Experiment on Style-Dependent Document Ranking ♣

© Pavel Braslavski, Andrey Tselischev

Institute of Engineering Science, UD RAS, Ekaterinburg
pb@imach.uran.ru, a.tselischev@foratec-com.ru

Abstract

The paper reports on experiments aimed at incorporating style-dependent parameters into ranking schemata in information retrieval tasks. We use ROMIP Web collection and ROMIP-2003 ad-hoc track results in the analysis. Factor analysis techniques have been used to extract factors that would reflect stylistic properties of documents. Comparison of the obtained style-dependent parameters and their derived ranks is conducted. A simple schema for rank aggregation is proposed. Evaluation of the results shows only moderate improvement of relevance ranking.

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. The principal attention in the field of information retrieval (IR) has been traditionally paid to the topical characteristics of documents and collections (consider the tasks of relevance ranking, document clustering and categorization). Although some topics determine strictly what style can be used, most topics allow their expression in various styles. Thus, style can be considered orthogonal to topic in a certain sense and represent a useful parameter in many text processing and information retrieval tasks.

In the years 2000-2002 we conducted a series of experiments aimed at developing automated procedures that enable text style recognition [3]-[5].

In the first series of experiments we adopted the theory of functional styles (see [10] for details). 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). Five functional styles are usually defined in Russian linguistics: *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). According to this classification, we collected a training set consisting of 305 documents in Russian. The initial feature set consisted of approximately 30 easily computable cues from surface, word-formation, morphological, lexical, and syntactic levels. After numerous optimization runs we obtained five linear classification functions based on only seven features. The resulting functions delivered reasonable quality for coherent Russian texts (average values lay in the range 0,7-0,8). An in-depth description of the experiment can be found in [3], [4].

Within the research framework we also applied canonical discriminant analysis and the principal components method to the experimental data. In case of correlated features these methods allow us to transform a linear space and subsequently shift 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 set in the first and second principal directions can be seen in Figure 1. It shows that the first factor describes fairly well the variations of features across different styles. Preliminary experiments have shown that the parameter corresponds well with the intuitive perception of text 'seriousness' or 'strictness'.

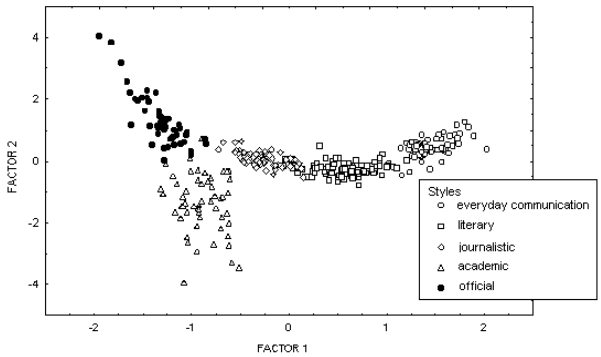

Figure 1: Two-dimensional scatterplot of the learning sample

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 [12]). Particularly, the linear combination of the initial features might serve for relevance ranking in information retrieval tasks. Our hypothesis is that this

Proceedings of the 7th Russian Conference on Digital Libraries RCDL'2005, Yaroslavl, Russia, 2005

parameter could be useful for search over heterogeneous collections. The research investigates into the possibilities of incorporating style-dependent ranking into ranking schemata used in the Web IR and digital library applications.

The paper is organized as follows: in the next section we present an overview of selected literature on automatic style recognition. Two subsequent sections describe data set used in the experiment and data processing, respectively. We conclude with results, evaluation, and discussion.

2 Related Work

There are many research areas related to the problem of automatic style recognition: e.g. authorship attribution (i.e. automatic recognition of *individual styles*) or quantitative methods in applied stylistics. A simple yet practical tool for writing skills teaching and preparing textbook material is diverse readability measures (see [7] for underlying theory, applications, formulae, etc.). Most readability measures estimate the difficulty of texts using simple parameters such as the average length of words and sentences (e.g. the well-known Flesch Reading Ease formula).

What follows is a short survey of selected papers related more closely to our study.

The paper by Jussi Karlgren and Douglas Cutting [8] 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) was used. A number of different features – surface cues along with e.g. part of speech (POS) and present participle counts – were used for classification. Discriminant function analysis was employed for data processing. The evaluation of the classification functions showed that the error rates grows rapidly with the number of classification categories.

