Cartograms of self-organizing maps to explore user-generated content

André Bruggmann, Marco M. Salvini, Sara I. Fabrikant

University of Zurich, Department of Geography – Zurich, Switzerland

Abstract. As the amount of user-generated content dramatically rises, the need to structure these data, to extract relevant semantic relationships buried in the data, and to visualize found relationships appropriately has significantly risen as well.

We suggest an innovative method to structure, visualize, and visually explore user-generated data using a cartogram of a self-organizing map. This distorted self-organizing map overcomes the cognitive limitations of the traditional self-organizing map by combining this neural network mapping approach with cartographic methods used to generate cartograms.

First, our novel mapping approach is put to a rigorous test in a case study aimed to uncover the latent semantic structure from text documents in the Wikipedia Encyclopedia. Second, the latent structure uncovered with the self-organizing map cartogram is systematically evaluated by comparing it to an established network visualization method and output.

The resulting self-organizing map cartogram reveals relevant structures in the considered Wikipedia data. The comparative evaluation confirms the validity and the stability of the found patterns, and therefore of our novel visualization solution.

This paper further contributes to the spatialization research line, by expanding the use of well-established and empirically evaluated cartographic depiction methods to the visualization of non-geographic data, such as, for example, user-generated data increasingly available in today's networked information society.

Keywords: geovisualization, visual analytics, self-organizing maps, cartogram, network visualization

1. Introduction

The availability and the amount of user-generated content on the WWW rise dramatically. Wikipedia and Twitter are two well-known examples of massive text and graphic-based, crowd-sourced, and freely available online databases. In order to systematically analyze and explore such massive semi-structured semantic databases with visual analytics methods, cartographers may contribute with perceptually salient and cognitively adequate mapping solutions, as presented in this paper.

Following the line of research at the interface of cartography and information visualization, Skupin & Fabrikant (2005, 2007) introduced the *spatialization framework* which suggests a systematic approach to transform high-dimensional data sets into lower-dimensional, spatial representations for facilitating data exploration and knowledge production using spatial metaphors. Their approach pays tribute to different traditional and empirically evaluated cartographic design principles, such as the theory of the visual variables (Bertin 1974, MacEachren 1995), and the cartographic generalization process (McMaster 1989, Buttenfield & McMaster 1991), for example, and integrates established dimension reduction techniques from information visualization, such as *self-organizing maps* (SOM) and network visualizations.

A self-organizing map, in essence a neural network, projects input data onto a two dimensional, topological space, typically represented by a regular tessellation (i.e., hexagon) including neurons (Kohonen 2001). The neurons in the SOM have the same attributes as the input data, and are placed near each other if they share similar attributes, and are therefore semantically similar (Skupin & Agarwal 2008). The original input data are mapped as points onto neurons with semantically most similar attributes. However, traditional SOMs have an important conceptual limitation: the proximity metric, defined by data similarity, is not uniform across the SOM space, and thus violates the distance-similarity metaphor, defined by Montello et al. (2003). We present an innovative approach to overcome this limitation which combines SOM with the long-standing cartographic tradition of cartograms. This new approach is an extension of the cartographically inspired spatialization approach presented by Skupin & de Jongh (2005), and Fabrikant & Salvini (2011) at prior ICC meetings. This extended framework is first put to a rigorous test in a case study spatializing more than 2,000 Wikipedia articles. Second, our approach is systematically evaluated in comparison with a well-established network visualization approach using the same data.

2. Methods

The used methods are inspired by previous work of Skupin & de Jongh (2005), and Fabrikant & Salvini (2011) who semantically explored and analyzed the ICC conference proceedings using the SOM and the network visualization techniques. In this study, we analyze a set of Wikipedia articles, and extend the spatialization framework to the neuronal network space in the SOM, applying the cartogram techniques. In doing so, we intend to enhance the cognitive plausibility and the visual saliency of the resulting visualization.

2.1. Data

As this proposed approach is a part of a larger project intended to uncover the functional structure of the regional organization in the Eastern parts of Switzerland, we considered only the German version of Wikipedia. The semantic corpus consists of the titles of 2,158 Wikipedia articles including the standard description of 8,812 related categories.

2.2. Towards the distorted self-organizing map

As a first step, we had to analyze the semantic content captured in the article titles, and the standard category descriptions in Wikipedia. To do so, we employed the *probabilistic topic model* (TM) method as described in Steyvers & Griffiths (2007), available in the *Text Visualization Toolbox* (TVT) in MATLAB (Hespanha & Hespanha 2011).

