DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **14**, No 3

Sofia • 2014

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2014-0030

Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields

Bohdan Pavlyshenko

Ivan Franko Lviv National University, Ukraine
Email: b.pavlyshenko@gmail.com

Abstract: *This paper describes the analysis of possible differentiation of the author's idiolect in the space of semantic fields; it also analyzes the clustering of text documents in the vector space of semantic fields and in the semantic space with orthogonal basis. The analysis showed that using the vector space model on the basis of semantic fields is efficient in cluster analysis algorithms of author's texts in English fiction. The study of the distribution of authors' texts in the cluster structure showed the presence of the areas of semantic space that represent the idiolects of individual authors. Such areas are described by the clusters where only one author dominates. The clusters, where the texts of several authors dominate, can be considered as areas of semantic similarity of author's styles. SVD factorization of the semantic fields matrix makes it possible to reduce significantly the dimension of the semantic space in the cluster analysis of author's texts. Using the clustering of the semantic field vector space can be efficient in a comparative analysis of author's styles and idiolects. The clusters of some authors' idiolects are semantically invariant and do not depend on any changes in the basis of the semantic space and clustering method.*

Keywords: *Text mining, text clustering, semantic fields.*

1. Introduction

In the analysis of author's texts it is efficient to use the methods of data mining, particularly the clustering methods. In clusterization of text arrays, a vector model of the text documents is used, according to which the documents are considered as vectors in some vector space and they are formed by quantitative characteristics of words P a n t e l and T u r n e y [6]. As quantitative characteristics the frequencies of keywords are widely used. One of the problems of such an approach is the great dimension of the text documents, caused by the size of the vocabulary of the analyzed text array. A promising trend is to use the vector space with the basis formed by quantitative characteristics of word associations, in particular semantic fields. A semantic field is a set of words that are united under some common concept, e.g., the field of motion, the field of communication, the field of perception, etc. The number of semantic fields is significantly smaller than the size of a word dictionary and it reduces the amount of necessary calculations. Similar objects are the semantic networks that describe the relationships among different concepts. An example of lexicographic computer system, which represents the semantic network of links between words, is a WordNet system, developed in Princeton University by F e l l b a u m [2]. This system is based on an expert lexicographic analysis of semantic structural relationships that describe the denotative and connotative characteristics of dictionary word composition. The paper of G l i o z z o and S t r a p p a r a v a [3] considered the concept of a semantic domain, which describes certain semantic areas of various issues under discussion, such as economics, politics, physics, programming, etc. The algorithms of clusterization and classification are often used in data mining (see S e b a s t i a n i [7], M a n n i n g, R a g h a v a n and S c h ü t z e [5]). Recording the text semantics in the problems of text documents clustering makes it possible to obtain the clustering of greater accuracy (S h e h a t a, K a r r a y and K a m e l [8]). In L a r s e n and A o n e [4] text clustering algorithms and the evaluation of their effectiveness are described.

In this paper, we investigate the clusterization of authors' texts in the space of semantic fields. In Section 2, we consider the vector model of the text documents in the space of semantic fields in terms of using this model in agglomerative clustering algorithms. We apply the singular decomposition of the matrices of semantic fields of the text documents to form an orthogonal semantic space in Section 3. We perform an experimental analysis of the authors' texts in English fiction using clustering algorithms in the space of semantic fields and in the semantic space with orthogonal basis (Section 4). In Section 5 we summarize our study and make conclusions.

2. The model of text documents clustering in the space of semantic fields

Here we consider a model based on a set theory, which describes a set of text documents and semantic fields. We describe a set of text documents as

(1) $$D = \{ d_j \mid j = 1, 2, ..., N_d \},$$

and introduce the set of semantic fields as

(2) $$S = \{ s_k \mid k = 1, 2, ..., N_s \}.$$

Then we form a matrix of a feature-document type, where the features are the frequencies of the semantic fields in the documents:

(3) $$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1, j=1}^{N_s, N_d}.$$

The frequencies of the semantic fields p_{kj}^{sd} are defined as sums of the word text frequencies that are included into these semantic fields. The values of these frequencies are normalized so that their sum for each document is equal to 1. The vector

(4) $$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right)^{\mathrm{T}}$$

displays the document d_j in N_s-dimensional space of the text documents. The introduction of the space of semantic fields not only reduces the size of the problem of texts analysis, but also introduces a new basis for text descriptions. In the semantic basis, qualitatively new clustering text documents can be observed.