In the research by Kessler et al. [9] the same corpus was used but the approach was quite different. As the authors stated, their goal has been "to prepare the ground for using genre in a wide variety of areas in natural language processing." The authors proposed a faceted genre classification including BROW, NARRATIVE, and GENRE facets. The first facet reflects the presumed readers' intellectual background with values *Popular*, *Middle*, *Upper-middle*, and *High*. The NARRATIVE facet is binary. GENRE has the values *Reportage*, *Editorial*, *SciTech*, *Legal*, *Nonfiction*, and *Fiction*. Logistic regression and neural networks were used for data classification. Only easily computable text features were employed in the experiments such as lexical cues, character-level cues, and their derivates. The experiments yielded reasonable results for NARRATIVE and GENRE facets and failed for BROW facet (though outperformed the baseline).

Text genre classification method based on word statistics revealed from the interplay of subject-related and genre-related tagging of the training data is described in paper by Lee and Myaeng [11]. Seven genre classes were used in the experiment: *reportage*, *editorial*, *technical paper*, *critical review*, *personal homepage*, *Q&A*, and *product specification*. The goodness of a term for classification depended on how evenly the term is distributed within a genre class (taking into account subject differences) and how well the term discriminated different genres. Text collection used in the experiment was gathered from the Web and consisted of both Korean and English documents (7000+ documents each). The documents were manually assigned to genre and subject classes. The collection was divided in equal subparts for training and testing. Naïve Bayesian classifier and vector similarity approach were used for classification. The method showed good results on the restricted set of genres in overall.

Incorporation of structural information of documents into a digital library navigation tool is introduced in [12]. In contrast to the studies mentioned previously, Rauber and Müller-Kögler adapted unsupervised technique for revealing genre-dependent similarities. The described analysis was based on a combination of various surface level features of texts, divided into four groups: (1) text complexity information and text statistics, (2) special character and punctuation counts, (3) characteristic words, and (4) format-specific mark-ups. The self-organizing map (SOM) was used to cluster documents according to their structural similarities. A news collection of 1000 documents was used for evaluation. The results of structural analysis were incorporated into a content-based representation through coloring individual documents according to their location on the resulting SOM. Though no extensive user study was performed, conducted tests produced encouraging results.

The presented survey of several works gives some idea of different approaches and applications of automatic genre detection from the IR perspective. An exhaustive up-to-date overview of the contemporary research in the field of automatic genre/style recognition can be found in [14]. To the best of our knowledge there are no publications in the IR realm on incorporating stylistic features of the documents into ranking schemata.

3 Experimental Data

For our experiment we use two fractions of ROMIP Web collection (see [13] for details). 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.

First, we use a randomly generated set of 500 Web pages longer than 50 sentences for 'global analysis' (see below). Second, we use Kodeks' information retrieval system (www.kodeks.ru) results obtained at ROMIP-2003 ad-hoc track, where Kodeks information retrieval system showed average results among other participating systems [6]. We adopted 51 of 54 evaluated queries of the ROMIP-2003 ad-hoc track. We excluded three result lists for different reasons (one contained no relevant results, another included only three items, and the other contained only few evaluated results). Each result

list is ranked according to relevance and contains from 6 to 60 pages. The subset contains ca. 2 700 Web pages in total. For the majority of this bulk we have available manual relevance assessments obtained using the pooling method. We use *weak relevant judgments* (i.e. at least one of the assessors considered document to be relevant, see [13] for details). There are 388 relevant documents in the subset used.

4 Data Processing

All HTML documents in our sample were converted into plain text files.

4.1 Style-Dependent Text Parameters

The initial feature set used for analysis was borrowed almost unchanged from our previous experiments [3]-[5] and included following parameters:

- Average word length (in characters);
- Direct speech sentence rate (based on simple template);
- Average sentence length (in words);
- Expressive punctuation mark (!, ?, …) per sentence rate;
- Morphology-related parameters (7 in total);
- First/second person pronoun rates;
- Particle *бы* rate (conjunctive mood cue);
- Particle *ну, вот, ведь* rate (everyday communication style cue);
- Genitive chain per sentence rate;
- Smiley *:) ;-)* per sentence rate;
- "Unknown" word rate (words absent in stemmer's dictionary);
- Acronym rate (based on simple acronym recognition rule);
- Punctuation mark (comma, semicolon, colon, dash) per sentence rate.

Morphology-related parameters are resolved using *mystem* stemmer developed by Yandex (see http://corpora.narod.ru/mystem) with minor post-processing of the output.

The parameters are calculated based on 100 sentences (or less, for short documents) from the document body, skipping ten leading and ten closing sentences. Our previous experiments have shown that automatic style recognition for short documents is weak. Therefore, we skipped documents shorter than 50 sentences (i.e. less than 30 processed sentences), which is less than a page using present RCDL'2005 layout. As a result we calculated stylistic parameters for 1824 documents which makes ca. 68% of the initial sample.

