This is a preprint of an article published in
Kirt, T., Liiv, I., & Vainik, E. (2007). Self-organizing map, matrix reordering and multidimensional scaling as alternative and complementary methods in a semantic study. In H. R. Arabnia, M. Q Yang, & J. Y. Yang (Eds.), Proceedings of the 2007 international conference on artificial intelligence (Vol 1., pp. 385–390), Las Vegas: CSREA Press.

Self-Organizing Map, Matrix Reordering and Multidimensional Scaling as Alternative and Complementary Methods in a Semantic Study

Toomas Kirt
Institute of Cybernetics at
Tallinn University of Technology
21 Akadeemia Rd.
12618 Tallinn, Estonia
tel. +372 620 4202
fax +372 620 4151
e-mail Toomas.Kirt@mail.ee

Innar Liiv
Tallinn University of Technology
15 Raja Str.
12618 Tallinn, Estonia
tel +372 620 2306
fax +372 620 2305
e-mail innar.liiv@ttu.ee

Ene Vainik
Institute of the Estonian Language
6 Roosikrantsi Str.
10119 Tallinn, Estonia
tel./fax +372 641 1443
e-mail

Abstract - *In this paper we compare the three methods as the Self-Organizing Map (SOM), Matrix Reordering (MR) and Multidimensional Scaling (MDS) to analyze and visualize multidimensional data. As a case study the survey of Estonian emotion concepts is analyzed by the three methods and the results are compared. We assess what are the common elements of the results of the analyses, how the methods complement each other and how the knowledge gained by different methods help us to interpret the results.*

Keywords: knowledge acquisition, neural networks, knowledge representation, natural language.

1 Introduction

To reveal the hidden structure of multidimensional data methods of data analysis are used. In this paper we compare three methods as the Self-Organizing Map (SOM), Matrix Reordering (MR) and Multidimensional Scaling (MDS) to analyze and visualize multidimensional data. Our aim is to bring out peculiarities and common elements of the results of the three methods and to assess how acquired knowledge could improve interpretation of the results separately. We have used a survey of Estonian emotion concepts as a case study.

The SOM and MDS are projection methods to cluster and to reduce dimensionality of multidimensional data. Some of the researchers have compared the SOM and MDS earlier and outlined both their similarities and dissimilarities [2], [5]. The two methods are quite similar in general in respect that both methods tend to reduce dimensionality of observed data and reveal its hidden structure but they differ in the strategy applied to the data. The SOM tries to preserve local neighborhood relations and the MDS the interpoint distances between sample vectors [5]. Both methods project data into a two-dimensional space and the SOM also uses color-coding to show the distances between the samples.

The MR method belongs to different class of data analysis methods – the goal is not dimensionality reduction, but reorganization of the initial data until relationships which lie within it can be perceived. The MR method applied in this paper reveals hidden clusters and orders the data matrix according to untypical and typical scale. The result can be regarded as projection into one-dimensional space.

In the first part of the paper three methods are introduced. In the second part the survey of Estonian emotion concepts is used as an example to demonstrate the similarities and differences between the methods.

2 The Self-Organizing Map

The self-organizing map [1] is an artificial neural network that uses an unsupervised learning algorithm – it means there is no prior knowledge how input and output are connected. The SOM is used project multidimensional data into a lower-dimensional space to visualize the hidden structure of the data. The output is a two-dimensional grid, named also as a map. The SOM is widely used in several areas of data analysis [6].

To describe how the process for creating the self-organizing map works let assume, that we have a set of input vectors x as samples. It is called also as an input space. The output of the self-organizing map is a grid of vectors m_i that that are with the same length as the input vector. At the beginning of the learning process all the output vectors are initialized randomly.

The algorithm of the SOM has two main basic steps that are repeated a number of times. First a random sample vector x(t) is chosen from the input space and similarity is measured all the output vectors m_i to find the closest vector c on the output grid. Second, this best matching or winning vector and its neighborhood are changed closer to the input vector. The formula for learning process is as follows:

$$m_i(t+1) = m_i(t) + a(t)\, h_{ci}(t)(x(t) - m_i(t)). \qquad (1)$$

Where a(t) is learning rate factor and $h_{ci}(t)$ – neighborhood function at the step t. During the learning process the learning rate and the neighborhood function are decreasing. As a result of the learning process similar samples will be located close map units and the map becomes ordered.

