Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends in Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

LINGUISTICS RESEARCH AND ANALYSIS OF THE BULGARIAN FOLKLORE. EXPERIMENTAL IMPLEMENTATION OF LINGUISTIC COMPONENTS IN BULGARIAN FOLKLORE DIGITAL LIBRARY

Konstantin Rangochev, Maxim Goynov,
Desislava Paneva-Marinova, Detelin Luchev

Abstract: *The observation of the lexical structure of the Bulgarian folklore is very important task for different science domains such as folkloristic, ethnology, linguistics, computational linguistics, Bulgarian language history, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what is the lexical structure of Bulgarian folklore works. First attempt for computational lexical analysis of the Bulgarian folklore and its constituents is made during the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" 1. During the project the Bulgarian folklore digital library (BFDL) is designed and developed. In its structure it is implemented linguistic components, whose aim is the realization of different types of analysis of folk objects from a text media type. Thus, we lay the foundation of the linguistic analysis services in digital libraries aiding the research of kinds, number and frequency of the lexical units that constitute various folk objects. This paper presents basic types of dictionaries needed to carry out such linguistic analysis. It describes the BDFL Linguistics Search in sets of folklore objects of text media type and a linguistic component for frequency analysis of the folklore vocabulary. Finally, a project for implementation of a dictionary - concordances of songs, prose, interviews, etc. is outlined.*

Keywords: *multimedia digital libraries, systems issues, user issues, online information services*

ACM Classification Keywords: *H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues.*

Introduction

The main component of the linguistic research of the Bulgarian folklore is the analysis of its lexical structure: How many and what token it contains? Is there and what is the domination or the lack of some groups of tokens, *etc*. Until today, such a linguistic analysis hasn't been made; it is unclear what is the lexical structure of Bulgarian folklore works. With a few exception (for Bulgarian heroic epoc [Rangochev, 1994] and for 'Veda Slovena", http://www.bultreebank.org/veda/index.html) lexical analysis for the Bulgarian folklore and its constituents is missing, the regional characteristics of the folklore lexical structure is unknown. Unfortunately, in 2010 the Bulgarian linguistics, folklore, ethnology, *etc*. cannot answer the question what are the lexical components of Bulgarian folklore (number, frequency, word forms, *etc*.) and so far, this type of research is carried out systematically and with a purpose.

[1] The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" is a national research project of the Institute of Mathematics and Informatics, supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03/2006. Its main goal is to build a multimedia digital library with a set of various objects/collections (homogeneous and heterogeneous), selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science. This research aims to correspond to the European and world requirements for such activities, and to be consistent with the specifics of the presented artefacts.

In the project, named "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" (FolkKnow) [Paneva-Marnova et al., 2009] [Luchev et al., '08b] the attention was directed to these researches in order to enrich both the content and functionality of the developed multimedia digital library of Bulgarian folklore (also called Bulgarian Folklore Digital Library or BFDL, FDL) [Rangochev et al., '07a] [Rangochev et al., '08]. Thus we aim to expand the target group of potential users of the library, covering not only those who are interested in Bulgarian folk music, but also narrow specialists in different fields of humanities (folklore, ethnology, linguistics, text linguistics, structural linguistics, *etc* .). Digital library with similar services are presented at [Pavlov and Paneva, 2007] [Pavlova-Draganova et al., 2007a] [Pavlov et al., 2006].

The Bulgarian folklore digital library has a flexible structure that involves the addition of linguistic components, whose main task is the realization of different types of analysis of folk objects from a text media type. This article presents the basic types of dictionaries needed to carry out such linguistic analysis. It describes the BDFL Linguistics Search in sets of folklore objects of text media type and a linguistic component for frequency analysis of the folklore vocabulary. Finally, a project for implementation of a dictionary - concordances of songs, prose, interviews, *etc.* is outlined.

Frequency Dictionaries and Concordance Dictionaries

The frequency dictionary presents the frequency of the lexemes in a definite corpus of texts. It is considered that the facts in one frequency dictionary are reliable enough if there are minimum 20 000 lexical units in it. The frequency dictionaries gave versatile information: presence/ absence of definite lexemes or group of lexemes in comparison with a standard frequency dictionary of the Bulgarian speech [Radovanova, 1968]; frequency of verbs (the so called “verb temperature” [Gerganov et al., 1978] (for the Bulgarian speech at least 21 % verbs in the examined corpus of texts); investigating of the paradigmatic relations in the vocabulary of the text corpus (river- stream- brook- rill…). The domination of group lexemes and respectively small number or absence of other group reveals the constituent characteristics of the text type and its originators.