Most of the proposed parameters cannot be computed absolutely accurately in a fully automatic mode. For instance, grammatical ambiguity was not resolved. Another crucial problem is sentence border recognition.

4.2 Factor Extraction

Factor analysis techniques have been employed for parameters aggregation. We used Factor Analysis Module of the STATISTICA software (www.statsoft.com). In particular, we used principal components method for factor extraction. In a nutshell, this method allows us to reduce the number of variables (i.e. parameters) that describe objects (documents in our case). Reduction is possible through combining multiple correlated variables into a single factor. Computationally, the task is equal to the eigenvalue problem for the correlation matrix.

We apply factor analysis (1) to the whole random sample ('global analysis') and (2) to each of the 51 ranked document lists ('local analysis') independently. In both cases a new score for each document is a linear combination of the same initial parameters. However, in the former case the scores are computed uniformly for every document in the same way. In the latter case every document subset corresponding to a query is processed separately, and the coefficients (so the resulting factor scores) may differ.

4.3 Readability Scores

Our supplementary goal was to compare the obtained parameters with well-established (at least for English) and (mostly) easily computable readability measures. Unfortunately, we failed to find any analytic descriptions of such measures for Russian. We had to compute 'blindly' two characteristics implemented in MS Word for each document: Reading Ease (value range: 0 – 100) and Grade Level (value range: 0 – 20).

5 Results

5.1 Feature Selection & Factor Extraction

Descriptive statistics for the random document sample are presented in the Table 1.

Table 1: Random Sample Parameters

	Mean	Min	Max	Std. Dev.
Expressive punctuation	0,12	0,00	0,95	0,15
Direct speech sentence rate	0,04	0,00	0,71	0,10
Average sentence length	14,29	2,45	35,21	5,63
Average word length	5,69	3,72	7,89	0,65
Noun rate	0,41	0,26	0,96	0,07
Neuter noun rate	0,23	0,01	0,45	0,07
Adverb rate	0,07	0,00	0,15	0,03
Verbal forms rate	0,16	0,00	0,27	0,04
Personal verb forms rate	0,10	0,00	0,24	0,04
Adjective rate	0,13	0,02	0,26	0,04
Short adjective rate	0,01	0,00	0,14	0,01
"Unknown" word rate	0,05	0,00	0,50	0,06
Acronym rate	0,01	0,00	0,27	0,02
Genitive chain rate	0,09	0,00	0,45	0,08
*Particle **бы** rate*	0,03	0,00	0,43	0,04
*Particle **ну**, **вот**, **ведь** rate*	0,03	0,00	0,46	0,05
First person pronoun rate	0,03	0,00	0,25	0,05
Second person pronoun rate	0,02	0,00	0,25	0,03
Smiley per sentence rate	0,01	0,00	0,41	0,04
Punctuation mark per sentence rate	1,86	0,05	6,33	0,91

Taking into account results of our previous experiments, after numerous trials on random sample, we selected four variables to combine into a single factor.

Those are average word length (x_1), personal verb forms rate (x_2), adjective rate (x_3), and first person pronoun rate (x_4). The correlations between variables are presented in Table 2.

Table 2: Correlations between selected variables

	x_1	x_2	x_3	x_4
x_1	1,00	-0,66	0,65	-0,59
x_2	-0,66	1,00	-0,61	0,55
x_3	0,65	-0,61	1,00	-0,44
x_4	-0,59	0,55	-0,44	1,00

The formula for style-dependent score (S_G) based on 'global analysis' looks like follows:

$$S_G = -0{,}32 \cdot x_{1S} + 0{,}31 \cdot x_{2S} - 0{,}30 \cdot x_{3S} + 0{,}28 \cdot x_{4S},$$

where $x_{.S}$ denotes the respective standardized value.

