Lecture Notes in Computer Science 7663, 160-167. Springer, Heidelberg, 2012

Self Organizing Maps for Visualization of Categories

Julian Szymański[1] and Włodzisław Duch[2,3]

[1] Department of Computer Systems Architecture, Gdańsk University of Technology, Poland,
julian.szymanski@eti.pg.gda.pl
[2] Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
[3] School of Computer Engineering, Nanyang Technological University, Singapore
Google: W. Duch

Abstract. Visualization of Wikipedia categories using Self Organizing Maps shows an overview of categories and their relations, helping to narrow down search domains. Selecting particular neurons this approach enables retrieval of conceptually similar categories. Evaluation of neural activations indicates that they form coherent patterns that may be useful for building user interfaces for navigation over category structures.

Keywords: categories visualization, Wikipedia, self organizing maps, documents categorization

1 Introduction

Hierarchical systems of categories allow for effective navigation over large datasets, presenting data at different levels of abstraction. It is one of the most effective approaches to data retrieval. However, tree structures cannot represent complex data in a faithful way. Many alternative structures for organizing data may reveal different aspects inherent in the data. Visualizations presented below for Wikipedia categories considered only one approach to compute similarity, but the same method can be used with specific, structured evaluation of proximities, revealing different aspects of the data.

Categories are initially organized in tree-like structures that form hierarchies, with general abstract concepts in the higher parts of the tree. As some concepts can be related to more than one parent the relations between categories form a graph. Thus its presentation with folder-like approach as it is used in file systems misses many aspects of the data. Additional issues related to the cycles should be also considered here. One way to solve these problems is offered by the Self Organizing Maps [9] that display overall similarities of the processed data. The extensions of the original SOM approach in the form of the WebSOM [6], Growing Hierarchical SOM (GHSOM [12]), DISCERN language understanding architecture [10], have already been used for documents organization. Despite many interesting aspects these methods have not found wider use, suffering from scaling problems and enforcing single static hierarchies of categories. In this paper SOM approach is used to visualize generalizations of document sets provided by categories. This allows for narrowing the number of processed documents and the number of possible inner similarities between them, resulting in clearer presentation of high-level information.

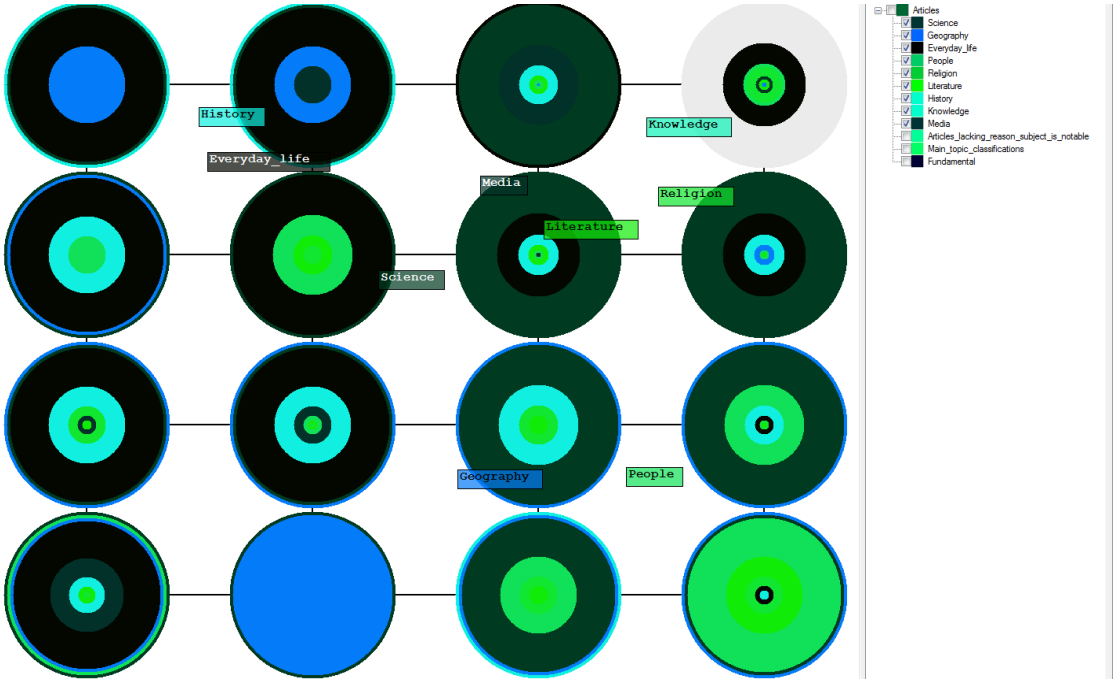

Fig. 1. SOM for top level Wikipedia categories

2 Representation of Categories

Representation of text data should facilitate computations that extract information using machine learning algorithms. This is especially crucial while large texts repositories are processed [13]. Two approaches to object representation use either features (descriptional approach) or inner objects similarities (referential approach) [8]. The referential approach represents each category in relation to all other categories, coding these relations in form of square similarity matrix that records proximities between categories. Embedding similarity matrix in high-dimensional space (equal or lower than the number of categories) allows for view all categories as network of points with distances representing their similarities.

