Document Style Recognition Using Shallow Statistical Analysis

Pavel Braslavski

Institute of Engineering Science UB RAS
Komsomolskaya 34, 620219 Ekaterinburg, Russia
+ 7 (343) 375 3579
pb@imach.uran.ru

1. INTRODUCTION

Documents differ not only in topic but also in style. Style is a very broad and ambiguous term used in arts, fashion, literary criticism, and linguistics. In case of text documents we can accept an intuitive understanding that style is mainly related to the form (*how*) whereas topic – to the content (*what*) of a document. Although some topics determine strictly the style can be used, most topics allow their expression in various styles. Thus, style can be considered to be orthogonal to topic in a certain sense and represent therefore a useful parameter in many text processing and information retrieval tasks.

The main goal of our research is to develop automated procedures that enable text style recognition in favor of Web information retrieval. The research is related partly to the formal methods in authorship attribution, i.e. *individual style* recognition. There are also several studies aiming theoretical and educational goals that investigate quantitative variations of textual parameters within different text styles (e.g. different readability indices).

The paper by Jussi Karlgren and Douglas Cutting [3] gave the initial impulse to our research. The paper reports on stylistic experiments based on the Brown corpus of English text samples. Three-level genre hierarchy (from the 'imaginative/informative' dichotomy on the top down to 15 genres on the bottom) is used. A number of different features – surface cues along with e.g. part of speech (POS) and present participle counts – are used for classification. Discriminant function analysis is employed for data processing.

Several publications on the topic have appeared recently. Genre classification based on word statistics revealed from the interplay of subject-related and genre-related tagging of the training data is described in [5]. Incorporation of structural information of documents into a digital library navigation tool is introduced in [6]. The latter paper includes a detailed survey of different approaches in automatic genre and style analysis.

This paper describes two series of stylistic experiments conducted 1999-2002 within my work towards PhD and ongoing related research. In the next two sections, we introduce these experiments. Section 4 concludes the papers and outlines the future research suggested by the obtained results.

2. EXPERIMENT I

2.1. Experimental setting

In the first series of experiments we adopted the theory of functional styles, which is well-established and well-founded in Russian linguistics. The main idea of the functionalist approach is the distinction between the language (as a symbolic system) and the speech (as the very process of discourse generation). According to the theory, the style of a text is determined mainly by the communication context. Five functional styles are usually defined, such as *official* style, *academic* style,

journalistic style, *everyday communication* style, and *literary* style (although some scholars consider literary style, or fiction, to be a special case that is able to incorporate all other styles). More details on the theory of functional styles can be found in [4].

According to this approach, a training sample was composed carefully. The sample contained 305 documents in Russian (50 federal laws, 54 scientific papers in natural sciences, 61 online news articles, 79 short stories by modern Russian authors, and 61 fragments of chat listings).

We opted for linear *classification functions* for text categorization. This approach differs from the one described in [3], where *discriminant functions* are used. Each classification function is a linear combination of the classification variables (features) returning classification scores for each case for each group. The case is classified as belonging to the group for which it has the highest classification score. The assumptions of the technique are weaker than those of discriminant function analysis (see [11] for details).

Since most stylistic studies operate with qualitative descriptions, the selection of quantitative features becomes a challenging task. The initial feature set was composed based on both analysis of previous stylistic studies and examination of the training sample. We opted for easily computable features only; no format-specific parameters (e.g. HTML tags) were employed. The initial set consisted of approximately 30 features and was intentionally redundant. An overview of the initial features is shown in table 1.

The main unit of examination was a single word (i.e. no complete surface syntactic parsing was performed), although sub-word-level features (specific prefixes) and sentence-level features (e.g. sentence length, expressive punctuation marks, and genitive noun chains) were presented. Numerous morphological parameters (POS and distinctive grammatical forms) were determined using the dictionary-based stemmer LINGUIST by Agama Company. Lists of general academic terms and official document names were also composed.

