Semantic Analysis of Web Pages Using Nonnegative Matrix Factorization

Václav Snášel[1], Hana Řezanková[2], Dušan Húsek[3],
Miloš Kudělka[1], Ondřej Lehečka[1]

[1]Department of Computer Science, VSB-Technical University of Ostrava,
17. listopadu 15, Ostrava, Czech Republic,
vaclav.snasel@vsb.cz
[2]Department of Statistics and Probability, University of Economics, Prague,
W. Churchill sq. 4, 130 67 Prague, Czech Republic,
rezanka@vse.cz
[3]Institute of Computer Science, Academy of Sciences of the Czech Republic,
Prague, Czech Republic,
dusan@cs.cas.cz

Abstract. In this paper, the web pages concerning products sale are analyzed with the aim to create groups of similar web pages and characterize these by GUI patterns. We applied nonnegative matrix factorization and methods of cluster analysis. Gradient descent constrained least squares method (GD-CLS) were found as a suitable technique for this purpose.

1 Introduction

One of the key tasks in a problem of orientation in large space of weak-structured data like web is to find out what information it is possible to find on the web page and which methods can be used. There are many approaches which lead more or less to the same destination. It is to offer effective orientation in big amount of weak-structured data. A deeper analysis must be performed to obtain detail page information [8]. This detail information is something what can be said about the page. Simple example can be occurrence of particular word (with its frequency and position). Next example can be fact that the page belongs to particular domain (for example, selling products). The field for evolving page analysis methods is very wide. One feature of all the approaches is independency on form which is the information written in. If we focus on methods based on the full-text approach then the key problem seems to be the loss of information about structure. On the other hand there are methods which prefer page structure and hence they use HTML code elements structure as one of the information source. Trouble with both approaches is changing web, markup language standards and evolution of requirements on web page design [6].

In this paper we introduce our perspective on mentioned problems which comes out from an interesting connection of two fields - information retrieval methods and semantic web on one side and on the other side web design practices and web patterns [13,15,18,17]. This perspective allows us to efficiently combine both mentioned approaches. The form in which the user can find information on web page is changing (but some features remains).

© Václav Snášel (Ed.): Znalosti 2008, pp. 234–243, ISBN 978-80-227-2827-0.
FIIT STU Bratislava, Ústav informatiky a softvérového inžinierstva, 2008.

The content is more or less still the same. This implies that the most effective approaches must handle on key aspects of the form and the content which is held in the form. In our perspective we are strictly focused on the way the user perceive web page. This sensation then motivates web page designers to create pages conforming user requirements. This never-ending and invisible interaction between web designers and users is projected to patterns [1]. The evidence of patterns on web pages with the same content but created by different designers is demonstrated by more or less the same look. This fact shows key feature of our approach and is essential for us. If we realize this fact then we can find a lot of methods in information retrieval and semantic web fields which can be used very efficiently for finding ways to fulfill user requirements. On the other hand this perspective opens new motivation for domain and GUI pattern experts. They formulate requirements on user interface with a view to web designers. We are binging new motivation to them because patterns can be found and described with regard to their usage in description of web page semantics.

Web pages concerning products sale can contain individual parts with different kinds of information (GUI patterns [4,15,16,18,17]). There are short description and price of the product, possibility to buy it, detail description (technical data), review, discussion etc. We can distinguish different types of these web pages according to degrees of detection of GUI patterns [6,7,12]. The aim of this paper is to classify such web pages and search typical patterns which are contained in them.

The paper is organized in the following way. In the chapter 2, we will show how the web pages were described and which data were used as an input for the further analyses. In the chapter 3, we will describe the analyses by methods of cluster analysis. We will mention both clustering GUI patterns and clustering web pages and characterization of obtained clusters by their centroids. Chapter 4 concerns the analysis by nonnegative matrix factorization. The GD–CLS method was applied for finding clusters of web pages.