Let us consider the document groupings by semantic features using the hierarchical clustering algorithm. Suppose there is a set of text documents D, which is described by the expression (1) and a set of clusters

(5) $$C = \{ c_m \mid m = 1, 2, ..., N_c \}.$$

It is necessary to build a mapping of the document set by a cluster set

(6) $$U_{DC} : D \to C.$$

The mapping U_{DC} specifies the data model, which is a solution of the clustering problem. Each element c_m of the set of clusters C consists of a subset of text documents that are similar to each other according to some quantitative similarity measure r

(7) $$c_m = \{ d_i, d_j \mid d_i \in D,\ d_j \in D,\ r(d_i, d_j) < \varepsilon \},$$

where ε defines a threshold for including the documents into the cluster. The value $r(d_i, d_j)$ is the distance between the elements d_i and d_j, and if it is less than some value, then the sample elements are considered as being similar and belonging to a common cluster. The distance $r(d_i, d_j)$ must match the following conditions: $r(d_i, d_j)>0$; $r(d_i, d_j)=0$ if $d_i=d_j$; $r(d_i, d_j)=r(d_j, d_i)$; $r(d_i, d_j) \le r(d_i, d_j)+ r(d_i, d_j)$. Since the concept of distance is introduced on the set of text documents, each document is represented as a point in N_s-dimensional space of N_s semantic features. In our studies we calculate the Euclidean distance. Now we consider a hierarchical agglomerative clustering method. At the first step the entire set of text documents is considered as a set of clusters. At the next step two documents, close to each other, are combined into one common cluster, a new set at this step is composed of N_d–1 clusters. Reiterating the steps, at which the clusters will be combined, we obtain a set of N_c clusters. The process of the cluster combining comes to an end at this step of the algorithm, when no pair of clusters meets the threshold of combining for the proximity measure of elements. There are different methods of forming and joining

clusters on the basis of distances between the objects within the cluster. One of the efficient methods of text documents clustering in the semantic fields space is Ward's method. This method calculates the squares of the Euclidean distances from the individual documents up to the center of each cluster. Then these distances are summed. If the combination of clusters gives the smallest increase in the sum of squares of the distances, then those clusters can be combined in a new one. The graphic representation of the hierarchical clustering result is a dendrogram, which indicates the process of agglomerative clustering aggregation. The numbers of clusters are on the abscissa axis and the distances between clusters are on the ordinate axis. At certain values of the distances the clusters begin to merge. With the increase of the intercluster distance the clusters are merging up to a complete union of clusters into one cluster. Therefore, in order to obtain an informative cluster structure, we must choose some threshold of intercluster distance, while forming the optimal cluster structure, from the point of view of the text arrays analysis.

3. Text analysis in the semantic space with orthonormal basis

The method of latent-semantic analysis, based on the singular decomposition of keywords frequencies matrix, allows to reduce significantly the dimension of the vector space of documents (Deerwester et al. [1]). Let us consider the singular decomposition of the matrix of semantic fields. Suppose there is a matrix of a "semantic_fields_frequencies-documents" type M_{sd}, which is described by formula (3). The vector V_j^s (4) displays the document d_j in N_s-dimensional space of text documents. The product of two vectors $(V_p^s)^{\mathrm{T}} V_q^s$ determines the quantitative measure of similarity of these vectors in N_s-dimensional semantic space of text documents. Accordingly, the product of two matrices $(M_{sd})^{\mathrm{T}} M_{sd}$ contains scalar products of vectors $(V_p^s)^{\mathrm{T}} V_q^s$ of all documents and it reflects their correlations in the semantic vector space. The singular matrix decomposition M_{sd} looks as

$$M_{sd} = U_{sd} \Sigma_{sd} Y_{sd}^{\mathrm{T}}. \tag{8}$$

The diagonal matrix Σ_{sd} contains singular numbers in a descending order. If we take K of the largest singular numbers of the matrix Σ_{sd} and, correspondingly, K of the singular vectors of the matrices U_{sd} and Y_{sd}, we will get the K-rank approximation of the matrix M_{sd}:

$$(M_{sd})_K = (U_{sd})_K (\Sigma_{sd})_K (Y_{sd})_K^{\mathrm{T}}. \tag{9}$$

The matrix $(Y_{sd})_K$ reflects the relations between the vectors of the documents $\hat{V}_j^s$ in the new combined K-dimensional orthogonal semantic space. The relations

between the vector V_j^s of the document in the original semantic space and the vector $\hat{V}_j^s$ in the orthogonal semantic space can be described as

$$V_j^s \approx (U_{sd})_K (\Sigma_{sd})_K \hat{V}_j^s,$$
$$\hat{V}_j^s \approx (\Sigma_{sd})_K^{-1} (U_{sd})_K^{\mathrm{T}} V_j^s. \quad (10)$$