Following Fabrikant & Salvini (2011) we chose 20 topics for the TM step. As a result we get a two-mode article-topic matrix, where each article is described as a vector associated with probability values for each of the 20 topics. The higher the probability, the higher the semantic similarity between an article and the respective topic is.

This article-topic matrix served as the input for the SOM calculation performed with the SOM Analyst toolbox in ArcGIS, developed by Lacayo-Emery (2011). We chose the SOM parameters proposed by Lacayo-Emery (2011) and Skupin & Esperbé (2011). Following SOM guidelines as suggested by Wendel et al. (2009), we decided to create a SOM of 30x30 neurons in size.

The initial SOM, based on the two-mode article-topic matrix, was trained first with 4,500 runs, using a neighborhood radius of 30, to establish broad, global structures. In a second stage, we trained the SOM with 40,000 runs, and applied a neighborhood radius of 6 to carve out regional and local structures. The trained SOM consists of 20 different *component planes* which represent the distribution of the data values for each of the 20 TM

input vectors. We then project the *best matching unit* (BMU) which consists of all considered Wikipedia articles onto the *component planes*. During this step, each input data point is assigned to the neuron in the *component planes* which fits best with its semantic attributes.

As a final step we calculated the *U-matrix* which contains a semantic similarity value for every neuron compared to all neighboring neurons in the space. An excerpt of the resulting *component planes*, the BMU and the neighborhood similarity is shown in *Figure 1*.

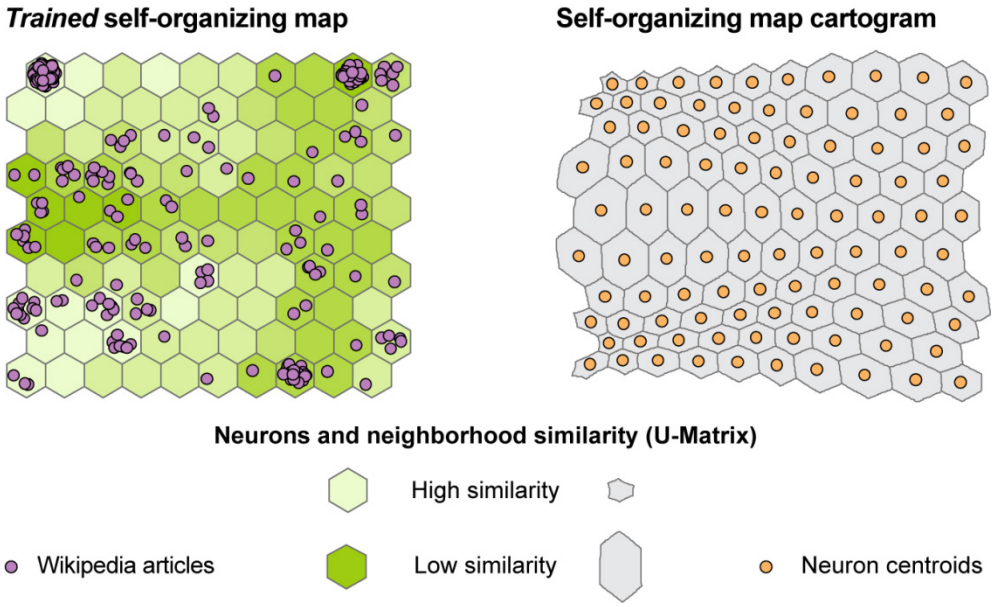

Figure 1. Excerpt of trained SOM including BMU (left) and SOM cartogram (right)

In order to distort the SOM, based on the distance-similarity metaphor, we first transferred the *component planes* in shapefile format to the *Scape Toad* cartogram software (Andrieu et al. 2008). In *Scape Toad* we selected the *U-matrix* values as variable to distort the SOM space. Following the distance-similarity principle neurons that are semantically less similar are pushed apart compared to neurons that are more similar to each other. The resulting cartogram is transferred back to ArcGIS for further processing. As a next step we calculate the center points of the distorted neurons. The re-

sulting SOM cartogram including the center points are shown in *Figure 1* (SOM cartogram).

So far we used size as a visual variable to distort the SOM according to the distance-similarity metaphor. Additionally to the distorted *component planes*, we get a point grid consisting of neuron center points. The distances between the points in the grid represent the semantic similarity between the neighboring neurons, according to the distance-similarity metaphor (Fabrikant et al. 2006), using location as an additional visual variable.

