AUTOMATIC RETRIEVAL WITHIN WATER-RELATED PICTURES DATABASE USING LATENT SEMANTIC INDEXING METHOD

AUTOMATICKÉ VYHLEDÁVÁNÍ V DATABÁZI FOTOGRAFIÍ S VODOHOSPODÁŘSKOU TÉMATIKOU METODOU LATENTNÍ SÉMANTICKÉ INDEXACE

Petr PRAUS [1], *Pavel PRAKS* [2]

[1] *Department of Analytical Chemistry and Material Testing,*

Faculty of Metallurgy and Material Engineering,

VSB-Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic

e-mail: petr.praus@vsb.cz

[2] *Department of Mathematics and Descriptive Geometry, Department of Applied Mathematics,*

Faculty of Electrical Engineering and Computer Science,

VSB-Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic

e-mail: pavel.praks@vsb.cz

Abstract

The Latent Semantic Indexing (LSI) method was used for the automatic retrieval of similar images within a database containing 232 water-related landscape images. The principle of LSI is the images dimensionality reduction and the computation of cosine similarity between a query image and other images from the database. The optimal image dimensionality reduction parameter k was determined by means of a scree plot, which is used in principal component analysis. Using $k = 8$, the photographs displaying similar-looking objects were found. Presented results indicate that low or too high k values can cause a loss of useful information or on the contrary a redundancy of information noise in the analysed images, respectively.

Abstrakt

Metoda Latent Semantic Indexing (LSI) byla použita pro automatické vyhledávání podobných obrázků v databází obsahující 232 fotografií krajiny. Principem LSI je redukce dimensionality jednotlivých obrázků a výpočet kosinové podobnosti mezi dotazovaným obrázkem (query) a ostatními obrázky v databázi. Optimální redukce dimensionality k byla určena pomocí grafu scree plot, který se používá v analýze hlavních komponent. Při $k = 8$ byly v databázi nalezeny podobné fotografie. Příliš vysoké nebo nízké k způsobuje ztrátu informace nebo naopak přebytek informačního šum.

Keywords: image retrieval, latent semantic indexing, dimensionality reduction, environmental monitoring.

1 INTRODUCTION

Environmental pollution has been mostly monitored using a lot of chemical, physical, and biological methods, which are mostly expensive and time consuming for regular analyses. Moreover, these methods can be performed in laboratories, where samples must be transported from distant sampling points. It takes time and costs of materials. In many cases only continual visual monitoring of landscape appearance is satisfactory. For instance, unusual objects on water surface (oil spots, tree branches, dead animals, etc) can signalise environmental problems, which consequently must be solved be means of the sophisticated methods mentioned above.

Contemporary information technologies enable a simple and low-cost acquisition of a large number of quality images, their fast transmission and computer processing. The methods of linear algebra enable automatic images evaluation within a very short time without the presence of human experts.

The area of multimedia information retrieval is systematically studied. For example, the work of Barnard et al. (2003) deals with new approach to modelling multimedia data sets focusing on segmented images with an associated text. The work of Oliva and Torralba (2001) deals with the modelling and recognition of a scene. Their procedure is based on a very low dimensional representation of scene by the set of perceptual dimensions (for example naturalness, openness, roughness, expansion) that represents the dominant spatial structure of scene. The effect of video events classification by the Time Interval Multimedia Event framework is presented in the dissertation of Snoek (2005). This dissertation also includes an overview of approaches for multimedia mining techniques.

The aim of this paper was to test the latent semantic indexing strategy, originally used for the semantic analysis in text documents, for environmental monitoring based on the image retrieval within a database of landscape photographs taken on various localities in different times. Such application should be very effective in the automatic retrieval of unusual observations in a large number of images. If some suspicious abnormality is automatically detected in the images operating staff would be alerted to analyse and solve the problem. For this purpose, the photographs related to the water environment including hydraulic engineering works were taken and used in this work.

