Self–Organizing Map representation for clustering Wikipedia search results

Julian Szymański

Department of Computer Systems Architecture,
Gdańsk University of Technology, Poland,
julian.szymanski@eti.pg.gda.pl

Abstract. The article presents an approach to automated organization of textual data. The experiments have been performed on selected sub-set of Wikipedia. The Vector Space Model representation based on terms has been used to build groups of similar articles extracted from Kohonen Self-Organizing Maps with DBSCAN clustering. To warrant efficiency of the data processing, we performed linear dimensionality reduction of raw data using Principal Component Analysis. We introduce hierarchical organization of the categorized articles changing the granularity of SOM network. The categorization method has been used in implementation of the system that clusters results of keyword-based search in Polish Wikipedia.

1 Introduction

The amount of information given in the form of documents written in a natural language requires researching methods for effective content retrieval. One way of improving retrieval efficiency is performing documents categorization which organizes documents and allows find relevant content easier.

In the article we present an approach to organization of a set of textual data through an unsupervised machine learning technique. We demonstrate how our method works on test dataset and describe the system that utilizes the method for categorization of the search results retrieved form Wikipedia.

Because of high dimensionality of the processed data (documents represented with terms as their features) we used a statistical method of **P**rincipal **C**omponent **A**nalysis [1]. The method identifies significant relations in the data and combines correlated features into one artificial characteristic which allows to reduce features space significantly. In the reduced feature space we construct Kohonen **S**elf-**O**rganising **M**ap [2] which allows 2D presentation of topological similarity relations between objects (here documents). Employing DBSCAN clustering we extract from the SOM groups of the most similar documents. Changing the SOM granularity we construct hierarchical categories that organize documents set.

2 Text representation

The documents, to be effectively processed by machines, require to be converted from the form readable to humans into a form processable by machines. The main problem is a drawback between text representation used by humans, and that of the machines. Humans, while reading the text, understand its content, and thus he or she is able to know what it is all about. Despite some promising projects, understanding of a text by machines is still unsolved [3], [4]. Because machines don't understand the language they use features for document description which allowes extraction of important relations between processed data. In Artificial Intelligence it is called knowledge representation, and it aims at presenting some aspects of the world in a computable form [5].

In Information Retrieval [6] a typical approach for text representation [7] is usage of Vector Space Model [8] where documents are represented as points in feature space. As features typically words or links are used.

In the experiments presented here we use document content to represent it. This text representation employs words which the article contains. The features set has a size near to the number of all distinctive words which appear in the processed repository of the documents. To reduce size of the set we perform text preprocessing which contains the following procedures:

- stop words removal – all words which appear in so-called stop words list are removed from the features set. This allows us to exclude words which are not very informative in terms of machine processing, and which bring noise to the data.
- stemming – this preprocessing procedure allows to normalize words, through bringing different inflections of the word into its basic form. As a result, different forms of the word are treated as the same term.
- frequency filtering – we remove terms that were related to only one document.

The words preprocessed in this way are called terms. The value, or descriptiveness of a term for a given document may be estimated by the strength w of association between the term and the text. Typically for n-th term and k-th document w value is calculated as a product of two factors: term frequency tf and inverse document frequency idf, given by $w_{k,n} = tf_{k,n} \cdot idf_n$. The term frequency is computed as the number of its occurrences in the document and is divided by the total number of terms in the document. The frequency of a term in a text determines its importance for document content description. If a term appears in the document frequently, it is considered as more important. The inverse document frequency increases the weight of terms that occur in a small number of documents. The idf_n factor describes the importance of the term for distinguishing documents from each other and is defined as $idf_n = \log(k/k_{term(n)})$, where k is the total number of documents, and $k_{term(n)}$ denotes the number of documents that contain term n.

3 The Data

To perform experiments which would validate our approach to organize search results in repositories of documents we find Wikipedia very useful. This large source of human knowledge contains articles referenced one to another and provides also a system of categories. Despite the fact that the Wikipedia category system is not perfect, it can be used as a validation set for algorithms which perform articles organization in an automated way.

Wikipedia off-line data is publicly available for download[1]. The data contain SQL-dumps which provide structural information – linkages between articles and category assignments. There are also available XML dumps which offer textual content of all articles. Importing the data into local database and building the application allowed to extract selected information from XML files and turn it into computationable form. The application we have implemented allows us also to select articles for which the representation will be generated. It allows us to select only a subset of Wikipedia which warrants that the experiments run on single PC will be performed in reasonable time.