2.3. Thematic clusters

As a next step, we applied a clustering algorithm in order to further generalize our input data, and to explore how clusters of semantically similar articles might be distributed within the SOM cartogram. We therefore transformed the two-mode article-topic matrix (see *Section 2.2*) to a one-mode article-to-article matrix which indicates the semantic similarity between the input articles. We employed the Blondel community detection algorithm (Blondel et al. 2008), an established social network cluster algorithm for this step as suggested by Fabrikant & Salvini (2011). Applying this algorithm, thirteen article clusters emerge. In order to automatically describe the semantic content of the thirteen clusters, we employed the tf-idf method which extracts the most relevant terms of every cluster (Manning et al. 2009) in *Figure 2*. Cluster membership is illustrated in the SOM using the color hue; the number of articles per neuron is visualized by scaling the neuron's center point using the visual variable size.

3. Results

In *Figure 2* the SOM cartogram including the thirteen Blondel clusters and the number of articles per neuron are depicted. The size of the center points in the neurons represents the number of articles which fits best to the corresponding neuron according to the *BMU*. The larger the center point, the higher the number of articles per neuron. Color hue depicts the cluster membership representing the majority of articles within a neuron. White center points depict neurons where none of the thirteen clusters represents more than 50% of the articles at that center point. The three most relevant terms are listed in the legend of *Figure 2*, below a general content description for each cluster.

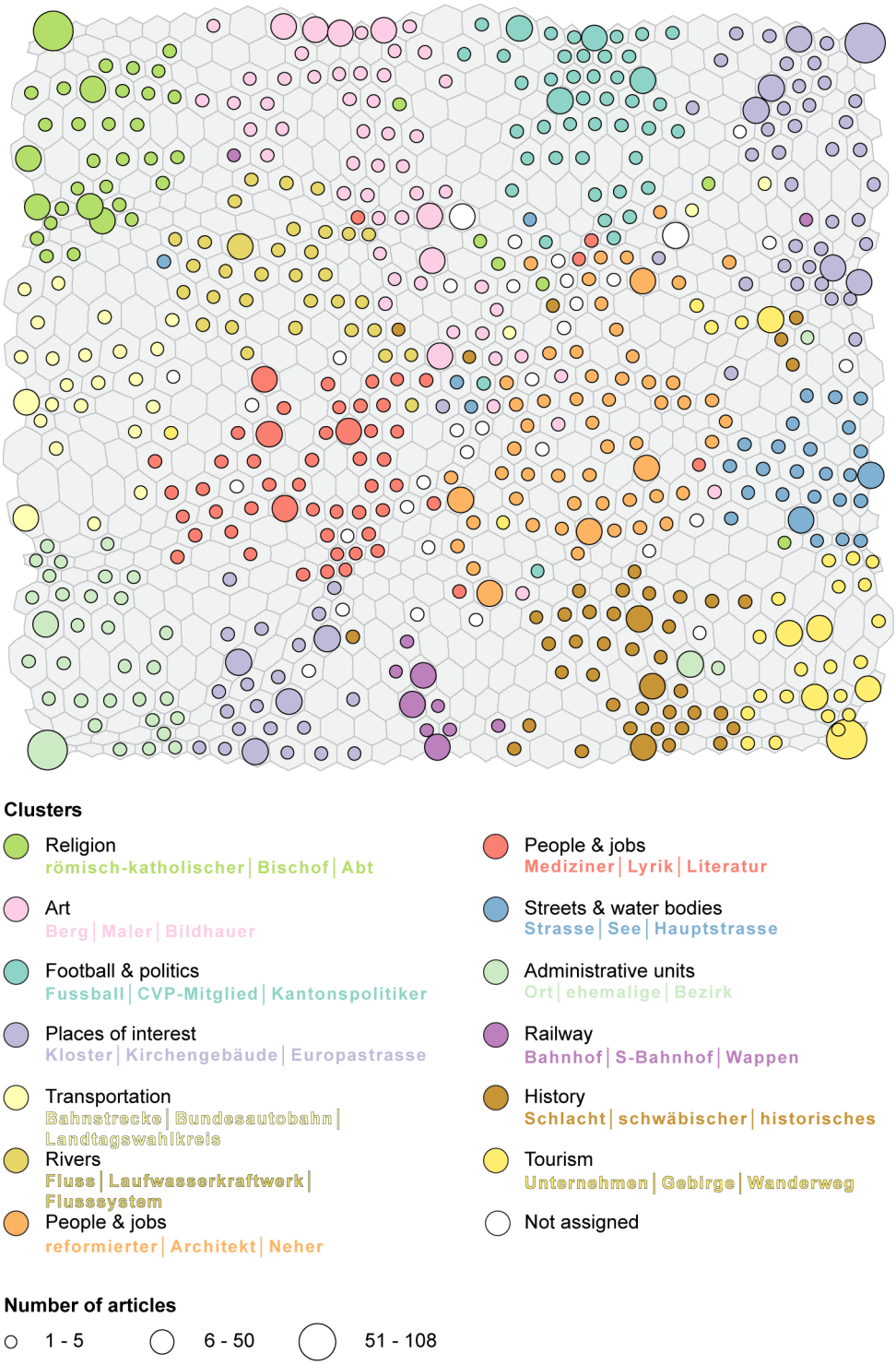

Figure 2. SOM cartogram including cluster membership and number of represented articles