2 METHODOLOGY

2.1 Basic Concept of Latent Semantic Indexing

Numerical linear algebra is used as a basis for information retrieval in indexing and retrieval strategy referred as the Latent Semantic Indexing (Berry et al., 1995 and 1999; Grossman et al., 2000). Originally, LSI was used for the semantic analysis of large amounts of text documents. The main reason is that more conventional retrieval strategies, such as vector space, probabilistic and extended Boolean, are not very efficient for real data because they retrieve information solely on the basis of keywords. There are two main problems in using the keywords as indexing units: polysemy (words having multiple meanings) and synonymy (multiple words having the same meaning). As a result, the keywords are not often effectively matched.

LSI can be viewed as a vector space model variant with a low-rank approximation of an original data matrix via numerical methods (Berry et al., 1999). The classical LSI procedure consists of the following steps: i) Singular Value Decomposition (SVD) of a term matrix using numerical linear algebra. SVD is used to identify and remove redundant information noise from data, ii) computation of similarity coefficients between transformed vectors and thus the revelation of some hidden (latent) structures in data.

2.2 Singular Value Decomposition

Let the symbol A is used to denote a term-document matrix, i.e., a $m \times n$ data matrix. The aim of SVD is to compute the decomposition

$$A = U\,S\,V^T \qquad (1)$$

where S is the $m \times n$ diagonal matrix with nonnegative diagonal elements called singular values, U and V^T are the $m \times m$ and $n \times n$ orthogonal matrices, for which the following conditions hold: $U^T = U^{-1}$, $V^T = V^{-1}$. The columns of the matrices U and V^T are called left singular vectors and right singular vectors, respectively. The SVD decomposition can be computed so that the singular values are sorted in decreasing order.

For large matrices the full SVD decomposition is memory and time-consuming operation. Moreover, our later experiments showed that the computation of very small singular values and associated singular vectors can damage the retrieval results. Due to these facts, only the k largest singular values of A and the corresponding left and right singular vectors are computed and stored in computer memory in practice. In this way the multi-dimensional space is reduced to the k-dimensional vector space according to

$$A_k = U_k\,S_k\,V_k^T \qquad (2)$$

where the symbol U_k denotes the $m \times k$ matrix derived from the matrix U by the selection of the k first columns, S_k is the $k \times k$ diagonal matrix with a diagonal including the first k singular values, and V_k is the $n \times k$ matrix acquired by the selection of the k first columns of the matrix V. The columns of the matrix V_k^T contain the transformed values of the original data. In other words, SVD approximates the matrix A with respect to its column vectors. The k-approximation (A_k) of the matrix A's rank is acquired by choosing only the k first singular values of the matrix S while the other ones are neglected.

The LSI algorithm was implemented in MATLAB (The Math Works, Ltd.). For the computation of a few singular values and vectors of the matrix A we used the standard Matlab command *svds(A,k)*. There is no exact routine for the selection of an optimal number of the computed singular values and vectors (Berry et al., 1995 and 1999). For this reason the number of the singular values and associated singular vectors used for the computation of SVD was estimated using a so called scree plot, which is often used in principal component analysis for determining components number (Cattel, 1966).

2.3 Computation of Similarity

The retrieval procedure based on LSI returns to a user the vector of similarity coefficients arranged in a *sim* vector. The *i*-th element of the *sim* vector contains the value, which indicates a measure of the semantic similarity between the *i*-th document and the query document. The increasing value of similarity coefficients indicates the increasing semantic similarity.

There are a lot of possibilities how to calculate the similarity between two vectors. We use the well-known cosine similarity, which measures cosine of the angle between two vectors in the vector space. Generally, the similarity of two documents can be interpreted as the angle between their vectors (Fig. 1):

$$\cos\varphi_j = \frac{(q, D_j)}{\sqrt{(q,q)}\sqrt{(D_j, D_j)}} \tag{3}$$

where q and D_j denote the transformed query and the transformed document vector, respectively, $1 < j < n$.