Table 1. The data used in the experiments.

Category name	Symbol and color	Number of articles
Biology	magenta □	66
Chemistry	green $+$	229
Mathematic	blue $\cdot$	172
Theology	black $*$	135

The experiments presented here we performed on test set of Wikipedia articles selected using categories. The articles in the set are relatively similar (they all belong to one super category). It enables to show usability of the presented method for introducing organization in documents set that contains elements that are conceptually similar. For test set we selected 602 articles which belong to 4 arbitrarily selected categories from one super category Science. In Table 1 we present categories used in the experiments, and the amount of the articles they contain. The initial features set consists of 37368 features that after preprocessing have been reduced to 12109.

It is comfortable to have a rough view of the data. To see how it is distributed we provide visualization in 2D using principal components computed with PCA (described in section 4.1). What can be seen in Figure 1 the articles from category *biology* described with magenta □ are not separable from other classes. This task can be performed using other components what presents Figure 2 where third and fourth principal components have been used.

4 Documents Organization Method

Having documents represented with text representation, we are able to process them and perform experiments aiming at research methods for organizing them. Our approach we based on categorization and perform it in three steps:

1. dimension reduction,
2. mapping the the articles into Kohonen map,
3. exploiting proximities extracted from the map creation of articles clusters.

[1] http://download.wikimedia.org

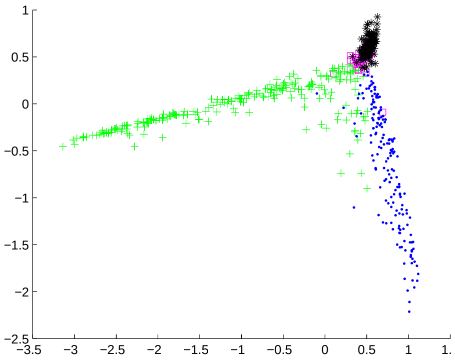


Fig. 1. View of the class distribution of the test dataset performed in 2D, created with two highest principal components.

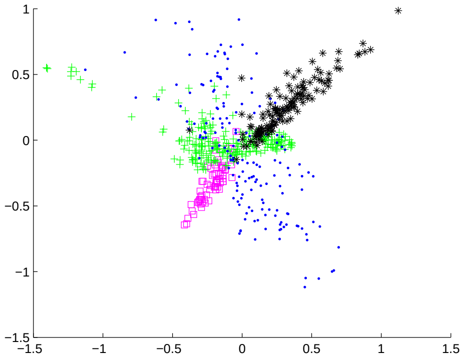


Fig. 2. View of the class distribution of the test dataset performed in 2D, where dimensions are created with third and fourth components.

4.1 Dimension reduction

The raw data we process are high dimensional (test set contains 12109 unique features). Some of the features which are used to describe processed objects are strongly correlated to one another. They can be replaced with artificial characteristic which is the combination of the original ones. One way of performing this a task is statistical method called **P**rincipal **C**omponent **A**nalysis [1]. The idea of the method is based on identification of principal components for the correlation matrix of the features. Selecting the most significant components is performed by computing eigenvectors of the correlation matrix and sorting them according to eigenvalues. A chosen number of eigenvectors which have the highest variance, can be used for representation of the original data. Multiplication of the truncated eigenvectors matrix by original data constructs lower dimensional space for representation of original objects. Selecting the number of eigenvectors used for reduction is crucial to obtain a good approximation of the original data. Very good approximation is to take eigenvectors that complete 99% of the data variance. In Figure 3 we present the % of variances for each of components we also provide information about the number of components whose cumulative sum completes 99% of the variance (136).

