Information retrieval in hydrochemical data using the latent semantic indexing approach

Petr Praus and Pavel Praks

ABSTRACT

The latent semantic indexing (LSI) method was applied for the retrieval of similar samples (those samples with a similar composition) in a dataset of groundwater samples. The LSI procedure was based on two steps: (i) reduction of the data dimensionality by principal component analysis (PCA) and (ii) calculation of a similarity between selected samples (queries) and other samples. The similarity measures were expressed as the cosine similarity, the Euclidean and Manhattan distances. Five queries were chosen so as to represent different sampling localities.

The original data space of 14 variables measured in 95 samples of groundwater was reduced to the three-dimensional space of the three largest principal components which explained nearly 80% of the total variance. The five most proximity samples to each query were evaluated.

The LSI outputs were compared with the retrievals in the orthogonal system of all variables transformed by PCA and in the system of standardized original variables. Most of these retrievals did not agree with the LSI ones, most likely because both systems contained the interfering data noise which was not preliminary removed by the dimensionality reduction. Therefore the LSI approach based on the noise filtration was considered to be a promising strategy for information retrieval in real hydrochemical data.

Key words | hydrochemistry, information retrieval, latent semantic indexing, principal component analysis, similarity

Petr Praus (corresponding author)
Department of Analytical Chemistry and Material Testing,
VSB-Technical University Ostrava,
17 listopadu 15, 708 33 Ostrava,
Czech Republic
Tel.:+420 59 732 3370
Fax: 420 59 732 3370
E-mail: *petr.praus@vsb.cz*

Pavel Praks
Department of Mathematics and Descriptive Geometry, Department of Applied Mathematics,
VSB-Technical University Ostrava,
17 listopadu 15, 708 33, Ostrava,
Czech Republic

ACRONYMS

COD	Chemical Oxygen Demand
EC	Electric Conductivity
LSI	Latent Semantic Indexing
PC	Principal Component
PCA	Principal Component Analysis

INTRODUCTION

Water quality is mostly characterized by many parameters forming an n-dimensional data space where each point represents the composition of a water sample taken at a specific locality at a specific time. Real hydrochemical data are mostly noisy, which means that they are not normally distributed, are often co-linear or autocorrelated, and containing outliers or errors. Retrieval of similarities among such data, in the case of water quality assessment, etc., can lead to incorrect findings. Principal component analysis is often applied for removal of the data noise by reduction of their dimensionality.

Latent semantic indexing is the method that has been successfully used for the semantic analysis of large amounts of text documents (Berry *et al.* 1995, 1999). The LSI retrieval algorithm consists of two procedures: (i) reduction of the data dimensionality by PCA and (ii) computation of the similarity measures between the transformed vectors of the jth document and a query document. The query is a term or the set of terms presented in the document. Besides the text retrieval, the LSI approach has been successfully used for image retrieval by Praks *et al.* (2003, 2006) and Labský *et al.* (2005) during the last several years.

doi: 10.2166/hydro.2007.003

The hydrochemical datasets are usually summarised in data matrices. Each column of these matrices represents a sample composition and can be expressed as a vector $\mathbf{x} = (x_1, x_2, x_3, \ldots x_n)$, where x_i is the ith chemical parameter and n is the total number of chemical parameters analysed in water. Such a "chemical" n-dimensional vector space is an analogy to the "document" vector space in which LSI has been currently applied. The document vectors correspond to the water samples and the query document could be a sample selected from a database.

The aim of this paper was to demonstrate the LSI approach for information retrieval in hydrochemical data. For this purpose the searching of samples with similar composition within the groundwater database was tested.

METHODOLOGY DESCRIPTION

Principal component analysis

Principal component analysis seeks abstract principal components (eigenvectors) which explain most of the data variance in a new coordinate system (Lavine 2000; Jolliffe 2002). Each principal component (PC) is a linear combination of the original variables and describes a different source of variance. The largest (the first) PC is oriented in the direction of the largest variance of the original variables and passes through the centre of the data. The second largest PC lies in the direction of the next largest variance, passes through the centre of the data and is orthogonal to the first PC. The third largest PC is directed towards the next largest variance, goes through the data centre and is orthogonal to the first and second PCs, and so forth. Classical PCA is based on the decomposition of a covariance/correlation matrix by eigenvalue decomposition or by the singular value decomposition of real data matrices.

The results of PCA are often interpreted by means of the two-dimensional (2D) and 3D scatter and component plots (or their combination-biplots) which are intended for sample mapping and recognition of relationships among variables.