Apparently, the number K can be significantly smaller than the N_s dimension of the initial semantic space. This reduces the dimension of the problem of the analysis of text documents similarity in the semantic vector space.

4. Experimental part

For the experimental study of text documents clustering in the space of semantic fields, we chose a text base containing 503 literary works of 17 authors (A. C. Doyle, A. Trollope, Ch. Dickens, E. Gaskell, E. Lytton, G. Meredith, H. Wells, J. Conrad, J. Galsworthy, Jack London, Mark Twain, R. Kipling, R. Stevenson, T. Hardy, W. Colllins, W. Scott, W. Thackeray). For the semantic space generation we chose the words grouped by the semantic fields of nouns and verbs in the semantic network WordNet (Fellbaum [2]). The semantic fields in the WordNet network (http://wordnet.princeton.edu) are represented as lexicographic files. In our studies we used the semantic fields of nouns and verbs. The semantic fields of nouns consist of 26 lexicographic files, out of which we have selected 54464 words. The semantic fields of verbs contain 15 lexicographic files, out of which we have selected 9097 words. The derivative forms of words were also included into the semantic fields. Lexicographic files WordNet for nouns and verbs have the names that define the semantic core of these fields: noun.tops, noun.act, oun.animal, noun.artifact, noun.attribute, noun.body, noun.cognition, noun.communication, noun.event, noun.feeling, noun.food, noun.group, noun.location, noun.motive, noun.object, noun.person, noun.phenomenon, noun.plant, noun.possession, noun.process, noun.quantity, noun.relation, noun.shape, noun.state, noun.substance, noun.time, verb.body, verb.change, verb.cognition, verb.communication, verb.competition, verb.consumption, verb.contact, verb.creation, verb.emotion, verb.motion, verb.perception, verb.possession, verb.social, verb.stative, verb.weather. We selected the agglomerative clustering method with Euclidean intercluster distance. For the formation of clusters we chose Ward's method. For the cluster analysis we will use a dendrogram, which is a tree diagram used to illustrate the cluster creating by hierarchical clustering. Fig. 1 shows the cluster dendrogram, which describes the formation of the cluster structure. This dendrogram represents the formation of the first 20 clusters. The clustering process is stopped as soon as the cluster structure contains 20 clusters. Fig. 2 shows the histograms of the distribution of authors' texts in the clusters. Each histogram corresponds to a particular cluster. The number of the column indicates the corresponding number of the author. These histograms reflect how the documents

of different groups are distributed in each cluster. There are the clusters where the texts of separate authors stand on the dominant position. As follows from the data given, some clusters contain the texts of wide semantic spectrum. Obviously, the area of these clusters in the semantic space is semantically homogeneous and it has a semantically low differentiating potential. However, there are also such clusters where the texts of one or more authors stand at the dominant position. Such clusters characterize the author's idiolect of individual authors. The semantic space areas of these clusters have differentiating potential for author's idiolect and they can be used while analyzing authors' texts as an additional factor in the analysis of author's lexicon. The areas of the semantic space, corresponding to the clusters where two or more authors dominate, can be considered as the fields of the semantic similarity of these authors. Fig. 3 provides a detailed distribution by authors for the clusters where one author dominates. Such clusters can be regarded as the areas of semantic space that can be used for differentiating author's idiolect in the tasks of analyzing author's style and authors' texts. In the analyzed array of text documents, the areas with the dominance of author's idiolect are characteristic for authors like A. Trollope, Ch. Dickens, Mark Twain, W. Collins. Fig. 4 shows the distribution of texts by authors in the cluster where the texts of several authors dominate. This cluster includes such authors, as J. Galsworthy, Jack London, Mark Twain, R. Kipling. Such cluster describes the area of the vector space of semantic fields characterizing the points of similarity of authors' idiolects.