For visualization of the self-organizing map a Unified distance matrix (U-matrix) is used. The U-matrix presents the distances between each map unit by color-coding. The light color indicates a small distance between two map units and dark color presents bigger difference between map units. The points on the output map that are on the light area belong to the same group or cluster and are separated by the darker areas marking the borders between the clusters.

The analysis is performed by the SOM toolbox ver 2.0 for Matlab [18].

3 Matrix Reordering

Matrix reordering (also known as seriation) methods have a long history of revealing the hidden structure of multidimensional data in several disciplines – anthropology, archaeology, biology, cartography, manufacturing, operations research and survey data analysis (extended review of literature and research issues is available in [6]). The main goal of such matrix analysis is to permutate rows and columns to maximize the similarity of the neighboring elements. The initial data is only reorganized and therefore always preserved to the full extent in the results, thus this technique is not classically considered as a dimensionality reduction method. However, some methods (e.g., [16]) allow the identification of cluster boundaries in the ranking of rows and columns, therefore resulting also two-mode clusters (for an introductory overview of two-mode clustering methods, see [13]). In this paper, we are applying the method initially used for market survey data [14]-[16]. In addition to traditional matrix reordering methods, it has an additional goal – to establish a scale of typicality in the data, describing the transformation of data and formed clusters from the least to most typical. The main difference from the Self-Organizing map (SOM) is the ability to visualize two-mode data. SOM allows to visualize either rows (objects) or columns (attributes), MR methods visualize simultaneously rows, columns and corresponding data points. Visualization is constructed from the order identified from the data that would minimize the distances between rows and columns and establish. Already such order itself without the visualization, gives us also good insights about typicality in the system.

4 Multidimensional Scaling

The multidimensional scaling is a set of related statistical techniques often used in data visualization for exploring proximities in data. The goal of the method is to project data points as points in some lower-dimensional space so that the distances between the points correspond to the dissimilarities between the points in original space as closely as possible. Such representation is valuable to gain insight into the structure of data. The MDS can be used as a method to reduce the dimensionality the data and reveal the dissimilarities between the samples.

The method of multidimensional scaling is said to be metrical if it is based on measured proximities and nonmetrical when the proximities are based on judgment [4]. The original MDS method was metric [11]. In the current paper the analysis is based on nonmetrical data and, therefore, nonmetrical MDS is used. Data is analyzed by the statistical software package SPSS. In our analysis the ALSCAL algorithm [10] is used.

There are n sample vectors $x_1,...,x_n$ and the distance between original samples i and j is g_{ij}. The y_i is the lower-dimensional representation of x_i and the distance between projected items i and j is d_{ij}. The aim of the MDS method is to find a configuration image points $y_1,...,y_n$ in a lower dimensional space for which the distances d_{ij} between the samples are as close as possible to the corresponding original distances g_{ij} so that the dissimilarities between the sample vectors are retained as well as possible Because it is impossible to find a configuration for which $d_{ij} = g_{ij}$ for all i and j, certain criteria is needed whether the result is good enough to finish the approximation process.

5 Analysis of The Estonian Terms of Emotions

The purpose of the case study was to discover the hidden structure of the Estonian emotion concepts according to the laypersons' intuitive knowledge.

5.1 Subjects and Method

The inquiry was carried out in written form during the summer months of 2003 in Estonia. The number of respondents was 100 (50 men and 50 women), aged from 14 to 76, all native speakers of the Estonian language. There were 24 concepts of emotion selected for the study that forms a small but representative set of the category, sharing the prototypical features of emotion concepts to various degrees. The selection is based on the results of tests of free listings [12] as well as on word frequencies in the corpora.

The participants of the survey had to evaluate the meaning of every single word against a set of seven bipolar scales, inspired by Osgood's method of semantic differentials [7]. The "semantic features" measured with polar scales drew qualitative (unpleasant vs. pleasant), quantitative (strong vs. weak emotion, long vs. short in duration), situational (increases vs. decreases action readiness, follows vs. precedes an event), and interpretative distinctions (felt in the mind vs. body, depends mostly on oneself vs. others). In the process of subsequent data handling the original bipolar scales were transformed from having +/- values into positive scales of 7–1, starting from 7 as the maximum value of the dominant or default feature, over 4 pointing to the irrelevance of the scale, and up to 1 as the minimum value (corresponding to the maximum of the opposite feature).

As a result of the answers a data pool of 33600 items was gathered (100 x 7 x 24) and subsequently searched for its hidden structure. To find out how emotion concepts are connected we generated a matrix from the original data pool so that each word is represented by 700 answers.

