Some relationships between Kleinberg's hubs and authorities, correspondence analysis, and the Salsa algorithm

François Fouss[1], Jean-Michel Renders[2], Marco Saerens[1]

[1]ISYS Unit, IAG – Université catholique de Louvain – Place des Doyens 1 – 1348 Louvain-la-Neuve – Belgique
[2]Xerox Research Center Europe, Chemin de Maupertuis 6 – 38240 Meylan (Grenoble) – France

{saerens,fouss}@isys.ucl.ac.be, jean-michel.renders@xrce.xerox.com

Abstract

In this work, we show that Kleinberg's hubs and authorities model is closely related to both correspondence analysis, a well-known multivariate statistical technique, and a particular Markov chain model of navigation through the web. The only difference between correspondence analysis and Kleinberg's method is the use of the **average** value of the hubs (authorities) scores for computing the authorities (hubs) scores, instead of the **sum** for Kleinberg's method. We also show that correspondence analysis and our Markov model are related to SALSA, a variant of Kleinberg's model. We finally suggest that the Markov model could easily be extended to the analysis of more general structures, such as relational databases.

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 Kleinberg, 1999); called the HITS algorithm) has been well-received in the community.

In this paper, we show that Kleinberg's hubs and authorities procedure (Kleinberg, 1999) is closely related to both correspondence analysis (see for instance Greenacre, 1984 ; Lebart *et al.*, 1995), a well-known multivariate statistical analysis technique, and a particular Markov chain model of navigation through the web. We further show that correspondence analysis and the Markov model are related to SALSA (Lempel and Moran, 2001), a variant of Kleinberg's model. This puts new lights on the interpretation of Kleinberg's procedure since correspondence analysis has a number of interesting properties that makes it well suited for the analysis of frequency tables. On the other hand, the Markov model can easily be extended to more general structures, such as relational databases

In section 2, we briefly introduce the basics of Kleinberg's procedure for ranking query's results. In section 3, we introduce correspondence analysis and relate it to Kleinberg's procedure while, in section 4, we introduce a Markov model of web navigation and relate it to both correspondence analysis and Kleinberg's procedure.

2. Kleinberg's procedure

In 1999, Kleinberg introduced a procedure for identifying web pages that are good hubs or good authorities, in response to a given query. The following example is often mentioned.

When considering the query "automobile makers", the home pages of Ford, Toyota and other car makers are considered as good authorities, while web pages that list these home pages are good hubs.

2.1. *The updating rule*

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 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. Let n be the number of nodes of G. We now briefly describe how these scores are computed. Let $\mathbf{W}$ be the adjacency matrix of the 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 vector scores corresponding to each node of the subgraph.

Kleinberg uses an iterative updating rule in order to compute the scores. Initial scores at $k = 0$ are all set to 1, i.e. $\mathbf{x}^h = \mathbf{x}^a = \mathbf{1}$ where $\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}}$ is a $n \times 1$ column vector made of 1. Then, the following mutually reinforcing rule is used: 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 hub scores of all nodes pointing to j. This corresponds to the following updating rule:

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

$$\mathbf{x}^a(k+1) = \frac{\mathbf{W}^{\mathrm{T}}\mathbf{x}^h(k)}{\|\mathbf{W}^{\mathrm{T}}\mathbf{x}^h(k)\|_2} \quad (2)$$

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

2.2. *An eigenvalue/eigenvector problem*

Kleinberg (1999) showed that when following this update rule, $\mathbf{x}^h$ converges to the normalized principal (or dominant) right eigenvector of the symmetric matrix $\mathbf{W}\mathbf{W}^{\mathrm{T}}$, while $\mathbf{x}^a$ converges to the normalized principal eigenvector of the symmetric matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$, provided that the eigenvalues are distinct.

