Combining Relevance and Genre-Related Rankings: an Exploratory Study

Pavel Braslavski
Institute of Engineering Science, RAS
Komsomolskaya 34
620219 Ekaterinburg, Russia
pb@imach.uran.ru

Abstract

In this paper, we examine whether it is possible to effectively incorporate document genre features into document relevance ranking. First, a method for extracting 'seriousness' score of a document using canonical discriminant analysis applied to a sample of functional styles is proposed. Second, effects of aggregating genre-related and text relevance ranks are considered. Evaluation of the results shows moderate positive effects.

Keywords
Relevance ranking, genre analysis, readability.

1. Introduction

Recent years have shown a growing interest to automatic genre analysis of Web documents, especially in the context of Web search.

One of the possible research tasks is automatic genre categorization, i.e. automatic classification of documents into predefined set of genres (see our early study [2] and a comprehensive survey of the field [13]). When the genre palette is not very fine-grained and controversial, the problem can be solved with acceptable quality. However, when thinking of applying genre categorization to a commercial general-purpose search engines, the main problem could be to adapt or invent a suitable genre palette that is intuitively clear, complete, and not ambiguous for the majority of users. Moreover, the appropriate interface should be presented. Meanwhile a simple search box and a sorted list of search results is a standard de facto for millions of search engine users. The experiments [11] show that though most users expect genre information to be helpful for their Web search tasks, a straightforward implementation of genre-related hints doesn't improve user search effectiveness significantly.

Research on readability has its roots in psycholinguistics but in fact is very similar to automatic genre analysis. The aim is to obtain a simple measure to compare the comprehension complexity of texts conveying similar meaning using surface cues [3].

The paper reports on ongoing experiments aimed at embedding genre information into relevance ranking, which makes use of genre analysis transparent for the end-user. The idea is to obtain a simple measure of document's genre (similar to readability score) and embed it into ranking. The idea can be seen in the context of static ranking: to incorporate diverse page-level features that are independent from query into ranking scheme [10].

In contrast to our previous study [1] when we used unsupervised approach, in the current study we employ supervised methods for extracting genre-related scores.

A related study is described in [6]: a 'familiarity classifier' is build upon several hundreds of documents manually tagged as 'introductory' or 'advanced'. However, the method doesn't consider topic relevance: top-20 documents returned by a search engine are all assumed to be relevant to the query, which seems to be a very strong assumption.

2. Data

In this study we used two datasets of Russian documents: 1) a small corpus of five functional styles as learning sample for extracting genre-related score and 2) a subset of reference ROMIP Web collection for experimentation and evaluation purposes. ROMIP stands for Russian Information Retrieval Evaluation Seminar which is a Russian TREC-like initiative [12].

2.1 Functional Styles Sample

We draw on the well-established in Russian linguistics concept of functional styles. There are five basic functional styles: *official style*, *academic style*, *journalistic style*, *literary style,* and *everyday communication style*. We use this sample only for building a genre-related score.

We re-use a sample of 305 documents in Russian that was employed in our previous experiments. This sample of documents consisted of 50 federal acts, 54 scientific papers in natural sciences, 61 online news articles, 79 short stories by modern Russian authors, and 61 fragments of online chats. More details on the sample can be found in [2].

2.2 ROMIP Collection

ROMIP Web collection contains about 600,000 HTML pages in Russian from the free Web hosting *narod.ru* and reflects well the diversity of Web genres. The collection is used in the ROMIP *ad hoc* retrieval track and is freely available upon request.

Along the documents the collection contains a list of about 20 000 queries taken from a real-life search engine query log. Each participating system performs the whole set of queries over ROMIP collection. A small selection of queries is evaluated manually using pooling method each year. To increase assessors' agreement each query is provided with an extended description which represent one of the possible query interpretations (Fig. 1). Most descriptions imply detailed and informative documents. This fact suggests that we could improve the overall search quality within the ROMIP framework by ranking 'serious' documents higher. We implement this approach in our experimental framework, however it won't suit all real-life information needs obviously. (For example, homepage of the *Sexologies* journal published by Elsevier wouldn't satisfy the vast majority of users asking one of the most frequent queries *sex*.)