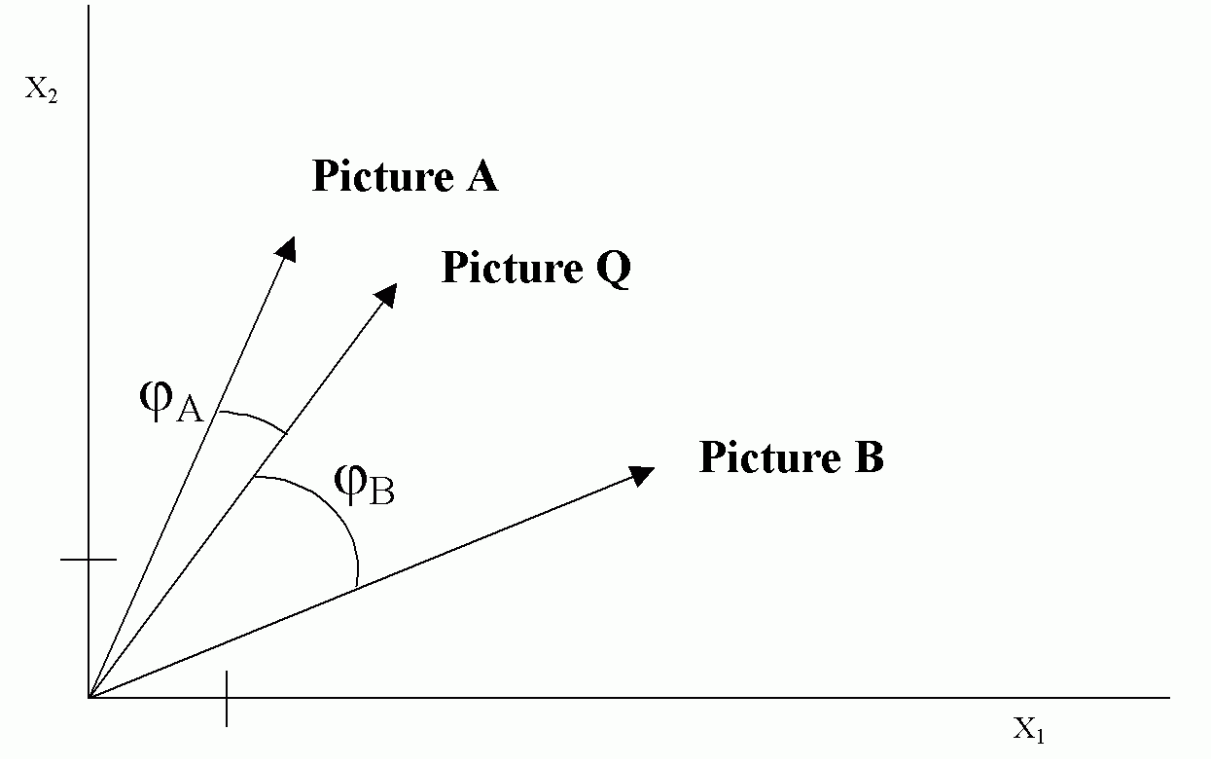

Fig. 1 Example of the cosine similarity measured in the 2-D vector space. A, Q, B are vectors representing the

samples A, B, and the query sample Q, the symbols φ_A and φ_B denote the angle between vectors A,Q and B,Q, respectively.

2.4 Image Retrieval Applications using Latent Semantic Indexing

LSI has become increasingly popular for the image and multimedia retrieval. In our approach a raster image is coded as a sequence of pixels (Praks et al., 2003, 2006 a,b). Then the coded image can be understood as the vector of the *m*-dimensional space where *m* denotes the number of pixels (attributes). Then the symbol *A* denotes the $m \times n$ term-document matrix related to *m* pixels in *n* images. The matrix construction is well described in ,e.g., the paper of Praks et al (2006). After SVD calculation, the document retrieval is done by the calculation of the cosine similarity between the query vector q_C and the vector *V(i),* which is the *i*-th row of the *V* matrix.