Wikipedia combines articles into categories forming a directed graph. To calculate similarity of categories without falling into the bag of words representation the minimum number of transitions required to get from A node (representing category) of this graph to another node C is stored in a vector $V_A(C) = \#(A \to C)$. The cosine distance (dissimilarity) between these vectors $D(A, B) = \cos(V_A, V_B)$ is then used to evaluate similarities. More sophisticated approaches based on spectral decompositions may be more accurate but for a very large repository like Wikipedia this is a useful first step worth investigation.

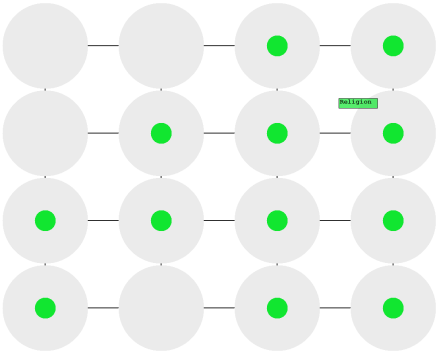

Fig. 2. Activation of the neurons for a single category "religion".

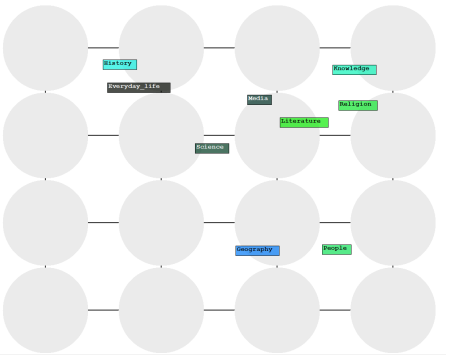

Fig. 3. Category labels displayed for the centroids of neuron activations.

Wikipedia Simple English `xml` dump, available on line[4], has been used in tests described below. First the `html` tags are removed and the information required for referential representation extracted. For that purpose we have created *Matrix'u* application that calculates similarity matrices using various methods of representation. It is available on-line[5] free for academic use.

3 Visualization of Categories with SOM

SOM maps present topographical relations between categories that come from the same conceptual level and may be used directly with the referential approach based on similarity (or distance) matrix. In Figure 1 9 top categories from the highest Wikipedia hierarchy level are shown in SOM using representation described in the previous section. As these categories are quite broad most SOM neurons are activated to some degree by several categories. This activation level is marked using color rings, with size representing activation level and color specific category. For example, category "religion" activates 12 of the 16 neurons in the map.

SOM map based on feature similarities may be easily changed introducing different ways for calculating the distance. Strengthening particular features will lead to different visualization that will show other aspects of the similarity of the categories [3]. With link-based representation we do not have such ambiguity.

Activation of the SOM with selected one category results with map neural activation. An example of single category activation has been presented in Figure 2. In the figure for a sample category *religion* we show activation of the SOM neurons. The strength of the neuron response has been marked with the size of color occupied particular output.

In practice displaying full map of neural activations with many labels will make visualizations illegible, especially when there is a high number of categories. Therefore

[4] http://download.wikimedia.org
[5] http://kask.eti.pg.gda.pl/CompWiki/

only a single label for each category is displayed, as shown in Figure 3. The label is placed in the center of the shape formed by the neural activation for a particular category (as in the example shown in Figure 2). The centroid has been calculated according the formula 1:

$$m_k = \frac{1}{|X_k|} \sum_{i \in X_k} z_i m_i \tag{1}$$

where m_k denotes position of the centroid for k category, X_i is the set of $|X_i|$ neurons at positions m_i with normalized activity $z_i \in [0, 1]$ for this category.

Disabling the SOM neuron activations and displaying only category labels related to the centroids activations is shown in Fig. 3. In comparison to the Fig. 1 this is more clear but here the categories related to the particular neuron are not seen. Using Fig. 1 visualization a particular category or a neuron can be selected. Selecting a category on the map will expand its sub-categories, as shown in Fig. 4 for the category *Religion*. This top-down navigation may be done until lowest level (leaf) categories are reached. Square maps are used a dimensions large enough to have more neurons than categories at a given level.