4.2 Self - Organizing Maps for topological representation of articles similarity

One of the methods for presenting significant relationships in the data is identification of similar groups of objects and, instead of the objects per se, presentation their representatives – prototypes. In our experiments we we use neural-inspired approach of the **S**elf-**O**rganizing **M**ap introduced by Kohonen [9]. The method is based on an artificial neural network which is trained in a competitive process, also called vector quantization [10]. The learning process uses the strategy **W**inner **T**akes **A**ll (in some algorithm versions **M**ost) which updates the weights of the neuron which is the most similar to the object used to activate the network. The neurons with strongest activations for the

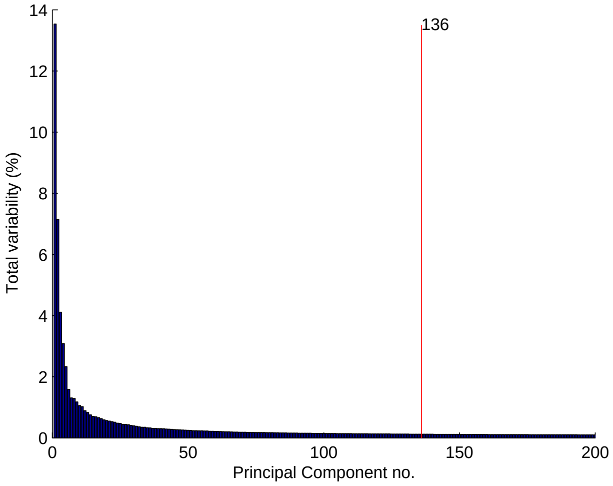


Fig. 3. The % of variability for succeeding principal components

objects that belong to the same class form prototypes [11] which can be used for representation of the particular set of objects.

The effect achieved after training the neural network, is functionally equal to nonlinear scaling from the n-dimensional objects to the smaller, here 2-dimensional space of their prototypes [12]. The advantage of the SOM method is its ability of graphical presentation. The results are visualized in 2D called maps where prototypes of the objects are presented. They keep topological distances according to the given object similarity measure which has been used during training of the neural network.

The SOM-based approach is known to be successfully applied in text processing [13] and it also found applications in web pages organization [14]. In this approach, the information retrieval process is accomplished with presentation of similarity of documents on the Kohonens map which aims at improving the searching process based on their proximity. The data presented in Table 1 and reduced with 136 highest principal components have been used to construct SOM. It presents topological relations between documents projected on 2D space where they are represented by their neural prototypes. SOM presented in (Figure 4) shows firing areas of the network while it was activated with articles that belong to different categories. We also provide joint SOM activation (figure 5) where areas of the network that overlaps have been marked with red.

4.3 Clustering the data

The neurons of SOM may be interpreted as prototypes for articles. While the network is activated by articles that belong to different categories it responds with firing neurons from different areas of SOM. If articles belong to same category they activate close SOM areas. This fact allows to capture proximities of the articles on higher level of abstraction that arises form the fact the comparison is performed in low dimensional space (2 dimensions of SOM). The proximities can be computed calculating average

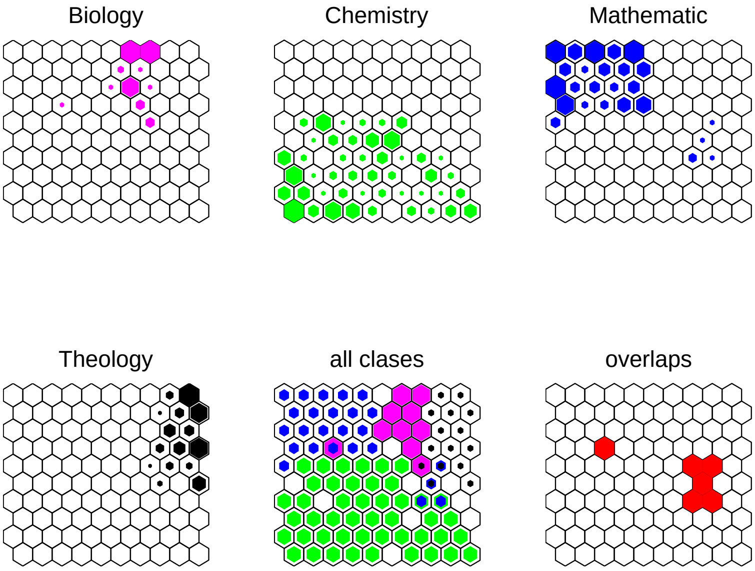


Fig. 4. Sample dataset presented on Self-Organized Map. Activations of different SOM areas for articles from different categories and their overlaps.

distances between neurons activated for each pair of the articles. It allows to construct articles similarity matrix where elements are calculated using formula 1.

$$sim(a_1, a_2) = \frac{1}{d(X,Y)} = \frac{1}{|X|}\sum_{i=1}^{|X|} d(x_i, Y) \tag{1}$$

where X and Y denote sets of neurons on SOM activated respectevely for article a_1 and a_2 and $d(x_i, Y)$ is calculated using formula 2.