Recently, the LSI approach has been also successfully used for the macroeconomic data analysis in economics (Dvořák et al., 2004), for the information retrieval from HTML product catalogues (Labský et al., 2005), and for the analysis of hydrological data (Praus and Praks, 2007).

2.5 Computer Implementation of the Latent Semantic Indexing Method

The function *lsi* returns the vector of the similarity coefficients *sim* to the user. The Latent Semantic Indexing procedure was written in MATLAB by the following way (Grossman et al., 2000):

```
function sim = lsi(A,q,k)
% Input:
% A ... the m × n document matrix
% q ... the query vector
% k ... compute k largest singular values ; % k << n
% Output:
% sim ... the vector of similarity coefficients
[m,n] = size(A);
```

1. step: Compute the co-ordinates of all documents in the *k*-dimensional orthogonal space by the partial SVD of the document matrix *A*:

```
[U, S, V] = svds(A, k);
```

2. step: Compute the co-ordinate of the query vector *q*

```
qc = q^T * U * pinv(S);
```

The vector *qc* includes the co-ordinate of the query vector *q*. The matrix pinv(s) contains the reciprocals of non-zeros singular values (for instance Moore-Penrose pseudoinverse).

3. step: Compute the similarity coefficients between the transformed query vector q_C and the vector *V(i,:)*

```
for i = 1:n %Loop over all documents
sim(i) = (qc*V(i,:)^T)/(norm(qc)*norm(V(i,:)));
end
```

Analysing the original LSI and using the observations of linear algebra, the new LSI procedure was derived by Praks et al. (2006). The derived LSI algorithm replaces the expensive SVD of the non-square matrix *A* by the symmetric partial eigenproblem of A^TA, where the symbol *T* denotes the transpose superscript. Of course, the solution of this partial symmetric eigenproblem using the Lanczos-based iterative method can be obtained very effectively. In addition, the size of the eigenproblem does not depend on the number of attributes.

3 RESULTS AND DISCUSSION

232 photographs of natural and artificial objects related to water such as brooks, rivers, ponds, lakes, bridges, basins, and water works were obtained by a digital camera in various localities of North Moravia, the Czech Republic. The image retrieval using the photos of water-lily and a pond as the queries was performed from all photographs. The retrieval results are demonstrated in Fig. 3 and 4. The dimensionality reduction performed by the symmetric partial eigenproblem was estimated at $k = 8$ by means of the scree plot (Fig. 2).

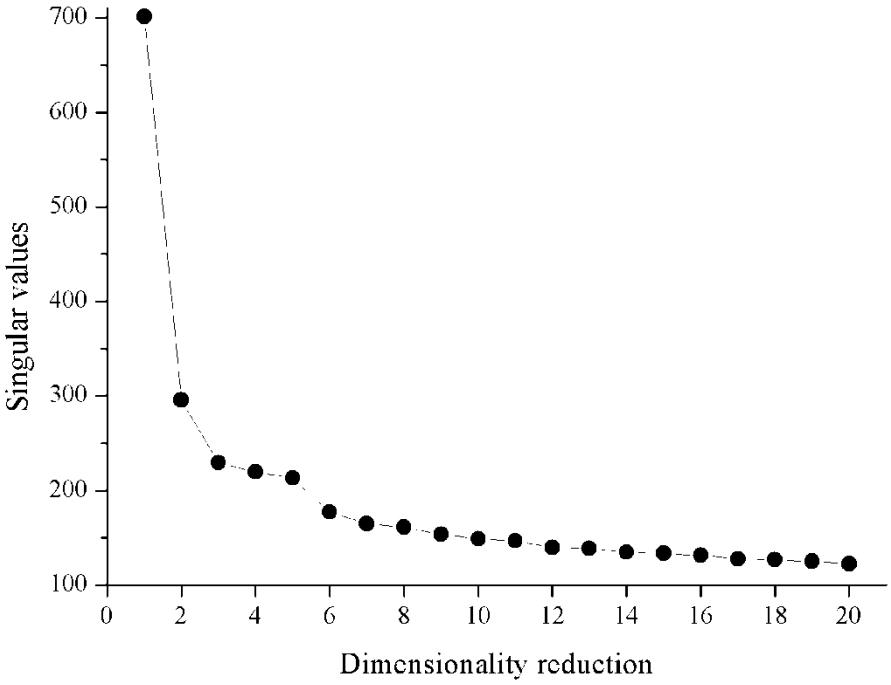

Fig. 2. Scree plot of the singular values

The LSI parameters are given in Tab. 1. In each figure nine images are arranged according to their decreasing similarity. The first image with the similarity equals to 1 always represents the self-retrieval in the image database.