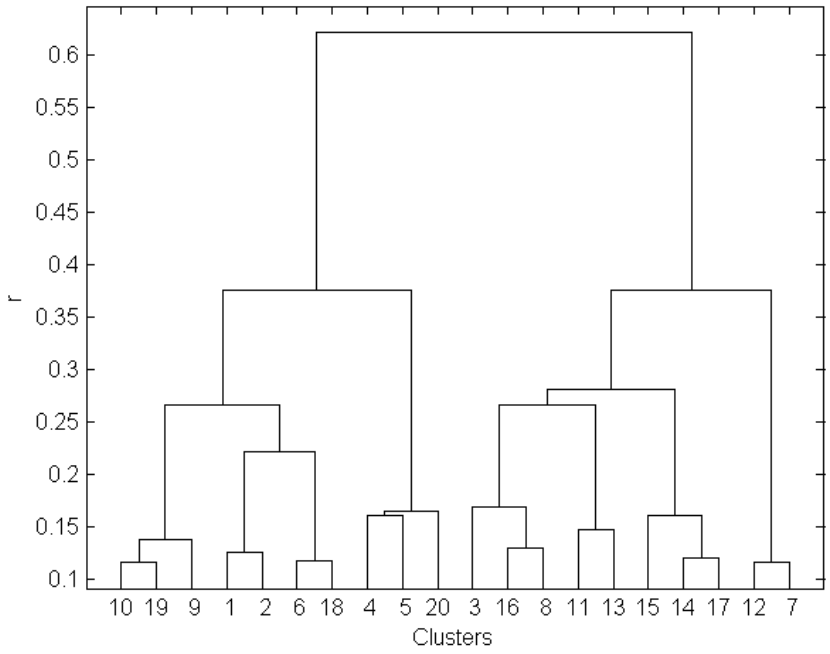

Fig. 1. The dendrogram of hierarchical clustering of authors' texts
in the space of semantic fields

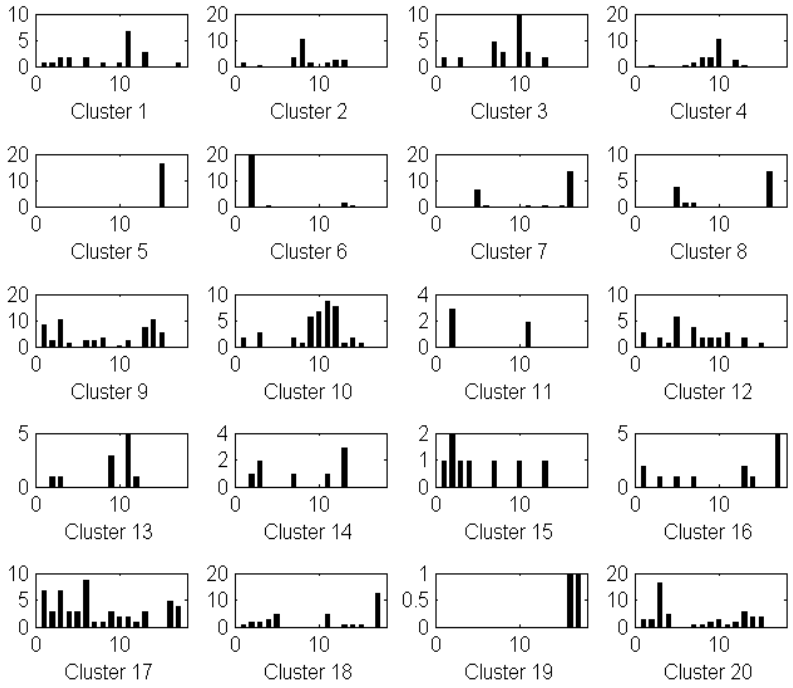

Fig. 2. The distribution of the authors by the clusters
in the space of semantic fields