$$d(x_i, Y) = \frac{1}{|Y|}\sum_{j=1}^{|Y|} \sqrt{(x_{i1} - y_{j1})(x_{i2} - y_{j2})} \tag{2}$$

The articles proximity matrix allows us to extract groups of the most similar articles. There are many methods and strategies to perform such a task [15]. We used here density-based approach that is known effective non parametric clustering technique suitable for textual data[16]. *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) [17] is a clustering algorithm based on densities of points in feature space. Its advantage is not very sensitive to noise and also it is able to find not only convex clusters, which is big limitation of typical clustering algorithms. The main idea of the algorithm is a concept of point neighborhood given by the radius ϵ (that is algorithm parameter) that must contain fixed, minimal number of other points (τ) belonging to the same cluster. The shape of neighborhood depends on proximity function. Eg.: for Manhattan distance it is rectangle. In our approach usage of formula 1 allows to build clusters of any shape. In DBSCAN there are three types of points: root (inside cluster), border and outlayers. Changing the parameters ϵ and τ we can minimize number of outlayers and thus tune the algorithm. Border points are interpreted as articles that belong to more than one cluster and thus multi-categorisation is introduced, which is closer to the real-word categorization, performed by humans.

The clustering quality Q have been evaluated comparing cardinality of clusters C that has been computed and articles categories K created by humans. For each category K_j the quality Q_j is calculated using the formula 3.

$$Q_j = \frac{1}{|K|} \sum_{j=1}^{|K|} \frac{|a| \in C_{max}}{|a| \in K_j} \tag{3}$$

where $|a| \in C_{max}$ denotes number of articles that belong to cluster with highest cardinality and $|a| \in K_j$ denotes cardinality of K_j category.

For the categories from the test dataset (described in table 1) we obtain qualities shown in table 2.

Table 2. Clustering qualities for each of the category from sample dataset

Quality \ Category name	Biology	Chemistry	Mathematic	Theology
Q_j	0.71	0.82	0.84	0.66

4.4 Hierarchical organization

Hierarchy is one of the most well-known ways for organization of large number of objects. It can also be built for a set of the documents using SOM [18], [19]. This task can be done changing the size of the SOM, which introduces different granularity of data organization. The organization is based on hierarchically layered prototypes that represent SOM neurons. The Figures 6 and 5 show bottom-up process of changing the size of the SOM which transforms articles to their more general prototypes. If we take prototypes that bind together sets of the articles, this process can be seen as generalizing the articles into more abstract categories. Overlaps between neuron activations induce the ability to introduce new relations between categories, as well as it show some other possible directions in which the categorization can be performed further.

5 Application and Future Directions

In the article we present the approach for documents categorization based on terms VSM used for representation of Wikipedia articles. We perform linear dimensionality reduction based on PCA. It allows to build Self-Organizing Map in effective way. Changing the SOM size we introduce hierarchical organization of the documents set. Using SOM representation for DBSCAN clustering we identify clusters of the most