2 Web page description

We analyzed more then 20 thousands Czech web pages. We searched for 10 GUI patterns [2] which could be contained in them. There were price, sale, discount, credit, description, opinion, discussion, login, bazaar and questionnaire. Therefore, we described each web page by a 10–dimensional vector of values which express degrees of detection of individual GUI patterns contained in it. In this way we obtained 23422 vectors $S = (v_1, \ldots, v_{10})$ characterizing individual web pages where $0 \leq v_j \leq 1$. Average values for individual patterns based on all web pages are the following:

	price	sale	discount	credit	description	opinion	discussion	login	bazaar	questionnaire
Average	0.482	0.342	0.235	0.077	0.113	0.135	0.135	0.215	0.081	0.049

For further analyses, the data were transformed so that $\sum_j^{10} v_j = 1$. In this case, average values for individual patterns based on all web pages are the following:

	price	sale	discount	credit	description	opinion	discussion	login	bazaar	questionnaire
Average	0.254	0.164	0.095	0.027	0.053	0.098	0.113	0.114	0.053	0.030

3 Analysis by traditional methods

The base image on web pages structures can be obtained by investigation of associations between GUI patterns (i.e. attributes). Because of common occurrence of higher values is more important than common occurrence of lower (especially zero) values, the cosine measure is better than correlation coefficient for this purpose [14,10,5]. Cosine measure matrix for the set of patterns is shown in Table 1. Cosine similarity measure is expresses by the formula

$$s_C(X_k, X_l) = \frac{\sum_{i=1}^{n} v_{ik} \cdot v_{il}}{\sqrt{\sum_{i=1}^{n} v_{ik}^2} \cdot \sqrt{\sum_{i=1}^{n} v_{il}^2}}$$

where X_k and X_l are k^{th} and l^{th} patterns, n is the number of web pages and v_{ik} is a value for the i_{th} web page and the k_{th} pattern.

The highest value is 0.416 for the pair price, discount; the second high value is 0.410 for the pair price, sale.

	price	sale	discount	credit	description	opinion	discussion	login	bazaar	questionnaire
price	1.000	0.410	0.416	0.184	0.142	0.039	0.037	0.151	0.140	0.040
sale	0.410	1.000	0.316	0.160	0.142	0.031	0.027	0.166	0.048	0.029
discount	0.416	0.316	1.000	0.169	0.100	0.023	0.019	0.115	0.077	0.026
credit	0.184	0.160	0.169	1.000	0.097	0.030	0.013	0.062	0.026	0.019
description	0.142	0.142	0.100	0.097	1.000	0.029	0.026	0.067	0.026	0.016
opinion	0.039	0.031	0.023	0.030	0.029	1.000	0.206	0.125	0.033	0.055
discussion	0.037	0.027	0.019	0.013	0.026	0.206	1.000	0.069	0.011	0.065
login	0.151	0.166	0.115	0.062	0.067	0.125	0.069	1.000	0.041	0.059
bazaar	0.140	0.048	0.077	0.026	0.026	0.033	0.011	0.041	1.000	0.010
questionnaire	0.040	0.029	0.026	0.019	0.016	0.055	0.065	0.059	0.010	1.000

Table 1. Cosine measure matrix for the set of 10 patterns (obtained by the SPSS system)

By complete linkage method, the distance between two different clusters is the greatest distance between two objects in the clusters. On the base of this similarity matrix, we can analyze the data structure by hierarchical cluster analysis. We applied different linkage methods and we obtained two main clusters by complete linkage (dissimilarity was computed as $1 - cosine\ measure$). The dendrogram is shown on Figure 1.

We compared results mentioned above with the results obtained by means described in [5,10,14]. We applied factor analysis and we used factor loadings for two components as an input for fuzzy cluster analysis in the S-PLUS system.

