Using Complex Numbers in Website Ranking Calculations: A Non-ad Hoc Alternative to Google's PageRank

Keita Sugihara*
Nanzan University, 18 Yamazato-cho, Showa-ku, Nagoya 466-8673, Japan.

* Corresponding author. Tel.: +81-52-832-3278; email: sugihara@nanzan-u.ac.jp

Manuscript submitted August 15, 2018; accepted December 9, 2018.
doi: 10.17706/jsw.14.2.58-64

Abstract: This paper presents an alternative to Google's PageRank, i.e., it presents an algorithm used to calculate the score for a webpage using complex numbers that overcomes the problems inherent in Google's method. This algorithm was inspired by eigenvector centrality in social network analyses and is designed to reproduce the ranking results of Google's PageRank and to satisfy the condition of soundness. This algorithm can be developed further to achieve more desirable outcomes.

Key words: Complex number, Hermitian adjacency matrix, Google, PageRank.

1. Introduction

This paper proposes a non-ad hoc alternative to Google's PageRank, an algorithm that is widely used by search engines on the Internet. PageRank gives high scores to popular sites and low scores to unpopular sites. Google's algorithm uses hyperlink relationships between Internet websites. PageRank models the hyperlink relationships between websites as a directed graph and expresses them as an adjacency matrix using real numbers. To produce a strongly connected directed graph, this algorithm requires a damping factor, which ranges in value from 0 to 1. However, there are practical problems associated with determining specific damping factor values. This paper proposes a method that does not require a damping factor and produces results similar to those of PageRank and can be developed systematically for a specific purpose. The Hermitian adjacency matrix is used to express the link relationships of the nodes in a directed graph. This method only requires that the directed graph be weakly connected and can be applied to a non-weakly connected graph.

2. PageRank

Definition 2.1 A semi path is a collection of distinct nodes, $v_1, v_2, \dots, v_n$, together with $n-1$ links, one from each v_1v_2 or v_2v_1, v_2v_3 or v_3v_2, ..., $v_{n-1}v_n$ or v_nv_{n-1}.

Definition 2.2 A path is a collection of distinct nodes, $v_1, v_2, \dots, v_n$, together with the links, v_1v_2, v_2v_3,..., $v_{n-1}v_n$.

Definition 2.3 A directed graph $G=(V,E)$ is called weakly connected if for all nodes $v_1, v_2 \in V$ there exists a semi path between v_1 and v_2.

Definition 2.4 A directed graph $G=(V,E)$ is called unilaterally connected if for all nodes $v_1, v_2 \in V$ there exists a path from v_1 to v_2 or from v_2 to v_1.

Definition 2.5 A directed graph $G=(V,E)$ is called strongly connected if for all nodes $v_1, v_2 \in V$ there exists a path from v_1 to v_2.

PageRank has three characteristics [1], [2] that can be summarized as follows. First, a page receives a high score when it has an inlink from a node with a high score. Second, a page receives a high score when it has many inlinks. Third, a page receives a high score when it has an inlink from a node with few outlinks. The PageRank scores of the nodes of a directed graph are defined as follows [3]. Let $|P_i|$ be the number of outlinks from a node i. We define the $n \times n$ matrix H_{ij} as follows: $H_{ij} = 1/|P_i|$ if there is a link from node i to node j and equals 0 otherwise. We define the matrix S as follows using e^T to designate a row vector of all 1s. $S = H + a((1/n)e^T)$, where $a_i = 1$ if node i has no outlink and 0 otherwise. We define the matrix G as follows: $G = \alpha S + (1-\alpha)(1/n)ee^T$. The PageRank scores are the elements of the normalized dominant left-hand eigenvector of G corresponding to the real dominant eigenvalue, 1. The dominant eigenvalue is the absolute maximum eigenvalue of a square matrix. The coefficient α in the equation is called the damping factor. We need this factor to ensure that the matrix G is a matrix of a strongly connected directed graph. A square matrix A is irreducible if and only if its directed graph is strongly connected [4]. According to the Perron–Frobenius theorem, if $A \geq 0$ is irreducible, $r = \rho(A) > 0, r \in \sigma(A)$, and the multiplicity of the eigenvalue is 1. Here, $\rho(A)$ is the spectral radius of A, and $\sigma(A)$ is the spectrum of A [4]. Therefore, a real positive dominant eigenvalue exists. In addition, this value is unique because the multiplicity is 1. Otherwise, a dominant eigenvalue cannot be determined. In this model, the damping factor, α, can be understood as a parameter that controls the proportion of time that a user follows the hyperlinks, as opposed to randomly jumping to new webpages. If, for example, $\alpha = 0.85$, then 85% of the time, the user follows the hyperlink structure of the Web, and the other 15% of the time, the user jumps to a random new page [3].