5.2 Results of the SOM Analysis

Fig. 1 presents the result of SOM analysis. The corresponding translations of words on the SOM are given in Table 1.

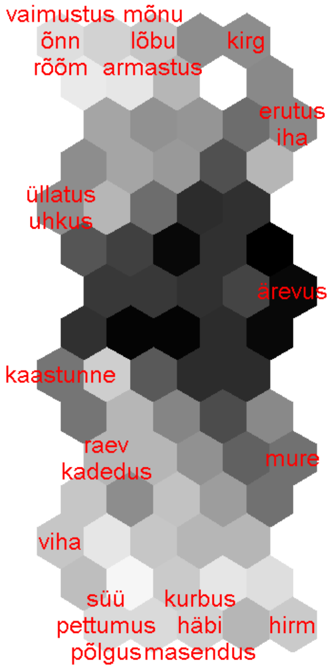

Fig. 1. The SOM of the emotion concepts.

Table 1. Location Of Words On The Som Of Emotion Concepts.

enthusiasm happiness joy	pleasure fun love	passion
		excitement desire
surprise pride		
		anxiety
pity		
rage envy		concern
anger		
guilt disappointment contempt	sadness shame oppression	fear

The SOM appears as a bilaterally symmetrical representation. The positive emotion concepts tend to gather to the upper part of the graph and the words referring to negative emotions to the lower part of the graph. The main organizing dimension of the representations appears to be the negativeness and positiveness of the concepts that extends the shape of the SOM map in one direction. As the anticipatory states (*hirm* 'fear', *erutus* 'excitement', *mure* 'concern'), gathered to the right edge of the graph, the scale follows vs. precedes an event seems to function as an additional dimension. The darker area in the middle clearly separates these two clusters. There is a concept *ärevus* 'anxiety' located outside of these two clusters. Apparently it is identifiable neither as positive nor negative or having conflicting specifications in respect of affiliation. This kind of structure has most in common with the results of Watson & Tellegen [17] who have found that 50—75 % of the semantics of emotion vocabulary in multiple languages is accounted for two unipolar dimensions of Positive and Negative affect. The extendedness of the graph also speaks for the preference of focus on valence over arousal [3].

5.3 Results of the Matrix Reordering Analysis

The results of the Matrix Reordering analysis are given on Fig. 2 and Table 2. The table gives an overview of the weight value and its change and the figure an overall view on the data table after reordering.

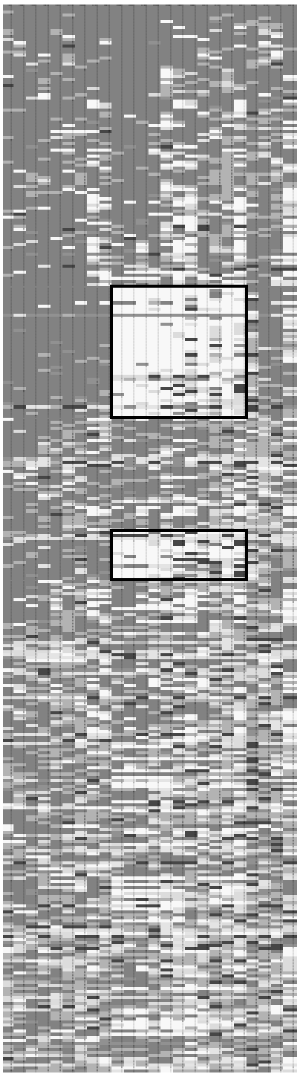

Fig. 2. The Matrix Reordering of the emotion concepts.

The value of weight calculated by the MR method and especially the change of the weight gives us an insight into the structure of the components. We use the change of the weight value as an indicator of group's borders. If the value is relatively close to zero or is negative, we mark this position as a border between groups. The MR ordered concept according to the scale of untypical-typical. The most untypical concept is at the bottom of the table and the most typical at the top. When we start to study the table from the most typical end of the scale we could see there is a group of negative emotion concepts starting from *viha* 'anger' and ends with *mure* 'concern'. After negative change of the weight value a new group begins (see *rõõm* 'joy') consisting of positive concepts and finally there are four concepts (*kaastunne* 'pity', *põlgus* 'contempt', *kadedus* 'envy', *ärevus* 'anxiety') that are the most untypical and form a separate group. The group of positive terms also divides into the two groups where *rõõm* 'joy', *õnn* 'happiness', *armastus* 'love', and *vaimustus* 'enthusiasm' belong to the more typical group and the rest to the more untypical group.