The results of fuzzy cluster analysis are memberships u^{jh} for each j^{th} pattern and each h^{th} cluster. Memberships have to satisfy the following conditions

$0 \leq u_{jh} \leq 1$ for all $j = 1, \ldots, 10$ and all $h = 1, \ldots, K$ (K is the number of clusters),
$\sum_{h=1}^{k} u_{jh} = 1$ for all $j = 1, \ldots, 10$.

In the S-PLUS system, the memberships are defined through minimization of function f:

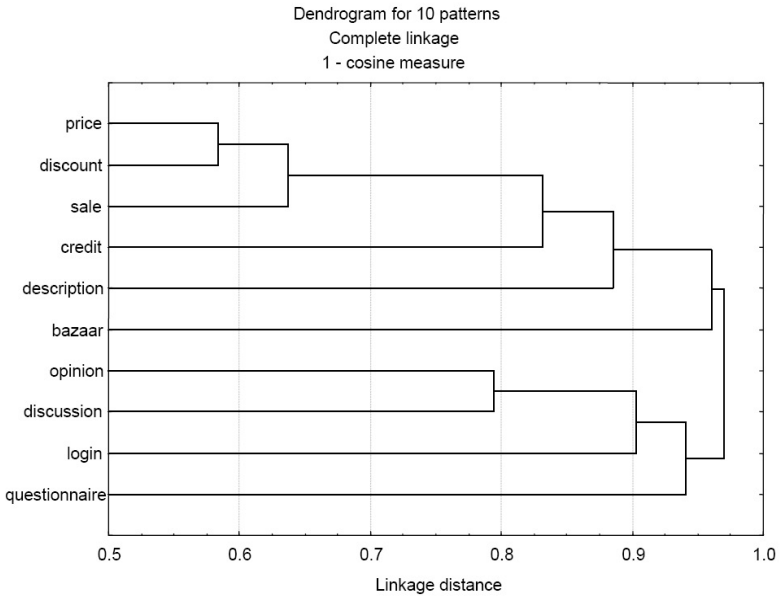

Fig. 1. The dendrogram for the set of 10 patterns (obtained by the STATISTICA system).

$$f = \sum_{h=1}^{K} \frac{\sum_{k=1}^{10} \sum_{l=1}^{10} u_{kh}^2 u_{lh}^2 d_{kl}}{2 \sum_{l=1}^{10} u_{lh}^2}$$

where dissimilarities d_{kl} are known and memberships u_{kh} an u_{lh} are unknown.

This technique was designed for binary data but it can be used also in other cases. We used it from two reasons. Firstly, interpretation of factor loadings is disputable and secondly, it is difficult to apply fuzzy cluster analysis to clustering attributes (in S-PLUS only objects can be clustered). For two clusters, the patterns were assigned in the way shown in Table 2.

	[Cluster 1]	[Cluster 2]	Closest hard clustering
price	0.736	0.264	1
sale	0.738	0.262	1
discount	0.805	0.195	1
credit	0.748	0.252	1
description	0.603	0.397	1
opinion	0.164	0.836	2
discussion	0.187	0.813	2
login	0.470	0.530	2
bazaar	0.444	0.556	2
questionnaire	0.288	0.712	2

Table 2. Membership coefficients for 2 clusters (obtained by the S-PLUS system)

Graphical output from fuzzy cluster analysis is the silhouette plot. For this graph, the value ψ_j is calculated for the j^{th} pattern in the cluster C_g on the base of average

distance of this pattern from other patterns in individual clusters, i.e.

$$\psi_j = \frac{\eta_j - \mu_j}{\max\{\eta_j, \mu_j\}}$$

where

$$\eta_j = \frac{\sum_{i \in C_g} d_{ij}}{m_g - 1} \qquad \mu_j = \min_{h \neq g} \left\{ \frac{\sum_{i \in C_g} d_{ij}}{m_h} \right\}$$

and m_g and m_h are the number of patterns in the g^{th} (h^{th}) cluster

It is shown in Figure 2 for 2 clusters. Patterns bazaar and login can be assigned both to cluster 1 and to cluster 2; therefore are displayed by different way.