Indeed, the equations (1), (2) result from the application of the power method, an iterative numerical method for computing the dominant eigenvector of a symmetric matrix (Golub and Loan, 1996), to the following eigenvalue problem:

$$\mathbf{x}^h \propto \mathbf{W}\mathbf{x}^a \Rightarrow \mathbf{x}^h = \mu\mathbf{W}\mathbf{x}^a \quad (3)$$

$$\mathbf{x}^a \propto \mathbf{W}^{\mathrm{T}}\mathbf{x}^h \Rightarrow \mathbf{x}^a = \eta\mathbf{W}^{\mathrm{T}}\mathbf{x}^h \quad (4)$$

where $\propto$ means "proportional to" and T is the matrix transpose. Or, equivalently,

$$x_i^h \propto \sum_{j=1}^{n} w_{ij}x_j^a \Rightarrow x_i^h = \mu\sum_{j=1}^{n} w_{ij}x_j^a \quad (5)$$

$$x_j^a \propto \sum_{i=1}^{n} w_{ij}x_i^h \Rightarrow x_j^a = \eta\sum_{i=1}^{n} w_{ij}x_i^h \quad (6)$$

Meaning that each hub node, i, is given a score, x_i^h, that is proportional to the sum of the authorities nodes scores to which it links to. Symmetrically, to each authorities node, j, we allocate a score, x_j^a, which is proportional to the sum of the hubs nodes scores that point to it. By substituting (3) in (4) and vice-versa, we easily obtain

$$\begin{aligned} \mathbf{x}^h &= \mu\eta\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}^h = \lambda\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}^h \\ \mathbf{x}^a &= \mu\eta\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}^a = \lambda\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}^a \end{aligned}$$

which is an eigenvalue/eigenvector problem.

2.3. *A variant of Kleinberg's procedure*

Many extensions of the updating rules (1), (2) were proposed. For instance, in Lempel and Moran (2001) (the SALSA algorithm), the authors propose to normalise the matrices $\mathbf{W}$ and $\mathbf{W}^{\mathrm{T}}$ in (3) and (4) so that the new matrices verify $\mathbf{W}'\mathbf{1} = \mathbf{1}$ and $\left(\mathbf{W}^{\mathrm{T}}\right)'\mathbf{1} = \mathbf{1}$ (the sum of the elements of each row of $\mathbf{W}'$ and $\left(\mathbf{W}^{\mathrm{T}}\right)'$ is 1). In this case, (3) and (4) can be rewritten as

$$x_i^h \propto \sum_{j=1}^{n} w'_{ij}x_j^a = \frac{\sum_{j=1}^{n} w_{ij}x_j^a}{w_{i.}} \text{, where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{7}$$

$$x_j^a \propto \sum_{i=1}^{n} w'_{ij}x_i^h = \frac{\sum_{i=1}^{n} w_{ij}x_i^h}{w_{.j}} \text{, where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{8}$$

This normalization has the effect that nodes (web pages) having a large number of links are not privileged with respect to nodes having a small number of links. In the next section,we will show that this variant of Kleinberg's procedure is equivalent to correspondence analysis.

3. Correspondence analysis and Kleinberg's procedure

Correspondence analysis is a standard multivariate statistical analysis technique aiming to analyse frequency tables (Greenacre, 1984; Mardia *et al.*, 1979; Lebart *et al.*, 1995).

3.1. *Correspondence analysis*

Imagine that we have a table of frequencies, $\mathbf{W}$, for which each cell, w_{ij}, represents the number of cases having both values i for the row variable and j for the column variable (we simply use the term "value" for the discrete value taken by a categorical variable). In our case, the records are the directed edges; the row variable represents the index of the origin node of the edge (hubs) and the column variable the index of the end node of the edge (authorities).

Correspondence analysis associates a score to the values of each of these variables. These scores

relate the two variables by what is called a "**reciprocal averaging**" relation (Greenacre, 1984):

$$x_i^h \propto \frac{\sum_{j=1}^{n} w_{ij} x_j^a}{w_{i.}}, \text{ where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{9}$$

$$x_j^a \propto \frac{\sum_{i=1}^{n} w_{ij} x_i^h}{w_{.j}}, \text{ where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{10}$$

which is exactly the same as (7) and (8). This means that each hub node, i, is given a score, x_i^h, that is proportional to the average of the authorities nodes scores to which it links to. Symmetrically, to each authorities node, j, we allocate a score, x_j^a, which is proportional to the average of the hubs nodes scores that point to it.