3. Problem Description

The following three problems associated with the damping factor have been noted. First, the choice of a damping factor value is eminently empirical, and in most cases, the value of 0.85 suggested by Brian and Page is used [5]. Second, a network has inconsistent rankings when using different damping factor values [6]. Third, a specific damping factor value could be used to create spam against a search engine [7]. To overcome these problems, we need a PageRank-like and non-ad hoc ranking score algorithm that is not dependent on a damping factor.

4. Hermitian Adjacency Matrix

Let a and b be real numbers, and let $i = \sqrt{-1}$. According to these definitions, $z = a + bi$ is a complex number [8]. The complex numbers $a + bi$ and $a - bi$ are conjugates of each other, where the conjugate of z is denoted $\overline{z}$. A Hermitian matrix is a square matrix $H = [h_{ij}]$, such that$\overline{H'} = H$. Thus, H is Hermitian, provided that $a_{ij} = \overline{a_{ji}}$ for all values of i and j. An eigenvalue of a Hermitian matrix is always a real number [9]. For a directed graph $X = (V, E)$, the Hermitian adjacency matrix is defined as follows [9] using uv to designate the ordered pair $< u, v >$:

$$H_{uv} = \{ \begin{matrix} 1\ if uv and vu \in E; \\ i\ if uv \in E and vu \notin E; \\ -i\ if uv \notin E and vu \in E; \\ 0\ otherwise. \end{matrix} \tag{1}$$

5. Hermitian Score

Definition 5.1 The node-induced subgraph of directed graph $G = (V, E)$ is a graph $G_s = (V_s, E_s)$, where

$V_s \in V$ and E_s includes all lines from E that are from one node to another node in V_s.

Definition 5.2 The adjacency matrix of a directed graph $G = (V, E)$ is defined as follows.

$$A_{uv} = \{ \begin{matrix} 1\ if uv \in E; \\ 0\ otherwise. \end{matrix} \tag{2}$$

Definition 5.3 The adjacency matrix of an undirected graph $G = (V, L)$, using (uv) to express the unordered pair (u, v), is defined as follows.

$$A_{uv} = \{ \begin{matrix} 1\ if (uv) \in L; \\ 0\ otherwise. \end{matrix} \tag{3}$$

Because (uv) is an unordered pair, the adjacency matrix of an undirected graph is always a symmetric matrix.

Definition 5.4 The startin g node of a unilaterally connected directed graph is defined as the origin of the longest path in the graph. If there is more than one longest path, label the longest-path origin with the most outlinks as the starting node. If there is more than one longest-path origin with the same number of most outlinks, add a node as a dummy to the graph, add the links from the dummy to the origins, and label the dummy the starting node.

Definition 5.5 In a directed graph $G = (V, E)$, if there is a path from $u \in V$ to $v \in V$, u is said to be reachable to v and v is said to be reachable from u.