Latent semantic indexing

LSI can be viewed as a variant of the vector space model with a low-rank approximation of the original data matrices. That is, the original matrix is replaced by another matrix which is as close as possible to the original matrix but whose column space is a subspace of the column space of the original matrix. Rank reduction is performed by PCA. These $t \times d$ (term-by-document) matrices are composed from d documents described by t terms. The d vectors represent the d documents from the columns of the matrix. Each matrix element is a weighted frequency at which the term i occurs in the document j. The details of this vector space construction are given in, for instance, the review of Berry *et al.* (1999). Numerical experiments pointed out that some kind of dimensionality reduction brings automatic noise filtering of the data.

The semantic similarity between a query document and the jth document of the reduced-rank matrices (vector spaces), as the second step of the LSI procedure, has been mostly interpreted as a cosine similarity (CS), i.e. the cosine of an angle between two vectors in the vector space:

$$\cos \varphi_j = \frac{qD_j}{|q||D_j|} \tag{1}$$

where q and D_j are the query and the document vector, respectively, $1 \leq j \leq n$. The geometrical meaning of the cosine similarity is demonstrated in the 2D vector space model shown in Figure 1. Increasing the absolute value of $\cos \phi_j$ (decreasing absolute value of the angle ϕ_j between vectors) indicates increasing the similarity between the query and the document. Computation of the similarity thus reveals some hidden (latent) structures of data.

In multivariate analysis the commonly used metrics of similarity between any points (x_k, x_l) in the n-dimensional

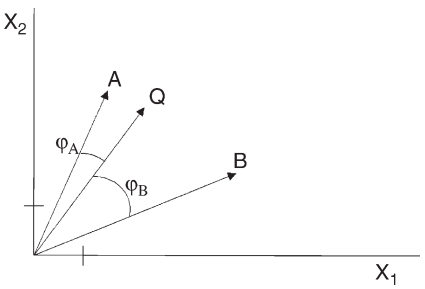

Figure 1 | An example of the cosine similarity measure in the 2D vector space. A, Q, B are the vectors representing samples A and B and the query sample Q, while the symbols ϕ_A and ϕ_B denote the angle between the vectors A, Q and B, Q, respectively.

space are the Euclidean distance

$$D_E = \sqrt{\sum_{j=1}^{n} (x_{k,j} - x_{k,l})^2} \quad (2)$$

and the Manhattan distance

$$D_M = \sum_{j=1}^{n} |x_{k,j} - x_{k,l}| \quad (3)$$

where n is the number of dimensions, e.g. variables characterising the water composition. All these metrics can be correctly used supposing that the coordinate system is orthogonal. In this work, they were tested for the retrieval of samples with proximity composition in the groundwater database.

Multivariate computations

Principal Component Analysis and other statistical calculations were performed by the Statgraphics Plus 5.0 software package (Statistical Graphics Corp.). The testing data matrix of 95 samples was prepared and processed by Excel 97. The rows were constructed from the 14 parameters (variables) measured in each sample of ground water. There were no missing values in the database dataset (14 × 95 values). Before PCA the data were standardized, i.e. mean (average) centred and scaled by the standard deviation of the original measurement variables, to avoid scaling effects.

APPLICATION

Hydrochemical data collection

The groundwater samples providing a regular monitoring of water quality were taken from five different localities on the region of Ostrava (Figure 2). The water quality in these localities has been very similar for a long time. Ostrava is an industrial city of about 300 000 inhabitants located in North Moravia, the Czech Republic. This region is located in the Odra River basin, whose area is about 6252 km^2 and the total watercourse is about 1360 km in length.

Water analyses, including sampling and sample handling, were carried out according to the actual standard ISO

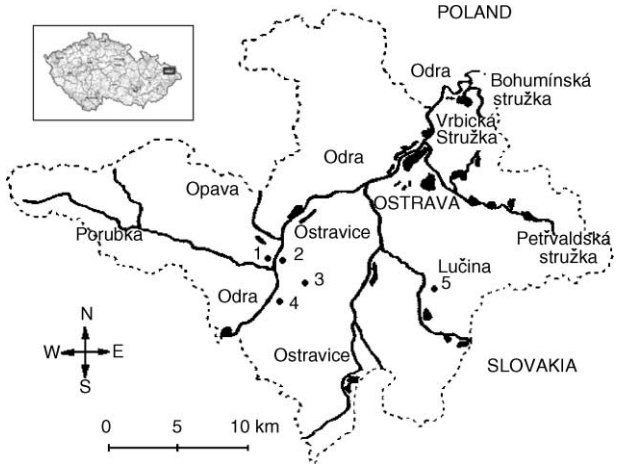

Figure 2 | Map of the Odra River basin. The sampling localities are denoted as 1–5.

methods: pH, ammonium, nitrate, chloride, sulfate, hardness, electric conductivity (EC), alkalinity, acidity, chemical oxygen demand by permanganate (COD-Mn), iron, manganese, dissolved oxygen and aggressive carbon dioxide. Summary statistics of these samples are given in Table 1.