3.2. *Links with Kleinberg's procedure*

Notice that (9) and (10) differ from (5) and (6) only by the fact that we use the **average value** in order to compute the scores, instead of the sum.

Now, by defining the diagonal matrix $\mathbf{D}^h = \mathrm{diag}(1/w_{i.})$ and $\mathbf{D}^a = \mathrm{diag}(1/w_{.j})$ containing the number of links, we can rewrite (9) and (10) in matrix form

$$\mathbf{x}^h \propto \mathbf{D}^h \mathbf{W} \mathbf{x}^a = \mu \mathbf{D}^h \mathbf{W} \mathbf{x}^a \tag{11}$$

$$\mathbf{x}^a \propto \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h = \eta \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h \tag{12}$$

In the language of correspondence analysis, the row vectors of $\mathbf{D}^h\mathbf{W}$ are the hub **profiles**, while the row vectors of $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ are the authorities **profiles**. These vectors sum to one.

Now, from (11), (12), we easily find

$$\mathbf{x}^h = \mu\eta \mathbf{D}^h \mathbf{W} \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h = \lambda \mathbf{D}^h \mathbf{W} \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h \tag{13}$$

$$\mathbf{x}^a = \mu\eta \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{D}^h \mathbf{W} \mathbf{x}^a = \lambda \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{D}^h \mathbf{W} \mathbf{x}^a \tag{14}$$

Correspondence analysis computes the subdominant right eigenvector of $\mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ and $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W}$. Indeed, it can be shown that the right principal eigenvector (the largest one) is a trivial one, $[1, 1, \ldots, 1]^{\mathrm{T}}$ with eigenvalue $\lambda = 1$ (all the other eigenvalues are positive and smaller that 1; see Greenacre, 1984) since the column values of $\mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ (respectively $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W}$) sum to one for each row. Therefore, correspondence analysis computes the second largest eigenvalue, called the **subdominant eigenvalue**, as well as the corresponding eigenvector.

In standard correspondence analysis, this subdominant right eigenvector has several interesting interpretations in terms of "optimal scaling", of the "best approximation" to the original matrix or of the linear combinations of the two sets of values that are "maximally correlated", etc. (see for instance Greenacre, 1984; Lebart *et al.*, 1995). With respect to this last interpretation, it can be shown that correspondance analysis computes $\max_{\mathbf{a},\mathbf{b}} \left[\mathrm{corr}\left(\mathbf{W}\mathbf{a}, \mathbf{W}^{\mathrm{T}}\mathbf{b}\right)\right]$.

The next eigenvectors can be computed as well; they are related to the proportion of chi-square computed on the original table of frequencies that can be explained by the m first eigenvectors. Correspondence analysis is therefore often considered as an "equivalent" of principal components analysis for frequency tables.

4. A Markov chain model of web navigation

We now introduce a Markov chain model of random web navigation that is equivalent to correspondence analysis, and therefore closely related to Kleinberg's procedure. It therefore provides a new interpretation for both correspondence analysis and Kleinberg's procedure. Notice that this random walk model is very similar to the one proposed in Lempel and Moran (2001) (for other random walk models, see also the PageRank system (Page *et al.*, 1998; Ng *et al.*, 2001), and has some interesting links with diffusion kernels (Kondor and Lafferty, 2002), or with a measure of similarity between graphs vertices that was proposed in Blondel and Senellart (2002), which we are currently investigating.