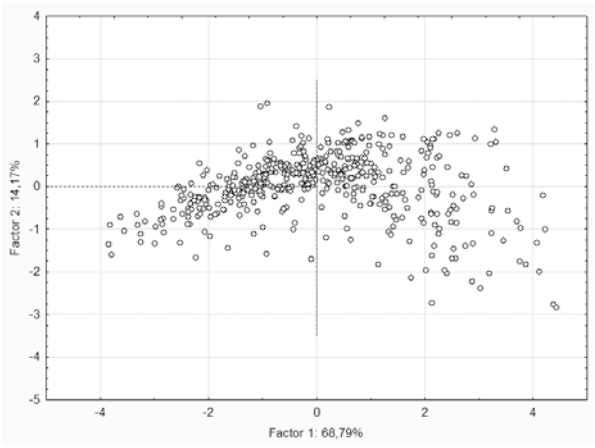

Figure 2: Projection of the random sample on the factor-plane

Formally, the obtained factor represents quite a good "information compression": it explains 68,8% of the sample variance (the scatterplot in Figure 2 illustrates this fact) and reproduces correlations between variables fairly well. Though the scatterplot is not as tight as the one in Figure 1, we have to bear in mind that the former sample was collected manually, whereas the latter generated randomly from the ROMIP collection.

Using the same factor analysis technique we computed style-dependent document scores based on 'local analysis' of each of the 51 result lists (S_L). The values of S_G and S_L for individual documents are comparable, which proves the intuition that search results returned to a query vary significantly in style. However, we should carefully consider factors obtained based on only six or eight cases.

The correlations between our style scores S_G, S_L and readability measures implemented in MS Word (*Reading Ease – RE*, *Grade Level – GL*) are shown in Table 3. As we can see, S_G and S_L are fairly interchangeable. Moreover, S_G is correlated with *GL*; the obstacle in using the latter parameter for our purposes is lack of its analytic description.

Table 3: Correlations between Style-Dependent Scores

	GL	*RE*	S_G	S_L
GL	1,00	-0,91	-0,73	-0,50
RE	-0,91	1,00	0,57	0,38
S_G	-0,73	0,57	1,00	0,81
S_L	-0,50	0,38	0,81	1,00

5.2 Comparison of Ranks

On the next stage of our investigation we ranked the Kodeks' result lists according to the obtained style-dependent scores. Short documents for those style-dependent parameters were not calculated were put to the bottom of the re-ranked lists.

We performed subjective evaluation for selected lists to figure out if the ranks comply with intuition of 'formality/informality' of documents' styles. The results were rather encouraging, except for the link lists, mixed-style documents (documents with extensive quotations or multi-part pages), and HTML pages with complex layout (menu items, navigation bars, advertising, etc. skewed the results).

Subsequently, we compared new ranks and initial Kodeks' relevance rank (R_K) with each other using *Spearman R* and *Kendall τ* statistics. Both statistics yielded similar results. If we take a look at the rank correlation matrix (Table 4), we can make two important observations. First, transition from style-dependent scores to the ranks smoothes the difference between approaches. Second, all correlations between relevance rank (R_K) and style-dependent ranks are low. This observation proves that (1) style is independent from topic of the document to a certain extent, and (2) the result lists may include documents of different styles.

Table 4: Spearman Rank Order Correlations

	R_K	R_{GL}	R_{RE}	R_{SG}	R_{SL}
R_K	1,0	0,11	0,09	0,18	0,18
R_{GL}	0,11	1,0	0,97	0,73	0,73
R_{RE}	0,09	0,97	1,0	0,67	0,68
R_{SG}	0,18	0,73	0,68	1,0	0,998
R_{SL}	0,18	0,73	0,68	0,998	1,00

5.3 Rank Aggregation

Next, we attempted to aggregate obtained style-dependent ranks with initial relevance rank. Since style-dependent ranks demonstrate similar behavior, we opted for single 'global' rank (R_{SG}) as a rank with a more practical computational scheme.

We used a straightforward approach to aggregation: new rank was computed as a linear combination of relevance and style-dependent ranks. This scheme can be referred to as a simple case of weighted Borda method that is widely used in different areas [1]. Note, that in contrast to a more generic problem definition of rank aggregation for metasearch, we had only two 'voters' (i.e. ranks) and they represented two different orders over the *same set of items*.

For evaluation of aggregated ranks we used several measures.

First, we employed *Rank Displacement Coefficient* (RDC) as proposed in [2] and its variants. RDC sums the ups and downs of individual documents in the new list in comparison to the original one. Since the new rank is merely a new permutation in our case, the total RDC for all documents in a list would be equal to zero. We compute RDC for relevant documents separately. For example, if in a list one relevant document dropped two spots in rank and another relevant document rose three positions, then RDC for relevant documents would be –2+3=1. After that we summed the coefficients over individual tasks (i.e. document lists), which yielded Absolute Rank Displacement Coefficient for relevant documents (D_R) over the whole set of lists. Moreover, we computed Averaged Rank Displacement Coefficient for both relevant (AD_R) and non-relevant documents (AD_N). These figures reflected average movements of documents in the new ranks.

