

HITS is PCA

Marco Saerens & Francois Fouss*

October 29, 2004

Abstract

In this short paper, we show that Kleinberg's hubs and authorities model (HITS) is simply Principal Components Analysis (PCA – maybe the most widely used multivariate statistical analysis method), albeit without centering, applied to the adjacency matrix of the graph of web pages. In addition to providing a new interpretation for HITS, this result suggests to rely on existing work, already published in the multivariate statistical analysis literature (extensions of PCA), in order to design new web pages ranking procedures.

1. Introduction

Exploiting the graph structure of large document repositories, such as the web environment, is one of the main challenges of computer science and data mining today. In this respect, Kleinberg's proposition to distinguish web pages that are hubs and authorities (see [8]; called the HITS algorithm) has been well-received in the community (see [1] for a review).

In this short paper, we show that Kleinberg's HITS procedure [8] is simply a Principal Component Analysis (PCA), albeit without centering, applied to the adjacency matrix of web pages. Although some links between HITS and multivariate statistics were highlighted in [3], [4], and PCA has often been mentioned as being an alternative to HITS, to our knowledge, nobody has shown the total equivalence between HITS and the uncentered version of PCA.

In section 2, we briefly introduce the basics of Kleinberg's procedure for ranking query results. In section 3, we introduce principal components analysis and relate it to Kleinberg's procedure.

2. Kleinberg's HITS procedure

In [8], Kleinberg introduced a procedure for identifying web pages that are good hubs or good authorities, in response to a given query.

To identify good hubs and authorities, Kleinberg's procedure exploits the graph structure of the web. Each web page is a node and a link from page a to page b is

*Marco Saerens and François Fouss are with the ISYS Unit (Information Systems Research Unit), IAG, Université catholique de Louvain, Place des Doyens 1, B-1348 Louvain-la-Neuve, Belgium. Email: {saerens, fouss}@isys.ucl.ac.be.

represented by a directed edge from node a to node b . When introducing a query, the procedure first constructs a focused subgraph G , and then computes hubs and authorities scores for each node of G (say n nodes in total). We now briefly describe how these scores are computed. Let $\mathbf{W}$ be the adjacency matrix of the directed subgraph G ; that is, element w_{ij} (row i , column j) of matrix $\mathbf{W}$ is equal to 1 if and only if node (web page) i contains a link to node (web page) j ; otherwise, $w_{ij} = 0$. We respectively denote by $\mathbf{x}^h$ and $\mathbf{x}^a$ the hubs and authorities $n \times 1$ column vectors containing the hubs and authorities scores corresponding to each node of the subgraph.

Kleinberg uses an iterative updating rule in order to compute these scores: The hub score for node i , x_i^h , is set equal to the normalized sum of the authority scores of all nodes pointed by i and, similarly, the authority score of node j , x_j^a , is set equal to the normalized sum of the hub scores of all nodes pointing to j . This corresponds to the following updating rule:

$$\mathbf{x}^h(k+1) = \frac{\mathbf{W}\mathbf{x}^a(k)}{\|\mathbf{W}\mathbf{x}^a(k)\|_2} \quad (2.1)$$

$$\mathbf{x}^a(k+1) = \frac{\mathbf{W}^T\mathbf{x}^h(k)}{\|\mathbf{W}^T\mathbf{x}^h(k)\|_2} \quad (2.2)$$

where $\|\mathbf{x}\|_2$ is the Euclidian norm, $\|\mathbf{x}\|_2 = (\mathbf{x}^T\mathbf{x})^{1/2}$.

Furthermore, Kleinberg [8] showed that when following this update rule, $\mathbf{x}^h$ converges to the normalized principal eigenvector of the symmetric matrix $\mathbf{W}\mathbf{W}^T$, while $\mathbf{x}^a$ converges to the normalized principal eigenvector of the symmetric matrix $\mathbf{W}^T\mathbf{W}$, provided that the eigenvalues are distinct. Indeed, the previous equations result from the application of the power method, an iterative numerical method for computing the dominant eigenvector of a symmetric matrix [5], to the following eigensystem problem:

$$\mathbf{x}^h = \lambda \mathbf{W}\mathbf{W}^T\mathbf{x}^h \quad (2.3)$$

$$\mathbf{x}^a = \lambda \mathbf{W}^T\mathbf{W}\mathbf{x}^a \quad (2.4)$$

3. Principal components analysis and HITS

Principal Components Analysis (PCA) is a standard multivariate statistical analysis technique aiming to analyse numerical data [9]. Imagine we have a $n \times n$ data matrix, $\mathbf{W}$, for which each cell, w_{ij} , contains the measurement of a numerical feature j taken on a record i . In our case, the records are web pages, and the features are web pages as well (the matrix is square). A measurement indicates how many links are pointing from page i (considered as a record) to page j (considered as a feature). We will consider that the data matrix, $\mathbf{W}$, is of full rank; in this case, the related matrix $\mathbf{W}^T\mathbf{W}$ is positive definite, is also of full rank, and has n distinct positive eigenvalues.