Figure 2 reveals that most clusters are homogeneous regions in the SOM. However, the two *People & jobs* clusters, both located in the center of the SOM, and the clusters *Art* and *Places of interest* are not as homogenous as the other clusters. Articles belonging to the cluster *Places of interest* are even separated into two isolated regions in the SOM. One of these regions is located in the upper-right, and the other in the lower-left corner of the map. It is also obvious that the clusters in the middle of the SOM are semantically more ambiguous, compared to the clusters at the corners or at the edges of the SOM. Due to the cartogram distortion of the SOM, the semantic similarity between neighboring neurons can be identified. For example, the distortions pattern within a yellow cluster called *Transportation* on the left of the SOM is interesting. While in the middle of the yellow cluster the neurons are quite small, and thus articles are semantically similar, the larger neurons at the edge of this cluster highlight lower semantic similarity of a group of articles even if the articles belong to the same cluster. Additionally, the semantic dissimilarity between different clusters can be observed, looking for example at the large neurons on the border between the yellow *Transportation* and the green *Administrative units* clusters in the lower-left corner of the map.

4. Evaluation

We evaluate our approach by comparing the structures and the distribution of the clusters in the SOM cartogram with a network visualization using the same data as input. Second, we quantitatively assess the consistency of the found clusters in the SOM cartogram and in the network visualization. Finally, we compare the distribution and the semantic content of one specific cluster in the two visualizations in more detail.

4.1. Comparison of the uncovered structures

In order to produce a network visualization of our data we input the one-mode article-to-article matrix (see *Section 2.3*) to Network Workbench (NWB Team 2006), following the methods presented in Fabrikant & Salvini (2011) to generalize and visualize a spatialization network display. To cluster the articles we again employ the Blondel community detection algorithm, as presented in *Section 2.3*. The resulting network is depicted in *Figure 3*.

In order to improve the legibility, and perceptual salience of this visualization, we applied empirically validated design principles to the network configuration (Fabrikant et al. 2004). In particular, we identified semantically central nodes, and aggregated less central nodes within a node cluster to its

closest center node using ArcGIS. Then, we visualized the aggregated points as graduated circles, as illustrated in *Figure 3*. The pie charts in *Figure 3* are scaled based on the number of articles that were aggregated. We depict the cluster membership of the aggregated nodes with pie chart segments, whereas the colored segments represent the proportion of articles which belong to a specific cluster. The edges in *Figure 3* represent the structural most salient linkages according to the semantic similarity relationships.

The uncovered latent structure in the network visualization, shown in *Figure 3*, fits with the latent structure depicted in the SOM cartogram (*Figure 2*). A first qualitative comparison of the distribution of the thirteen clusters shows a noteworthy pattern: *Figure 3* illustrates that there is at least one pie chart for every cluster where the proportion is at least 75% for this cluster. The orange cluster labeled *People & jobs* is the only exception. Although this cluster appears in many of the charts, it never reaches such a large proportion. It is also interesting to see that the red *People & jobs* cluster and the pink cluster named *Art* are jointly distributed across the network and in some charts even reach more than 20%. This pattern is also visible in the SOM cartogram in *Figure 2*, as where the clusters *People & jobs* and *Art* are located in the middle of the SOM cartogram thus indicating that they are semantically vague and close to many other clusters in the space.

Looking at the cluster *Places of interest* in *Figure 3* one can notice that it splits into three different pie charts, where the cluster reaches a proportion over 75%. A comparable pattern is visible in the SOM cartogram in *Figure 2*, as *Places of interest* is the only cluster with a high number of clustered articles in two different regions on the map. This could be an evidence for the semantic diversity of this cluster.

In comparison to the network visualization, the SOM cartogram provides a more nuanced picture about the document similarity. The SOM cartogram allows not only to recognize the semantic similarity between different clusters, but also the structure within clusters. On the other hand, because of the higher degree of generalization in the network visualization, the similarity between articles within the graduated pie chart is not easily accessible.