- A general frequency dictionary – it contains the all lexical units which are in the BFDL (songs, proverb and descriptions of the rites…);
- A regional frequency dictionary – it contains all the text units which come of a definite folklore region or of a concrete settlement (if there are enough texts). Practically, this is a dialect dictionary of the region/ settlement as far as the folklore regions coincides with the dialect areas.
- A functional frequency dictionary – it contains all the text units which have identical functions: descriptions of the rites, various types of songs, narratives *etc*. This kind of dictionary would describe some genre specifics of the different parts of the Bulgarian folklore;
- Another dictionary – by user’s wish.

The advantage of creating of frequency dictionaries is the possibility to make comparisons between the different types of texts and it can be also followed the tendencies in the dynamics of the lexis – presence/ absence of various group of lexemes, *etc.*

The following table illustrates the comparison of the Bulgarian folklore and spoken languages based on data available in frequency dictionaries.

Rank list	
Bulgarian spoken language[1]	Bulgarian heroic epoc[2]
1. съм – 4 041	1. съм – 1342
2. и – 3764	2. да – 1 247
3. да – 3 148	3. си – 548
4. аз – 2 433	4. Марко – 1 036
5. той – 2 288	5. се – 828
6. не – 1 956	6. на – 801
7. се – 1 928	7. и – 796
8. този – 1 701	8. па – 657
9. на – 1 669	9. у – 582
10. ти – 1 249	10. я – 553
11. ще – 1 183	11. та – 526
12. един – 1 131	12. не – 412
13. в – 1 099	13. юнак – 396
14. си – 1065	14. го – 338
15. казвам – 1 045	15. му – 320
16. тя – 1044	16. че – 318
17. викам – 1 031	17. а – 286
18. те – 1014	18. кон – 276
19. какъв – 938	19. от – 272
20. за – 913	20. ми – 233
21. че – 874	21. ти – 225
22. с – 809	22. що – 222
23. имам – 768	23. по – 218
24. така – 742	24. добър – 201
25. от – 731	25. три – 201

Table 1: Comparison of the Bulgarian folklore and spoken languages.

Concordance dictionaries are these which show the lexeme with/ in her context – it is present the previous one (or more than one) lexeme and the following lexeme according to the examined lexeme. Example: “Fifty heroes are drinking wine” – the underlined lexeme is the examined and the lexemes in italic are her context. Of course, about the songs this could be concordance dictionary of their verses, about the narrative texts (descriptions of the rituals, *etc.*) – sentences in which they are contained (from point to point…). The creating and using of

[1] The frequency dictionary is made of texts of the Bulgarian spoken language and the corpus contains 100000 lexemes [Nikolova, 1987].

[2] The frequency dictionary is made of 100 song from [Romanska, 1971] and the texts of the songs contains 7871 verses while there are in it 40042 lexemes.

concordance dictionaries of the texts from BFDL would give good possibilities for folklorists and ethnologists to solve a series of problematic areas as presence/ absence of formulas in the folklore songs and epics, the structure of the folklore text, *etc.*

Linguistic Search in Set of Folklore Objects of Text Media Type

This type of a search has for aim to supply the needs of linguists and explorers of the Bulgarian dialects for researching the language of the "folklore song", the "folklore prose", *etc*. The variants for searching of folklore objects of text media type are the following:

- Search of a word in the different types of dictionaries;
- Search of two or more words – searching of verbal formulas in the folklore lexis: "Drinking wine", "Marko seated".
- Search of a group of words – this has for aim to investigate the paradigmatic relations in the folklore lexis (river- stream- brook- rill…) – for example, the frequency of the lexemes, verses/ sentences in which they are, number, numbering in the song, *etc*. of the verses/ sentences.
- Search for a root of a word for studying the folklore word-formation: 'drink" (I am drinking, I have drunk, they have drunk…).

A Frequency Dictionary in BFDL. A Project for Concordance Dictionary about Songs, Prose, Interviews, etc. in BFDL

In the process of the primary testing of BFDL come into being the necessity of insurance of resources for linguistic analysis of the folklore knowledge. For this aim it was projected and worked out a frequency dictionary with the following functional specification:

- Linguistic analysis of the available multitude of folklore objects of text media type in BFDB;
- Determination of the frequency of meeting the lexemes in text folklore objects;
- Creating of lists of the lexemes,
    - in frequency order
    - in alphabetical order.
- Taking the number of the lexical units;
- Taking the number of the repeats of the lexical units.

Figure 1 depicts the sequence of actions that has to be executed in order to be generated a frequency dictionary. Standard step is the passing through BFDL search service and its sub-functions: 1) user searches by some criteria; 1.1) service performs search in metadata repository, 1.1.1) service gets media data for the found objects, 1.1.1.1) service returns all found media objects by the search criteria, and 1.1.1.1.1) result sent to user. When the result set is generated the user could choose to generate a functional dictionary (step 2). Dictionary generation is performed and the result is shown by frequency or alphabetically.

