PageRank and Interaction Information Retrieval

Sándor Dominich

Department of Computer Science, University of Veszprém

8200 Veszprém, Egyetem u. 10, Hungary,

Tel.: +36 88 422022/4711

Email: dominich@dcs.vein.hu

ABSTRACT

The PageRank method is used by the Google Web search engine in computing the importance of Web pages. Two different views have been developed for the interpretation of the PageRank method and values: (i) stochastic (random surfer): the PageRank values can be conceived as the steady state distribution of a Markov chain, and (ii) algebraic: the PageRank values form the eigenvector corresponding to eigenvalue 1 of the Web link matrix. The Interaction Information Retrieval (I^2R) method is a non-classical information retrieval paradigm, which represents a connectionist approach based on dynamic systems. In the present paper, a different interpretation of PageRank is proposed, namely a dynamic systems viewpoint, by showing that the PageRank method can be formally interpreted as a particular case of the Interaction Information Retrieval method; and thus, the PageRank values may be interpreted as neutral equilibrium points of the Web.

1 INTRODUCTION

The PageRank method (Brin and Page, 1998) is an important component of the Google Web Information Retrieval (IR) engine (Google), and triggered research into and application of link analysis to the Web. It allows the calculation of a priori importance measures for Web pages. The measures are computed offline, and are independent of the search query. The PageRank values form the eigenvector corresponding to eigenvalue 1 of the Web matrix. At query time, the measures are combined with query-specific scores used for ranking Web pages.

The PageRank method uses the metaphor of an easily bored Web surfer. The PageRank value of a page is conceived as being the probability that the surfer reaches that page by following forward links. The PageRank values form a probability distribution over the Web. This stochastic approach is treated within the wider framework of Markov chains in (Breyer, 2002), where many interesting properties are established and discussed, and based on which techniques are suggested to enhance the PageRank computation. Also, the view based on link analysis was used in enhancing the indexing of Web pages (Henzinger et al., 1999).

Two different interpretations have been developed for PageRank thus far: stochastic (random surfer view) and algebraic (eigenvector of eigenvalue 1).

In the present paper, a different view on PageRank is proposed. It is shown that it can be formally conceived as a particular case of the Interaction Information Retrieval (I^2R) (Dominich, 1994) method. Thus, PageRank can be looked at from a connectionist dynamic systems standpoint, too, which may open up new perspectives (and thus perhaps on methods based on link analysis in general).

2 PAGERANK

The Google search engine exploits the citation graph of the crawled portion of the publicly accessible World Wide Web (briefly Web), and calculates a measure of the relative importance of Web pages using a stochastic process view.

2.1 Place of PageRank in Google's Retrieval and Ranking

The retrieval and ranking of Web pages follows an usual IR scenario, and is performed in several steps as follows:

(a) Find the Web pages containing the query terms.

(b) Compute a relative importance of Web pages.

(c) Rank the Web pages according to their relative importance.

The relative importance of Web pages is calculated taking into account several factors such as:

— 'on page factors', i.e., terms occurring in title, anchor, body, proximity of terms,

— appearance of terms: small font, large font, colour,

— frequency of occurrence of terms,

— PageRank values,

— other factors.

Although not known publicly exactly, based on (Brin and Page, 1998) it can be assumed that two numeric vectors (for query and Web page) are defined, whose dot product gives an intermediate score for that Web page, which is then combined with the PageRank value of that Web page to obtain its final score (importance).

2.2 The PageRank Method

The PageRank method is considered to be one of the factors used by Google in computing the relative importance of Web pages. The PageRank value of a Web page depends on the PageRank values of pages pointing to it and on the number of links going out of these pages.

2.2.1 The Principle

The starting point for the principle of PageRank is citation analysis, which is concerned with the study of citation in the scientific literature, and dates back, in its present form, about fifty years (Garfield, 1955). The underlying idea — which is well-known (and has a tremendous literature) — of citation analysis reads as follows: citation counts are a measure of importance (Garfield, 1972). This idea was used in (Carriere and Kazman, 1997) for Web retrieval. In the PageRank method (Page et al., 1998), this idea is refined in that citation counts are not absolute values anymore, rather relative ones and mutually dependent (as will be seen below). The principle on which PageRank is based can thus be referred to as an extended citation principle, and can be formulated as follows: a Web page's importance is determined by the importance of Web pages linking to it.