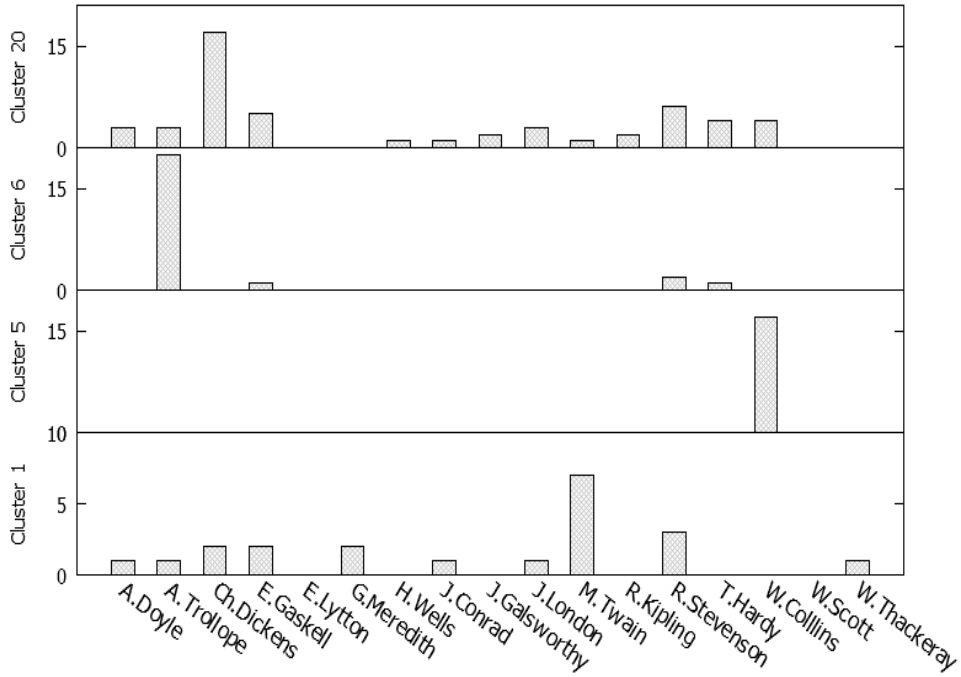

Fig. 3. The distribution of texts in the clusters where one author dominates

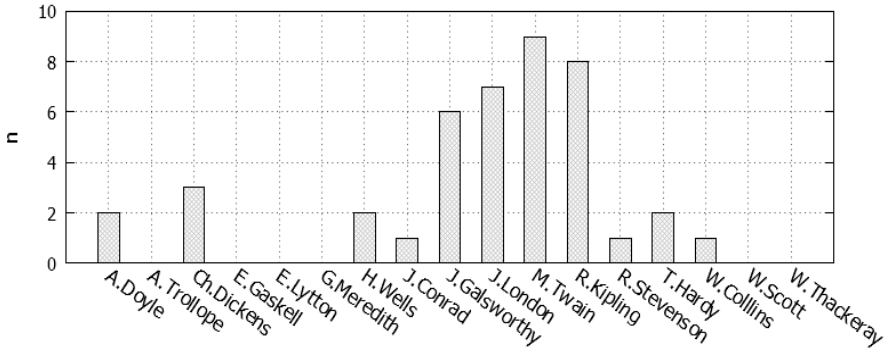

Fig. 4. The distribution of texts in the clusters where several authors dominate

Our next step is to consider the clustering of authors' texts in orthogonal low-dimensional space of secondary semantic fields, generated by SVD factorization of the semantic fields matrix. Fig. 5 shows the first 10 singular values of the semantic frequencies matrix M_{sd}. Here we observe a significant decrease of the values of singular numbers. For the formation of orthogonal semantic subspace we took the coordinates of the secondary semantic fields that correspond to the first 10 singular numbers of the matrix M_{sd}. On the basis of the generated low-dimensional orthogonal space, we have conducted similar calculations of the dendrogram (Fig. 4) and the distribution of authors in clusters (Fig. 7). As follows from the data obtained, the clusters with predominance of individual authors are also present. These clusters also characterize the semantic area of author's idiolect in low-dimensional semantic space with orthogonal basis. We have also conducted the studies for the orthogonal subspace with the dimension of 3. In this case the clusters, where the texts of a certain author dominate, are not observed. Fig. 8 shows the distribution of texts in the clusters with predominance of one author and hierarchical clustering in the orthogonal space. In these clusters the following authors dominate: A. C. Doyle, A. Trollope, Mark Twain, W. Collins, W. Scott. While comparing Figs 3 and 8, we can observe that such authors as A. Trollope, Mark Twain, W. Collins have areas where their author's idiolects dominate both in the space of semantic fields and in the orthogonal semantic space. Fig. 9 shows an example of text distribution by authors in the cluster where several authors dominate. Such a cluster describes the area of contact of author's idiolects of different writers. Along with hierarchical clustering, we have also conducted the clustering using *k*-means. Fig. 10 shows examples of the text distribution by authors in the clusters where one author dominates in the clustering with the use of *k*-means method in the 10-dimensional orthogonal semantic space. In the obtained clusters authors like A. Trollope, Jack London, Mark Twain, and W. Thackeray dominate. As follows from the results obtained, the formation of clusters, where the idiolect of only one author dominates, is defined by both the choice of the basis of the semantic space and the method of clustering. While analyzing the clusters with