4.1. *Definition of the Markov chain*

We first define a Markov chain in the following way. We associate a state of the Markov chain to every hub and every authority node ($2n$ in total); we also define a random variable, $s(k)$, representing the state of the Markov model at time step k . Moreover, let s^h be the subset of states that are hubs and s^a be the subset of states that are authorities. We say that $s^h(k) = i$ (respectively $s^a(k) = i$) when the Markov chain is in the state corresponding to the i^{th} hub (authority) at time step k. As in Lempel and Moran (2001), we define a random walk on these states by the following single-step transition probabilities

$$\mathrm{P}(s^h(k+1) = i | s^a(k) = j) = \frac{w_{ij}}{w_{.j}} \text{ , where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{15}$$

$$\mathrm{P}(s^a(k+1) = j | s^h(k) = i) = \frac{w_{ij}}{w_{i.}} \text{ , where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{16}$$

$$\mathrm{P}(s^a(k+1) = j | s^a(k) = i) = \mathrm{P}(s^h(k+1) = j | s^h(k) = i) = 0 \text{ , for all } i, j \tag{17}$$

In other words, to any hub page, $s^h(k) = i$, we associate a non-zero probability of jumping to an authority page, $s^a(k+1) = j$, pointed by the hub page (equation 16), which is inversely proportional to the number of directed edges leaving $s^h(k) = i$. Symmetrically, to any authority page $s^a(k) = i$, we associate a non-zero probability of jumping to a hub page $s^h(k+1) = j$ pointing to the authority page (equation 15), which is inversely proportional to the number of directed edges pointing to $s^a(k) = i$. We suppose that the Markov chain is irreducible, that is, every state can be reached from any other state. If this is not the case, the Markov chain can be decomposed into closed sets of states which are completely independent (there is no communication between them), each closed set being irreducible. In this situation, our analysis can be performed on these closed sets instead of the full Markov chain.

Now, if we denote the probability of being in a state by $x_i^h(k) = \mathrm{P}(s^h(k) = i)$ and $x_i^a(k) = \mathrm{P}(s^a(k) = i)$, and we define $\mathbf{P}^h$ as the transition matrix whose elements are $p_{ij}^h = \mathrm{P}(s^h(k+1) = j | s^a(k) = i)$ and $\mathbf{P}^a$ as the transition matrix whose elements are $p_{ij}^a = \mathrm{P}(s^a(k+1) = j | s^h(k) = i)$, from equations (15) and (16),

$$\begin{aligned} \mathbf{P}^h &= \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \\ \mathbf{P}^a &= \mathbf{D}^h \mathbf{W} \end{aligned}$$

The Markov model is characterized by

$$\begin{aligned} x_i^h(k+1) &= \mathrm{P}(s^h(k+1)=i) = \sum_{j=1}^{n} \mathrm{P}(s^h(k+1)=i|s^a(k)=j)\, x_j^a(k) \\ &= \sum_{j=1}^{n} p_{ji}^h\, x_j^a(k) \\ x_i^a(k+1) &= \mathrm{P}(s^a(k+1)=i) = \sum_{j=1}^{n} \mathrm{P}(s^a(k+1)=i|s^h(k)=j)\, x_j^h(k) \\ &= \sum_{j=1}^{n} p_{ji}^a\, x_j^h(k) \end{aligned}$$

Or, in matrix form,

$$\mathbf{x}^h(k+1) = (\mathbf{P}^h)^{\mathrm{T}}\mathbf{x}^a(k) \quad (18)$$
$$\mathbf{x}^a(k+1) = (\mathbf{P}^a)^{\mathrm{T}}\mathbf{x}^h(k) \quad (19)$$