Level	Parameter
Surface	equals sign per sentence rate smiles *:) ;-)* per sentence rate average word length (in characters) average sentence length (in words) expressive punctuation mark (*!*, *?*, ...) per sentence rate
Word formation	scientific prefix (*aqua-*, *aero-*, etc.) per word ratio
Morphology	POS rates (13 in total) neuter noun rate reflexive verb rate acronym rate first person pronoun rate second person pronoun rate particle *бы* rate (conjunctive mood cue) particle *ну, вот, ведь* rate (everyday communication style cue)
Lexis	word from *general academic term* list (37 words, statistically selected) rate word from *official document name* list (43 words, manually composed) rate
Syntax	genitive chain per sentence rate subordinating conjunction per sentence rate

Table 1: Features overview

For each text the first 1000 words (or the whole text if shorter) plus the words until the end of the sentence containing the 1000th word were processed. It is clear that some of the proposed features could not be computed absolutely accurately in full automatic mode. For instance, grammatical ambiguity was not resolved. Inaccuracy was caused also by end-of-sentence and end-of-word (due to invisible characters and hyphens) errors.

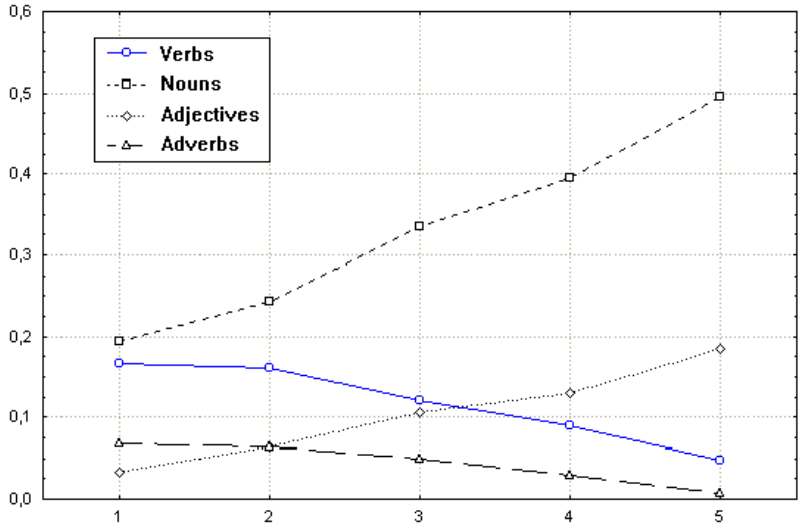

Figure 1: POS average rates across five styles
(1 – *everyday communication style*, 2 – *literary style*, 3 – *journalistic style*, 4 – *academic style*, 5 – *official style*)

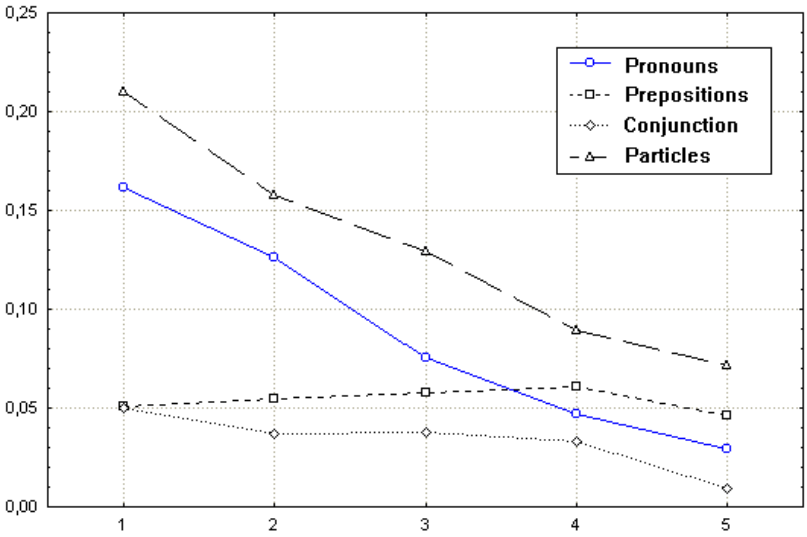

Figure 2: POS average rates across five styles
(1 – *everyday communication style*, 2 – *literary style*, 3 – *journalistic style*, 4 – *academic style*, 5 – *official style*)

2.2. Morphological characteristics

The morphological characteristics of the training sample appeared to be of interest to theoretical stylistics, since there are only few quantitative up-to-date comparative studies on of the functional styles. Moreover, our research, although not absolutely exact because of fully automatic processing, dealt with a reasonably large text collection. The amount of data processed was 239 696 words, the morphological characteristics were determined for 227 257 of them.