> *Query rb4095*: common spruce
> *Description*: A relevant page must contain information about common spruce – e.g. main characteristics of this tree type, natural habitat, industrial use, etc.
>
> *Query arw13494*: memory training
> *Description*: Documents containing advices for human memory improvement, diverse techniques for memory training. Documents containing recipes of food supplements are useful. Especially important are documents containing detailed and precise instructions for those who want to train their memory.
>
> *Query arw19003*: are we alone in the universe?
> *Description*: The page must contain information on extraterrestrial intelligence research, existing hypotheses as well as different opinions on this issue.

Figure 1. Sample ROMIP tasks: query and its description (originally in Russian, descriptions are used for evaluation purposes only)

For our experiment, we took the results of one of the ROMIP'2006 participating systems which utilizes only text relevance methods [4]. All ROMIP documents were converted to *plain text*; no other pre-processing was performed.

Our ROMIP subset contains 6,906 documents corresponding to 70 evaluated search tasks. The majority of these documents have relevance judgments: 5393 documents (420 relevant + 4973 non-relevant) with so-called 'strong' judgments (i.e. all assessors agreed on judgment) and 5416 documents (1105 relevant + 4311 non-relevant) with 'weak' relevance judgments (i.e. at least one assessor judged a document as 'relevant'). Some tasks have no corresponding relevant documents (13 in case of 'strong' relevance and three in case of 'weak' relevance). The rest of the documents have tag 'can't be judged' or didn't fall into the evaluated document pool.

We re-ranked ROMIP subset in different ways to find if we can achieve a better relevance ranking if we take genre-related document score into account.

3. Methods and Results

3.1 Functional Styles and Genre Score Extraction

Unfortunately, there is no widely accepted and use-proven readability score for Russian that would be appropriate for our aims. So we opted for building a 'seriousness' score based on our previous research.

We employed the concept of functional styles, which is well-established 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: *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 [5]. Functional styles have been subject of an early study on automatic stylistic analysis [8].

We consider five functional styles as text classes of gradually decreasing 'seriousness'. The quantitative characteristics of the functional styles sample confirm our intuition (Fig. 2). Such features as average word length (one of the most commonly used features in different readability formulae) and POS distribution change monotonically over five styles.

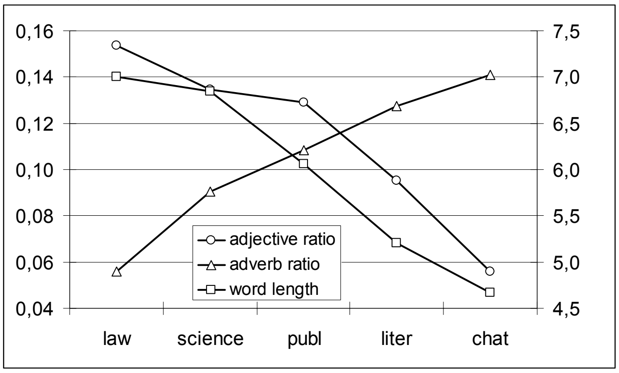

Figure 2. Selected characteristics of the functional styles sample

Our approach is rather operational. Though it is not quite correct we don't make distinction between *genre* and *style* assuming that both concepts relate to a higher-level notion of "*how* a given piece of information is presented" [9].

We use *canonical discriminant analysis* to extract the 'seriousness' score. The method is illustrated in Fig. 3: we perform feature space transformation in order to find a direction (a weighted sum of initial features) with the best separating ability between classes. The method is similar to *principal components analysis* (PCA); the difference is that we take class structure into account.

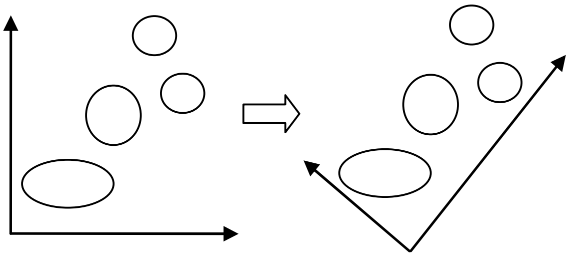

Figure 3. The idea of canonical discriminant analysis

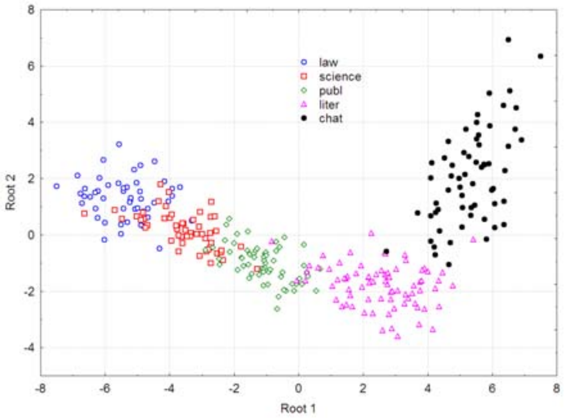

Figure 4. Scatter-plot of the learning sample in the 1st and 2nd canonical roots

We carried out trails with different genre-related easily computable text features. After a series of trials we opted for a combination of nine features. The formula for the first canonical root that we treat as 'seriousness' score is as follows (standardized values):