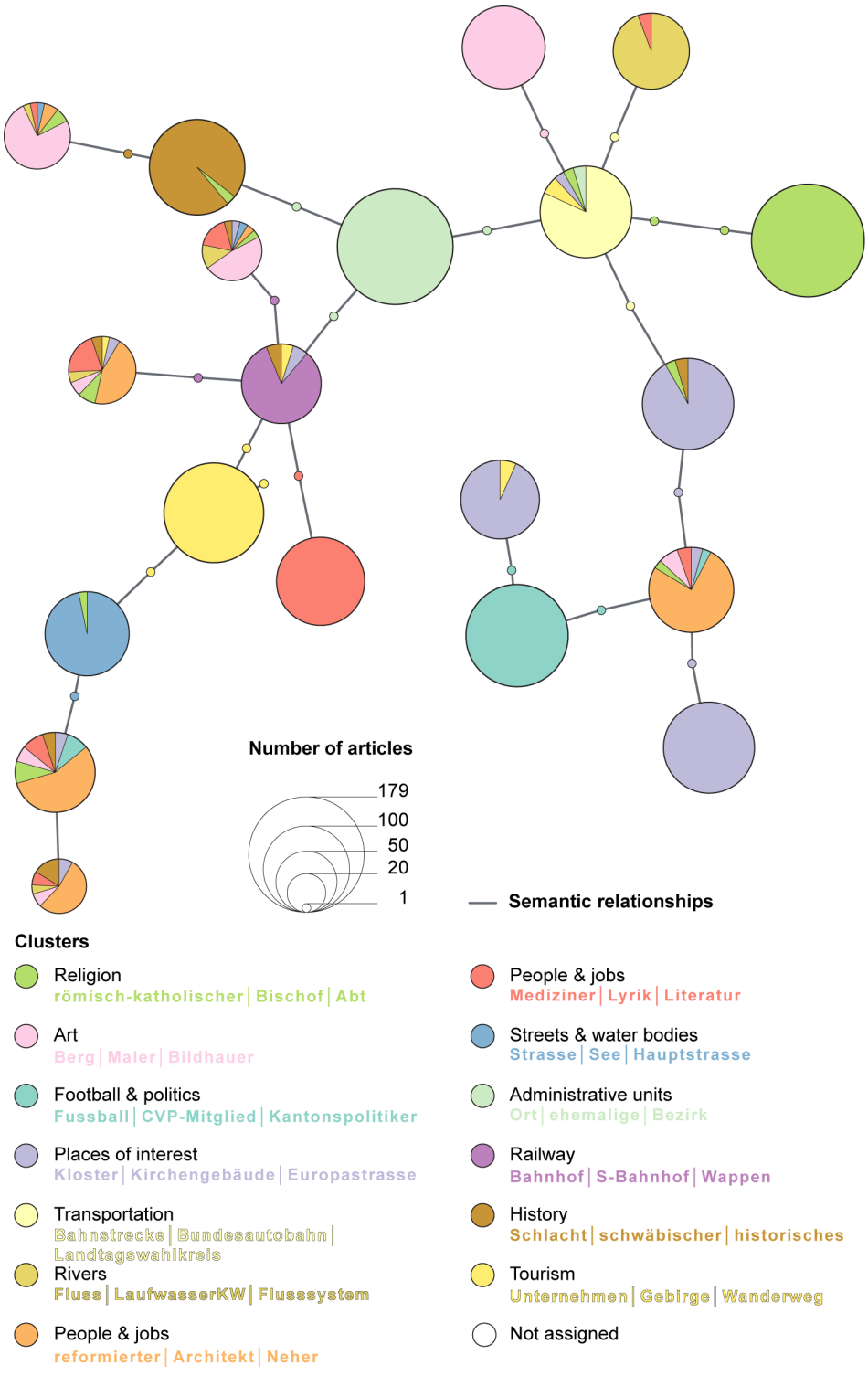

Figure 3. Network spatialization including cluster memberships and the number of articles in each cluster

To quantitatively assess the consistency of the uncovered latent structures in the SOM cartogram, we compare it statistically with the network visualization. In particular, we are interested to identify whether articles grouped in the same graduated pie charts in *Figure 3* would also be grouped into the same regions in the SOM cartogram. We employed the k-means-algorithm to identify 20 regions in the SOM, as there are also 20 graduated pie charts in the network visualization. The small charts that appear between the large graduated pie charts in *Figure 3* were ignored for this comparison, as they only consist of one article. As a next step, we calculated a square matrix including the co-occurrence of articles in the k-means-regions of the SOM and of the graduated pie charts in the network visualization. This co-occurrence matrix provides the basis to quantitatively assess how well the uncovered latent structures match in the two considered visualizations. The consistency between the two latent structures is defined with the Cohen's Kappa coefficient (Cohen 1960), and the hypergeometric test (Kos & Psenicka 2000).