5.1. The Basic Idea of the Hermitian Score from Eigenvector Centrality

Let us focus on social network analysis [10] and centrality analysis. Eigenvector centrality is defined for a node of a connected undirected graph [11]. Eigenvector centrality suggests that the score of node i, denoted by x_i, is the sum of all of the scores of all of the nodes that have undirected links to that node. Based on this concept, we can establish the following matrix equation, where A denotes the adjacency matrix of the undirected graph. We know that multiplying the right-hand term by a real coefficient does not change the interest of this equation.

$$X = AX, \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{4}$$

$$X = \frac{1}{\lambda} AX, \lambda \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, AX = \lambda X \tag{5}$$

The equation is the eigenvalue and eigenvector problem of the matrix, where λ is an eigenvalue of A. Because A is the adjacency matrix of an undirected graph, it is a real symmetric matrix. Therefore, λ is always a real number. In addition, there exists a positive dominant eigenvalue for A with a multiplicity of 1 [4], [11]. The eigenvector centrality score of the ith node is the ith element of the dominant eigenvector for the positive dominant eigenvalue. We would like to apply the idea of eigenvector centrality for an undirected graph to a directed graph to reproduce the three characteristics of PageRank using the Hermitian adjacency matrix described above.

The results of trials suggest that, if a directed graph is weakly connected, the dominant eigenvalue of the graph's Hermitian adjacency matrix, H, is a positive number with a multiplicity of 1, a negative number with a multiplicity of 1, or a positive number with a multiplicity of 1 and a negative number with a multiplicity of

1. According to the results of the trials, these conditions are satisfied when we derive the Hermitian matrix from $H^{'}$ from H using the method described below and when we create the Hermitian matrix $H^{''}$ from $H^{'}$ using the method described subsequently in this paper.

5.2. Algorithm for the Hermitian Score

The algorithm for Hermitian scores of nodes a directed graph G is defined as follows. There are four stages with sub-stages included for Stage 3 and sub-sub-stages included for Stage 3-3.

The algorithm for Hermitian scores of nodes of a directed graph: Stage1to Stage 4

Stage 1: For the entire directed graph G, which may not be weakly connected, let each maximal weakly connected directed graph $G_1, G_2, \dots, G_i, \dots G_N$.

Stage 2: For each G_i, label the number of the nodes of the graph $n(G_i)$, and, designate the maximum $n(G_i)$ as $max(n(G_i))$.

Stage 3: For each G_i, apply the following algorithm to determine the Hermitian scores of nodes of a weakly connected graph (Stages 3-1 to 3-3).

The algorithm for Hermitian scores of nodes of a weakly connected graph: Stage 3-1 to 3-3

Stage 3-1: For each G_i, label nodes with more than one inlink, without counting the inlinks of pairs of mutual links, as the separating node $s_{i1}, s_{i2}, \dots, s_{ip}, \dots, s_{iS}$.

Stage 3-2: For each separating node s_{ip}, designate the number of other separating points that are reachable to separating and that are reachable from a node without an inlink as $n(s_{ip})$: 0,1,2 …, k, ….

Stage 3-3: Repeat the following algorithm for Hermitian scores of nodes of a unilaterally connected subgraph, in the ascending order of $n(s_{ip})$ (Stage 3-3-1 to 3-3-11).

The algorithm for Hermitian scores of nodes of a unilaterally connected subgraph: Stage 3-3-1 to 3-3-11

I: For the nodes that are reachable from the node without an inlink and that are reachable to a separating node, complete Stages 3-3-1–3-3-10.

Stage 3-3-1: For the separating node s_{ip} for $n(s_{ip}) = k$, induce the maximal subgraph(s) using nodes that are reachable to the separating node and that are reachable from the separating node(s) for $n(s_{ip}) \leq k$, excluding the separating node. For $k = 0$, consider the node without an inlink as the previous separating point. Designate the set of the subgraph(s) as $S(s_{ip})$.

Stage 3-3-2: For each induced subgraph in $S(s_{ip})$, if (tentative) the starting node has mutual links with a node, add a dummy and an outlink from the dummy to the starting node and recognize the dummy as the starting node.