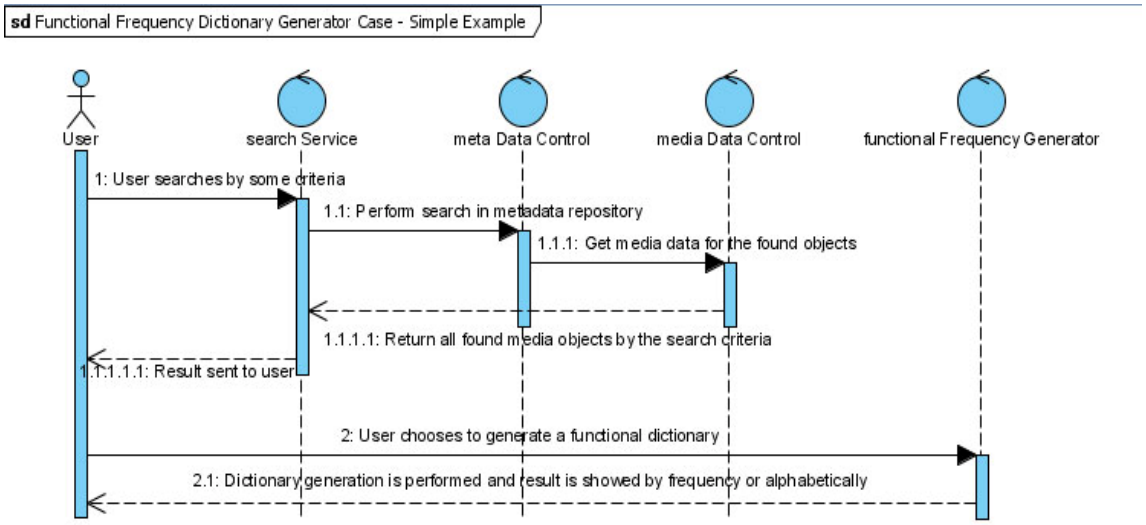

Figure 1: Sequence Diagram

Figure 2 depicts analysis class diagram for the BFDL linguistic component.

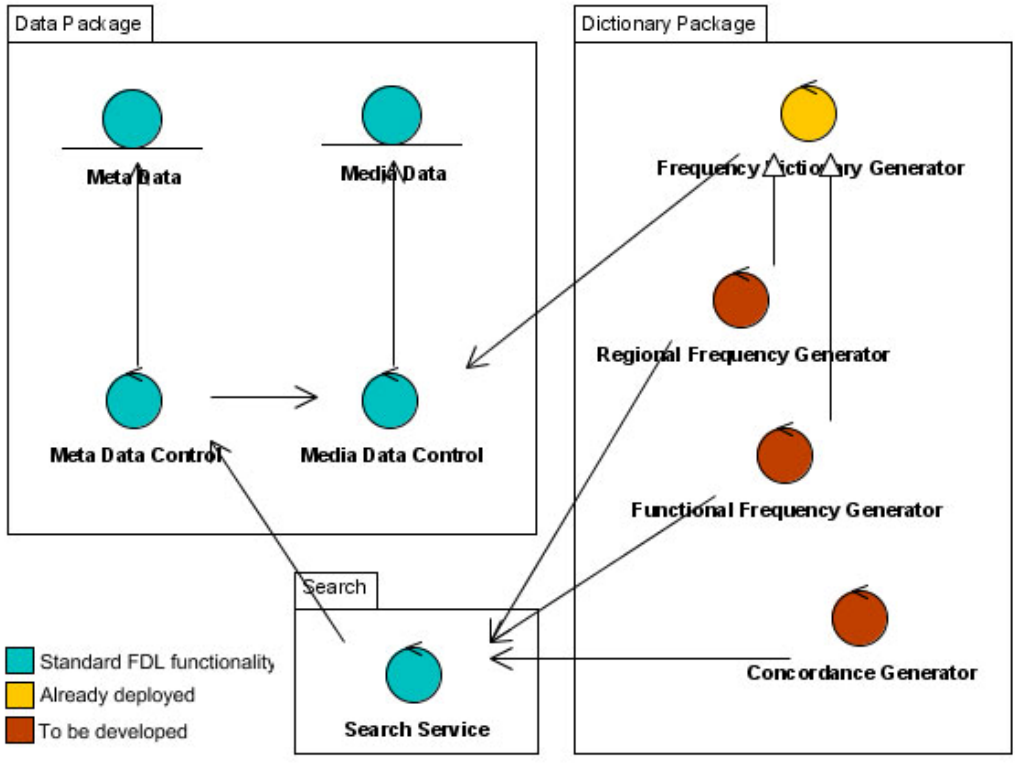

Figure 2: Analysis class diagram

The diagram shows the relations between the data package, the dictionary package and the search service. In the dictionary package there are clearly illustrated different types of generators for frequency dictionary, regional frequency dictionary, functional frequency dictionary and dictionary-concordance.