dominance of only one author that are obtained by the method of agglomerative clustering in the space of semantic fields (Fig. 3) in the orthogonal semantic space (Fig. 8) and by the method of *k*-means in the orthogonal space (Fig. 10), we can see that there are some authors whose texts dominate in all these cases. These authors are A. Trollope and Mark Twain. Semantic clusters with dominance of author's idiolect of these writers can be regarded as semantically invariant and independent on the considered semantic spaces and clustering methods.

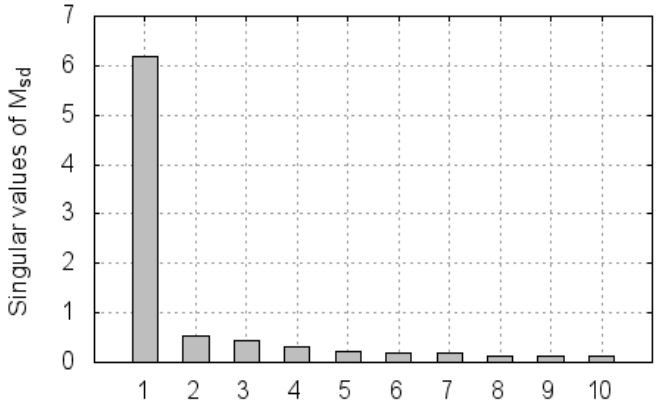

Fig. 5. Singular values for the matrix of semantic fields

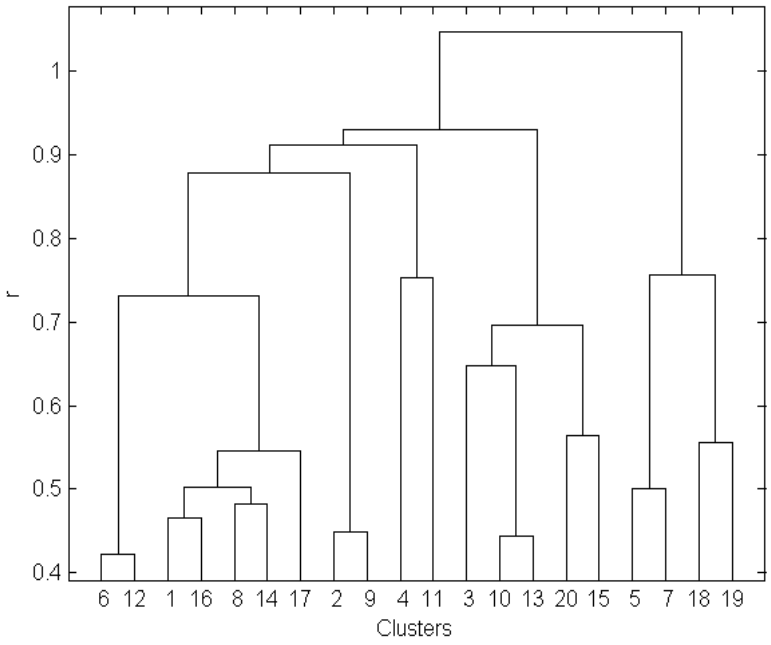

Fig. 6. The dendrogram of hierarchical clustering of authors' texts in the semantic space with orthogonal basis