Selecting particular neuron creates a map with categories related to that neuron. In this way two types of similarities are captured - the first is a hierarchical one, the second captures proximities provided by the specified similarity measure used for building SOM.

The interactive visualization helps to narrow down search results according to the particular interest. The example of such visualization is shown in Fig. 5 where user selects to visualize categories related to neuron [4:2] from the Figure 1.

Selecting neurons instead of labels allows for viewing categories related to this neuron. Rules that lead to the aggregation of categories in a single neural activation may be added and related to the similarity evaluation. Such rules should help to organize categories in particular direction of interest. This may be easier with SOM build with feature-based or concept-based representations, as rules may be easier expressed using concepts, although similarities between categories may also be used in such rules.

4 Evaluation of the visualization consistency

The distribution of the categories over the SOM maps plays a key role in their legibility. Categories should form coherent patterns on the map, activating connected subset of neurons. Positive and negative example of such cases is shown in Fig. 6 and 7. Distribution of neural activity should be proportional to the number of separate connected groups of neurons g_k activated by k-th category, and to the number of weakly active nodes. It is estimated using a dispersion-like heuristic measure:

$$R_k = \tau_1 g_k \frac{\sum_{x \in X_k} (1 - z_{xk})}{|X_k|} + \frac{g_k - 1}{\tau_2} \tag{2}$$

where R_k is the activity dispersion, X_k is the set of neuron activated for this category, and $z_{xk} \in [0, 1]$ is the activation degree of particular neuron x for category k. τ_1 and τ_2 are parameters set after some experimentation to 0.6 and 6.0, respectively.

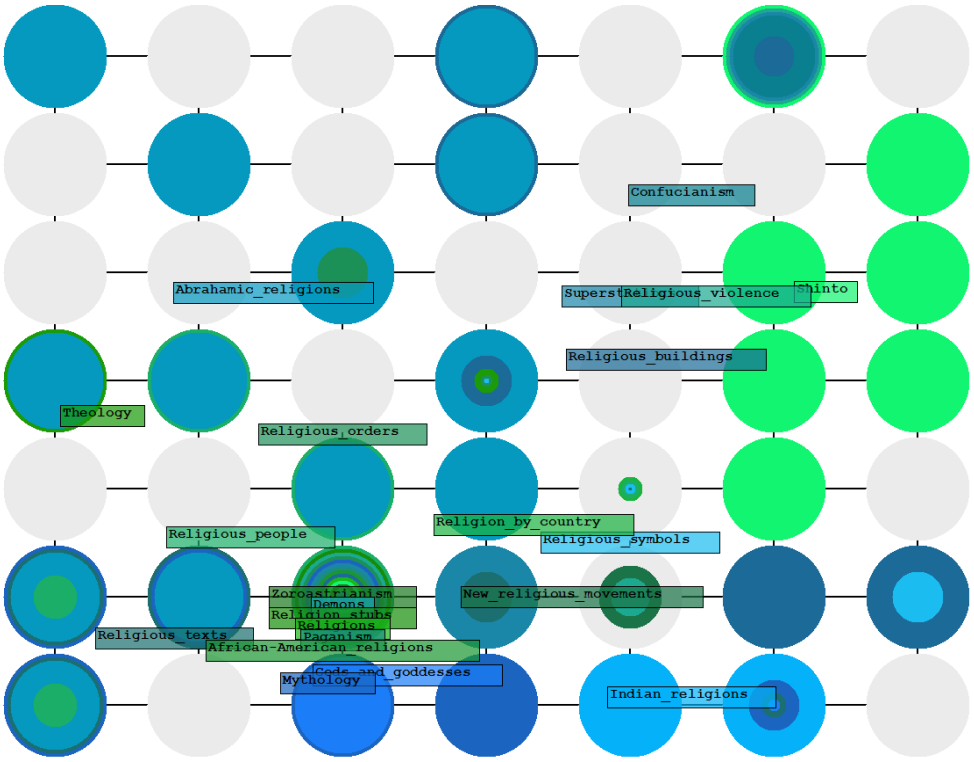

Fig. 4. Activations of subcategories for *Religion*

This measure evaluates the extent to which a particular category occupies the map in terms of neural neighborhood activation. To describe particular map average values for all categories should be computed. The average value of dispersion describes the degree of inner cohesion of categories pattern activation.

Evaluation of dispersion measure for all possible maps that can be generated for Wikipedia categories is almost impossible in practice. Dispersion estimations have been approximated running random walks around the SOM hierarchy. The walk starts with random selection of a neuron from the top SOM level (see Fig. 1), creating a map for categories related to that neuron (see as an example activation given in Fig. 5). For such map an average dispersion measure is calculated and the walk is continued repeating the procedure at the lower level, until leaf categories are reached. Such random walk is then started again from the top.