Tab 1. Image retrieval using the partial eigenproblem: properties of the document matrix and the partial eigen-problem parameters

Properties of the document matrix *A*	
Number of keywords: Number of documents: Size in memory:	320×240= 76 800 232 135.93 MB
Partial eigen-problem parameters	
Dim. of the original space	232
Dim. of the user defined space	$k = 8$
Time for A^TA operation	1.297 seconds
Results of the eigensolver	0.656 seconds
Total time	1.953 seconds

Fig. 3. Example of the partial eigen-problem based image retrieval results with the query of water-lily, k = 8. The query image is situated in the upper left corner (sim = 1)

In Fig. 3 the second image with the similarity of 0.91165 shows the same water-lily taken from other angle. There were only the two pictures of water-lilies in the whole image database: the first one was selected as the query and the second one was retrieved as the most similar picture. The shape similarity between the water-lily leafs and objects occurring in the following six photographs such as stones, leaves and flowers is probably the reason for their selection as the another most similar images. The ninth picture with the similarity of 0.7598 is thematically quite different from the others. Probably some leaf-like objects such as stones making up a mole, pipe discharge or reflexes on water surface were recognised by LSI and therefore this picture was added to the other similar ones.

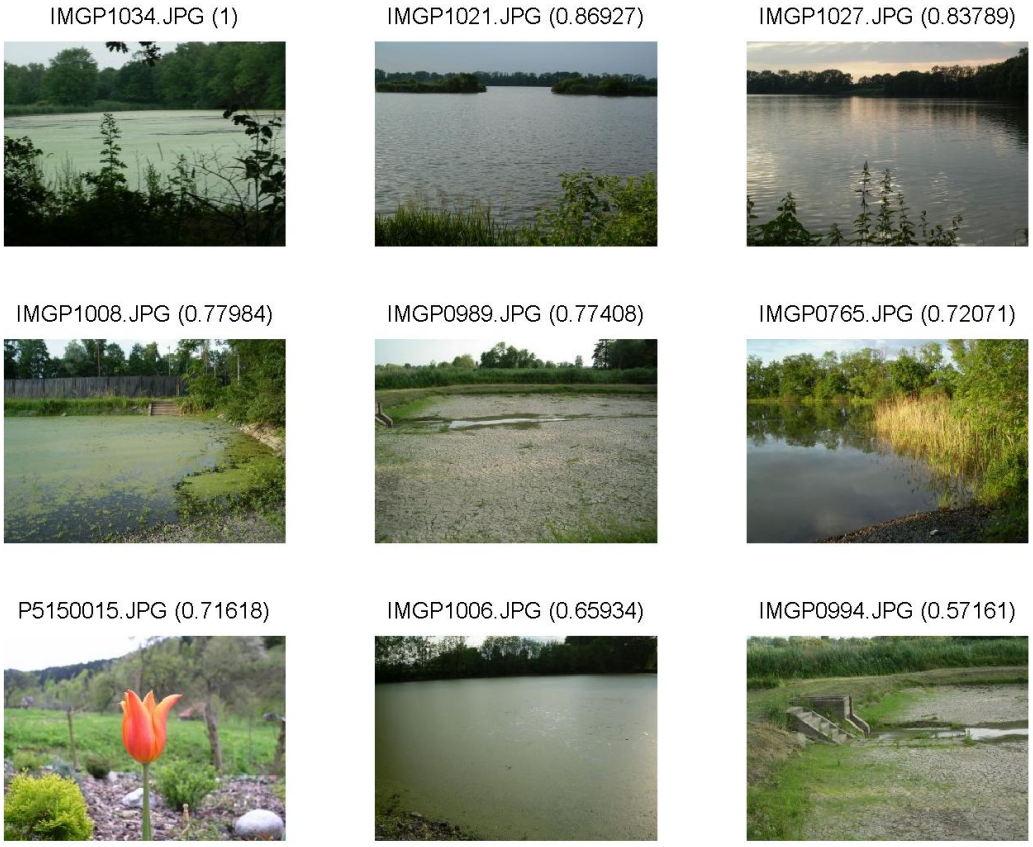