Figure 1 and figure 2 show monotonous decline of average verb, adverb, pronoun, and particle rates from the everyday communication style to the official style. Nouns and adjectives demonstrate an inverse behavior. The auxiliary POS rates are approximately the same across the styles.

Table 2 represents a portion of correlation matrix for POS rates across the whole training sample. The results show that knowing for example the noun count in a text we can predict the adverb count with a high degree of confidence.

Part of speech	1	2	3	4	5	6	7
1. Noun	1,00	0,85	-0,87	-0,85	-0,88	0,77	-0,86
2. Adjective	0,85	1,00	-0,81	-0,75	-0,85	0,67	-0,79
3. Pronoun	-0,87	-0,81	1,00	0,70	0,79	-0,78	0,77
4. Adverb	-0,8	-0,75	0,70	1,00	0,80	-0,69	0,76
5. Verb	-0,88	-0,85	0,79	0,80	1,00	-0,75	0,75
6. Participle	0,77	0,67	-0,78	-0,69	-0,75	1,00	-0,77
7. Particle	-0,86	-0,79	0,77	0,76	0,75	-0,77	1,00

Table 2: Correlation coefficient of the most correlated POS.

The results related to the morphological structure of the functional styles are not surprising but give a good overview and are based on experimental work. The morphological characteristics of the training sample are introduced in [1] in detail.

2.3. Results and Evaluation

Discriminant analysis (DA) module of the STATISTICA system [10] was employed for classification function generation.

After numerous optimization runs we obtained five linear classification functions based on only 7 features. The classification functions can be presented in the form $\bar{s} = A\bar{x} + \bar{b}$, where $A = \begin{pmatrix} 458{,}75 & 318{,}61 & 37{,}47 & 0{,}09 & 8{,}95 & -252{,}01 & -238{,}23 \\ 471{,}56 & 313{,}31 & 40{,}60 & 0{,}25 & 23{,}62 & -292{,}64 & -244{,}43 \\ 408{,}92 & 275{,}27 & 44{,}98 & 0{,}29 & 67{,}34 & -393{,}37 & -197{,}99 \\ 367{,}12 & 173{,}79 & 50{,}59 & 0{,}11 & 436{,}48 & 157{,}17 & -201{,}34 \\ 287{,}77 & 122{,}34 & 48{,}41 & 0{,}40 & 190{,}42 & -423{,}40 & 160{,}86 \end{pmatrix}$, $\bar{b} = \begin{pmatrix} -139{,}79 \\ -158{,}25 \\ -173{,}39 \\ -211{,}02 \\ -192{,}68 \end{pmatrix}$, $\bar{x}$ is a feature vector (x_1 – verb rate, x_2 – adverb rate, x_3 – average word length, x_4 – average sentence length, x_5 – scientific prefix rate, x_6 – general scientific term rate, and x_7 – official document name rate), and $\bar{s}$ contains resulting scores (s_1 – everyday communication style, s_2 – literary style, s_3 – journalistic style, s_4 – academic style, and s_5 – official style). As mentioned above, the case is classified as belonging to the group for which it has the highest classification score. The optimization does not decrease the computational cost significantly since the classification functions are linear and the most consumptive operation is the obtaining of the morphological features. The optimization serves the purpose of the style description clarity and ease.

There was no free and representative Russian corpus that could be used for the evaluation at the moment the experiment was conducted. Besides, we wanted to test the obtained procedure within the proposed Web information retrieval framework. For this purpose we selected a number of fairly generic queries from the Yandex search engine [13] log, which could be of interest for stylistic categorization of search results (i.e. we rejected queries containing specific technical abbreviations, exact institutions' names, etc.). Then, we downloaded the documents from the top of the result lists (up to 130 for each query). After that, we removed documents in Ukrainian, short documents (up to

500 words), and non-coherent text documents (rejection was made using verb rate). On the next stage we had to attribute the documents to the five functional styles. It was difficult in many cases, as there were numerous texts that fall poorly in the five-style system (advertising and religious texts, mixed-content and therefore mixed-style documents, translations of ancient authors, etc.), as well as intermediate style documents (popular science, scientific news, etc.). We marked the texts of the former type as 'undefined style'. Tables 3 and 4 show evaluation of the stylistic categorization of the documents returned in response to queries '*небесные тела*' and '*расход воды в нагревательных печах*', respectively. The columns reflect the automatic categorization, whereas the rows – the manual assessment.