$$S = -0.49x_1 + 0.27x_2 + 0.46x_3 + 0.04x_4 + + 0.24x_5 + 0.32x_6 - 0.48x_7 + 0.32x_8 - 0.11x_9,$$

where

x_1 – average word length;

x_2 – smiley count;

x_3 – finite verb count;

x_4 – adjective count;

x_5 – first person pronoun count;

x_6 – expressive punctuation count;

x_7 – neuter noun count;

x_8 – adverb count;

x_9 – genitive chain count.

The first canonical root explains 84% of sample's variance (Fig. 4).

The obtained index is similar to a readability score. Such corpus-based approach is low-cost, flexible, and easily adjustable compared to traditional methods for building readability scores based on reading tests. Our approach is similar to one described in [14]. However, we don't utilize lexical features, which leads to much lower computational costs.

3.2 Ranks Aggregation

We calculated genre-related scores for our ROMIP subset. Relevant documents appeared to be somewhat 'more serious': averaged normalized genre scores for our ROMIP subset are 0.62 and 0.59 for relevant and non-relevant documents, respectively (the difference is significant at $p<0.005$). Then, we ranked all evaluated documents that were longer than five sentences according to the genre-related score ('serious' documents on the top); all other documents preserved their initial positions.

Next, we aggregated the obtained genre-related ranks (R_G) with the initial keyword-relevance ranks (R_Y) (see [4] for details on R_Y). We used a straightforward approach to aggregation: new rank was computed as a linear combination of text relevance and genre-related ranks, i.e. $R_Y + \alpha R_G$. This scheme can be referred to as a simple case of weighted Borda method that is widely used in different areas, including rank aggregation for metasearch.

It is important to note that we didn't aim at finding an optimal α for the rank combination. Although the number of processed documents is big enough, the number of document sets (57 and 67 for 'strong' and 'weak' relevance, respectively) didn't allow us to test our results properly and generalize well.

For evaluation of the aggregated ranks we used rank displacement of relevant documents (D_R). D_R sums the ups and downs of relevant documents in the new list in comparison to the original one (Fig. 5). Furthermore, we counted up tasks with positive and negative values of D_R.

The most illustrative results were obtained on weak relevance judgments. Fig. 6 shows both macro- and micro-averaged D_R values depending on genre rank's weight. The best macro-averaged value (0.48) is achieved at α=0.18, while micro-averaged D_R reaches its maximum (0.26) at α=0.14. Fig. 7 shows the absolute numbers of tasks in the sample with positive vs. negative changes depending on α. The best relation (42 vs. 18) is achieved at α=0.07.

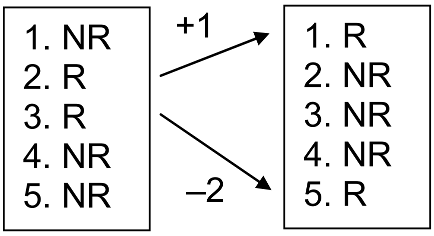

Figure 5. Rank displacement of relevant (R) documents (for this example $D_R = -1$)

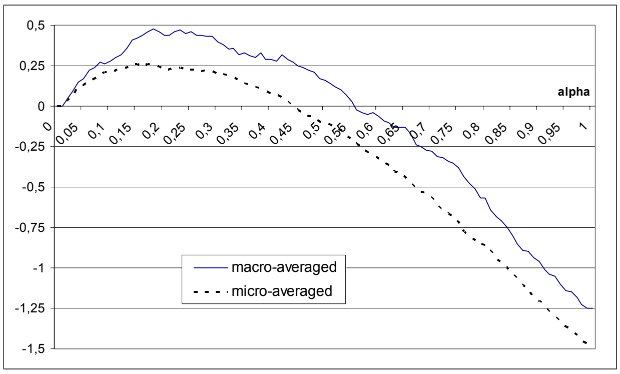

Figure 6. Averaged rank displacement of relevant documents (weak relevance judgments, 4846 documents in 67 sets processed)