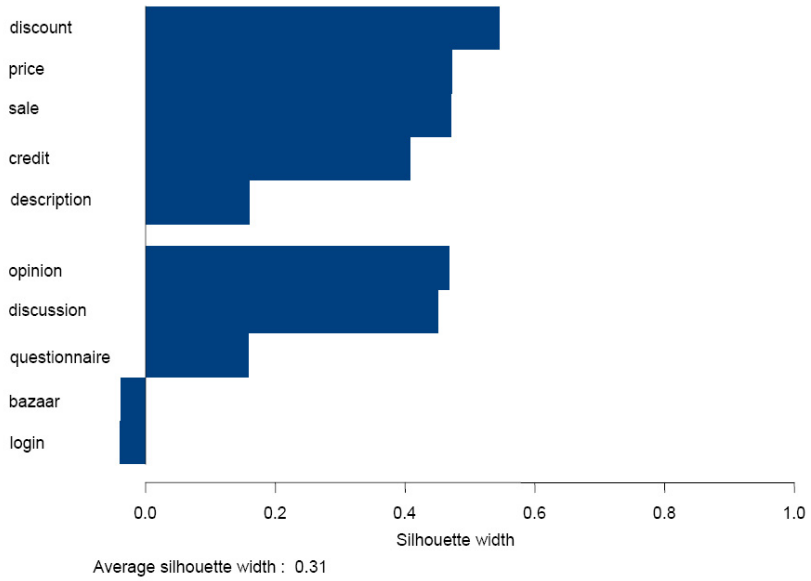

Fig. 2. The silhouette plot for the set of 10 patterns (obtained by the S-PLUS system).

However, our main aim was the identification of clusters of web pages and characterization of these clusters by mean values of attributes. First, we used two-step cluster analysis (in the SPSS system) with log-likelihood dissimilarity measure. Log-likelihood distance between clusters C_g and C_h is

$$d_{gh} = \zeta_g + \zeta_h - \zeta_{<g,h>}$$

where $< g, h >$ denotes a cluster created by joining objects from clusters C_g and C_h, and

$$\zeta_g = -n_g\Big(\sum_{j=1}^{10}\frac{1}{2}\ln(s_j^2 + s_{gj}^2)\Big)$$

where n_g is the number of web pages in the g^{th} cluster, s_j^2 is the sample variance of the j^{th} pattern and s_{gl}^2 is the sample variance of the j^{th} pattern in the g^{th} cluster.

This method is suitable for large data files. The results correspond from 1 to 5 clusters are shown in Table 3 and 4. In Table 4, the values greater than 0.09 are considered as significant. The procedure in SPSS determined 3 clusters as an optimal number (both by Schwarz's Bayesian and Akaike's information criterions). BIC (Schwarz's Bayesian Information Criterion) is expresses by the formula

$$BIC(k) = -2\sum_{h=1}^{k}\zeta_h + w_k\ln(n)$$

where ζ_h is computed according the formula described above, n is the number of web pages and w_k is computed according the formula $w_k = k \cdot 2 \cdot 10$. AIC (Akaike Information Criterion) is computed according the formula

$$AIC(k) = -2\sum_{h=1}^{k}\zeta_h + 2w_k$$

From the variants from 1 to 15 clusters, the variant of 10 clusters (with the sizes from 1322 to 5152 web pages) was determined as an optimal.