	1	**2**	**3**	**4**	**5**	**Total**	**Recall**
1. Everyday Communication	-	-	-	-	-	-	-
2. Literary	-	2	1	-	-	3	0,67
3. Journalistic	-	2	44	8	-	54	0,81
4. Academic	-	-	10	17	-	27	0,63
5. Official	-	-	-	-	7	7	1,0
Undefined	1	1	7	3	-	12	-
Total	1	5	62	28	7	103	
Precision	0,0	0,2	0,71	0,61	1,0		0,74

Table 3: Search results categorization. Query: '*небесные тела*' ('*celestial bodies*')

	1	**2**	**3**	**Total**	**Recall**
1. Journalistic	9	1	1	11	0,82
2. Academic	3	7	2	12	0,58
3. Official	-	-	5	5	1,0
Undefined		1	-	1	-
Total	12	9	8	29	
Precision	0,75	0,78	0,63		0,72

Table 4: Search results categorization.
Query: '*расход воды в нагревательных печах*' ('*water consumption in heating furnaces*')

As the tables show, both the variety of styles presented in the search engine responses and the categorization quality are query-sensitive to a certain extent. However, the average quality of stylistic categorization for coherent Russian texts lay in the range 0,7-0,8.

3. EXPERIMENT II

3.1. Experimental setting

The goal of the second series of experiments was to develop an automated procedure for maintaining Yandex Web directory [12], which is built using faceted classification (FC). A FC contains a number of independent classifications that allow users to see the Web resources from the different points of view and keep the topical taxonomy reasonably simple at the same time. The FC schema used in Yandex directory includes a *genre* facet, among others [2]. In this case, *genre* is a property of the whole website; individual documents inherit the genre label from respective site roots.

There were 11 genres presented at the moment the research began. It is to be noticed that the genre set was pretty vague and ill-defined. So we limited the experimental set to four genres: *scientific papers*, *official documents*, *fiction*, and *guidance* (the last genre was still fairly imprecise). Since the selected genres were far from covering all the variety of genres, an additional rejection procedure had to be developed.

The training sample consisted of 285 documents in Russian (*fiction* – 77, *scientific papers* – 48, *official document*s – 94, *guidance* – 66). A set of initial features was constructed similarly to the previous experiment. Three parameters were based on the word lists statistically built upon official documents, scientific papers, and guidance subsets of the training sample. The morphological guesser *mystem* by Yandex was used for obtaining morphological features in this experiment. The advantage of *mystem* is that it works faster, and the morphological features of even unknown words can be resolved (see [9] for details).

3.2. Results and Evaluation

The same methods for building classification functions and similar optimization procedures were employed as in the previous experiment. Additionally, a procedure for type I and II error probability estimation was developed based on the Mahalanobis distance.

The resulting classification function employed again 7 features and had the following parameters:

$$A = \begin{pmatrix} 28{,}11 & -31{,}97 & 85{,}66 & -217{,}40 & 488{,}38 & -46{,}81 & 319{,}67 \\ 32{,}04 & -30{,}84 & 138{,}48 & 93{,}24 & 471{,}69 & -53{,}24 & 168{,}15 \\ 33{,}96 & 67{,}75 & 133{,}35 & -172{,}96 & 426{,}42 & -81{,}72 & 131{,}97 \\ 28{,}98 & -26{,}55 & 305{,}66 & -182{,}24 & 617{,}48 & 11{,}26 & 202{,}92 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} -105{,}17 \\ -121{,}30 \\ -131{,}84 \\ -107{,}96 \end{pmatrix}, \quad \text{and}$$

vector $\bar{x}$ elements are as follows: x_1 – average word length, x_2, x_3, x_4 – rates of the *OfficialDocument*, *Guidance*, and *Science* word lists items, respectively, x_5 – adverb rate, x_6 – template '*{можно| нужно} + Infinitive*'* occurrence rate, and x_7 – verb rate. Vector $\bar{s}$ elements are as follows: s_1 – fiction, s_2 – scientific papers, s_3 – official documents, and s_4 – guidance.