In the framework of PCA (see for instance, [9]), it is well-known that the k th *unit vector*, $\mathbf{v}_k$ (the eigenvectors are ordered by decreasing importance of corresponding eigenvalues), on which the data are projected in order to obtain the scores related to the k th principal component, is provided by the eigenvalue/eigenvector equation

$$\mathbf{W}^T\mathbf{W}\mathbf{v}_k = \lambda_k \mathbf{v}_k \quad (3.1)$$

where $\mathbf{W}^T\mathbf{W}$ is the sample “sums of squares and products” matrix. $\mathbf{v}_k$ is called the k th principal axis. Usually, the data matrix, $\mathbf{W}$, is centered; that is, the mean value of the corresponding column is subtracted to each column. However, in the case of HITS, no centering is computed; as will be shown in the next paragraph, this is the only difference between HITS and standard PCA.

Here, we exploit the duality relations of PCA in order to prove the equivalence between HITS and uncentered PCA. The developments are largely inspired by [2] and [6]. By pre-multiplying the equation (3.1) by $\mathbf{W}$, we obtain

$$\mathbf{W}\mathbf{W}^T(\mathbf{W}\mathbf{v}_k) = \lambda_k(\mathbf{W}\mathbf{v}_k) \quad (3.2)$$

Thus, each eigenvector $\mathbf{v}_k$ of $\mathbf{W}^T\mathbf{W}$ corresponds to an eigenvector $\mathbf{W}\mathbf{v}_k$ of the matrix $\mathbf{W}\mathbf{W}^T$, associated to the same eigenvalue λ_k . If we denote by $\mathbf{u}_k$ the corresponding *unit eigenvector* ($\|\mathbf{u}_k\| = 1$) of $\mathbf{W}\mathbf{W}^T$ (that is, $\mathbf{W}\mathbf{W}^T\mathbf{u}_k = \lambda_k\mathbf{u}_k$ and $\|\mathbf{u}_k\| = 1$), from Equation 3.2, we must have $\mathbf{W}\mathbf{v}_k = c\mathbf{u}_k$, where c is some constant. Since $\mathbf{u}_k$ is a unit vector, we have $1 = \mathbf{u}_k^T\mathbf{u}_k = c^{-2}\mathbf{v}_k^T\mathbf{W}^T\mathbf{W}\mathbf{v}_k = c^{-2}\lambda_k$; hence $c = \sqrt{\lambda_k}$.

Now, we easily observe that the column vector $\mathbf{W}\mathbf{v}_k$ precisely represents *the projection of the data (the records) on the k th principal axis*, $\mathbf{v}_k$, and therefore contains the coordinate k of the records (the web pages) in the principal components coordinate system. Thus, the k th eigenvector of $\mathbf{W}\mathbf{W}^T$ contains precisely *the scores of the records (the web pages) on the k th principal component*, up to a scaling factor. Since equation (3.2) is exactly the same as equation (2.3), the hub scores $\mathbf{x}^h = \mathbf{W}\mathbf{v}_k$, are simply the principal components scores of the uncentered data matrix $\mathbf{W}$ (up to a scaling factor).

Now, exactly the same reasoning applies to the dual data matrix $\mathbf{W}' = \mathbf{W}^T$, where a measurement, w'_{ij} , now indicates how many links are pointing to page i (considered as a record) from page j (considered as a feature). The PCA eigensystem is $\mathbf{W}'^T\mathbf{W}'\mathbf{v}'_k = \mathbf{W}\mathbf{W}^T\mathbf{v}'_k = \lambda_k\mathbf{v}'_k$ and, as before, we find $\mathbf{x}^a = \mathbf{W}'^T\mathbf{v}'_k = \mathbf{W}'\mathbf{v}'_k$, which precisely represents *the projection of the data on the k th principal axis*, $\mathbf{v}'_k$, and therefore contains the coordinate k of the records in the principal components coordinate system. Thus, the authorities scores, $\mathbf{x}^a$, are simply the principal components scores of the uncentered data matrix $\mathbf{W}' = \mathbf{W}^T$ (up to a scaling factor).

4. Conclusion

We showed that Kleinberg’s method for computing hubs and authorities scores is closely related to principal components analysis, maybe the most widely used multivariate statistical analysis method. This provides a new interpretation of Kleinberg’s method in terms of percentage of explained variance, etc. It also suggests to have a close look to existing work already published in the multivariate statistical analysis literature and, in particular, extensions of principal components analysis [7], in order to design new algorithms for web pages scoring.

References

- [1] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. John Wiley and Sons, 2003.
- [2] T. Cox and M. Cox. *Multidimensional scaling, 2nd ed.* Chapman and Hall, 2001.
- [3] F. Fouss, J.-M. Renders, and M. Saerens. Links between kleinberg’s hubs and authorities, correspondence analysis, and markov chains. *Proceedings of the third IEEE International Conference on Data Mining*, pages 521–524, 2003.
- [4] F. Fouss, J.-M. Renders, and M. Saerens. Some relationships between kleinberg’s hubs and authorities, correspondence analysis, and the salsa algorithm. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data (JADT 2004)*, pages 445–455, 2004.
- [5] G. H. Golub and C. F. V. Loan. *Matrix Computations, 3th Ed.* The Johns Hopkins University Press, 1996.
- [6] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer-Verlag, 2003.
- [7] I. Jolliffe. *Principal components analysis, 2th ed.* Springer-Verlag, 2002.
- [8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [9] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

Acknowledgments

We thank Prof. Vincent Blondel, from the “Département d’Ingénierie Mathématique” of the Université catholique de Louvain, for insightful discussions.