Furthermore, we counted up tasks with positive (⊕), null (∅), and negative (⊗) values of Rank Displacement Coefficients for relevant documents.

Evaluation results are summarized in Table 5.

Table 5: Aggregated Ranks Evaluation

	D_R	AD_R	AD_N	⊕	∅	⊗
R_{SG}	-1377	-3,55	0,62	16	0	35
$R_K + R_{SG}$	-95	-0,24	0,04	21	1	29
$R_K + 0{,}5 * R_{SG}$	73	0,19	-0,03	22	0	29
$R_K + 0{,}25 * R_{SG}$	57	0,15	-0,03	22	6	23
$R_K + 0{,}125 * R_{SG}$	54	0,14	-0,02	22	11	18

6 Discussion

Evaluation of the aggregated ranks shows that the proposed method yields only moderate improvements. Though we can find a combination of ranks that would produce positive Rank Displacement Coefficients, the number of individual tasks with improved ranking order is quite discouraging. The mentioned issue with mixed-style documents can explain the moderate gain. Probably, we played too safe setting the low margin for documents to be processed on 50 sentences, since many documents judged as relevant turned to be shorter.

It can be noted that different tasks behave differently when exposed to re-ranking. About 20 tasks showed definite improvements under all combination schemata. At the beginning of the experiment we marked some tasks as potential candidates for relevance rank improvement based on examination of extended task descriptions for the assessors. Interestingly that the tasks with positive effect were not necessarily the ones we expected to be. This fact implies a possible direction for the future work: query analysis in respect of potential suitability for style-dependent ranking.

Another option for getting more promising results would be incorporating stylistic analysis *inside* the information retrieval system to allow a subtler interplay between relevance and style-dependent scores.

Although the use of random sample for factor extraction is very attractive due to low efforts, we are going to perform further experiments with manually collected document samples or tagged corpus.

Acknowledgements

This work is supported by Yandex Company (www.yandex.ru).

We would like to thank Maxim Gubin (www.kodeks.ru) for providing us with data and helpful bibliographic references.

We also thank Nadezhda Shalamova for reading the draft and making useful comments and remarks on the paper.

References

[1] Aslam, J. A., Montague, M. Model for Metasearch. In *Proceedings of the SIGIR'01*, September 9-12, 2001, New Orleans, USA, p. 275-284.

[2] Beitzel, S. M., et al. Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. In *JASIST*, vol. 55(10), 2004, p. 859-868.

[3] Braslavski, P. Document Style Recognition Using Shallow Statistical Analysis. In *Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, Nancy, France, 2004, p. 1-9.

[4] Braslavski, P. Experiments on Automatic Text Classification According to Style (Case Study on Web Documents) (in Russian). [Opyt avtomatičeskoi klassifikacii tekstov po stilyam (na materiale dokumentov Internet)]. In *Russian Language on the Internet*. Ed. Valery Solovyev. Kazan, 2003, p. 6-15.

[5] 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]. In *Proceedings of the International Workshop "Dialogue-2002. Computational Linguistics and Intelligent Technologies"*, Moscow, 2002, vol. 2, p. 83-93. Available online: http://company.yandex.ru/articles/article8.html

[6] Gubin, M. Information Retrieval System 'Kodeks' at ROMIP-2003 (in Russian). [Opyt učastiya IS "Kodeks" v ROMIP'2003]. In *Proceedings of ROMIP'2003*, St.-Petersburg, 2003. Available online: http://romip.narod.ru/romip2003/2_gubin.pdf

[7] DuBay, W. H. The Principles of Readability. Available online: http://www.nald.ca/fulltext/readab/readab.pdf

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

[9] Kessler, B., Nunberg, G., and Schütze, H. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 7-12 July, Madrid, Spain, 1997, p. 32-38.

[10] Kožina, M.N. Foundations of Functional Stylistics (in Russian). [K osnovaniyam funkcional'noi stilistiki], Perm, 1968.
[11] Lee, Y.-B. and Myaeng, S. H. Text Genre Classification with Genre-Revealing and Subject-Revealing Features. In *Proceedings of the SIGIR'02*, August 2002, Tampere, Finland, p. 145-150.
[12] 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.
[13] Russian Information Retrieval Evaluation Seminar, http://romip.narod.ru
[14] Santini, M. State-of-the-Art on Automatic Genre Identification. Technical Report ITRI-04-03, Information Technology Research Institute, Univ. of Brighton, 2004. Available online: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf

♣ This research is supported through Yandex grant program (http://company.yandex.ru/grant), grant #102604.