	2 clusters	3 clusters	4 clusters	5 clusters
Cluster				
1	17354	10909	10310	5262
2	6068	6695	4669	4533
3		5818	2733	2806
4			5710	5679
5				5142
Total	23422	23422	23422	23422

Table 3. The size of clusters for 2 – 5 clusters (obtained by the SPSS system)

	Price	Sale	Discount	Credit	Description	Opinion	Discussion	Login	Bazaar	Questionnaire
1 / 2	0.333	0.216	0.128	0.036	0.070	0.019	0.018	0.074	0.068	0.039
2 / 2	0.029	0.014	0.001	0.000	0.004	0.324	0.384	0.231	0.009	0.004
1 / 3	0.411	0.281	0.161	0.000	0.028	0.009	0.011	0.089	0.009	0.002
2 / 3	0.200	0.107	0.071	0.093	0.137	0.038	0.034	0.058	0.163	0.100
3 / 3	0.022	0.011	0.001	0.000	0.002	0.333	0.394	0.227	0.008	0.002
1 / 4	0.418	0.283	0.164	0.000	0.023	0.008	0.012	0.089	0.001	0.002
2 / 4	0.241	0.125	0.106	0.133	0.046	0.030	0.010	0.055	0.250	0.005
3 / 4	0.143	0.100	0.015	0.000	0.287	0.058	0.074	0.076	0.008	0.239
4 / 4	0.022	0.011	0.001	0.000	0.001	0.334	0.399	0.227	0.005	0.000
1 / 5	0.313	0.209	0.330	0.000	0.038	0.010	0.012	0.076	0.007	0.005
2 / 5	0.239	0.123	0.103	0.137	0.046	0.031	0.010	0.055	0.251	0.006
3 / 5	0.148	0.107	0.011	0.000	0.287	0.057	0.071	0.081	0.008	0.230
4 / 5	0.021	0.010	0.001	0.000	0.001	0.335	0.400	0.226	0.005	0.000
5 / 5	0.522	0.354	0.000	0.000	0.003	0.006	0.013	0.100	0.002	0.000

Table 4. Clustering by two-step cluster analysis for 2 – 5 clusters (centroids of clusters)

4 Analysis by nonnegative matrix factorization

The nonnegative matrix factorization (NMF) method for text mining is a technique for clustering that identifies semantic features in a document collection and groups the documents into clusters on the basis of shared semantic features [3]. A collection of documents can be represented as a term-by-document matrix. Since each vector component is given a positive value if the corresponding term is present in the document and a zero value otherwise, the resulting term-by-document matrix is always nonnegative. This data non-negativity is preserved by the NMF method as a result of constraints that produce nonnegative lower rank factors that can be interpreted as semantic features or patterns in the text collection.

4.1 NMF method

With the standard vector space model a set of documents S can be expressed as an $m \times n$ matrix V, where m is the number of terms and n is the number of documents in S. Each column V_j of V is an encoding of a document in S and each entry v_{ij} of vector V_j is the value of i-th term with regard to the semantics of V_j, where i ranges across the terms in the dictionary. The NMF problem is defined as finding an approximation of V in terms of some metric (e.g., the norm) by factoring V into the product WH of two reduced-dimensional matrices W and H [11]. Each column of W is a basis vector. It contains an encoding of a semantic space or concept from V and each column of H contains an encoding of the linear combination of the basis vectors that approximates the corresponding column of V. Dimensions of W and H are $m \times k$ and $k \times n$, where k is the reduced rank. Usually k is chosen to be much smaller than n. Finding the appropriate value of k depends on the application and is also influenced by the nature of the collection itself.

Common approaches to NMF obtain an approximation of V by computing a (W, H) pair to minimize the Frobenius norm of the difference $V - WH$. The matrices W and H are not unique. Usually H is initialized to zero and W to a randomly generated matrix where each $W_{ij} > 0$ and these initial values are improved with iterations of the algorithm.

4.2 The NMF problem

GD-CLS is a hybrid method that combines some of the better features of other methods. The multiplicative method, which is basically a version of the gradient descent optimization scheme, is used at each iterative step to approximate the basis vector matrix W. H is calculated using a constrained least squares (constrained least squares - CLS) model as the metric.