Performing 100 random walks through the Wikipedia maps of categories the average category tree depth is estimated to be 6.93. In Fig. 8 the average dispersion of SOM maps as a function of the number of categories is displayed. Surprisingly dispersion is almost stable, it does not dependent on the number of categories displayed on the map. Concentration of neural activation within one connected region of SOM linked to the category is desired because it allows to keep visualization clear. The distribution measure (2) or quantization errors [4] may be used as error functions in SOM or fuzzy

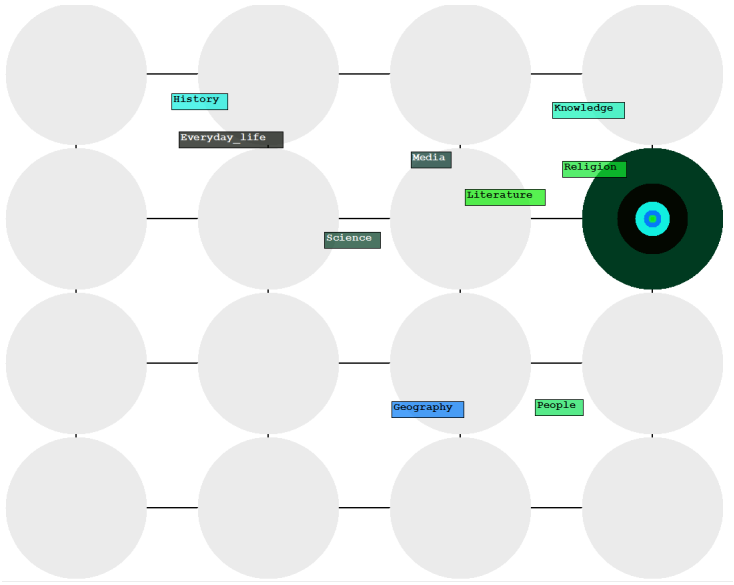

Fig. 5. Activation of categories related to selected neuron [4:2] from map given in Fig. 1

semisupervised approaches [7]. Minimization of functions to achieve maximally connected unsupervised maps should provide legible presentations. Good maps are created when SOM training starts from large neighborhoods with slow reduction of their size, but similarity of text categories has to be expressed in high dimensional space, so one should not expect complete connectivity.

5 Discussion and future directions

Visualization and navigation over Wikipedia categories using SOM approach presented here should capture similarities between categories that can be used for browsing large repositories of texts at various category levels using proximity representation. It will be especially useful when other categories related to the selected one are looked for. As the 'related' can be defined in various ways different similarity measures may be introduced showing different aspects of the data – here mainly various associations between categories. We shall modify similarity measures to strengthen sets of correlated features, facilitating organization of categories along particular conceptual direction.

The dispersion measure calculated for random walks on succeeding maps indicates that the method forms acceptably coherent areas of neuron activations, but training the map using supervised mode, or maximizing some index of maximum connectivity, should allow to build more clear end-user interfaces. Instead of classical SOM one of competitive constructive approach with growin cell structures or neural gas may be used. Usability of the proposed approach will be evaluated using technology acceptance model [15] [1].

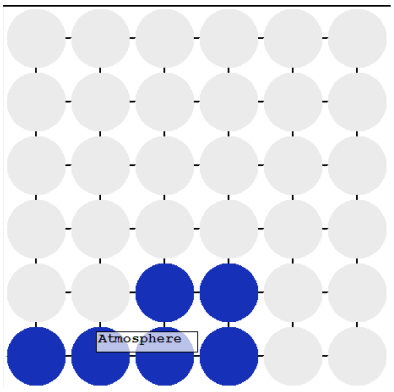

Fig. 6. Example of proper neurons activity having low dispersion

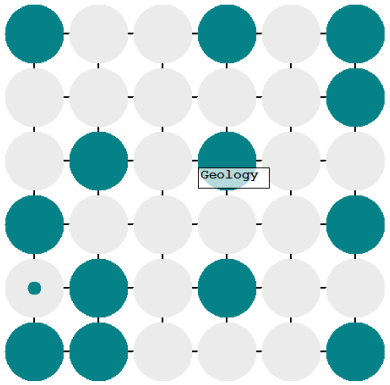

Fig. 7. An example of wrong distribution of neurons activity with high dispersion

The Wikipedia categories for directed graph, therefore non-symmetric hierarchical approach should be introduced [11]. To capture differences between general and specific categories similarity can be estimated using modified formula that promotes stronger associations from detailed categories to more general ones simply by multiplying them with a scaling parameter. Thus the distance from *math* to *algebra* will be larger than from *algebra* to *math*. This simple modification of the proximity should allow to present hierarchical relations on flat SOM maps.