2.2.2 The Model

In order to apply the extended citation principle in practice, an appropriate model of a system of Web pages has been constructed as follows. Let (Figure 1)

(i) $\Omega = \{W_1, W_2, \ldots, W_i, \ldots, W_N\}$ denote a set of Web pages under focus,

(ii) $\Phi_i = \{W_k \mid k = 1, \ldots, n_i\}$ denote the set of Web pages W_i points to, $\Phi_i \subseteq \Omega$,

(iii) $B_i = \{W_j \mid j = 1, \ldots, m_i\}$ denote the set of Web pages that point to W_i, $B_i \subseteq \Omega$,

It can be seen that this view models the Web as a directed graph denoted by, say, *G*.

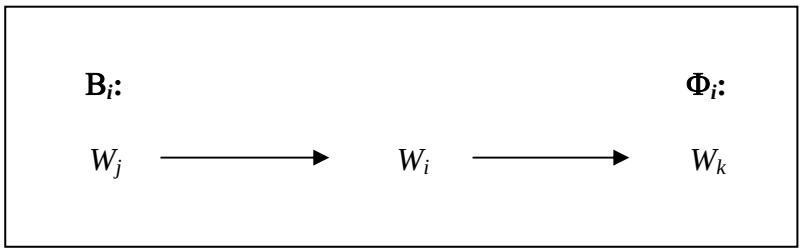


Figure 1. Web pages and links are viewed as a graph *G* in PageRank, where pages W_1, W_2, …, W_i, …, W_N are its nodes. $B_i = \{W_j \mid j = 1, \ldots, m_i\}$ denotes the set of Web pages that point to W_i, $\Phi_i = \{W_k \mid k = 1, \ldots, n_i\}$denotes the set of Web pages W_i points to.

2.2.3 The Equation

Based on the extended citation principle and using the graph model, Haveliwala (1999) and Page (2001) define the PageRank value of a Web page W_i, denoted by R_i, using the following equation:

$$R_i = \sum_{W_j \in B_i} \frac{R_j}{L_j} \qquad (1)$$

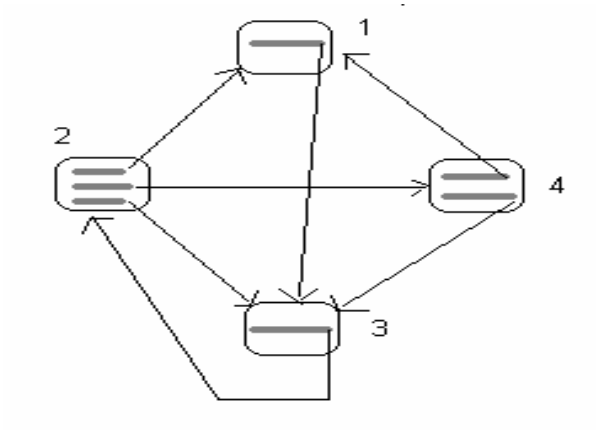
where L_j denotes the number of outgoing links from the page W_j. Equation 1 is a homogenous and simultaneous system of linear equations, which, as it is well-known, always has trivial solutions (the null vector), but which has nontrivial solutions too if and only if its determinant is equal to zero.

Another commonly used technical definition of PageRank is as follows. Let $G = (V, A)$ denote the directed graph of the Web, where the set $V = \{W_1, W_2, \ldots, W_i, \ldots, W_N\}$ of vertices denotes the

set of Web pages. The set A of arcs denotes the links (given by URLs) between pages. Let $M = (m_{ij})_{N\times N}$ denote a square matrix attached to graph G such that $m_{ij} = {}^{1}\!/_{L_j}$ if there is a link from W_j to W_i, and 0 otherwise. Because the elements of the matrix M are the coefficients of the right hand side of equation 1, this can be re-written in a matrix form as $M \times R = R$, where R denotes the vector of PageRank values, i.e., $R = [R_1, \ldots, R_i, \ldots, R_N]^{\mathsf{T}}$.