Stage 3-3-3: For each subgraph in $S(s_{ip})$, if a node has no outlink, add a loop link to the node.

Stage 3-3-4: For each subgraph in $S(s_{ip})$, create the Hermitian adjacency matrix H of the subgraph.

Stage 3-3-5: For H of the subgraph, convert each element of i to $s(t+i)$ and convert each element $-i$ to $s(t-i)$, thereby deriving $H^{'}$, where $s = sin\{\pi/(2 \times n)\}$, $t = cos\{\pi/(2 \times n)\}/sin\{\pi/(2 \times n)\}$ and $n = max(n(G_i))$.

Stage 3-3-6: For $H^{'}$ of the subgraph, in each row, divide each $s(t+i)$ by the number of appearances of $s(t+i)$ and 1 in the row and divide each diagonally corresponding $s(t-i)$ by the same number, creating $H^{''}$.

Stage 3-3-7: For each $H^{''}$ of the subgraph, label the dominant eigenvalue λ_1 and label each element of the dominant eigenvector $x_1, x_2, \dots, x_s, \dots, x_n$, where x_s denotes the element corresponding to the starting node of the subgraph. If the dominant eigenvalues include a positive and a negative real value, choose the positive eigenvalue.

Stage 3-3-8: For each $H^{''}$, divide each term of the series $x_1, x_2, \dots, x_s, \dots, x_n$ by x_s, deriving the series

$x_1/x_s, x_2/x_s, \dots, x_s/x_s, \dots, x_n/x_s$.

Stage 3-3-9: For $H^{''}$, locate each term of $x_1/x_s, x_2/x_s, \dots, x_s/x_s, \dots, x_n/x_s$ on a complex plane, and consider each term to be a two-dimensional vector on the plane, and define θ as the angle to the vector from the real axis in the clockwise direction and l to be the length of that vector.

Stage 3-3-10: For the Hermitian scores of nodes of the subgraph in $S(s_{ip})$ having $H^{''}$, rotate each of the vector $x_1/x_s, x_2/x_s, \dots, x_s/x_s, \dots, x_n/x_s$ in the clockwise direction by θ of s_{ip} for $n(s_{iq}) \leq k$, driving $\theta^{'}$ of the element of the term $x_1/x_s, x_2/x_s, \dots, x_s/x_s, \dots, x_n/x_s$. For $k = 0$, set θ to 0. For each node, excluding the separating node of $n(s_{ip}) = k$, the Hermitian score of the node is defined as $\theta \times l$ on the complex plane. For the Hermitian score of the separating node s_{ip} for $n(s_{ip}) = k$, compose all the 2-dimensional vectors of s_{ip} for $n(s_{ip}) = k$ by setting all the angles of the 2-dimensional vectors to the the maximal angle of all of the vectors.

II: For the nodes that do not appear in I, including nodes which are reachable from the separating node and which are reachable to the node without an outlink, complete Stage 3-3-11.

Stage 3-3-11: For the Hermitian scores of these nodes, induce the maximal subgraph(s) induced by the nodes that are reachable to the node and reachable from the separating node(s) and apply the above method(3-3-1 to 3-3-10).

Stage 4: To determine the Hermitian score of a node in the entire directed graph G, use the Hermitian score of that node calculated for the weakly connected graph G_i as the score.

Note that using the value $max(n_i)$ in Stage 3-3-5 for all weakly connected directed graphs with different numbers of nodes makes it possible to compare and rank the Hermitian scores of nodes belonging to more than one weakly connected graph.

Let us confirm that the algorithm for Hermitian score of a directed graph satisfies the three characteristics of PageRank: the first characteristic of PageRank is satisfied by Stage 3-3-7. The second characteristic of PageRank is satisfied by Stage 3-3-7 and Stage 3-3-10. The third characteristic of PageRank is realized by Stage 3-3-6.