RESULTS AND DISCUSSION

LSI of the hydrochemical data

PCA, performed in the first step of LSI, creates a new coordinate system of the independent (orthogonal) transformed variables. In particular, real hydrochemical parameters are often co-linear because they correlate to each other (see Table 1), such as electric conductivity and salt concentrations, ammonia and nitrate, hardness and sulfate, pH and alkalinity, etc. In order to remove the data noise the data dimensionality has to be reduced by determination of the number of principal components.

For this purpose the Cattel scree plot (Figure 3) and the Kaiser criterion of eigenvalue greater or equal 1 were used. All eigenvalues and their variances were summarised in Table 2. The three largest principal components explaining nearly 79.7% of the data variance, were evaluated. These PCs define the reduced (3D) data space for the retrieval of proximity samples.

Table 1 | Summary statistics of groundwater samples

	Average	Standard deviation	Minimum	Maximum	Standard skewness	Standard kurtosis
Ammonia	0.74	0.98	0.014	3.62	5.178.74	1.41348
Chloride	34.9	15.5	12.3	90	4.733.63	3.2008
COD-Mn	0.84	0.52	0.21	2.36	2.698.93	−0.73901
CO2 aggressive	44.1	24.3	0.21	91.3	−0.735.33	−1.99214
Nitrate	17.1	18.2	0.50	81.7	4.210.18	1.23388
Iron	5.76	7.57	0.06	27.8	4.191.33	−0.18692
Alkalinity	1.50	0.93	0.25	4.1	1.6625	−1.69405
Manganese	0.463	0.490	0.06	1.76	3.092.99	−1.39442
pH	6.33	0.35	5.63	7.01	−0.177.51	−2.34029
Dissolved oxygen	4.04	3.23	0.49	11.9	3.690.17	−0.78504
Sulfate	147	74.4	37.7	367	3.057.74	−0.09845
Hardness	2.20	0.83	0.83	4.4	1.496.23	−1.30413
Conductivity	50.3	17.4	24.5	95.8	2.263.73	−0.53798
Acidity	1.20	0.45	0.25	2.25	2.318.62	−0.86485

The next step of the LSI procedure was computation of the similarity measures between the query vectors and the vectors representing other groundwater samples. The five samples denoted as 1 (1), 53 (3), 62 (5), 90 (2) and 91 (2), representing different sampling localities, were selected as the queries (notations of the localities are given in the parentheses). Sample 91 was also taken in locality 2 but it was indicated to have a very different composition. Therefore it

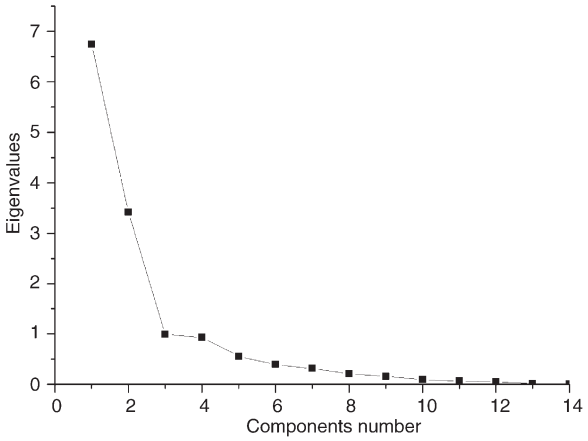

Figure 3 | Scree plot of the eigenvalues.