If the graph G is strongly connected, the column sums of the matrix M are equal to 1 (stochastic matrix). Because the matrix M has only zeroes in the main diagonal, the matrix $M - I$ has zero column sums (I denotes the unity matrix); i.e., the matrix obtained by subtracting 1 from the main diagonal of the matrix M). Let D denote its determinant, i.e., $D = |M - I|$. If every element of, say, the first line of D is doubled we get a new determinant D', and we have $D' = 2D$. We add now, in D', every other line to the first line. Because the column sums in D are null, it follows that $D' = 2D = D$, from which we have $D = 0$. The matrix $M - I$ is exactly the matrix of equation 1, hence it has nontrivial solutions too (of which there are an infinity). The determinant $|M - I|$ being equal to 0 is equivalent to saying that the number 1 is an eigenvalue of the matrix M.

The PageRank values are computed in practice using some numeric approximation procedure to calculate the eigenvector R corresponding to the eigenvalue 1 or to solve equation 1. Computational details as well as convergence considerations in the numeric algorithms used are presented and discussed in (Arasu, 2002; Haveliwala, 1999, 2002; Kim and Lee, 2002). Figure 2 shows an example.



Matrix M PageRank

0	1/3	0	1/2		0.325
0	0	1	0		0.651
1	1/3	0	1/2		0.651
0	1/3	0	0		0.217

Figure 2. A small Web its graph *G* with four pages: 1, 2, 3 and 4. The bars within each page symbolise links to other pages as shown by the arrows. The elements of the matrix *M* are also shown, they were computed as follows: $m_{ij} = 1/L_j$ (see text or eq. 1). The PageRank values, i.e., the eigenvector corresponding to the eigenvalue 1, were computed using the Mathcad command 'eigenvec(*M*,1)'.

As it was initially suggested (Brin and Page, 1998), the PageRank value of a Web page can be interpreted using an easily bored random surfer metaphor: the surfer clicks on links at random with no regard towards content. The random surfer visits a web page with a certain probability which derives from the page's PageRank. The probability that the random surfer clicks on one

link is solely given by the number of links on that page. This is why one page's PageRank is not completely passed on to a page it links to, but is divided by the number of links on the page. The probability for the random surfer reaching one page is the sum of probabilities for the random surfer following links to this page; his probability is reduced by a damping factor *d*, set between 0 and 1. The justification is that the surfer does not click on an infinite number of links, but gets bored sometimes and jumps to another page at random. The higher *d* is, the more likely will the random surfer keep clicking links. Since the surfer jumps to another page at random after he stopped clicking links, the probability therefore is usually implemented as a constant (1–*d*) into the numeric approximation algorithm. A PageRank of zero means the page has no reputation at all. All PageRanks sum up to 1, since after *n* clicks, the surfer must be on exactly one web page. Thus, all the PageRank values may be interpreted as forming a probability distribution over the Web, so that the probabilities sum up to unity. The solutions of equation 1 can be so scaled as to satisfy this condition. Thus, in Figure 2, the values that comply (within an inherent numeric approximation error) with this probabilistic interpretation are as follows: 0.163, 0.326, 0.326, 0.109 (this is an eigenvector, too, and they are roots of equation 1 as well, and proportional to the eigenvector corresponding to eigenvalue 1). A stochastic view of PageRank based on Markov chains is detailed in (Breyer, 2002).

The existence of isolated vertices (Web pages without outgoing and incoming links), sink vertices (Web pages with incoming links but without outgoing links), and source vertices (Web pages with outgoing links but without incoming links) cannot be excluded in the real Web. Based on equation 1, a number of equations have been worked out and published to compute PageRank values (Brin et al., 1998; Brin and Page, 1998; Page et al., 1998) in order to cope with WWW realities (e.g., loops, sink pages, etc.). As an interesting formal aspect, we note that these versions can be written in a unified form introducing formal parameters and using just one equation as follows:

$$R_i = \alpha \cdot (1-d) + \beta \cdot d \cdot \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j} + \gamma \cdot E \tag{2}$$

such that different versions are obtained using the following values of the parameters α, β and γ as follows (Table 1):

Table 1. Table of parameters, their values, and different versions of PageRank equations used in practice as reported in (Brin et al., 1998; Brin and Page, 1998; Page et al., 1998).