On Fig. 2 the reordered data matrix is visible. On the vertical scale 24 emotion concepts and on the horizontal scale 700 answers are given. The dominant answers are colored darker. After reordering of the data matrix some untypical clusters appear. Those clusters are connected with positive emotion concepts. We could identify that positive concepts are described as non-dominant edge of the bipolar scale. The four most typical positive concepts tend to have more homogeneous light color than the rest.

Table 2. Ordering Of The Features.

	Weight	Change	Word
typical ↑	700	0	anger
	1063	363	rage
	1295	232	guilt
	1618	323	shame
	1925	307	oppression
	2178	253	disappointment
	2448	270	sadness
	2702	254	fear
	2813	111	concern
	2580	-233	joy
	2909	329	happiness
	3184	275	love
	3398	214	enthusiasm
	3389	-9	pleasure
	3633	244	passion
	3861	228	excitement
	3975	114	surprise
	4159	184	pride
	4372	213	fun
	4577	205	desire
	4583	6	pity
	4701	118	contempt
	4821	120	envy
↓ untypical	4976	155	anxiety

5.4 Results of the MDS analysis

The MDS represents concepts on the circle (Fig. 3). By shape it resembles the circumplex model proposed by Russell [8], [9]. The MDS presents also a clear distinction between the positive and negative concepts on the horizontal scale – the more negative the concepts the more left they are situated and the positive concepts are situated on the right-hand side, accordingly.

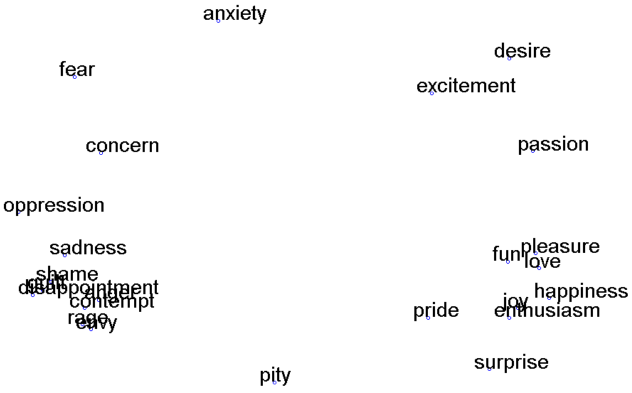

Fig. 3. The MDS of the emotion concepts.

There is another dimension that distinguishes the concepts on vertical scale: the states perceived as event preceding are situated on the upper part of the circle and the states perceived as following some event are situated in the bottom.

6 Discussion

In previous section three methods accessed to semantics of Estonian emotion terms were compared. All the methods revealed somehow the distinction between the positive and negative concepts, but there were some differences in the clustering of the concepts that were assessed more variably.

If we compare the results of the methods we could see that the words *ärevus* ‘anxiety’ and *kaastunne* ‘pity’ are separating the clusters in the results of the SOM and MDS. Words *põlgus* ‘contempt’, *kadedus* ‘envy’ are not so clearly separated on the SOM and MDS plots as on the MR scale. On the other hand the words that are on the borders of the clusters of the MR like *mure* ‘concern’ and *iha* ‘desire’ are also separated in the SOM and MDS analysis. The concepts *rõõm* ‘joy’, *õnn* ‘happiness’, *armastus* ‘love’ and *vaimustus* ‘enthusiasm’ form a core of the most typical positive concepts of all results.

It can be seen, that as the concepts were addressed through their relation to the individual perceptions of episodes of emotional experience, the SOM and MDS result in very similar layouts, except the orientation of the dimensions and the way of discriminating the groups.

Comparing the results of analysis of linguistic data the SOM formed clearly separable clusters and the MDS projected data on the circle. The MR method projects the concept on a scale untypical-typical. There might be that the MDS presented the overall distances between the sample vectors and therefore the extremity of dominant positive negative scale became dominant and the overall look of the results is the same - circular. At the same time the SOM gives an overview of local relations between concepts and forms local clusters, but even projection of local relationships between the sample vectors gave insight that there is clear division between the positive and negative concepts. The MR method also grouped negative concepts on the area of the most typical concepts and between them and the most untypical concepts is a group of positive emotion concepts.

The MR distribution of concepts shows correspondence with the characteristics of emotions. In general the reordering places the dominant answers on top of the table and most untypical on the bottom. We could even identify distinction between more typical and less typical positive concepts. The more typical positive concepts are also located together on the SOM and MDS. The most typical concepts on the SOM form a group on the edge of the map but on the MDS as a core of the positive concepts.

