JOURNAL OF THE TENNESSEE ACADEMY OF SCIENCE 82(1–2):42–45 JANUARY–APRIL 2007

NATURAL NETWORKS AND THE *GOOGLE*™ SEARCH ENGINE

TERRANCE J. QUINN

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN, 37132

ABSTRACT—The *Google* internet search engine is well known for its success. In addition to it being of intrinsic interest, the underlying algorithm can be applied to any matrix Markov process and so has broad interdisciplinary applications. This article is directed to readers across the sciences and technology, and is intended to give a generally accessible introduction to the mathematical basis of the *Google* search algorithm.

Early search engines for the internet gave out long lists of unordered "title matches". In 1998, Brin and Page published their results (Brin and Page, 1998) on *PageRank*, a new approach to ranking websites, which in part led to the success of *Google*™. The approach appeals directly to the natural graph structure of the internet. In an efficient way, it uses the citation (link) pattern of the network to rank sites. The result, therefore, does not give some absolute rank of a document, but measures how a document fares in the "citation competition" relative to other documents in the same network. It is commonly supposed that once a month or so *Google* runs a crawl and selection from the internet to allow for an update of their database and website rankings.

The success of the Google search engine now is common knowledge; and mathematical analysis of the *PageRank* algorithm is based on well-known (but non-trivial) results that reach back to the 20th century. The mathematician may recall the fixed-point theorems of Brower, Hahn-Banach (Istratescu, 1981) and the matrix theorem of Perron-Frobenius (Berman and Plemmons, 1994; Meyer, 2000). There is a body of literature on search methods. An excellent survey article that includes a discussion of *PageRank* (as well as other eigenvector methods for web information retrieval) is (Langville and Meyer, 2005). A further bibliographical source can be found in (Kamvar et al., 2004; Sankaralingam et al., 2003). Specialized articles take the defining equation for *PageRank* as a starting point, and go on to consider: Aspects of how to use the defining equation (Craven acc. 2004; Rogers, 2002); Theoretical results on convergence rates (Kamvar et al., 2004); and Illustrations of non-trivial implementation of algorithms, and their refinements (Sankaralingam et al., 2003). There is a very good article for mathematics students (Bryan and Leise, 2006). The derivation of the equation, though, does not follow the Brin and Page approach as such, but is based on the interesting linear "continuity equation" given in (Kleinberg, 1999). Ultimately, the two approaches are mathematically equivalent. The Bryan and Leise article also provides proofs regarding linear algebraic quantities involved in the *PageRank* formula.

In addition to being of intrinsic interest, the algorithm has broad interdisciplinary significance. Indeed, the Google algorithm can be applied to any matrix Markov process. This short article therefore is expository in nature, and is intended for a general audience in science and technology. Having a modest familiarity with matrix multiplication would be helpful background. For the purposes of illustration, we use only 2 D or 3 D examples. All results extend to the general case.

CITATION RANK OF A WEBSITE

Suppose we have a collection of only three documents A, B and C. Of course, for actual internet searches, a typical collection of documents (websites) will have millions of sites. Suppose that document A cites B, but not C; B cites both A and C; C cites only A. This citation pattern can be represented by the directed graph in Fig. 1. The problem now is to obtain a relative ranking for the documents, using only the citation pattern. The discussion in Brin and Page (1998) provides the following framework: Suppose it is known that document C is not significant in the field, while A is an authority. In terms of citations, the rank of A should not be significantly affected by the reference to it in C; while the authoritative A that cites B can be taken to add importance to document B; and obviously a document cannot increase its rank by citing itself.