α	β	γ	**Version**
0	1/*d*	0	$R_i = \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j}$
0	1	0	$R_i = d \cdot \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j}$
1	1	0	$R_i = (1-d) + d \cdot \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j}$
1/*N*	1	0	$R_i = \frac{1}{N} \cdot (1-d) + d \cdot \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j}$
0	*d*/*c*	1	$R_i = \frac{d}{c} \cdot d \cdot \sum_{W_j \in \mathrm{B}_i} \frac{R_j}{L_j} + E$

where $0 \le d \le 1$ is the damping factor (usually set to 0.85 in practice), N denotes the total number of Web pages, E is a parameter, and c is a normalisation factor. Note that, quite understandably, few details are given exactly as regards d, c and E.

The PageRank method has proved to be useful for other purposes, too, for example as an importance metric in crawling the Web pages (Chakrabarti et al., 1999; Chao et al., 2002).

3 INTERACTION INFORMATION RETRIEVAL

The Interaction Information Retrieval (I^2R) paradigm was proposed in (Dominich, 1994). I^2R exploits the changing nature of links between objects, and calculates their relative importance as activity levels. In (Dominich, 1997) a more detailed description of the theoretical and practical aspects of I^2R is given, whereas (Dominich, 2001) presents and treats I^2R within a wide formal context of IR models.

3.1 Retrieval and Ranking

The retrieval and ranking of objects follows an usual IR scenario, and is performed in several steps as follows:

(a) Find the object-documents containing terms of the object-query, and link the object-query with these object-documents.

(b) Compute the importance of objects.

(c) Rank the objects according to their importance.

The object-documents form an interconnected network of artificial neurons. The query is integrated into this network as any other object would. The importance of objects is expressed by their activity levels. Retrieval and ranking are based on activation spreading starting from object-query according to a winner-takes-all strategy. The object-documents belonging to reverberative circles are retrieved (local memories evoked by the object-query).

3.2 I^2R Method

The I^2R method uses a connectionist approach based on dynamical systems, and it provides a way to compute the relative importance of objects as activity levels based on a qualitative model from which quantitative (implementable, numeric) models can be obtained.

3.2.1 Qualitative Model of I^2R

The qualitative model of I^2R consists of a principle, model and generic equation (from which specific computational models can be derived).

3.2.1.1 The Principle

The underlying idea of ANNs (Artificial Neural Networks) goes back to (James, 1890), where it is stated, as a fundamental law or principle, that the amount of activity of any artificial neuron depends on its weighted input, on the activity levels of artificial neurons connecting to it, and on inhibitory mechanisms. This idea gave birth to a huge literature and many applications, especially due to the results obtained in, e.g., (Feldman and Ballard, 1982; Grossberg, 1976; Hopfield, 1984).

The qualitative model of I^2R is based on the above principle which it applies to IR. Inhibitory mechanisms are not assumed, and the principle of I^2R can be formulated as follows: the activity level of an object is determined by the activity levels of objects which are linked to it.

3.2.1.2 The Model

In order to apply the I^2R method in practice, objects (e.g., documents, Web pages) are assigned (or modelled as) artificial neurons, these form an ANN. Let (Figure 3)

(i) $\Delta = \{O_1, O_2, \ldots, O_i, \ldots, O_N\}$ denote a set of objects; each object O_i is assigned an artificial neuron $\aleph_i$, $i = 1, \ldots, N$; thus we may consider $\Delta = \{\aleph_1, \aleph_2, \ldots, \aleph_i, \ldots, \aleph_N\}$,

(ii) $\Phi_i = \{\aleph_k \mid k = 1, \ldots, n_i\}$ denote the set of artificial neurons that are being influenced (i.e., synapsed) by $\aleph_i$, $\Phi_i \subseteq \Delta$,

(iii) $B_i = \{\aleph_j \mid j = 1, \ldots, m_i\}$ denote the set of artificial neurons that influence (i.e., synapse to) $\aleph_i$, $B_i \subseteq \Delta$.