Algorithm

1. Initialize W and H with nonnegative values, and scale the columns of W to unit norm.
2. Iterate until convergence or after l iterations:

- $W_{ic} = W_{ic} \frac{(VH^T)_{ic}}{(WHH^T)_{ic}+\epsilon}$, for c and i [$\epsilon = 10^{-9}$]
- Rescale the columns of W to unit norm
- Solve the constrained least squares problem where $min_{H_j}\{||V_j - WH_j||_2^2 + \lambda||H_j||2^2$ the subscript j denotes the j-th column, for $j = 1, \ldots, m$. Any negative values in H_j are set to zero. The parameter k is a regularization value that is used to balance the reduction of the metric $||V_j - WH_j||_2^2$ with the enforcement of smoothness and sparsity in H.

For any given matrix V, matrix W has k columns or basis vectors that represent k clusters, matrix H has n columns that represent n documents. A column vector in H has k components, each of which denotes the contribution of the corresponding basis vector to that column or document. The clustering of documents is then performed based on the index of the highest value of k for each document. For document i ($i = 1, \ldots, n$), if the maximum value is the j-th entry ($j = 1, \ldots, k$), document i is assigned to cluster j. We used the GD-CLS method for searching $k = 2, 3, 4, 5$ clusters. Results are in Table 6. We can see that in the table are very readable results. For example in the first row is a vector which represents sale - Price, Sale, Discount, Credit, Description, and Login. In the second row is a vector which describes information cluster - Opinion, Discussion, Login, and Questionnaire. Limit for a successful pattern we set up to 0.05. The method was unstable for 6 and more clusters. In the table there are clusters-vectors with only one higher value (for example row 5). These clusters do not have a good information value because we expect at least two patterns on each page.

	Price	Sale	Discount	Credit	Description	Opinion	Discussion	Login	Bazaar	Questionnaire
1 / 2	0	0.01	0	0.01	0.03	0.29	0.27	0.337	0	0.06
2 / 2	0.33	0.25	0.19	0.06	0.07	0	0	0.05	0.03	0.01
1 / 3	0.35	0.26	0.21	0.07	0.07	0	0	0	0.04	0.01
2 / 3	0.02	0	0	0.01	0.01	0.41	0.46	0	0.02	0.06
3 / 3	0	0.08	0	0	0.08	0.04	0	0.76	0	0.04
1 / 4	0.42	0	0.41	0.04	0	0	0	0	0.13	0.01
2 / 4	0.01	0	0	0.02	0.02	0.42	0.47	0	0.01	0.06
3 / 4	0	0	0.01	0	0.05	0.04	0	0.85	0	0.05
4 / 4	0.26	0.49	0	0.09	0.15	0.01	0	0	0	0
1 / 5	0	0	0	0.02	0.02	0.42	0.47	0	0	0.06
2 / 5	0.25	0.50	0	0.09	0.16	0	0	0	0	0
3 / 5	0.31	0.01	0.57	0.06	0	0	0	0	0.04	0.01
4 / 5	0.62	0	0	0	0	0	0	0	0.36	0.02
5 / 5	0.01	0	0	0	0.04	0.04	0	0.85	0	0.05

Table 6. Clustering by NMF

5 Conclusion

From cluster analysis of GUI patterns, we found the most similar (from the point of view of degree of detection) price, discount and sale, and further opinion and discussion. These patterns appear important together in the characterizations of web page clusters obtained both by two-step cluster analysis and NMF. The NMF method reflects better the fact that the similarity of price and discount is a litter higher than the similarity of price and sale. By two-step cluster analysis, we did not obtain any combination of price and discount without sale. By NMF, we found such a combination in the cases

of 4 and 5 clusters. Further, by NMF we found a combination opinion, discussion, login and questionnaire as important (for 2 clusters). Contained patterns correspond with members of the second cluster obtained by hierarchical cluster analysis of attributes. On the other hand, we found all clusters obtained by two-step cluster analysis with average membership degree with the value 0.1 and higher for 3 and more patterns. Experimental results on Web pages suggest the effectiveness of our approach. In future, we will also provide a unified view on binary clustering [5,9,10,14] by establishing the connections among various clustering approaches.