Table 2 | Principal component analysis

Component number	Eigenvalue	Percent of variance	Cumulative percentage
1	6.7485	48.203	48.203
2	3.4133	24.381	72.584
3	0.998 94	7.135	79.719
4	0.933 17	6.665	86.385
5	0.559 69	3.998	90.382
6	0.394 27	2.816	93.199
7	0.316 03	2.257	95.456
8	0.215 52	1.539	96.995
9	0.161 82	1.156	98.151
10	0.096 367	0.688	98.84
11	0.070 438	0.503	99.343
12	0.052 860	0.378	99.72
13	0.022 748	0.162	99.883
14	0.016 415	0.117	100

Table 3 | LSI retrievals in the reduced space of the three largest principal components

Query	CS	Sample	ED	Sample	MD	Sample	Final retrievals
1	1	1	0	1	0	1	1
	0.982 37	4	0.575 14	4	0.983 33	4	4
	0.973 68	2	0.705 61	2	0.985 77	2	2
	0.938 04	92	1.111 83	82	1.502 61	82	82
	0.937 61	82	1.145 76	92	1.935 77	92	92
	0.896 48	9	1.489 80	95	2.128 01	95	95
53	1	53	0	53	0	53	53
	0.994 89	52	1.072 21	52	1.327 09	52	52
	0.985 76	48	1.174 89	50	1.643 88	50	50
	0.982 02	43	1.225 94	51	1.644 09	51	51
	0.976 10	50	1.308 16	54	1.931 42	54	54
	0.971 28	51	1.427 70	49	2.025 18	49	49
62	1	62	0	62	0	62	62
	0.999 44	58	0.117 08	63	0.173 56	58	58
	0.999 39	63	0.139 53	58	0.178 89	63	63
	0.998 14	61	0.208 13	61	0.333 46	61	61
	0.996 60	60	0.280 80	60	0.408 99	60	60
	0.994 86	57	0.346 85	57	0.592 30	57	57
90	1	90	0	90	0	90	90
	0.991 15	93	0.692 04	7	1.138 07	7	7
	0.988 32	7	0.905 73	93	1.478 57	93	93
	0.976 90	80	1.137 50	72	1.671 11	80	80
	0.976 08	72	1.173 52	80	1.817 13	72	72
	0.929 48	9	1.640 25	15	2.531 76	9	9
91	1	91	0	91	0	91	91
	0.917 20	6	1.681 50	6	2.392 94	6	6
	0.891 33	84	1.968 19	84	2.964 95	84	84
	0.878 37	94	2.029 71	94	3.221 16	81	94
	0.873 75	73	2.108 16	73	3.419 20	73	73
	0.855 27	81	2.186 09	81	3.431 20	94	81

was used to test LSI for the detection of outlaying data. According to the cosine similarity (CS), the Euclidean (ED) and Manhattan distances (MD), the five most similar samples (to each query) were selected and summarised in Table 3. They were ordered in compliance with their similarity to the queries. The best query matchings should have the highest CS (close to 1) and the lowest ED/MD (close to 0).

It is obvious that the CS values in each group differ mutually much less than the ED or MD ones because of the behaviour of cosine function. The significantly low CS and high ED/MD values were computed for the samples close to the queries 53 and 91 which are likely of very different compositions from the others. On the other hand, the best similarities were obtained for the samples matching query 62.

The most proximity samples were evaluated by comparison of the three partial retrievals placed in the rows of Table 3. The coincidence of at least two of them is necessary to recognize the final retrievals that are arranged in the column Final Retrievals.

PCA clustering of the hydrochemical data

The LSI findings can be roughly demonstrated on the PCA scatter plot of the two largest principal components (Figure 4). These two largest PCs explain 72.6% of the total variance and that is why this plot approximates well the 3D space (79.7% of the variance) used by LSI. The four large groups of samples with similar composition are obvious in the plot quadrants. The queries and the matching samples retrieved by LSI can also be identified on this map. Some of the samples are further from or closer to their queries in comparison with the computations in Table 3. This disagreement is caused by the approximative 2D projection of the 3D data structure. However, queries 53 and 91 are also well indicated as the outliers.

Similarity computations in the non-reduced and the standardised original data space

In order to verify the LSI results, the direct computations of CS, ED and MD corresponding to the same queries were performed in the systems of all transformed (Table 4) and all standardised original variables (Table 5). The former system of orthogonal axes contains 100% of the data variance along with the data noise. The latter one is also noisy including the co-linearity of some standardised variables.