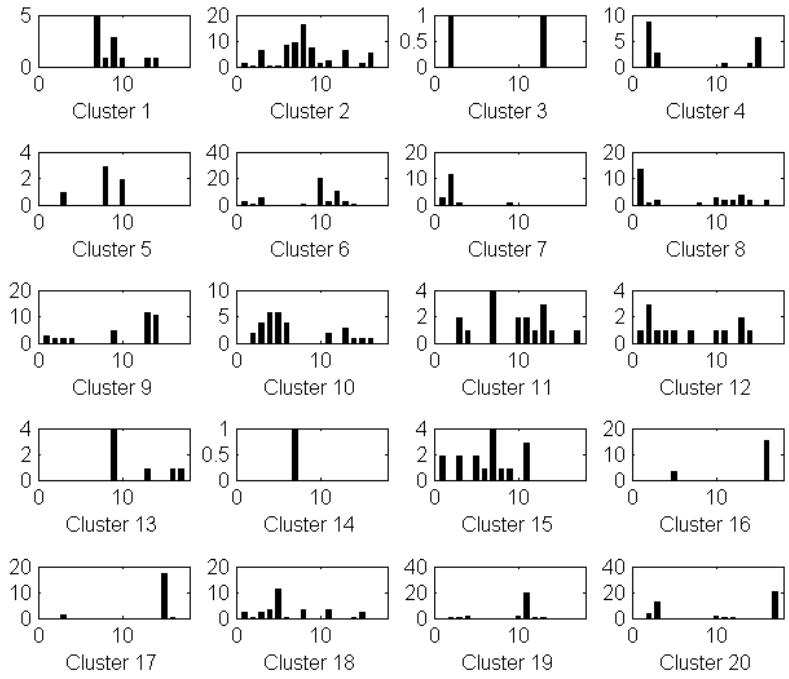

Fig. 7. The distribution of the authors by clusters
in the semantic space with orthogonal basis

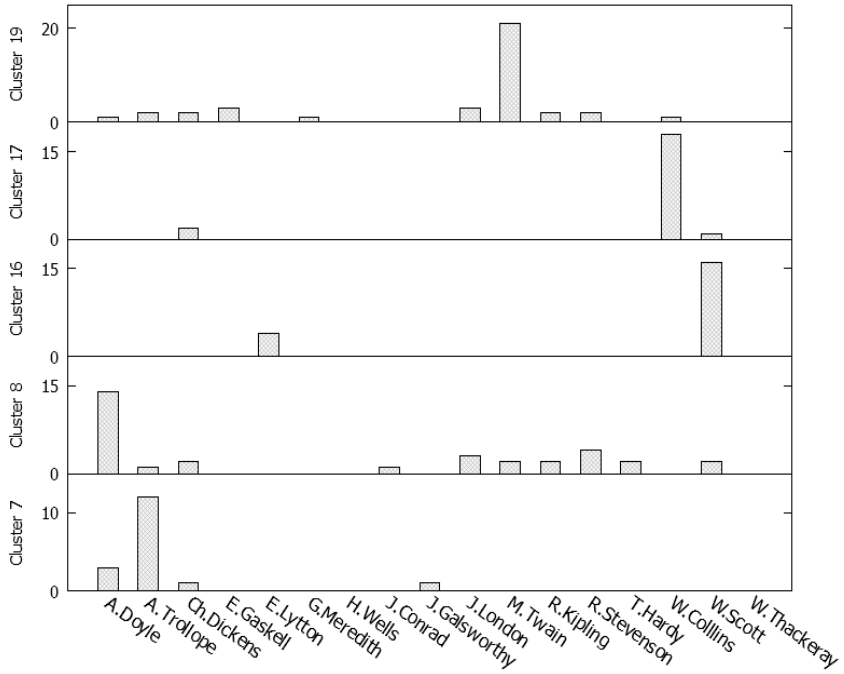

Fig. 8. The distribution of texts in the clusters in orthogonal space,
where one author dominates

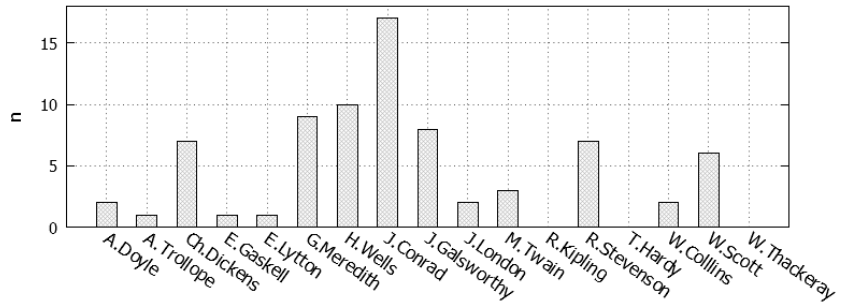

Fig. 9. The distribution of texts in the clusters in orthogonal space,
where several authors dominate