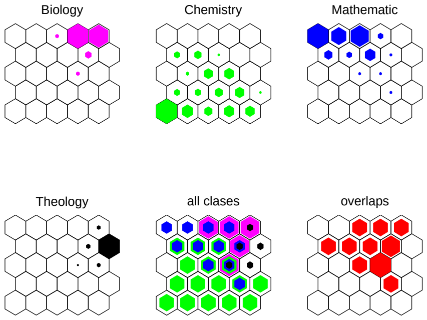


Fig. 5. SOM 5 $times$ 5, for creating hierarchical categories

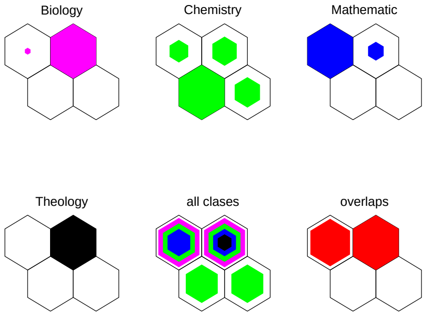


Fig. 6. SOM 2 $times$ 2, for creating hierarchical categories

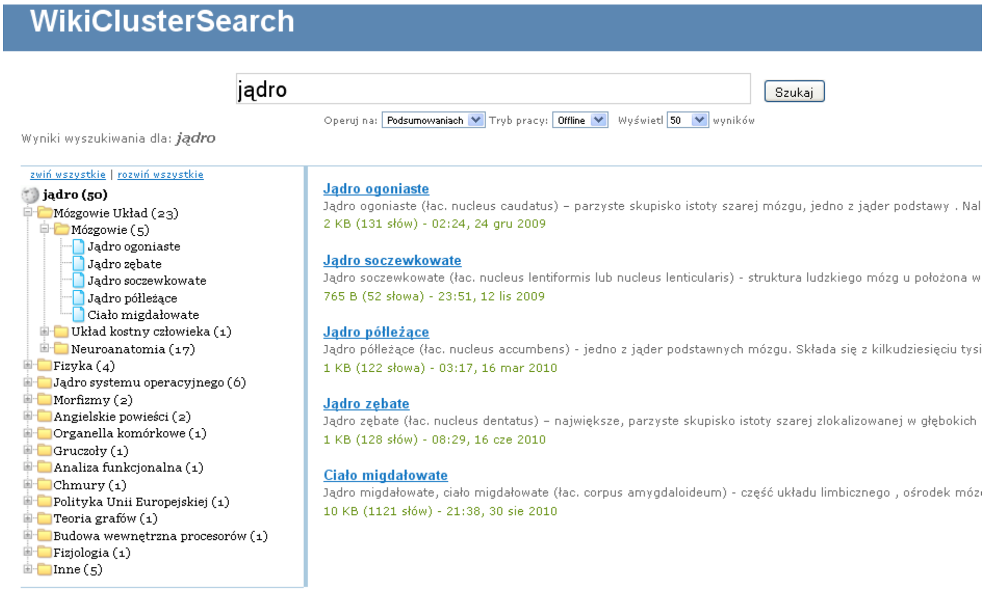
similar articles. We present how our method works for arbitrary selected categories of articles and how it allows to separate them.

We implemented the method presented here in the form of the system that provides clusters for keyword-based search within Polish Wikipedia. The prototype of the system is accessible on-line under url http://swn.eti.pg.gda.pl/UniversalSearch. The screenshot of the application have been presented in Figure 7. The system using the method forms in the fly clusters for the articles that has been selected from Wikipedia using keyword search. In the in Figure 7 we present sample clusters formed for the results returned from Wikipedia for Polish word *jądro* (kernel). In future we plan to implement searching based on clustering for English Wikipedia.

The application shows that forming clusters for Wikipedia pages is useful for organizing the search results. In future we plan to evaluate the clustering results according to the human judgments. Introducing human factor makes this task hard because it requires reviewing the search results manually and for each cluster decide whether the articles are related to proper cluster correctly or not.

Experiments shown here were performed on a limited set of articles. We plan to perform clustering computations on the whole Wikipedia, to introduce for this source of knowledge new, automated category system. It requires us to take into account some additional issues related to efficiency and requires reimplementation of algorithms to be run on clusters instead of single PC. Create machine-made system of the Wikipedia categories for articles will allow to improve the existing one through finding missing and wrong assignments. Application of this method is also possible for non-categorized documents repository. It allows users to find information using similarity and associations between textual data which is a different approach to the paradigm based on keyword-search.

We plan to research other methods of text representations. We will examine approach to the representation of documents based on algorithmic information [20]. In this approach the similarity between two articles is based on information complexity and is calculated from the size differences of the compressed files [21]. We also plan to research representations based on text semantics. The main idea is to map articles into a proper place of the Semantic Network and then calculate distances between them. As



jądro

Operuj na: Podsumowaniach Tryb pracy: Offline Wyświetl 50 wyników

Wyniki wyszukiwania dla: *jądro*

jądro (50)
- Mózgowie Układ (23)
  - Mózgowie (5)
    - Jądro ogoniaste
    - Jądro zębate
    - Jądro soczewkowate
    - Jądro półleżące
    - Ciało migdałowate
  - Układ kostny człowieka (1)
  - Neuroanatomia (17)
- Fizyka (4)
- Jądro systemu operacyjnego (6)
- Morfizmy (2)
- Angielskie powieści (2)
- Organella komórkowe (1)
- Gruczoły (1)
- Analiza funkcjonalna (1)
- Chmury (1)
- Polityka Unii Europejskiej (1)
- Teoria grafów (1)
- Budowa wewnętrzna procesorów (1)
- Fizjologia (1)
- Inne (5)