Applying Cohen's Kappa coefficient we get a value of 0.91 which means that the graduated pie charts in the network visualization and in the k-means-regions of the SOM cartogram are very consistent. Employing the hypergeometric test to our data, a probability value of 0.00 indicates that the network visualization reproduces the clusters in the SOM cartogram very well and statistically significant, at a significance level of 5%.

4.2. Distribution of semantic content within clusters

In the previous sections we observed that the articles of the *Places of interest* cluster in both visualizations split into different graduated pie charts in the network visualization, and are distributed across different regions in the SOM cartogram. We therefore expect to see different semantic content in this cluster. For this reason we further analyzed the semantic content of this specific cluster.

First, we selected the k-means-regions as described in *Section 4.1* in the SOM cartogram, containing at least five articles from the cluster *Places of interest*. The result is depicted in *Figure 4*. The dark grey boundaries indicate the borders of the selected k-means-regions. These regions are labeled with the most relevant words appearing in the respective cluster for each of the analyzed k-means-regions, extracted by the tf-idf method.

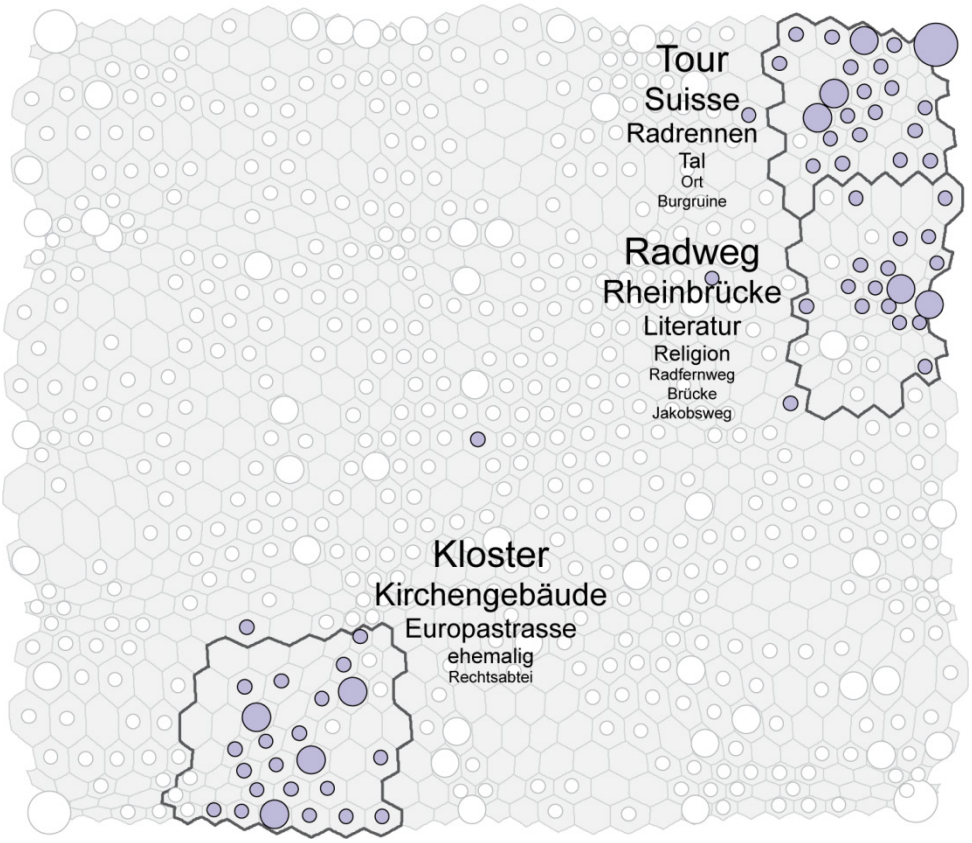

Figure 4. Places of interest cluster in the SOM cartogram

Figure 5 illustrates the same network as in *Figure 3* but only with the pie chart segments of the *Places of interest* cluster highlighted. Again we label the charts that contain at least five articles belonging to the cluster *Places of interest*, with the most relevant terms.