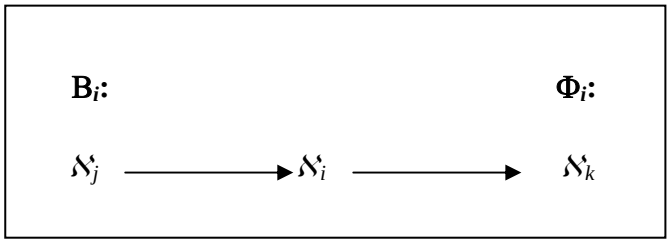


Figure 3. Objects and links as viewed in I^2R. $\aleph_1, \aleph_2, \ldots, \aleph_i, \ldots, \aleph_N$ form an artificial neural network, $\Phi_i = \{\aleph_k \mid k = 1, \ldots, n_i\}$ denotes the set of artificial neurons that are being influenced (i.e., synapsed) by $\aleph_i$, $B_i = \{\aleph_j \mid j = 1, \ldots, m_i\}$ denote the set of artificial neurons that influence (i.e., synapse to) $\aleph_i$.

3.2.1.3 The Equation

The equation of the I^2R method is based on the general network equation:

$$\mu_i \frac{dz_i(t)}{dt} = I_i(t) - z_i(t) + \sum_{\aleph_j \in B_i} f_j(z_j(t), w_{ij}, z_i(t)) \qquad (3)$$

where

- t denotes time,
- $z_i(t)$ denotes the activity level of the ith artificial neuron,
- w_{ij} denotes the weight of a link from the jth to the ith artificial neuron,
- $I_i(t)$ denotes external input to the ith artificial neuron,
- $f_j(z_j(t), w_{ij}, z_i(t))$ denotes the influence of jth artificial neuron upon the ith artificial neuron,
- μ_i coefficient.

Equation 3 is a simultaneous system of differential equations of the first degree. This is a generic equation, and can have different forms depending on the choice of I_i, f_j, w_{ij} and μ_i corresponding to the particular case or application where the ANN is being used. For example, when applied to neurons μ_i denotes a membrane time constant, z_i denotes membrane voltage, I_i means an external input, w_{ij} is interpreted as a weight associated to the synapse, whereas f_j takes the from of a product between the weight and z_j. For analogue electric circuits, the time constant μ_i is a product between resistance and capacitance, z_i denotes the potential of a capacitor, the left hand side of the equation is interpreted as a current charging a capacitor to potential z_i, whereas the summed terms mean potentials weighted by conductance (DeWilde, 1996).

From a practical point of view, the existence of solutions of equation 3 can be answered positively due to the Cauchy-Lipschitz theorem (Arrowsmith and Place, 1990), which is well-known in the theory of differential equations, and is stated here without proof:

THEOREM. Let

$$F(t,z) = \frac{1}{\mu_i}(I_i(t) - z_i(t) + \sum_{\aleph_j \in \mathrm{B}_i} f_j(z_j(t), w_{ij}, z_i(t)))$$

and consider the initial condition $z(t_0) = t_0$. If

a) the function $F(t, z)$ is continuous in a region $\Omega \subset \mathbb{R}^2$ ($\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^2$ denotes the plane),

b) the function $F(t, z)$ is a local Lipschitz contraction, i.e., $\forall P \in \Omega \; \exists K \subset \Omega$ and $\exists L_K > 0$ constant such that $|F(t, z_1) - F(t, z_2)| \leq L |z_1 - z_2|$, $\forall (t, z_1), (t, z_2) \in K$,

then there exists a vicinity $V_0 \subset \Omega$ of the point (t_0, z_0) in which the equation 3 has a unique solution satisfying the initial condition $z(t_0) = t_0$, which can be obtained by successive numeric approximations. ♦

3.2.2 Quantitative Model of I^2R

A quantitative model of I^2R is a computational model obtained from the qualitative model, as a possible interpretation of the latter, by introducing additional assumptions, and can be numerically implemented by an algorithm. One possible quantitative model, which has proved useful in several applications[1], is briefly described below.