It is easy to observe that the Markov chain is periodical with period 2: each hub (authority) state could potentially be reached in one jump from an authority (hub) state but certainly not from any other hub (authority) state. In this case, the set of hubs (authorities) corresponds to a subset which itself is an irreducible and aperiodic Markov chain whose evolution is given by

$$\mathbf{x}^h(k+2) = (\mathbf{P}^h)^{\mathrm{T}}(\mathbf{P}^a)^{\mathrm{T}}\mathbf{x}^h(k) = (\mathbf{Q}^h)^{\mathrm{T}}\mathbf{x}^h(k) \quad (20)$$
$$\mathbf{x}^a(k+2) = (\mathbf{P}^a)^{\mathrm{T}}(\mathbf{P}^h)^{\mathrm{T}}\mathbf{x}^a(k) = (\mathbf{Q}^a)^{\mathrm{T}}\mathbf{x}^a(k) \quad (21)$$

where $\mathbf{Q}^h$ and $\mathbf{Q}^a$ are the transition matrices of the corresponding Markov models for the hubs and authorities. This Markov chain is aperiodic since each link (corresponding to a transition) can be followed in both directions (from hub to authority and from authority to hub) so that, when starting from a state, we can always return to this state in two steps. Hence, all the diagonal elements of $\mathbf{Q}^h$ and $\mathbf{Q}^a$ are non-zero and the Markov chain is aperiodic.

Therefore, the transition matrices of the corresponding Markov chains are

$$\mathbf{Q}^h = \mathbf{P}^a\mathbf{P}^h = \mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}} \quad (22)$$
$$\mathbf{Q}^a = \mathbf{P}^h\mathbf{P}^a = \mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W} \quad (23)$$

The matrices appearing in these equations are equivalent to the ones appearing in (13), (14). We will now see that the subdominant right eigenvectors of these matrices (which are computed in correspondence analysis) have an interesting interpretation in terms of the distance to the "steady state" of the Markov chain.

4.2. *Interpretation of the subdominant right eigenvector of* $\mathbf{Q}^h$, $\mathbf{Q}^a$

Now, it is well-know that the subdominant right eigenvector of the transition matrix, $\mathbf{Q}$, of an irreducible, aperiodic, Markov chain measures the departure of each state from the "equilibrium position" or "steady-state" probability vector (see for instance Stewart, 1994; a proof is provided in appendix 6), π, which is given by the first left eigenvector of the transition matrix $\mathbf{Q}$:

$$\mathbf{Q}^{\mathrm{T}}\pi = \pi \text{ subject to } \sum_{i=1}^{n} \pi_i = 1 \tag{24}$$

with eigenvalue $\lambda = 1$. This principal left eigenvector, π, is unique and positive and is called the "steady state" vector. This "steady state" vector, π, is the probability of finding the Markov chain in state $s = i$ in the long-run behavior:

$$\lim_{k\to\infty} \mathrm{P}(s(k) = i) = \pi_i \tag{25}$$

and is independent of the initial distribution of states at $k = 0$. It represents the probability of being in a particular state in the long-run behaviour.

In appendix 6, we show that the elements of the subdominant right eigenvector of $\mathbf{Q} = \mathbf{Q}^h$ or $\mathbf{Q}^a$ can be interpreted as a "distance" from each state to its "steady-state" value, provided by π. The states of the Markov chain are often classified or clustered by means of the values of this subdominant eigenvector as well as the few next eigenvectors.

Notice that we compute the subdominant eigenvector because the right dominant eigenvector is $\mathbf{1}$, since the sums of the columns of $\mathbf{Q}$ is one ($\mathbf{Q1} = \mathbf{1}$), so that $\mathbf{1}$ is in fact the right dominant eigenvector of $\mathbf{Q}$ with eigenvalue 1. Indeed, from the theory of nonnegative matrices, this eigenvalue is simple and is strictly larger in magnitude than the remaining eigenvalues.

4.3. *Hubs and authorities scores*

Lempel and Moran (2001), in the SALSA algorithm, propose, as hubs and authorities scores, to compute the steady-state vectors, π^h and π^a, corresponding to the hubs matrix, $\mathbf{Q}^h$, and the authorities matrix, $\mathbf{Q}^a$. We propose instead or, more precisely, in addition, to use the subdominant right eigenvector, which produces the same results as correspondence analysis, and which is often used in order to characterise the states of the Markov chain.