Jądro ogoniaste
Jądro ogoniaste (łac. nucleus caudatus) – parzyste skupisko istoty szarej mózgu, jedno z jąder podstawy . Nal
2 KB (131 słów) - 02:24, 24 gru 2009

Jądro soczewkowate
Jądro soczewkowate (łac. nucleus lentiformis lub nucleus lenticularis) - struktura ludzkiego mózg u położona w
765 B (52 słowa) - 23:51, 12 lis 2009

Jądro półleżące
Jądro półleżące (łac. nucleus accumbens) - jedno z jąder podstawnych mózgu. Składa się z kilkudziesięciu tysi
1 KB (122 słowa) - 03:17, 16 mar 2010

Jądro zębate
Jądro zębate (łac. nucleus dentatus) – największe, parzyste skupisko istoty szarej zlokalizowanej w głębokich
1 KB (128 słów) - 08:29, 16 cze 2010

Ciało migdałowate
Jądro migdałowate, ciało migdałowate (łac. corpus amygdaloideum) - część układu limbicznego , ośrodek móz
10 KB (1121 słów) - 21:38, 30 sie 2010

Fig. 7. User interface of the application for clustering Wikipedia search results

the Semantic Network we plan to use WordNet dictionary [22]. We will use word disambiguation techniques [23] that allow to map words to their proper synsets to perform such a mappings. We made some research in this direction and the first results are very promising [24].

ACKNOWLEDGEMENTS

The work has been supported by the Polish Ministry of Science and Higher Education under research grant N519 432 338.

References

1. Jolliffe, I.: Principal component analysis. Springer verlag (2002)
2. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing **21** (1998) 19–30
3. Hayes, P., Carbonell, J.: Natural Language Understanding. Encyclopedia of Artificial Intelligence, Wiley Interscience Publication, John Wiley and Sons, New York (1987)
4. Allen, J.: Natural language understanding. Benjamin-Cummings Publishing Co., Inc. Redwood City, CA, USA (1995)
5. Russell, S., Norvig, P., Canny, J., Malik, J., Edwards, D.: Artificial intelligence: a modern approach. Prentice hall Englewood Cliffs, NJ (1995)
6. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval. Addison-Wesley Reading, MA (1999)

7. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34** (2002) 1–47
8. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
9. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78** (1990) 1464–1480
10. Gersho, A., Gray, R.: Vector quantization and signal compression. Kluwer Academic Pub (1992)
11. Blachnik, M., Duch, W., Wieczorek, T.: Selection of prototype rules: context searching via clustering. (Artificial Intelligence and Soft Computing–ICAISC 2006) 573–582
12. Duch, W., Naud, A.: Multidimensional scaling and Kohonen's self-organizing maps. In: Proceedings of the Second Conference of Neural Networks and their Applications. (Volume 1.) 138–143
13. Merkl, D.: Text classification with self-organizing maps: Some lessons learned. Neurocomputing **21** (1998) 61–77
14. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: Websom – self-organizing maps of document collections. In: Proceedings of WSOM. Volume 97., Citeseer (1997) 4–6
15. Berkhin, P.: A survey of clustering data mining techniques. Grouping Multidimensional Data (2006) 25–71
16. Jian, F.: Web text mining based on DBSCAN clustering algorithm. In: Science Information. Volume 1. (2007)
17. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of 2nd International Conference on Knowledge Discovery and. (1996) 226–231
18. Rauber, A., Merkl, D., Dittenbach, M.: The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks **13** (2002) 1331
19. Koikkalainen, P., Oja, E.: Self-organizing hierarchical feature maps. In: 1990 IJCNN International Joint Conference on Neural Networks, 1990. (1990) 279–284
20. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and its Applications. Springer (3rd ed, 2008)
21. Bennett, C., Li, M., Ma, B.: Chain letters and evolutionary histories. Scientific American **288** (2003) 76–81
22. Miller, G.A., Beckitch, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University Press (1993)
23. Voorhees, E.: Using WordNet to disambiguate word senses for text retrieval. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA (1993) 171–180
24. Szymański, J., Mizgier, A., Szopiński, M., P., L.: Ujednoznacznianie słow przy użyciu słownika WordNet. Wydawnictwo Naukowe PG TI 2008 **18** (2008) 89–195