In general, the similarities summarised in Tables 4 and 5 are worse in comparison with those of Table 3. It is likely

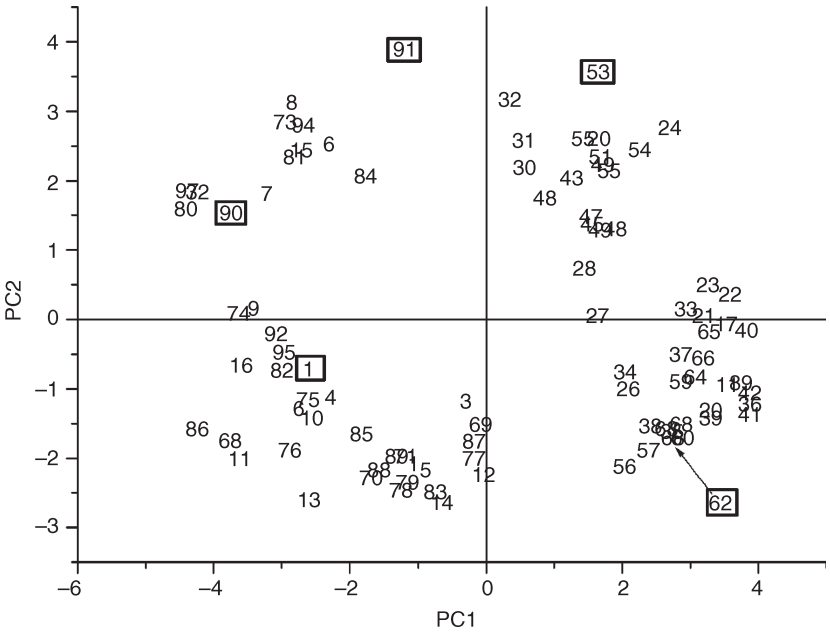

Figure 4 | Principal components scatter plot of the groundwater samples.

Table 4 | Retrievals in the non-reduced space of the 14 principal components

Query	CS	Sample	ED	Sample	MD	Sample	Final retrievals
1	1	1	0	1	0	1	1
	0.916 83	4	1.496 93	4	3.804 29	2	4
	0.906 00	2	1.542 86	2	4.590 18	4	2
	0.764 27	82	2.377 83	82	6.458 49	10	82
	0.720 31	10	2.613 99	10	6.848 25	82	10
	0.695 07	95	2.697 96	95	7.647 11	88	95
53	1	53	0	53	0	53	53
	0.947 77	50	1.701 15	50	4.664 42	55	50
	0.933 95	51	1.712 18	54	4.841 98	54	54
	0.931 29	52	1.810 84	51	4.935 59	50	?
	0.928 90	54	1.848 22	52	5.259 73	51	?
	0.916 28	55	1.853 78	55	5.277 29	52	55
62	1	62	0	62	0	62	62
	0.976 98	63	0.792 29	63	1.789 03	63	63
	0.975 23	61	0.918 68	61	2.282 58	61	61
	0.946 43	58	1.151 58	58	3.030 39	58	58
	0.943 47	60	1.233 13	60	3.327 78	60	60
	0.916 03	57	1.544 07	57	3.707 19	57	57
90	1	90	0	90	0	90	90
	0.944 55	7	1.583 93	7	4.721 22	7	7
	0.923 90	80	2.058 24	80	5.862 47	92	80
	0.922 67	72	2.069 15	72	6.296 08	72	72
	0.910 35	93	2.162 16	9	6.318 14	80	?
	0.892 69	9	2.225 97	93	6.366 58	9	9
91	1	91	0	91	0	91	91
	0.641 00	8	4.067 52	8	12.91 45	8	8
	0.598 44	15	4.275 09	43	13.49 69	15	15
	0.595 36	43	4.342 04	15	13.56 30	43	43
	0.529 25	44	4.504 89	44	13.63 90	44	44
	0.526 69	6	4.603 08	48	13.94 43	53	?

Table 5 | Retrievals in the standardised original data

Query	CS	Sample	ED	Sample	MD	Sample	Final retrievals
1	1	1	0	1	0	1	1
	0.916 83	4	1.477 36	4	4.259 65	4	4
	0.906 00	2	1.574 65	2	4.927 36	2	2
	0.764 27	82	2.383 79	82	6.258 77	82	82
	0.720 31	10	2.596 02	10	6.604 02	10	10
	0.695 07	95	2.710 54	95	6.980 73	95	95
53	1	53	0	53	0	53	53
	0.947 77	50	1.635 16	50	3.733 75	54	50
	0.933 95	51	1.647 77	54	4.047 03	50	?
	0.931 29	52	1.749 94	51	4.140 97	55	?
	0.928 90	54	1.803 03	52	4.173 07	51	?
	0.916 28	55	1.839 75	55	4.685 45	49	55
62	1	62	0	62	0	62	62
	0.976 98	63	0.805 35	63	2.186 49	61	63
	0.975 23	61	0.842 30	61	2.364 55	63	61
	0.946 43	58	1.234 11	58	3.554 07	60	58
	0.943 47	60	1.249 84	60	3.844 75	57	60
	0.916 03	57	1.511 95	57	3.994 67	58	57
90	1	90	0	90	0	90	90
	0.944 55	7	1.570 99	7	4.454 54	7	7
	0.923 89	80	2.018 73	80	4.611 09	80	80
	0.922 67	72	2.021 48	72	5.432 15	72	72
	0.910 35	93	2.181 58	9	6.037 04	93	93
	0.892 69	9	2.205 40	93	6.928 30	9	9
91	1	91	0	91	0	91	91
	0.641 00	8	4.292 07	8	7.619 12	8	8
	0.598 44	15	4.391 39	43	8.713 74	15	15
	0.953 60	43	4.424 21	15	9.781 78	94	?
	0.529 25	44	4.618 00	44	10.41 19	81	44
	0.526 69	6	4.802 05	48	11.33 19	7	?