Indeed, in the appendix, we show that the behavior of the Markov model can be expanded into a series of left/right eigenvectors (see equation (32)). The first term is related to the steady-state vector, while the second to the time needed to reach this steady state. Eventually, the next eigenvector/eigenvalue could be computed as well; it corresponds to second-order corrections. Of course, only experimental results could confirm this choice; we are currently performing experiments investigating this issue.

4.4. *Computation of the steady-state vector*

In Lempel and Moran (2001), it is shown that, in our case, the steady-state vectors π^h and π^a are given by

$$\pi_i^h = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = \frac{w_{i.}}{w_{..}} \tag{26}$$

$$\pi_j^a = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = \frac{w_{.j}}{w_{..}} \tag{27}$$

which are the marginal proportions of the frequency table $\mathbf{W}$. It is therefore not necessary to compute the eigenvector. This is in fact a standard result for a random walk on a graph (see for instance Ross, 1996). From the theory of finite-state irreducible and aperiodic Markov chains (Cinlar, 1975; Bremaud, 1999), we know that there exists exactly one limiting density obeying (26, 27), given by (24).

We saw that the subdominant eigenvector of $\mathbf{Q}^h$ and $\mathbf{Q}^a$ represents the departure from the marginal proportions of the frequency table. This is quite similar to correspondence analysis where the same eigenvectors are interpreted in terms of a chi-square distance to the "independence model", for which we assume $\mathrm{P}(s^a(k) = i, s^h(k) = j) = (w_{i.}w_{.j})/(w_{..}^2)$ (Greenacre, 1984; Lebart *et al.*, 1995).

Moreover, the markov chain is reversible (see Ross, 1996). This has important implications. For instance, it is known that all the eigenvalues of the transition matrix of a reversible Markov chain are real and distinct (see for instance Bremaud, 1999).

4.5. *The full Markov chain model*

Finally, notice that (18), (19) can be rewritten as

$$\begin{bmatrix} \mathbf{x}^h(k+1) \\ \mathbf{x}^a(k+1) \end{bmatrix} = \begin{bmatrix} 0 & (\mathbf{P}^h)^{\mathrm{T}} \\ (\mathbf{P}^a)^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(k) \\ \mathbf{x}^a(k) \end{bmatrix} \tag{28}$$

so that the transition matrix of the full Markov chain with state vector $\left[(\mathbf{x}^h)^{\mathrm{T}}, (\mathbf{x}^a)^{\mathrm{T}}\right]^{\mathrm{T}}$ is

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{P}^a \\ \mathbf{P}^h & 0 \end{bmatrix}$$

If $\mathbf{u}^h$ and $\mathbf{u}^a$ are respectively eigenvectors of $\mathbf{Q}^h$ and $\mathbf{Q}^a$, it is easy to show that the transition matrix $\mathbf{P}^2$ has eigenvectors $\left[(\mathbf{u}^h)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\left[\mathbf{0}^{\mathrm{T}}, (\mathbf{u}^a)^{\mathrm{T}}\right]^{\mathrm{T}}$.

5. Possible extensions of Kleinberg's model, based on correspondence analysis and the Markov chain model

The relationships between Kleinberg's procedure, correpondence analysis and Markov models suggest extensions of Kleinberg's HITS algorithm in three differents directions:

1. The proposed random walk procedure can easily be extended to the analysis of more complex structures, such as relational databases. Here is a sketch of the main idea. To each record of a table, we associate a state; each table corresponding to a subset of states. For each relation between two tables, we associate a set of transitions. This way, we can define a random walk model and compute interesting information from it, such as the distance to the steady state, the average first passage time from one state to another, etc, with the objective of computing similarities between states. This would allow us to compute similarities between records (states) of a same table as well as records (states) belonging to different tables of the database. These developments, as well as an experimental validation of the method, will be the subject of a forthcoming paper. In the same vein, extensions of Kleinberg's procedure were also proposed in Blondel and VanDooren (2002), with the objective of computing similarities between nodes of a graph. We also intend to generalize correspondence analysis to multiple correspondence analysis (correspondence analysis applied to multiple tables) by using the concept of stochastic complementation (Meyer, 1989).