Comparing *Figure 4* and *Figure 5*, the cluster *Places of interest* splits into three groups (i.e., k-means-regions, and graduated pie charts) with different semantic content. Still, the groups in the two visualizations have a high semantic correspondence. In both solutions, one group includes articles about religious buildings, and the most relevant terms are *Kloster* (abbey) and *Kirche/-ngebäude* (church / buildings). The second group is about bicycle paths and religion, with the most important words *Radweg* (bicycle path), *Rheinbrücke* (bridge over the Rhine), and *Religion* (religion). The third group is about the Tour de Suisse, a well-known bicycle race in Switzerland, with the most important words *Tour* (race), *Suisse* (Switzerland), and *Radrennen* (bicycle race).

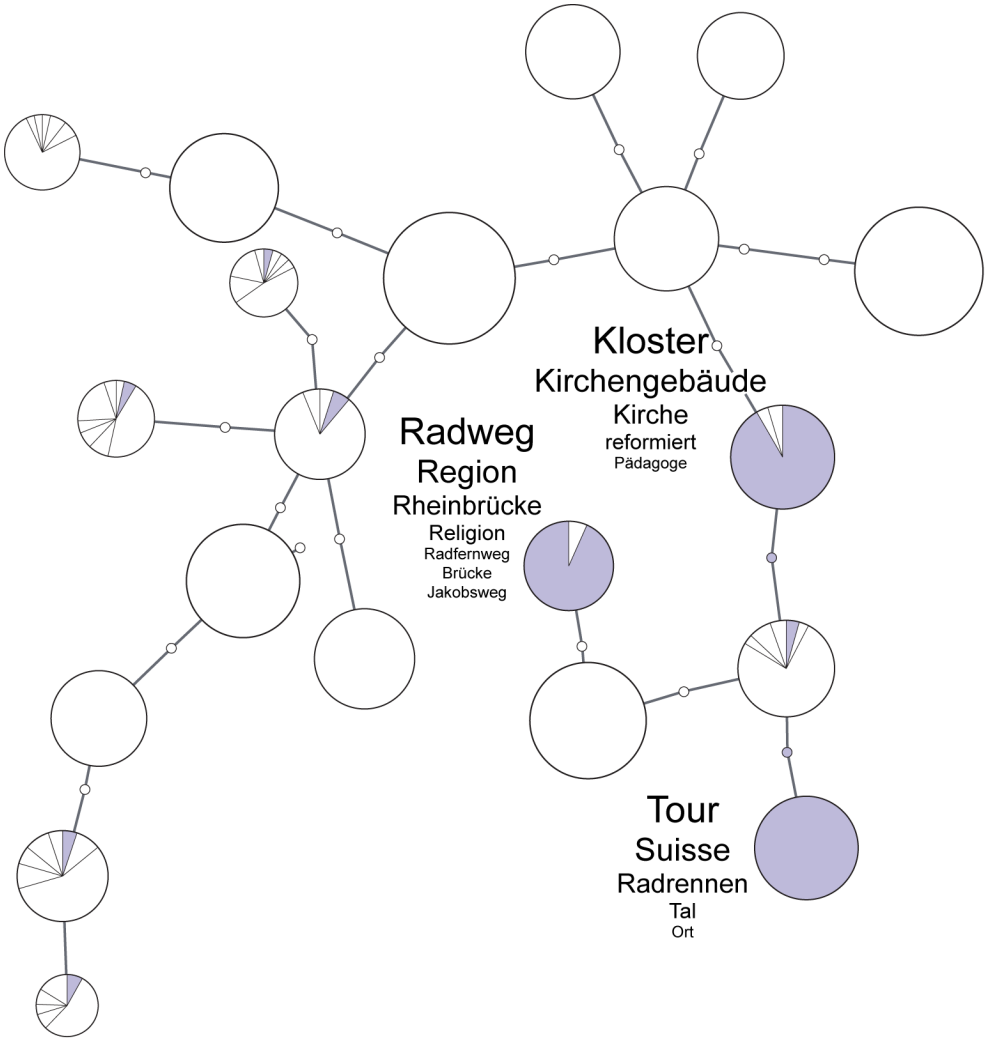

Figure 5. Places of interest cluster in the network visualization

This systematic comparison suggests that not only the global uncovered latent structure corresponds well in the two visualizations, but also the semantic content of the found local structures are similar in the two visualizations.

5. Conclusion

In this paper, we first proposed an innovative approach to combine self-organizing maps with the long-standing cartographic tradition of cartograms when analyzing user-generated content. Second, we analyzed and systematically evaluated the uncovered latent structure in the self-organizing map cartogram by comparing it to a well-established network visualization method and output.