Acknowledgements

This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03-03/2006 “Development of Digital Libraries and Information Portal with Virtual Exposition “Bulgarian Folklore Heritage”” from the project “Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage”.

Bibliography

[Rangochev et al., ‘07a] Rangochev K., D. Paneva, D. Luchev. Bulgarian Folklore Digital Library, In the Proceedings of the International Conference on Mathematical and Computational Linguistics „30 years Department of Mathematical Linguistics”, 6 July, 2007, Sofia, Bulgaria, pp. 119-124.

[Rangochev et al., ‘08] Rangochev, K., D. Paneva, D. Luchev, Data and Functionality Management in a Folklore Digital Library, In the Proceedings of the International Conference - Slovo: Towards a Digital Library of South Slavic Manuscripts, 21-26 February, 2008, Sofia, “Boian Penev” Publishing Centre, pp. 246 – 250.

[Paneva-Marnova et al., 2009] Paneva-Marnova, D., R. Pavlov, K. Rangochev, D. Luchev, M. Goynov (2009), Toward an Innovative Presentation and Creative Usage of the Bulgarian Folklore Wealth, International Journal „Information Technologies & Knowledge”, vol. 3, 2009 (in print)

[Luchev et al., '08b] Luchev D., D. Paneva, K. Rangochev, Approaches for Utilization of the Semantic Web Technologies for Semantic Presentation of the Bulgarian Folklore Heritage. In the Proceedings of the national conference “Bulgarian Museums in the circumstances of the country membership in the European Union”. Sliven, 2008, pp.271-281.

[Gerganov et al., 1978] Gerganov, E., A. Mateeva, Experimental Research of the Frequency of the Bulgarian Language, In the proceedings “Contemporary problems of the native language education, Sofia, 1978

[Rangochev, 1994] Rangochev, K., “Structural particularities of the epic text (using material of the Bulgarian heroic epos)”, Канд. дис., СУ „Св. Кл. Охридски”, София.

[Radovanova, 1968] Radovanova, V., „Representative frequency dictionary of text with length 500 000 tokens”, Master thesis, University of Sofia ‘St. Kl. Ohridski”, Sofia, 1968.

[Nikolova, 1987] Nikolova, Cv., A frequency dictionary of the Bulgarian spoken language. Sofia.

[Romanska,1971] Romanska Cv. (Ed.), Сборник за народни умотворения, col.53, "Bulgarian heroic epos", Sofia, 1971

[Pavlov and Paneva, 2007] Pavlov R., D. Paneva. Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, International Journal “Cybernetics and Information Technologies”, vol. 6 (2007), № 3, pp. 51-62.

[Pavlova-Draganova et al., 2007a] Pavlova-Draganova L., V. Georgiev, L. Draganov. Virtual Encyclopaedia of Bulgarian Iconography, Information Technologies and Knowledge, vol.1 (2007), №3, pp. 267-271.

[Pavlov et al., 2006] Pavlov R., L. Pavlova-Draganova, L. Draganov, D. Paneva, e-Presentation of East-Christian Icon Art, In the Proceedings of the Open Workshop “Semantic Web and Knowledge Technologies Applications”, Varna, Bulgaria, 12 September, 2006, pp. 42-48.

Authors' Information

Konstantin Rangochev *– PhD in Philology, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str.., bl. 8, Sofia 1113, Bulgaria; e-mail: krangochev@yahoo.com*

Major Fields of Scientific Research: Ethnology, Folklore studies, Culture Anthropology, Linguistics, Computational Linguistics, Digital Libraries.

Maxim Goynov *– Programmer, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str.., bl. 8, Sofia 1113, Bulgaria; e-mail: maxfm@abv.bg*

Major Fields of Scientific Research: Multimedia Digital Libraries and Applications.

Desislava Paneva-Marinova *– PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str.., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg*

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.

Detelin Luchev *– PhD in Ethnology, MA in Informatics, MA in History; Research Fellow at Ethnographic Institute and Museum, BAS; 6A, Moskovska Str., Sofia 1000, Bulgaria; e-mail: luchev_detelin@abv.bg*

Major Fields of Scientific Research: Communities and Identities, Ethno-statistics, Museums and Archives, Digital Libraries.