Fig. 4 Example of the partial eigen-problem based image retrieval results with the query of a pond, k = 8. The query image is situated in the upper left corner (sim = 1)

In case of the pond picture, the most similar images of other ponds including two emptied ones were found (Fig. 4). Even one picture of the scenery with a tulip was selected as the image similar to the pond. A common feature of all these images is their composition mostly containing a central oval, which represents the pond water surface or the pond bottom. The picture details such as the tulip were filtered out by SVD as information noise.

For comparison, the same LSI retrieval was obtained with the higher and lower dimensionality reduction at $k = 3$ and $k = 20$, respectively, as demonstrated in Fig. 5 and 6. It is obvious that the LSI results very little correspond to the query. It can be explained by the removal of useful information at $k = 3$ and, on the contrary, by the presence of confusing redundant information noise at $k = 20$. These facts indicate that the choice of the optimal dimensionality reduction k is the crucial problem for the correct LSI retrieval.

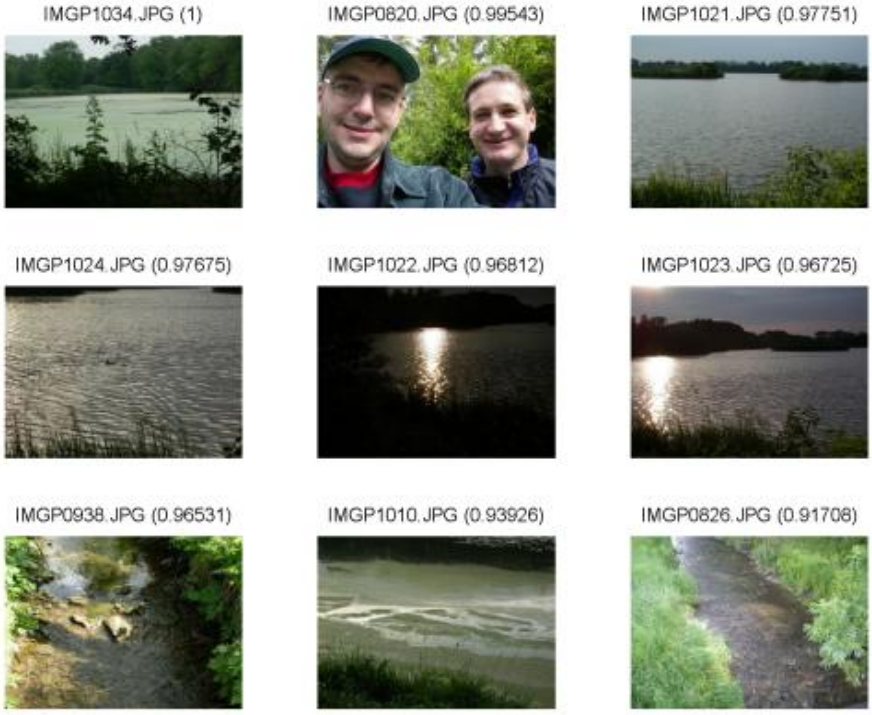

Fig. 5 Example of the partial eigen-problem based image retrieval results with the query of a pond, k = 3. The query image is situated in the upper left corner (sim = 1)

Fig. 6 Example of the partial eigen-problem based image retrieval results with the query of a pond, k = 20. The query image is situated in the upper left corner (sim = 1).