The most typical scales are unpleasant, strong, long, depends mostly on oneself, felt in the mind, follows an event and increases action readiness. The word *viha* ‘anger’ is like a prototype of the emotions it is described by the most typical qualities. The result is comparable with the previous finding from free listings experiment where the participants had to count emotion concepts where the most often mentioned word was *viha* ‘anger’ [12]. We could conclude that the scales describe the qualities of the emotions in the optimal way. The word *ärevus* ‘anxiety’ differs most from the prototype and is described by untypical or unclear characteristics.

7 Conclusions

We compared the results of the analysis of Estonian emotion concepts by three methods—the self-organizing map, matrix reordering and multidimensional scaling. All the methods revealed the hidden structure of the data in their peculiar way. The SOM stresses more on the local similarities and clearly distinguishes groups within the data. The MDS method reveals global dissimilarities between the samples. The MR method reveals what sample vectors are the most typical and what untypical and also some grouping is identifiable. As a result of the case study the main distinction between the emotion concepts is division into the positive and negative group. But all the methods also revealed more detailed and unique grouping information. There was a group of terms that did not belong directly to any of those two main groups. To find out those special terms the results of all three methods are compared. Knowledge about the results of different methods helps us to interpret each method separately and to identify peculiarity of the method. Such knowledge helps us to improve interpretation rules for every method.

8 Acknowledgement

This study was supported by Estonian Science Foundation, grant No 7149.

9 References

[1] T. Kohonen, “Self-organising maps” 3rd ed. Berlin: Springer, 2000.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification” 2nd ed. New York: John Wiley & Sons, 2001.

[3] L. Feldman Barrett, “Valence focus and arousal focus: Individual differences in the structure of affective experience” ; Journal of Personality and Social Psychology, 69, 153–166, 1995.

[4] J. D. Jobson, “Applied multivariate Data Analysis”. Vol. II. New York: Springer, 1992.

[5] S. Kaski, “Data exploration using self-organizing maps”. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82. Helsinki University of Technology, Finland, 1997.

[6] I. Liiv, “Czekanowski-Bertin Learning Paradigm: A Discussion“ ; In Proceedings of The International Conference on Artificial Intelligence (ICAI’07), CSREA Press, Las Vegas, USA, June 25–28, 2007.

[7] C. E. Osgood, G.. J. Suci, and P. H. Tannenbaum, “The Measurement of Meaning”. University of Illinois Press: Urbana and Chicago, 1975.

[8] J. A. Russell, “A circumflex model of affect”; Journal of Personality and Social Psychology, 39, pp. 1161–1178, 1980.

[9] J. A. Russell, M. Lewicka, and T. Niit, “A cross-cultural study of a circumplex model of affect” ; Journal of Personality and Social Psychology, 57, pp. 848–856, 1989.

[10] Y. Takane, , F. W., Young, and J. de Leeuw, “Nonmetric individual differences multidimensional scaling: an alternating least square method with optimal scaling features”; Psychometrika, 42, pp. 7–67, 1977.

[11] W. S. Torgerson,. “Theory and methods of scaling”. London: Chapman & Hall, 1958.

[12] E. Vainik, “Emotions, emotion terms and emotion concepts in an Estonian folk model”; Trames, 6(4), pp. 322–341, 2002.

[13] I. Van Mechelen, , H.-H., Bock, and P. De Boeck, “Two-mode clustering methods: A structural overview”; Statistical Methods in Medical Research, 13, pp. 363–394, 2004.

[14] L. Võhandu, “Some problems with data analysis”; Transactions of Tallinn Technical University, 366, pp. 3–14, 1974.

[15] L. Võhandu, “Rapid Data Analysis Methods”; Transactions of Tallinn Technical University, 464, pp. 21–39, 1979.

[16] L. Võhandu, “Some Methods to Order Objects and Variables in Data Systems”; Transactions of Tallinn Technical University, 482, pp. 43–50, 1980.

[17] D. Watson and A. Tellegen, “Toward a consensual structure of mood”; Psychological Bulletin, 98, pp. 219–235, 1985.

[18] E. Alhoniemi, J. Himberg, J. Parhankangas, and J. Vesanto, “SOM Toolbox” (Version 2.0). [Computer software and manual]. Retrieved November 11, 2005, from http://www.cis.hut.fi/projects/somtoolbox/