Because the objects to be searched are IR objects, e.g, documents, no external input is assumed, so we take $I_i(t) = 0$. One way to define f_j explicitly is to conceive the influences of object j upon object i as being determined by the strengths of the connections which convey this influence, i.e., weights w_{ij} of the links between them, whereas the coefficient μ_i can be taken as being equal to unity. Equation 3 thus reduces to the following equation:

$$\frac{dz_i(t)}{dt} = -z_i(t) + \sum_{\aleph_j \in B_i} w_{ij} \qquad (4)$$

The following computation is applied for the weights w_{ij} (but other methods may also be applied). Each $\aleph_i$ is associated an n_i-tuple of weights corresponding to its identifiers (e.g., keywords) t_{ik}, $k = 1, \ldots, n_i$. Given now another $\aleph_j$. If identifier t_{jp}, $p = 1, \ldots, m_j$, occurs f_{ijp} times in O_i then there is a link from $\aleph_i$ to $\aleph_j$, and this has the following weight:

[1] See http://dcs.vein.hu/CIR

$$w_{ijp} = \frac{f_{ijp}}{n_i} \tag{5}$$

Formula 5 can be applied in a binary (i.e., f_{ijp} = 1 or 0) or non-binary form. If identifier t_{ik} occurs f_{ikj} times in O_j, and df_{ik} denotes the number of objects in which t_{ik} occurs, then there is a link from $\aleph_i$ to $\aleph_j$, and this has the following weight (inverse document frequency):

$$w_{ikj} = f_{ikj} \cdot \log \frac{2N}{df_{ik}} \tag{6}$$

The total input to $\aleph_j$ is then given by

$$\sum_{k=1}^{n_i} w_{ikj} + \sum_{p=1}^{n_j} w_{ijp} \tag{7}$$

It can be shown that (i) the solutions of equation 4 are of the form $K \cdot e^{-t} \cdot (e^t - 1)$, where K is a constant, and that (ii) when the network operates for retrieval (i.e., activation spreading according to WTA), the activity level of an object is directly proportional to its total input. (The quantitative model described above can be deduced from other forms of the fundamental equations, too.)

An interesting property of this model is that it is able to return important documents even if they do not contain any of the query terms, but which are strongly linked with documents containing query terms. This property also appears in recent Web searching models (Kleinberg, 1999).

4 PAGERANK: QUANTITATIVE MODEL OF I^2R

It will be shown that the PageRank method can be formally conceived as another quantitative interpretation of the I^2R model by making specific assumptions.

4.1 The Principles

The principles on which PageRank and I^2R are based are equivalent with each other due to the following reasons:

(a) In PageRank, the importance of a Web page is expressed by its citation level, whereas in I^2R the importance of an object is given by its activity level, and thus PageRank's concepts of citation level and I^2R's concept of activity level correspond to each other, they can be paralleled.

(b) In PageRank, the citation level of a Web page depends on the citation levels of the pages pointing to it (and thus implicitly also on their number), whereas in I^2R the activity level of an object depends on the activity levels of the objects linking to it (and thus implicitly also on their number). Thus the importance of pages/objects depends on equivalent factors in the two models.

4.2 The Models

The models used in both PageRank and I^2R are the same, see parts 2.2.2 and 3.2.1.2).

4.3 The Formulas

In I^2R, the activity levels are conceived as dynamical quantities which vary with time during operation. In PageRank, the citation levels, once computed, remain constant while being used (in

the retrieval process). If I^{2}R's activity level is viewed as a particular case, namely constant in time, then I^{2}R's fundamental equation 3 has a null in its left hand side (the derivative of a constant is zero), and is thus asking for finding the equilibrium as a solution; it becomes:

$$0 = I_i(t) - z_i(t) + \sum_{\aleph_j \in \mathrm{B}_i} f_j(z_j(t), w_{ij}, z_i(t)) \tag{8}$$