2. Instead of computing only the first eigenvalue/eigenvector, we could additionally compute the next eigenvevalues/eigenvectors (as done in correspondence analysis) that could also provide additional information, as already suggested by Kleinberg himself. With this respect, notice that in correspondence analysis, each eigenvalue is directly related to the amount of chi-square accounted for the corresponding eigenvector. This way, we extract as many eigenvectors as there are significant eigenvalues.

3. The various interpretations of correspondence analysis put new light to Kleinberg's procedure. Moreover, extensions of correspondence analysis are available (for instance multiple correspondence analysis); these extensions could eventually be applied in order to analyse the graph structure of the web.

6. Conclusions and further work

We showed that Kleinberg's method for computing hubs and authorities scores is closely related to correspondence analysis, a well-known multivariate statistical analysis method. This allows us to give some new interpretations to Kleinberg's method. Also, this suggests to extract the next eigenvalues/eigenvectors as well, since these could provide some additional information (as is done in correspondence analysis).

We then introduce a Markov random walk model of navigation through the web, and we show that this model is also equivalent to correspondence analysis. This random walk procedure has an important advantage: it can easily be extended to more complex structures, such as relational databases. These developments will be the subject of a forthcoming paper. In the same vein, extensions of Kleinberg's procedure were also proposed in Blondel and VanDooren (2002), with the objective of computing similarities between nodes of a graph.

Acknowledgments

We thank Prof. Vincent Blondel, from the "Département d'Ingéniérie Mathematique" of the Université catholique de Louvain, Prof. Guy Latouche, Prof. Guy Louchart and Dr. Pascal Franck, both from the Université Libre de Bruxelles, for insightful discussions.

APPENDIX: PROOF OF THE MAIN RESULTS

Appendix: Distance to the steady state vector

In this appendix, we show that the entries of the subdominant right eigenvector of the transition matrix $\mathbf{Q}$ of a finite, aperiodic, irreducible, reversible, Markov chain can be interpreted as a distance to the "steady-state" probability vector, π. From (22), (23), we can easily show that $\mathbf{Q}$ has a positive real spectrum so that all its eigenvalues are positive real and its eigenvectors are real. Moreover, since $\mathbf{Q}$ is stochastic nonnegative, all the eigenvalues are ≤ 1, and the eigenvalue 1 has multiplicity one. The proof is adapted from Papoulis and Pillai, 2002; Stewart, 1994; Bremaud, 1999.

Let $\mathbf{e}_i = [0, 0, \ldots, 1, \ldots, 0, 0]^{\mathrm{T}}$ be the column vector with i^{th} component equal to 1, all others being equal to 0. $\mathbf{e}_i$ will denote that, initially, the system starts in state i. Since $\mathbf{Q}$ is aperiodic, irreducible, and reversible, we know that $\mathbf{Q}$ has n simple, real, distinct, nonzero eigenvalues.

After one time step, the probability density of finding the system in one state is (see (18),(19))

$$\mathbf{x}(1) = \mathbf{Q}^{\mathrm{T}}\mathbf{e}_i$$

After k steps, we have

$$\mathbf{x}(k) = \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i$$

The idea is to compute the distance

$$d_i(k) = \left\| \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i - \pi \right\|_2 \tag{29}$$

in order to have an idea of the rate of convergence to the steady state when starting from a particular state $s = i$.

Let $(\lambda_i, \mathbf{u}_i)$, $i = 1, 2, \ldots n$ represent the n right eigenvalue-eigenvectors pairs of $\mathbf{Q}$ in decreasing order of importance (of modulus). Thus

$$\mathbf{Q}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \tag{30}$$

where $\mathbf{U}$ is the $n \times n$ matrix made of the column vectors $\mathbf{u}_i$ which form a basis of $\Re^n$: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$.