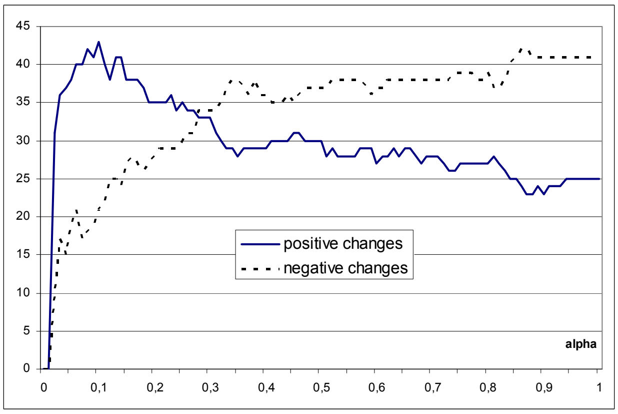

Figure 7. Number of tasks with positive and negative D_R

4. Conclusion and Future Work

Our evaluation of the aggregated ranks shows that we can achieve moderate improvements on our experimental data set in average by mixing in a small fraction of genre-related rank. Notably, there is a subset of queries that is very receptive to mixing genre ranks with traditional keyword-relevance ranks. This fact poses a much more difficult problem of how to define the expected genre (or *genre range* when thinking of continuous genre index) of the answer based on query preprocessing. To the best of our knowledge the sole study on predicting user's education level based on a query is [7]. We will address the problem in the future.

5. Acknowledgements

We would like to thank Mikhail Ageev and Andrei Tselishchev for their help with data processing. We also thank Yandex (www.yandex.ru) for providing us with the experimental data.

We thank anonymous reviewers whose comments helped us to improve the paper.

6. References

[1] Braslavski, P. and Tselishchev, A. Style-Dependent Document Ranking. In *Proc. of the 7th Russian Conference on Digital Libraries RCDL'2005*, Yaroslavl, Russia, 2005. Available online: http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7_1_paper.pdf

[2] Braslavski, P. Document Style Recognition Using Shallow Statistical Analysis. In *Proc. of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, Nancy, France, 2004. Available online: http://esslli2004.loria.fr/content/readers/36.pdf

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

[4] Gulin, A., Maslov, M., and Segalovich, I. Yandex' Algorithm for Text Relevance Ranking at ROMIP'2006 [Algoritm tekstovogo ranğirovanija Jandeksa na ROMIP'2006] (In Russian). In *Proc. of ROMIP'2006*, Suzdal, Russia, 2006. Available online: http://www.romip.ru/romip2006/03_yandex.pdf

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

[6] Kumaran, G., Jones, R. and Madani, O. Biasing Web Search Results for Topic Familiarity. In *Proc. of CIKM'05*, 2005.

[7] Liu, X., Croft, W. B., Oh, P. and Hart, D. Automatic Recognition of Reading Levels from User Queries. In *Proc. of SIGIR'2004*.

[8] Michos, S., Stamatatos, E., Fakotakis, N. and Kokkinakis, G. Categorizing Texts By Using a Three Level Functional Style Description. In A. M. Rasmsay (Editor): Artificial Intelligence: Methodology Systems, Applications, Frontiers in Artificial Intelligence and Applications, Vol. 35, 1996. Available online: http://slt.wcl.ee.upatras.gr/papers/michos2.pdf

[9] Rauber, A. and Müller-Kögler, A. Integrating Automatic Genre Analysis into Digital Libraries. In *Proc. of the JCDL'2001*.

[10] Richardson, M., Prakash, A., and Brill, E. Beyond PageRank: Machine Learning for Static Ranking. In *Proc. of WWW '2006*.

[11] Rosso, M.A. *Using Genre To Improve Web Search*. PhD thesis, University of North Carolina, Chapel Hill, 2005.

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

[13] Santini, M. *State-of-the-Art on Automatic Genre Identification*. Technical Report ITRI-04-03, Information Technology Research Institute, Univ. of Brighton, UK, 2004. Available online: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf

[14] Si, L. and Callan, J. A Statistical Model for Scientific Readability. In *Proc. of CIKM '2001*.