4 CONCLUSION

The latent semantic indexing approach was used for the retrieval of similar images from 232 photographs of natural or artificial water-related objects. LSI was based on the efficient symmetric partial eigenproblem and cosine similarity computation. For its correct performance the selection of the optimal image dimensionality reduction k is essential. There is no exact routine for this selection and thus we used the scree plot of the singular values. Using $k = 8$ the images displaying very similar objects were found. The behaviour of LSI is close to the classification by human experts. The results in Tab. 1 indicate the possibility of real-time analysis. We also chose $k = 8$ for the large-scale NIST TRECVID 2006 video data (Wilkins et al., 2006). In general, it seems that $k < 10$ is suitable for the intelligent image retrieval.

The obtained results of LSI indicate that this technique could be used for the environmental monitoring based on the automatic computer processing of large landscape photographs databases. In the next research, LSI will be tested for environmental pollution monitoring, control of continual technological processes, etc.

ACKNOWLEDGEMENT

The support of the Ministry of Education, Youth and Sports of the Czech Republic (1M06047–CQR.CZ) is gratefully acknowledged. The work leading to this contribution has been partially supported by the European Commission under the contract FP6-027026-K-SPACE.

REFERENCES

[1] Barnard K., Duygulu P., de Freitas N., Forsyth D., Blei D., Jordan M. I. *Matching Words and Pictures. Journal of Machine Learning Research.* 3, 2003, 1107-1135.

[2] Berry W.M., Dumais S.T., O'Brien G.W. *Using linear algebra for intelligent information retrieval*. SIAM Review, 37 (4), 1995, 573-595.

[3] Berry W.M., Drmač Z., Jessup J.R. Matrices, *Vector Spaces, and Information Retrieval. SIAM Review*, 41 (2), 1999, 336-362.

[4] Cattel R.D. The scree test for the number of factors. Multivariate Behaviour Research. 1, 1966, 245-276.

[5] Dvořák P., Střižík M., Praks P., Pudil P., Šumpíková M., Lešetický O. *The feasibility of using special quantitative methods for prediction of currency crises*, Proc. of the International Scientific Conference of European Finance – Theory, Politics And Practice. Banská Bystrica, September 8–9, 2004, 1-26.

[6] Grossman D.A., Frieder O. *Information retrieval: Algorithms and heuristics.* Kluwer Academic Publishers (second edition), 2000.

[7] 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*. Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence; Campiegne, Sep. 19-22, 2005, 401-404.

[8] Praks P., Dvorský J., Snášel V.. *Latent Semantic Indexing for Image Retrieval Systems". Proc. of the SIAM Conference on Applied Linear Algebra,* Williamsburg, July 15-19, 2003. http://www.siam.org/meetings/la03/proceedings/Dvorsky.pdf (accessed April 23, 2009)

[9] Praks P., Dvorský J., Snášel V., Černohorský J. *On SVD-free Latent Semantic Indexing for Image Retrieval for application in a hard industrial environment*. Proceedings of the IEEE International Conference on Industrial Technology – ICIT 2003; Maribor, Dec. 10-12, 2003, 466-471.

[10] Praks P., Machala L., Snášel V. *On SVD-free Latent Semantic Indexing for iris recognition of large databases.* In V. A. Petrushin and L. Khan, Multimedia Data Mining and Knowledge Discovery, Springer, 2006a.

[11] Praks P., Machala L., Snášel V., Střižík M.: *Intelligent Information retrieval using numerical linear algebra and applications*. International Symposium GIS Ostrava 2006 Proceedings., Ostrava, Jan 23-25,2006b,1-13. http://gis.vsb.cz/GISEngl/Conferences/GIS_Ova/GIS_Ova_2006/Proceedings/Referaty/praks.pdf (accessed April 23, 2009)

[12] Praus P., Praks P. *Information Retrieval in Hydrochemical Data Using the Latent Semantic Indexing Approach. Journal of Hydroinformatics*, 9 (2), 2007, 135-143.