From (30),

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} \tag{31}$$

where we set $\mathbf{V} = \mathbf{U}^{-1}$. We therefore obtain $\mathbf{V}\mathbf{Q} = \mathbf{\Lambda}\mathbf{V}$, where $\mathbf{V} = \mathbf{U}^{-1} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix}$, so that the column vectors $\mathbf{v}_i$ are the left eigenvectors of $\mathbf{Q}$, $\mathbf{v}_i^{\mathrm{T}}\mathbf{Q} = \lambda_i \mathbf{v}_i^{\mathrm{T}}$. Moreover, since $\mathbf{V}\mathbf{U} = \mathbf{I}$, we have $\mathbf{v}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}$.

Hence from (31),

$$\begin{aligned}
\mathbf{Q}^k &= \mathbf{U}\mathbf{\Lambda}^k\mathbf{V} \\
&= \sum_{i=1}^{n} \lambda_i^k \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \\
&= \mathbf{1}\pi^{\mathrm{T}} + \sum_{i=2}^{n} \lambda_i^k \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \\
&= \mathbf{1}\pi^{\mathrm{T}} + \lambda_2^k \mathbf{u}_2 \mathbf{v}_2^{\mathrm{T}} + O\left((n-2)\lambda_3^k\right)
\end{aligned} \tag{32}$$

since $\lambda_i < 1$ for $i > 1$ and the eigenvalues/eigenvectors are ordered in decreasing order of eigenvalue magnitude. This development of the transition matrix is often called the spectral decomposition of $\mathbf{Q}^k$.

Let us now return to (29)

$$\begin{aligned}
d_i(k) &= \left\| \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \left\| \left(\pi\mathbf{1}^{\mathrm{T}} + \lambda_2^k \mathbf{v}_2 \mathbf{u}_2^{\mathrm{T}}\right) \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \left\| \pi + \lambda_2^k \mathbf{v}_2 \mathbf{u}_2{}^{\mathrm{T}} \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \lambda_2^k \left\| \mathbf{v}_2 \right\|_2 u_{2i}
\end{aligned}$$

where u_{2i} is i^{th} component of $\mathbf{u}_2$. Since the only term that depends on the initial state, i, is u_{2i}, the eigenvector $\mathbf{u}_2$ can be interpreted as a distance to the steady-state vector.

References

Blondel V.D. and Dooren P.V. (2002). A measure of similarity between graph vertices, with application to synonym extraction and web searching. *Technical Report UCL 02-50*. Université catholique de Louvain.

Blondel V.D. and Senellart P.P. (2002). Automatic extraction of synonyms in a dictionary. In *Proceedings of the SIAM Text Mining Workshop*, Airlington, Virginia.

Bremaud P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag.

Cinlar E. (1975). *Introduction to Stochastic Processes*. Prentice-Hall.

Golub G.H. and Loan C.F.V. (1996). *Matrix Computations* (3th Ed.). The Johns Hopkins University Press.

Greenacre M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press.

Kleinberg J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. (46/5): 604-632.

Kondor R.I. and Lafferty J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*.

Lebart L., Morineau A. and P.M. (1995). *Statistique Exploratoire Multidimensionnelle*. Dunod.

Lempel R. and Moran S. (2001). Salsa: The stochastic approach for link-structure analysis. In *ACM Transactions on Information Systems*, vol. (19/2): 131-160.

Mardia K., Kent J. and Bibby J. (1979). *Multivariate Analysis*. Academic Press.

Meyer C. (1989). Stochastic complementation, uncoupling markov chains, and the theory of nearly reducible systems. *SIAM Review*, vol. (31): 240-272.

Ng A.Y., Zheng A.X. and Jordan M.I. (2001). Link analysis, eigenvectors and stability. In *International Joint Conference on Artificial Intelligence (IJCAI-01)*.

Page L., Brin S., Motwani R. and Winograd T. (1998). The pagerank citation ranking: Bringing order to the web. *Technical Report, Computer System Laboratory*. Stanford University.

Papoulis A. and Pillai S.U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Ross S. (1996). *Stochastic Processes* (2nd Ed.). Wiley.

Stewart W.J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.