References

1. Alexander, Ch.: A Pattern Language: Towns, Buildings, Construction, Oxford University Press, New York 1977.
2. Dearden, A., Finlay, J.: Pattern Languages in HCI: A critical review, Human Computer Interaction, Vol. 21, No. 1, Pages 49–102. January 2006.
3. Ding, C., Li, T., Peng, W., and Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM Press, New York, NY, Pages 126–135.
4. Van Duyne, D. K., Landay, J. A., Hong, J. I.: The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience. Addison-Wesley Professional, 2002.
5. Húsek, D., Moravec, P., Snášel, V., Frolov, A. A., Řezanková, H., Polyakov, P.: Comparison of Neural Network Boolean Factor Analysis Method with Some Other Dimension Reduction Methods on Bars Problem,PReMI 2007: LNCS 4815 Springer 2007, Pages 235–243
6. Ivory, M. Y., Megraw, R.: Evolution of Web Site Design Patterns. ACM Transactions on Information Systems, Vol. 23, No. 4 (2005) Pages 463–497.
7. Kudělka, M. Snášel, V., Lehečka, O., El-Qawasmeh, E.: Semantic Annotation of Web Pages Using Web Patterns. IEEE/WIC/ACM conference WI-2006, Hong Kong 2006. Pages 329–333.
8. Labský, M., Svátek, V., Šváb, O., Praks, P., Krátký, M., Snášel, V.: Information extraction from HTML product catalogues: from source code and images to RDF, Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on 19-22 Sept. 2005 Pages 401–404.
9. Li, T.: A general model for clustering binary data. In Proceeding of the Eleventh ACM SIGKDD international Conference on Knowledge Discovery in Data Mining (Chicago, Illinois, USA, August 21 - 24, 2005). KDD '05. ACM Press, New York, NY, Pages 188–197.
10. Řezanková, H., Húsek, D., Frolov, A. A.: Overlapping Clustering of Binary Variables. Anacapri. In: Knowledge Extraction and Modelling. Italy: TILAPIA Edizion, 2006, Pages 1–7.
11. Shahnaz, F., Berry, M., W., Pauca, P.V., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing and Management, Volume 42, Issue 2 (March 2006). Pages 373–386.
12. Snášel, V., Řezanková, H., Húsek, D., Kudělka, M., Lehečka, O.: Semantic Analysis of Web Pages Using Cluster Analysis and Nonnegative Matrix Factorization. 5th Atlantic Web Intelligence Conference 2007, AWIC, Springer, Advances in Soft Computing, Pages 328–336.
13. Snášel, V.: GUI Patterns and Web Semantics. CISIM 2007: IEEE, Elk, Poland, Pages 14–19.

14. Snášel, V., Húsek, D., Frolov, A. A., Řezanková, H., Moravec, P., Polyakov, P.: Bars Problem Solving - New Neural Network Method and Comparison. MICAI 2007: LNCS 4827 Springer 2007, Pages 671–682
15. Tidwell, J.: Designing Interfaces: Patterns for Effective Interaction Design. O'Reilly Media, Inc. 2006.
16. Van Welie M., van der Veer G. C.: Pattern Languages in Interaction Design: Structure and Organization. Proceedings of Interact '03, Zürich, Switzerland. IOS Press, Amsterdam 2003.
17. http://www.welie.com/patterns/ (October 10, 2007)
18. Wellhausen, T.: User Interface Design for Searching. A Pattern Language. http://www.tim-wellhausen.de/papers/UIForSearching/UIForSearching.html (October 10, 2007)

Acknowledgement This work was partially supported by grant 201/05/0079 awarded by the Grant Agency of the Czech Republic.