No external (i.e., from outside the Web) inputs to Web pages are assumed in PageRank, hence we take $I_i = 0$. It can be seen that, from a numerical point of view, in PageRank, the citation level of a Web page is inversely proportional with the number of links that pages pointing to that page have (if the Web pages have one link each then the matrix M has binary values, the inverse proportionality becomes visible if at least one Web page has more than one link, and L_j appears in denominators). Let the identifier associated to an object be an URL as such. In this case the binary version of formula 5 is identical with the term $1/L_j$ of PageRank's fundamental equation 1: the link from object j to object i has weight $w_{jik} = f_{jik}/n_j$, and $t_{jik} = \mathrm{URL}_i$ (i.e., object j contains the URL address of object i), $f_{jik} = 1$, $n_j = L_j$. Thus, I^{2}R's general equation re-writes as follows:

$$0 = -z_i(t) + \sum_{\aleph_j \in \mathrm{B}_i} f_j(z_j(t), \frac{1}{n_j}, z_i(t)) \tag{9}$$

Taking into account the principle of PageRank (or, equivalently, of I^{2}R), the function f_j does not depend on z_i (it can be seen from equation 1 that the citation level R_i does not occur on the right hand side; in other words, the citation level of a Web page does not depend on its own citation level, self-links are not assumed in the graph model used for equation 1), but it depends on z_j (the citation level of a Web page does depend on the citation levels of the pages linking to it). As it is common to take in ANNs, the function f_j is taken as the dot product of the vector of activity levels

and corresponding weights of the objects pointing to it, i.e., $f_j = z_j w_{ij}$ (linear combination). The equation of I^2R re-writes now as follows:

$$0 = -z_i(t) + \sum_{\aleph_j \in \mathrm{B}_i} z_j(t) w_{ij} \tag{10}$$

Because z_j does not depend on time (see above), equation 10 becomes:

$$z_i = \sum_{\aleph_j \in \mathrm{B}_j} \frac{z_j}{n_j} \tag{11}$$

which is the same as equation 1 expressing the extended citation principle used in PageRank.

Different version of PageRank's equation can also be obtained.

If the influence function f_j of equation 9 is defined as $f_j = d \cdot z_j \cdot w_{ij}$, where d is a coefficient, $0 < d \leq 1$, then the following version of PageRank's equation is obtained (second row in Table 1):

$$z_i = d \cdot \sum_{O_j \in \mathrm{B}_j} \frac{z_j}{n_j} \tag{12}$$

If an external input I_i is assumed in equation 8 and defined as $I_i = 1 - d$, $0 < d < 1$, and the influence function f_j as $f_j = d \cdot z_j \cdot w_{ij}$, another version of PageRank's equation is obtained (third row in Table 1):

$$z_i = 1 - d + d \cdot \sum_{O_j \in \mathrm{B}_j} \frac{z_j}{n_j} \tag{13}$$

5 CONCLUSIONS

5.1 PageRank: another quantitative interpretation of I^2R

After describing the PageRank and I^2R methods (the principles on which they are based, the formal models they use, and the equations they are built on) it was shown that the PageRank method can be formally conceived as a particular interpretation of the I^2R method: their principles and models are the same, and PageRank's equation can be formally obtained as a particular case of I^2R's equation.

5.2 PageRank: 'equilibrium state' of the Web

The result that PageRank may be formally viewed as a specific interpretation of the I^2R method by making certain assumptions makes it possible to view PageRank from a different perspective (beside the usual algebraic and stochastic ones), namely as a particular connectionist (dynamic) system.

The PageRank values may be viewed as equilibrium points of the system of the interlinked Web pages, whose Jacobian matrix is as follows:

$$J = \left(\frac{\partial f}{\partial z_j} \right)_i = M - I$$

Because the number 1 is an eigenvalue of the matrix *M*, the Jacobian matrix *J* is singular. The Hessian of the system is singular, too (*J* has constant elements), hence all the second partial derivatives are zero. Thus, we may say that the PageRank values constitute neutral equilibrium points of the Web. It is well-known that from a neutral equilibrium the system may jump to

another such equilibrium. So, we may say that each re-computation of PageRank values means seeking or finding another neutral equilibrium of the Web that it has been developed since the last computation of PageRank.