The main advantage of distorting SOMs compared to traditional SOMs is that internal structures of clusters of similar documents and the direct semantic relationships between found clusters can be identified more easily as the distortion pays tribute to the distance-similarity metaphor (Fabrikant et al. 2006). The visual variable location is ranked by Bertin (1967) as the most salient of the visual variables. It expresses the similarity relationships amongst the Wikipedia articles cognitively more plausibly and perceptually more saliently as other types of spatializations, for example, as compared to the network visualization, discussed in this paper.

By comparing the uncovered latent structure in the novel SOM cartogram with an already well-established network visualization approach, we were able to statistically assess the correspondence of the depicted patterns. This systematic evaluation provides validity to our visualization solutions, and more generally for using self-organizing maps and network visualizations as spatialization methods for the scientific investigation of unstructured text data.

References

Andrieu, D, Kaiser, C & Ourednik, A (2008) Scape Toad 1.1.

Bertin, J (1967) Sémiologie Graphique: Les Diagrammes - les Réseaux - les Cartes, Paris: Mouton.

Bertin, J (1974) Graphische Semiologie. Diagramme, Netze, Karten, Berlin, New York.

Blondel, V D, Guillaume, J-L, Lambiotte, R & Lefebvre, E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment P10008.

Buttenfield, B P & McMaster, R B (1991) Map generalization. Making rules for knowledge representation.

Cohen, J (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20 37-46.

Fabrikant, S I, Montello, D, Ruocco, M & Middleton, R (2004) The distance-similarity metaphor in network display spatializations. Cartography and Geographic Information Science 31 237-52.

Fabrikant, S I, Montello, D R & Mark, D M (2006) The distance similarity metaphor in region-display spatialization. IEEE Computer Graphics and Applications 26 34-44.

Fabrikant, S I & Salvini, M M (2011) Charting the ICA World of Cartography 1999-2009. 25th International Cartographic Conference. International Cartographic Association, Paris, France.

Fabrikant, S I & Skupin, A (2005) Cognitively plausible information visualization. in Dykes, J, MacEachren, A M & Kraak, M J eds Exploring Geovisualization. Amsterdam.

Hespanha, S R & Hespanha, J (2011) Text Visualization Toolbox - a MATLAB toolbox to visualize large corpus of documents. http://www.ece.ucsb.edu/~hespanha (accessed March 2013).

Kohonen, T (2001) Self-organizing maps Springer, Berlin.

Kos, A J & Psenicka, C (2000) Measuring cluster similarity across methods. Psychological Reports 86 858-62.

Lacayo-Emery, M (2011) An integrated toolset for exploration of spatiotemporal data using self-organizing maps and GIS. Departement of geography. San Diego State University, San Diego.

MacEachren, A M (1995) How maps work. Representation, visualization, and design, New York.

Manning, C D, Raghavan, P & Schütze, H (2009) Scoring, term weighting and the vector space model. Introduction to Information Retrieval - Online edition. Cambridge University Press, Cambridge, England. http://nlp.stanford.edu/IR-book/pdf/06vect.pdf (accessed March 2013).

McMaster, M B (1989) Introduction to "Numerical Generalization in Cartography". Cartographica 26 1-6.

Montello, D R, Fabrikant, S I, Ruocco, M & Middleton, R S (2003) Testing the First Law of Cognitive Geography on Point-Display Spatializations. in Kuhn, W, Worboys, M F & Timpf, S eds Conference on Spatial Information Theory (COSIT '03), Lecture Notes in Computer Science 2825. Springer Verlag, Berlin, Germany 316-31.

NWB Team (2006) Network Workbench Tool 1.0.0. Indiana University, Northeastern University and University of Michigan, http://nwb.slis.indiana.edu (accessed March 2013).

Skupin, A & Agarwal, P (2008) Introduction. What is a Self-Organizing Map? in Agarwal, P & Skupin, A eds Self-Organising Maps: Application in Geographic Information Science. Chicester.

Skupin, A & de Jongh, C (2005) Visualizing the ICA – A Content-based Approach. Proceedings of the 22nd International Cartographic Conference. La Coruña, Spain.

Skupin, A & Esperbé, A (2011) An alternative map of the United States based on an n-dimensional model of geographic space. Journal of Visual Languages & Computing 22 290-304.

Skupin, A & Fabrikant, S I (2007) Spatialization. in Wilson, J P & Fotheringham, A S eds The Handbook of Geographic Information Science. Reston.

Steyvers, M & Griffiths, T (2007) Probabilistic Topic Models. in Landauer, T K, McNamara, D S, Dennis, S & Kintsch, W eds Handbook of Latent Semantic Analysis. Erlbaum, Lawrence.

Wendel, J, Buttenfield, B P & Smith, J (2009) Spatializing a GIS Software Toolbox. Kartographische Nachrichten 5 257-63.