[13] Wilkins P., Adamek T., Ferguson P., Hughes M., Jones G., Keenan G., McGuinness K., Malobabic J., O'Connor N., Sadlier D., Smeaton A. F., Benmokhtar R., Dumont E., Huet B., Merialdo B., Spyrou E., Koumoulos G., Avrithis Y., R. Moerzinger., P. Schallauer., W. Bailer., Zhang Q., Piatrik T., Chandramouli K., Izquierdo E., Goldmann L., Haller M., Sikora T., Praks P., Urban J., Hilaire X. and Jose J. (2006). *K-Space at TRECVid 2006*. Proc. of the TRECVid 2006–Text Retrieval Conference, Maryland, 13-14 November, 1-12. http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/k-space.pdf (accessed April 23, 2009).

[14] Oliva A., Torralba A. *Modeling the shape of the scene: a holistic representation of the spatial envelope.* International Journal of Computer Vision. 42(3), 2001, 145-175.

[15] Snoek C. (2005). *The Authoring Metaphor to Machine Understanding of Multimedia.* . Ph.D. Dissertation, University of Amsterdam. http://staff.science.uva.nl/~cgmsnoek/pub/snoek-thesis.pdf (accessed April 23, 2009)

RESUMÉ

Metoda Latentní sémantické indexace (LSI) se úspěšně používá při vyhledávání podobností v rozsáhlých textových dokumentech. Byla využita i při analýze ekonomických a hydrochemických dat a při vyhledávání podobných obrázků v různých multimediálních databázích.

V této práci byla LSI testována pro vyhledávání v databázi ekologicky zaměřených fotografií. Cílem bylo identifikovat anomálie v podobě změn na fotografiích stejných nebo podobných objektů. V praxi by takto mohly být identifikovány změny způsobené např. ekologickou havárií nebo znečištěním životního prostředí. V případě, že by havárii potvrdila i lidská kontrola, následovalo by její standardní řešení. Současné informační technologie umožňují rychlý přenos velkých objemů dat na dlouhé vzdálenosti a jejich automatické zpracování. Na rozdíl od běžného monitoringu založeného na pravidelném odběru a analýze vzorků, by touto screeningovou kontrolou bylo ušetřeno značné množství času a finančních prostředků.

LSI využívá standardních metod lineární algebry. Z pixelů každé fotografie se tvoří vektor a vektory všech fotografií tvoří matici, jejíž dimensionalita se redukuje metodou singulárního rozkladu (Singular Value Decomposition). Redukcí se z datové matice odstraňují redundantní informace (šum). Podobnost dvou fotografií resp. podobnost dvou redukovaných vektorů, které tyto fotografie reprezentují, modelujeme kosinovou podobností, kterou lze chápat jako korelační koeficient: čím větší je velikost koeficientu podobnosti, tím jsou si vektory a tedy i příslušné fotografie podobnější a naopak.

Testovaná databáze obsahovala 232 digitálních fotografií krajiny a přírodních nebo uměle vytvořených objektů ve vztahu k vodnímu hospodářství: fotografie hladin vodních nádrží, řek, potoků, mostních konstrukcí, výustí, hladin aktivačních nádrží biologických čistíren odpadních vod apod. Z výsledků práce je zřejmé, že rozhodující pro správné přiřazování podobných fotografií je míra redukce na vektory, kterou lze odhadnou pomocí grafu „scree plot“ často používaného v analýze hlavních komponent. Je-li redukce dat příliš velká, dochází k odstranění důležitých informací a naopak při menší redukci obsahují redukované vektory nadbytečný informační šum, který způsobuje nesprávné výsledky vyhledávání. Numerické experimenty ukazují, že metodu LSI lze využít také pro zpracování rozsáhlých databází obrázků. V další fázi testování může být proto metoda LSI využita pro analýzu kontinuálně snímaných fotografií (tzv. video surveillance) z určitých lokalit v přírodě, provozu technologického zařízení apod.