6 ACKNOWLEDGEMENTS

The author would like to thank the anonymous Reviewers for their precious comments, and acknowledges the support of grants OTKA T 037821 (National Foundation fro Scientific Research, Hungary), NKFP OM-00359/2001 (Ministry of Education, National Research and Development Program, Hungary), and AKP 2001-140 (Hungarian Academy of Sciences).

7 REFERENCES

Arrowsmith, D.K., and Place, C.M. (1990). *An introduction to dynamical systems*. Cambridge University Press.

Arasu, A. (2002). PageRank Computation and the Structure of the Web: Experiments and Algorithms. *Proceedings of the World Wide Web 2002 Conference*, Honolulu, Hawaii, USA, 7-11 May, http://www2002.org/CDROM/poster (visited: 4 Nov 2002)

Breyer, L.A. (2002). Markovian Page Ranking Distributions: Some Theory and Simulations. http://www.lbreyer.com/preprints/googlerip.pdf (visited: 4 Nov 2002)

Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). What can you do with a Web in your Pocket?. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, http://www.n3labs.com/pdf/brin98what.pdf (visited: 4 Nov 2002)

Brin, S., and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia, 14-18 April, pp: 107-117

Carriere, J.,and Kazman, R. (1997). Webquery: Searching and Visualising the Web Through Connectivity. Proceedings of the World Wide Web Conference, Santa Clara, California, April, pp: 701-711

Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J. (1999). Mining the Link Structure of the World Wide Web. *IEEE Computer*, Vol. 32, No. 8, pp: 60-67

Cho, J., Garcia-Molina, H., and Page, L. (2002). *Efficient Crawling Through URL Ordering*. Stanford University, http://www-db.stanford.edu/~cho/crawler-paper (visited: 4 Nov 2002)

De Wilde, Ph. (1996). *Neural Network Models*. Springer.

Dominich, S. (1994). Interaction Information Retrieval. *Journal of Documentation*, vol. 50, no. 3, pp: 197-212

Dominich, S. (1997). The Interaction-based Information Retrieval Paradigm. In: Kent, A., and Williams, G.W. (eds.) *Encyclopedia of Computer Science and Technology*. Marcel Dekker, Inc., New York – Basel – Hong Kong, vol. 37, suppl. 22, pp: 175-193

Dominich, S. (2001). *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers, Dordrecht, Boston, London.

Feldman, J.A., and Ballard, D.H. (1982). Connectionist models and their properties. *Cognitive Science*, 6, pp: 205-254

Garfield, E. (1955). Citation indexes for science. *Science*, p. 108

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, pp: 471-479

Google, Inc., http://www.google.com

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybertnetics*, 23, pp: 121-134

Haveliwala, T.H. (1999). *Efficient Computation of PageRank*. Stanford University, http://dbpubs. stanford.edu:8090/pub/1998-31 (visited: 4 Nov 2002)

Haveliwala, T.H. (2002). Topic-Sensitive PageRank. *Proceedings of the World Wide Web Conference 2002*. May 7-11, Honolulu, Hawaii, USA, http:// www2002.org/CDROM/ (visited: 4 Nov 2002)

Henzinger, M.R., Heyden, A., Mitzenmacher, M., and Najork, M. (1999). Measuring Index Quality Using Random Walks on the Web. *Proceedings of the 8th International World Wide Web Conference,* Toronto, Canada, pp: 213-225

Hopfield, J.J. (1984). Neurons with graded response have collective computational properties like those of two-states neurons. *Proceedings of the National Academy of Sciences*, 81, pp: 3088-3092

James, W. (1890). *Psychology (Briefer Course)*. New York: Holt, Chapter XVI, "Association", pp: 253-279

Kim, S.J., and Lee, S.H. (2002). An Improved Computation of the PageRank Algorithm. In: Crestani, F., Girolamo, M., and van Rijsbergen, C.J. (eds.) *Proceedings of the European Colloquium on Information Retrieval*. Springer LNCS 2291, pp: 73-85

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. http://www.cs.cornell.edu/ home/kleinber/auth.pdf (visited: 4 Nov 2002)

Page, L. (2001). United States Patent 6,285,999. September 4 (visited: 4 Nov 2002)

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford University, http://dbpubs. stanford.edu:8090/pub/1998-66 (visited: 4 Nov 2002)