6. Soundness of the Algorithm

Let us focus on soundness of the algorithm used to calculate Hermitian scores. Soundness of an algorithm is proposed as the following [12]. Let us focus on vote by committee.

Definition 6.1. Let A be some set. A relation $R \subseteq A \times A$ is called an ordering on A if it is reflexive, transitive, complete and anti-symmetric. Let $L(A)$ denote the set of ordering on A.

Notation: Let $\preceq$ be an ordering, then $\simeq$ is the equality predicate of $\preceq$. Formally, $a \simeq b$ if and only if $a \preceq b$ and $b \preceq a$.

Definition 6.2. Let $\mathbb{G}_V$ be the set of all weakly connected graphs with node set V. A ranking system F is a functional that, for every finite node V, maps every weakly connected graph $G \in \mathbb{G}_V$ to an ordering $\preceq_G^F \in L(V)$.

Definition 6.3. Let $G = (V, E)$ be a directed graph, and let $v \in V$ be a node in G. Then, the successor set of v is $S_G(v) = \{u|(v,u) \in E\}$ and the predecessor set of v is $P_G(v) = \{u|(v,u) \in E\}$.

Definition 6.4. Let $G = (V, E)$ be some weakly connected directed graph, and assume $V = \{v_1, v_2, \dots, v_n\}$. The Hermitian score $H_G(v_i)$ of a node $v_i \in V$ is defined as the score of the node using the algorithm defined above. The Hermitian ranking system is a ranking system that, for the node set V, maps G to $\preceq_G^H$, where $\preceq_G^H$ is defined as follows: for all $v_i, v_j \in V$: $v_i \preceq_G^H v_j$ if and only if $H_G(v_i) \preceq H_G(v_j)$.

6.1. Vote by Committee

Axiom 6.1. Let F be a ranking system. F satisfies vote by committee if, for every node set V, for every node $v \in V$, for every graph $G = (V, E) \in \mathbb{G}_V$, for every $v_1, v_2 \in V$, and for every $m \in \mathbb{N}$: let

$G^{'} = (V \cup \{u_1, u_2, \cdots, u_m\}, E \setminus \{(v, x) | x \in S_G(v)\} \cup \{(v, u_i)\}, | i = 1, \cdots, m\} \cup \{(u_i, x) | x \in S_G(v), i = 1, \cdots, m\}\})$

, where $\{u_1, u_2, \cdots, u_m\} \cap V = \emptyset$. Then, $v_1 \preceq_G^F v_2 iff v_1 \preceq_{G^{'}}^F v_2$.

6.2. Soundness

Proposition 6.1 The Hermitian ranking system H satisfies vote by committee. This is shown in the appendix.

7. Conclusions

This paper presents an algorithm for calculating the Hermitian scores of the nodes of a directed graph. The Hermitian rankings of the nodes of the directed graphs is designed to reproduce the three characteristics of Google's PageRank. The primary advantage of this method for determining Hermitian scores is that it does not require a damping factor, whereas PageRank requires a damping factor with a specific value. The Hermitian scores determined by the proposed method are therefore unaffected by the problems with PageRank that are associated with the damping factor. The algorithm for the Hermitian scores satisfies soundness for an algorithm. We foresee the following future research efforts as being necessary to further develop the proposed method. We need to develop methods to achieve a desired score for a node by adding a loop link of a specific width to the target node. We can develop these methods theoretically and systematically using the Hermitian adjacency matrix because these methods do not require the use of a damping factor. We also need to show that the algorithm satisfies a variety soundness conditions.

Appendix

The following is the proof for the vote by committee axiom.

Let $V = \{v_1, v_2, \dots, v_n\}$ be a node set, let $G = (V, E) \in \mathbb{G}_V$ be a graph, and let $v_s, v_t \in V$ be nodes, and let $m \in \mathbb{N}$ be a natural number. Assume $v_s \preceq_G^H v_t$.