A test sample was gathered from the Yandex document repository. The sample consisted of 'long' (longer than 500 words) Russian documents belonging to the sites both with and without genre labels. The sample was randomly sifted and presented to 5 assessors for manual processing. A document was assigned to a genre if the opinions of three of five assessors agreed. The final test sample consisted of 291 documents, 135 of them were attributed to one of four mentioned genres. In the first test we used the smaller subset of the "known" genres. In the second test the whole sample was categorized. In the third test the rejection procedure was applied. The results can be seen in table 5.

	Test 1		**Test 2**		**Test 3**	
	P	**R**	**P**	**R**	**P**	**R**
Fiction	0.857	0.913	0.506	0.913	0.709	0.848
Science	0.912	0.912	0.553	0.724	0.682	0.517
OffDoc	0.950	0.864	0.452	0.864	0.842	0.727
Guidance	0.750	0.727	0.267	0.727	0.423	0.333
Other	-	-	-	-	0.633	0.705
Total	0.859		0.436		0.649	

Table 5: Test sample classification (P – precision, R – recall)

The majority of rejected documents could be attributed to news articles. This fact supported the decision to change the genre schema used in the directory. The renewed schema consists of 6 genres that tend to conform to the functional styles: *fiction*, *scientific and technical documents*, *official documents*, *guidance*, and *news articles*.

* English equivalents: '*{should|can} + Infinitive*' (guidance genre cue).

	P	R
Fiction	0.788	0.565
Science	0.447	0.500
OffDoc	0.783	0.818
Guidance	0.618	0.636

Table 6: Site genre vs. document genre (P – precision, R – recall)

Regardless of automated categorization results, the comparison of manual assessments and inheritance of genre labels from site roots to documents allows for making some interesting remarks. As table 6 shows, relatively high precision is observed in case of *fiction* and *official documents*, whereas only one genre – *official documents* – demonstrates high recall. According to those data we can assume that *official documents* are presented on the Russian Web as compact collections for the most part. *Fiction* is accessible both in the form of collections and isolated texts. Both *guidance* and *scientific papers* are pretty scattered across the Russian Web.

3. CONCLUSIONS AND ONGOING WORK

Our experiments have shown that easily computable text features are feasible for effective stylistic categorization. Some obtained results could also be of interest for theoretical stylistics. However, the second experiment series has proven a simple truth that the statistical methods cannot help without a solid conceptual basis.

Unfortunately we failed to implement the approach into a real-word application yet. The commercial Web search services believe that style, as an additional search feature, is unnecessary for the majority of users. We should probably turn to the more exacting realm of digital libraries as [6] suggests. At the moment we are going to make another attempt and research whether the style-related parameters could improve relevance ranking.

In the framework of the fist series of experiments we applied canonical discriminant analysis and the principal components method [11] to the experimental data. In case of correlated features the methods allow for a linear space transformation and subsequent shifting to a lower space dimension with minimal information loss (the fewer coordinates would explain the greater part of the overall variance).

The scatterplot of the training sample in the first and second principal directions can be seen in figure 3. It shows that the first component describes fairly well the variations of features across different styles. The lexical parameters contribute for distinction of academic and official styles (the second component).

This fact suggested the idea to reduce the description of styles to a single continuous parameter (a similar idea – understanding genres in terms of structural similarity rather than as a predefined set of classes – is expressed in [6]). Particularly, the linear combination of the initial features might serve for relevance ranking in information retrieval tasks. An additional advantage is that we do not need the time-consuming phase of training sample building in this kind of experiment.

Now we are planning to conduct an experiment on the ROMIP/RIRES data [7]. This test collection represents a 7+ Gb subset of the narod.ru domain including 600 000+ HTML pages in Russian from more than 20 000 websites. An important part of the collection is 54 evaluated queries for the classical ad-hoc task.