In experiments presented above referential representation based on graph node distances between the categories has been used. Similarity of categories may of course be calculated using many other approaches, for example by estimating similarities between sets of articles related to particular categories. Preliminary research in this direction using bag of words (BOW) combined with links gives promising results [14]. In the domain of medical articles categorization BOW approach has been enhanced with introduction of additional general concepts [5, 2] linked trough various relations. This approach has improved results of clustering considerably so it should also be very useful here.

Acknowledgments: The work has been supported by the Polish Ministry of Science and Higher Education under research grant N N516 432338. Authors would like to thank Tomasz Pilarski for his contribution in the application development.

References

1. Chuttur, M.: Overview of the technology acceptance model: Origins, developments and future directions. Indiana University, Working Papers on Information Systems (2009)
2. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks 21(10), 1500–1510 (2008)
3. Duch, W., Szymański, J.: Semantic web: Asking the right questions. In: Proceedings of the 7 International Conference on Information and Management Sciences. pp. 1–8. California Polytechnic State University Press (2008)

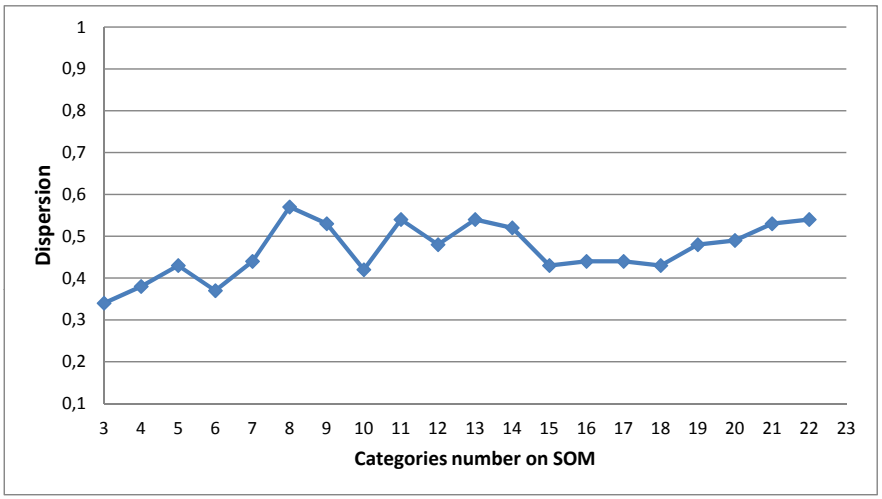

Fig. 8. Average dispersion of categories as a function of categories number

4. Hulle, M.M.V.: Faithful Representations and Topographic Maps: Faithful Representations and Topographic Maps: From Distortion- to Information-based Self-organization. J. Wiley (New York) (2001)
5. Itert, L., Duch, W., Pestian, J.: Influence of a priori knowledge on medical document categorization. In: Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on. pp. 163–170. IEEE (2007)
6. Kaski, S., Honkela, T., Lagus, K., Kohonen, T.: Websom-self-organizing maps of document collections. Neurocomputing 21(1-3), 101–117 (1998)
7. Kästner, M., Villmann, T.: Fuzzy supervised self-organizing map for semi-supervised vector quantization. In: Lecture Notes in Computer Science, Artificial Intelligence and Soft Computing. pp. 256–265. Springer (2012)
8. Kintsch, W.: The representation of meaning in memory. Lawrence Erlbaum (1974)
9. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing 21(1-3), 19–30 (1998)
10. Miikkulainen, R.: Subsymbolic Natural Language Processing: An Integrated Model Of Scripts, Lexicon, And Memory. MIT Press, Cambridge, MA (1993), `http://nn.cs.utexas.edu/?miikkulainen:subsymbolic`
11. Olszewski, D.: An experimental study on asymmetric self-organizing map. Intelligent Data Engineering and Automated Learning-IDEAL 2011 pp. 42–49 (2011)
12. Rauber, A., Merkl, D., Dittenbach, M.: The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks 13, 1331–1341 (2002)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
14. Szymanski, J.: Mining relations between wikipedia categories. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) Networked Digital Technologies - 2nd International Conference, Prague, Czech Republic. Proceedings, Part II. Communications in Computer and Information Science, vol. 88, pp. 248–255. Springer (2010)
15. Venkatesh, V., Davis, F.: A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science pp. 186–204 (2000)