Let

$G^{'} = (V \cup \{v_{n+1}, v_{n+2}, \cdots, v_{n+m}\}, E \setminus \{(v_1, x) | x \in S_G(v_1)\} \cup \{(v_1, v_{n+j}) | j = 1, \dots, m\}, \cup \{(v_{n+j}, x) | x \in S_G(v_1), j = 1, \cdots, m\}\})$. In G, each of $v: v \in S_G(v_1)$ has an inlink from v_1. In $G^{'}$, all of the inlinks to $v_{n+1}, v_{n+2}, \dots v_{n+m}$ are only from v_1 and each $v: v \in S_G(v_1)$ only has an inlink from each of $v_{n+1}, v_{n+2}, \dots v_{n+m}$. Therefore, based on Stage:3-3-6 of the algorithm, in G and $G^{'}$, the rankings between $v: v \in S_G(v_1)$ are same. Therefore, $v_s \preceq_G^H v_t$ implies $v_s \preceq_{G^{'}}^H v_t$.

Acknowledgment

This research was supported by Nanzan University Pache Research Subsidy I-A-2 for the 2018 academic year. This paper includes a patent pending invention.

References

[1] Lin, J. & Dyer, C., (2010). *Synthesis Lectures on Human Language Technologies Data-Intensive Text Processing with MapReduce*. Morgan & Claypool Publishers.

[2] Page, L. & Brin, S. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab Publication Server*. Retrieved August 5, 2018, from http://ilpubs.stanford.edu:8090/422/.

[3] Langville A. & Meyer, C. D. (2006). *Google's PageRank and Beyond, The Science of Search Engine*

Rankings. Princeton University Press, Princeton, NJ, USA.

[4] Meyer, C. D. (2000). Matrix analysis and applied linear algebra. *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA.

[5] Boldi, P., Santini, M., & Vigna, S. (2007). A deeper investigation of pagerank as a function of the damping factor. *Web Information Retrieval and Linear Algebra Algorithms*, 2007.

[6] Fu, H.-H., Dennis, K. J. L., & Tsai, H.-T. (2006). Damping factor in Google page ranking. *Applied Stochastic Models in Business and Industry*, 22, 431–444.

[7] Boldi, P. (2005). TotalRank: Ranking without damping. *Proceedings of the 14th International Conference on the World Wide Web*, 2005, 898–899.

[8] Ayres, F., Jr. (1962). *Theory and Problems of Matrices*. Schaum Publishing.

[9] Guo, K. J. (2015). Simple eigenvalues of graphs and digraphs: Ph. D. Dissertation, Simon Fraser University.

[10] Wasserman, S., & Faust, K. (2007). *Social Network Analysis Methods and Applications*. New York. Cambridge University Press.

[11] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*, 113–120.

[12] Altman, A., & Tennenholtz, M. (2004). Ranking systems: The pagerank axioms. Retrieved October 1, 2018, from https://ie.technion.ac.il/~moshet/pagerank.pdf

[13] Liu X., Lin, H., & Zhang, L. (2012). An attractive force model for weighting links in query-dependent web page ranking. *Journal of Computers*, 124–129.

[14] Langville A., & Meyer C. D. (2012). *The Science of Rating and Ranking Who's #1?* Princeton University Press, Princeton, NJ, USA.

[15] Altman, A., & Tennenholtz, M. (2005). Ranking systems: The pagerank axioms. EC '05 *Proceedings of the 6th ACM Conference on Electronic Commerce*.

[16] Suzuki, T. (2017). *R-de-Manabu-Data-Science 8 Network-Analysis* Dai-2-Han (Data Science with R 8 *Network Analysis* 2nd Edition). Kyouritsu-Shuppan.

Keita Sugihara received his Ph.D. degree in arts and science from Nagoya University, Aichi, Japan, in 2007. He is an assistant professor in the Faculty of Science and Engineering of Nanzan University, Nagoya, Aichi, Japan. His research interests include social network analysis and computational social science.