caused by the higher content of the noise presented in this data space. All cosine similarities are identical while the Euclidean and Manhattan distance are slightly different. It means that the cosine similarity is not sensitive to the co-linearity between standardised parameters.

The partial findings within these pentads correspond well to those of Table 3 with an exception of query 91 which indicates that this groundwater sample possesses a very different composition. However, most of these final retrievals do not agree with the final retrievals of LSI. Even in several cases no coincidence among all three metrics was reached.

It is hard to decide which of the retrieval strategies mentioned above provide the most accurate results. Very likely the values of computed similarities/distances should be the suitable criterion for this decision: the better similarity measures were computed, the more reliable retrievals were selected. Of course, there exist lot of mathematical representations of distance between two points in the n-dimensional space, for instance the Euclidean distance and the Manhattan distance. That is why several similarity metrics should be used in order to reach the ultimate results in real data analysis.

CONCLUSION

The LSI approach was tested for the retrieval of similar groundwater samples in the hydrochemical database. The original space composed from the 14 measured parameters was reduced by PCA to the space consisting of the three principal components. Using the cosine similarity, Euclidean and Manhattan distances the five most proximity samples to the selected queries were arranged.

The LSI findings were compared with the retrievals found in the system of all transformed variables (explaining 100% of the variance) and of all standardised variables. The obtained results mostly did not correspond to the LSI ones because of the interfering data noise which was not removed by the dimensionality reduction.

Unlike the commonly used multivariate methods, such as hierarchical clustering analysis, the benefits of LSI are (i) filtration of the noisy data, (ii) direct similarity calculation independent of any clustering mechanism, (iii) treatment of the large data sets, and (iv) easy implementation for the automated pattern recognition. It can be concluded that the LSI strategy is suitable for information retrieval not only in text documents but also in hydrochemical/chemical data.

ACKNOWLEDGEMENTS

This work was partially supported by the Ministry of Education, Youth and Sport of the Czech Republic (MSM 6198910016 and 1M06047).

REFERENCES

Berry, W. M., Drmač, Z. & Jessup, J. R. 1999 Matrices, vector spaces and information retrieval. *SIAM Rev.* **41** (2), 336–362.

Berry, W. M., Dumais, S. T. & O'Brien, G. W. 1995 Using linear algebra for intelligent information retrieval. *SIAM Rev.* **37** (4), 573–595.

Jolliffe, I. T. 2002 *Principal Component Analysis*, 2nd edn. Springer-Verlag. New York.

Labský, M., Svátek, V., Šváb, O., Praks, P., Krátký, M., Snášel V. 2005 Information extraction from HTML product catalogues: from source code and images to RDF. In: *WI '05: Proceedings of the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, Washington, DC. pp. 401–404. doi: http://dx.doi.org/10.1109/WI.2005.78.

Lavine, B. K. 2000 Clustering and classification of analytical data. In *Encyclopedia of Analytical Chemistry* (ed. R. A. Meyers), pp. 9689–9710. John Wiley & Sons, Chichester.

Praks, P., Dvorský, J. & Snášel, V. 2003 *SIAM Conference on Applied Algebra, 15–19 July, Williamsburg, USA*, pp. 1–8. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, Available online: http://www.siam.org/meetings/la03/proceedings/Dvorsky.pdf.

Praks, P., Machala, L. & Snášel, V. 2006 On SVD-free latent semantic indexing for iris recognition of large databases. In *Multimedia Data Mining and Knowledge Discovery* (ed. V. A. Petrushin & L. Khan). Springer, London.