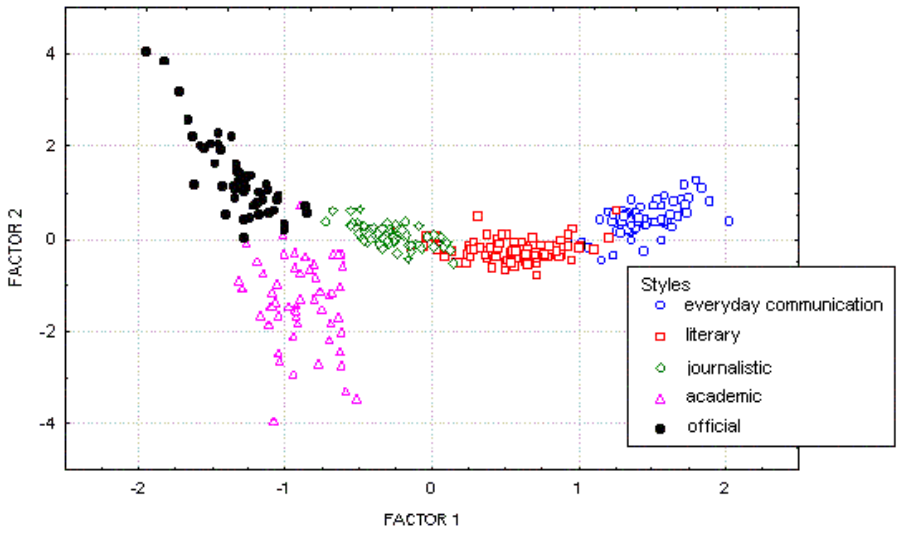

Figure 3: Two-dimensional scatter plot of the learning sample

We are going to conduct both global and local document analysis. In the former case we will obtain the new principal directions based on processing a reasonably large subset of the collection. Then 'stylistic scores' can be computed for every document in the same way. In the latter case every document subset returned in response to a query will be processed separately. In other words, the principal directions will be resolved for each evaluated query independently. We are going also to investigate the interplay between traditional relevance scores and stylistic features.

In addition, we examine the possibilities to make use of the Russian National Corpus [8] for the future stylistic experiments. The project was launched on April 27, 2004. The total amount of the collection is supposed to reach 100 million words by the end of 2005.

4. ACKNOWLEDGEMENTS

I would like to thank Mikhail Shchekotilov and Andrei Tselishchev for helping to develop the software, as well as Mikhail Maslov and Ilya Segalovich from the Yandex team for co-operation and support at the second stage of the experiments.

I am especially grateful to the ESSLLI 2004 Local Organization for a grant that made possible my participation in the summer school.

I would also thank the anonymous reviewers, whose comments helped me to improve the paper.

5. REFERENCES

[1] Braslavski, P. Morphological Structure of the Functional Styles: Case Study on Internet Documents. (in Russian) [Morfologičeskaya struktura funkcional'nyh stilei (na materiale dokumentov Internet)]. Proceedings of the Ural State University, Ekaterinburg, 2001, vol. 21, p. 8-17. Available at: http://proceedings.usu.ru/proceedings/N21_01/win/03.html

[2] Braslavski, P., Maslov, M., and Vovk, E. Facet-Based Internet Directory Design and Automated Genre Classification of Documents (in Russian). [Fasetnaya organizaciya internet-kataloga i avtomatičeskaya žanrovaya klassifikaciya documentov]. Proceedings of the International Workshop "Dialogue-2002. Computational Linguistics and Intelligent

Technologies", Moscow, 2002, vol. 2, p. 83-93. Available at: http://company.yandex.ru/articles/article8.html

[3] Karlgren, J. and Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, 1994, vol. 2, p. 1071-1075.

[4] Kožina M.N. Foundations of the Functional Stylistics (in Russian). [K osnovaniyam funkcional'noi stilistiki], Perm, 1968.

[5] Lee, Y.-B. and Myaeng, S. H. Text Genre Classification with Genre-Revealing and Subject-Revealing Features. Proceedings of the SIGIR'02, August 2002, Tampere, Finland, p. 145-150.

[6] Rauber, A. and Müller-Kögler, A. Integrating Automatic Genre Analysis into Digital Libraries. In Proceedings of the JCDL'01, June 2001, Roanoke, Virginia, USA, p. 1-10.

[7] Russian Information Retrieval Evaluation Seminar, http://romip.narod.ru

[8] Russian National Corpus, http://www.ruscorpora.ru

[9] Segalovich, I. A Fast Morphological Algorithm with Unknown Word Guessing Induced By a Dictionary for a Web Search Engine. In Proc. of the MLMTA-2003, Las Vegas, June 2003. Available at: http://company.yandex.ru/articles/iseg-las-vegas.html

[10] Statistica, http://www.statsoft.com

[11] StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft, http://www.statsoft.com/textbook/stathome.html

[12] Yandex Directory, http://yaca.yandex.ru

[13] Yandex Search Engine, http://www.yandex.ru