1992). *MUC-4 Evaluation Metrics* . Proceedings of the Fourth Message Understanding Conference, 22–29.

Duwairi, R. (2007). *Arabic text categorization* The International Arab Journal of Information Technology Vol4. No 2

Harris A. C. (2009). *Georgian Syntax: A Study in Relational Grammar* , Cambridge University Press, Apr 30, 2009 - Language Arts and Disciplines - 352 pages.

Khachidze M, Tsintsadze M, Archuadze M, Besiashvili G (2014). *Concept pattern formation in semantic search problems* . GESJ: Georgian Electronic Scientific Journals, Computer Sciences and Telecommunications. 13–20.

Khachidze M, Tsintsadze M, Archuadze M, Besiashvili G (2014). *Complex System State Generalized Presentation Based on Concepts* . Proceedings of Application of Information and Communication Technologies (AICT), IEEE 8th International Conference. 1–4

Liu, M, Yang, J. (2012). *An improvement of TFIDF weigthing in text categorization* .IPCSIT Vol 47, 44–48.

Lu, S.H. , Chiang, D.A. , Keh, H.C., Huang, H.H (2010). *Chinese text classification by the Nave Bayes Classifier and the associative classifier with multiple confidence threshold values* Knowledge-Based Systems,Volume 23, Issue 6, Pages 598604

Manning, C.D, Raghavan, P, Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press

Porter, M.F. (1980). *An algorithm for suffix stripping* , Program, Vol 14, no. 3, 130–137

Powers David M W (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation* . Journal of Machine Learning Technologies 2 (1), 37–63.

Sawaf, H., Zaplo, J., Ney, H. (2001). *Statistical classification methods for Arabic news*, Article, Natural Language Processing in ACL2001, Toulouse, France

Sebastiani F. (1999). *Evaluation of text classifiers* . Machine Learning in Automated Text Categorization, 1– 46.

Shumeyko, A, Sotnik, S. L. (2009). *Using Genetic Algorithms for Texts Classification Problems* Collected papers. Seria Informatic. Vol. VII fasc. 1

Torunolu, D., akrman, E., Ganiz, M.C, Akyoku, S and Zahid Grbz, M. (2011). *Analysis of Preprocessing Methods on Classification of Turkish Texts* Collected papers. Innovations in Intelligent Systems and Applications (INISTA). 112–117

Wei Zhao (2010). *A New Feature Selection Algorithm in Text Categorization* . International Symposium on Computer, Communication, Control and Automation.

Yang, Y., Liu X. (1999). *An re-examination of text categorization* . Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 42–49

Yiming Yang (1999). *An Evaluation of Statistical Approaches to Text Categorization* . Journal of Information Retrieval, Vol. 1 Issue 1-2, 69–90

Yu, B., Xu, Z., (2008). *A comparative study for content-based dynamic spam classification using four machine learning algorithms* Knowledge-Based Systems.Volume 21, Issue 4, 355–362

Yuan, P., Chen, Y., Jin, H(2008). *Combining SVM and k-NN for Multi-class Text Classification* . IEEE International Workshop on Semantic Computing and Systems.WSCS'08 (2008), 133–140

Authors' information

M. Khachidze was awarded the candidate of technical sciences degree at Technical University of Georgia in 1998 (Authomatic Systems of Control). She is a full professor of Iv. Javakhishvili Tbilisi State University Georgia, faculty of exact and natural sciences, computer sciences chair. Her research interests include Databases, IR, Quantum Computation, Problems of Medical Informatics etc. She was a head of project on Medical Image processing Theoretical Bases and Technological Aspects INTAS Ref. No : 04-77-7067 INTAS in2005-2007 and coordinator of project - Creation of Modeling Software of Molecular Systems Materials of Molecular Nanotechnology and Spinelectronic GNSF(Georgian National Scientific Foundation) - in 2010-2013.

M. Tsintsadze was awarded the candidate of Phys.Math Sciences at Iv. Javakhishvili Tbilisi State University Georgia in 2006 (Math.Cybernetics). At present she is an associate professor at Iv. Javakhishvili Tbilisi State University Georgia, faculty of exact and natural sciences, computer sciences chair. Her main research interests include the fuzzy logic, Information Retrieval, web programming and databases etc. She was awarded INTAS grant for YS in 2004, President Grant in 2008, and Fulbright grant in 2011.

M. Archuadze is a research associate of the Technical Informatics Chair of Iv. Javakhishvili Tbilisi State University. The main research interest is connected with the problem of quantum calculations, data retrieval issues, databases etc. Nowadays she holds a position of professor's assistant at Iv. Javakhishvili Tbilisi State University, Georgia. In 2010-2013 was a participant of the project - Creation of Modeling Software of Molecular Systems Materials of Molecular Nanotechnology and Spinelectronic GNSF(Georgian National Scientific Foundation).

G. Besiashvili was awarded the candidate of technical sciences degree at Technical University of Georgia in 2003. His research interests are connected with information theory and coding, cryptology, artificial intelligence, genetic algorythms etc. He is an assistant professor at Iv. Javakhishvili Tbilisi State University Georgia. In 2010-2013 he was a participant of the project - Creation of Modeling Software of Molecular Systems Materials of Molecular Nanotechnology and Spinelectronic GNSF(Georgian National Scientific Foundation).

Received September 24, 2015 , revised December 22, 2015, accepted December 22, 2015