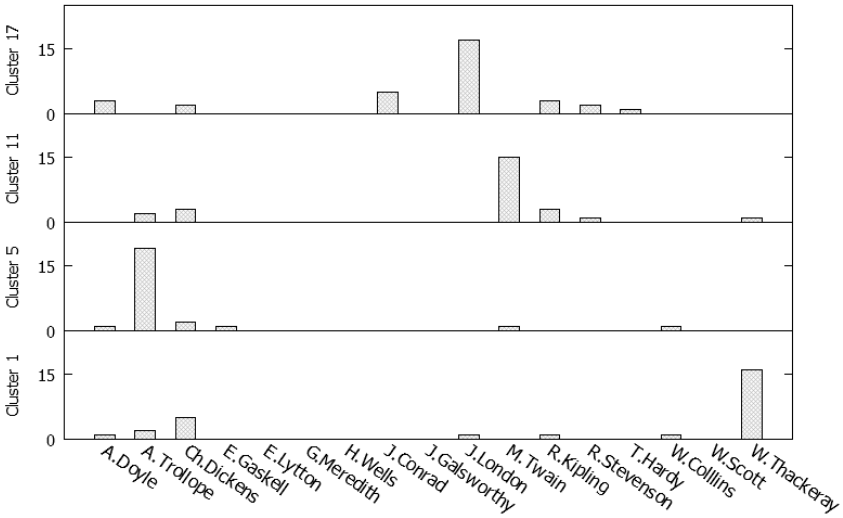

Fig. 10. The distribution of texts in the clusters in orthogonal space, where one author dominates,
using *k*-means clusterization

5. Summary and conclusions

In this paper we investigated the hypothesis of the possibility to differentiate the author's idiolect in the space of semantic fields. In the paper the clustering of text documents in the vector space of semantic fields and in the semantic space with orthogonal basis is considered. The dimension of the vector space basis of semantic fields is significantly lower in comparison with the clustering methods by keywords. The orthogonal semantic basis is formed on the basis of SVD factorization of the matrix of the semantic fields in text documents. Using the vector space model on the basis of semantic fields is efficient in the cluster analysis algorithms of authors' texts in English fiction. The frequency characteristics of the semantic fields were considered as semantic features. The analysis of the distribution of the authors' texts in the cluster structure showed the presence of the areas of semantic space that represent the author's lexicon of the individual authors.

The clustering of authors' texts in the space of semantic fields allows to detect the semantic areas of author's idiolect that are identified by the clusters with dominant text authors. The clusters, where the texts of several authors dominate, can be considered as areas of semantic similarity of the author's style. SVD factorization of the semantic fields matrix makes it possible to reduce significantly the dimension of the semantic space in the cluster analysis of authors' texts. Using the clustering of the text documents in the semantic fields vector space can be efficient in comparative analysis of the author's style and idiolect. The clusters of some authors' idiolects are semantically invariant and do not depend on any changes of the basis of the semantic space and clustering method.

References

1. Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by Latent Semantic Analysis. – Journal of the American Society for Information Science, Vol. **41**, 1990, 391-407.
2. Fellbaum, C. WordNet. An Electronic Lexical Database. Cambridge, MA, MIT Press, 1998.
3. Gliozzo, A., C. Strapparava. Semantic Domains in Computational Linguistics. Springer, 2009.
4. Larsen, B., C. Aone. Fast and Effective Text Mining Using Linear-Time Document Clustering. – In: Proc. of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), ACM, New York, 1999, 16-22.
5. Manning, C. D., P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
6. Pantel, P., P. D. Turney, From Frequency to Meaning: Vector Space Models of Semantics. – Journal of Artificial Intelligence Research, Vol. **37**, 2010, 141-188.
7. Sebastiani, F. Machine Learning in Automated Text Categorization. – ACM Computing Surveys. Vol. **34**, 2002, 1-47.
8. Shehata, S., F. Karray, M. Kamel. Enhancing Text Clustering Using Concept-Based Mining Model. – In: Proc of Data Mining, ICDM'06, 6th International Conference, Waterloo, 2006, 1043-1048.