The Brin and Page (1998) citation ranking therefore involves both the relative numbers of citations and the initial rankings of the documents. A key step in the derivation is not the initial rankings as such, but noting how the rankings change by virtue of the internal citation pattern within the collection. This is analogous to, say, the undergraduate exponential law that is obtained for the growth of a cell population. In that case, one supposes an initial population, but goes on to derive the rule for how the population changes. Following this same approach with the citation pattern problem, and invoking the Brin and Page framework mentioned above, we get the following: Suppose initial ranks R_A, R_B, R_C, then the citation pattern adds rank to documents A, B and C by the equations

$$\text{Citation rank added to A} = (0)\,R_A + (1/2)\,R_B + (1)\,R_C$$

$$\text{Citation rank added to B} = (1)\,R_A + (0)\,R_B + (0)\,R_C$$

$$\text{Citation rank added to C} = (0)\,R_A + (1/2)\,R_B + (0)\,R_C$$

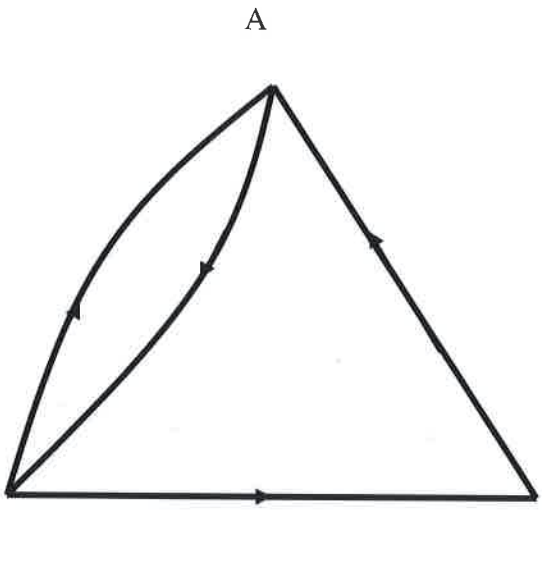

FIG. 1. Arrows indicate that source document cites terminal document.

In matrix notation, we have

$$\mathbf{Y} = \begin{bmatrix} \text{Citation rank added to A} \\ \text{Citation rank added to B} \\ \text{Citation rank added to C} \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} R_A \\ R_B \\ R_C \end{bmatrix} = \mathbf{MR},$$

where **Y** is a 3 × 1 column vector that represents the citation ranks added to each document, **M** is the 3 × 3 link matrix (of relative citation frequencies), and **R** is the 3 × 1 column vector of initial ranks. Each coordinate axis of **R** represents the corresponding document.

This is a 3-D weight distribution rule, where the weight attributed to document A is represented by the resulting first coordinate of **MR**, the weight attributed to document B is represented by the resulting second coordinate of **MR**, and so on. A classical way to define a reference scale for a weight distribution is to use one of its "principle axes" (Marion and Thornton, 1995). For example, the longest axis of a (uniform density) ellipsoid is a principal axis. Each document has its weight given by the components of **MR**. The classical result is that to calculate a major axis we identify a direction **R** that is parallel to **MR**. Finding the alignment of these two vectors reduces to solving the algebraic eigenvalue-eigenvector equation $\mathbf{MR} = a\mathbf{R}$, for some real number a. In this situation, a major axis is found when there is largest positive $a > 0$. Why it is that for a citation pattern there exists an eigenvector with non-negative components with a positive eigenvalue $a > 0$, will be discussed in the section below. For the citation problem, a non-negative component of an eigenvector indicates the weight of the corresponding document along the scale determined by the eigenvector direction. In the example given above with documents A, B and C, we first find the eigenvalues by solving the characteristic equation $\det(\mathbf{M} - x\mathbf{I}) = (x - 1)^2\left(x + {}^1\!/_2\right) = 0$. Here there is only positive eigenvalue $x = 1$; substituting into the eigenvector equation, a corresponding eigenvector is found to be $\mathbf{R} = [2, 2, 1]$. Recall that the objective is to rank the documents relative to each other. Along the direction of $\mathbf{R} = [2, 2, 1]$, the first and second document tie, each with weight 2, while the third receives weight 1. This of course corresponds directly to the original citation pattern seen in Fig. 1, where each of A and B both receive two citations, while C only receives 1. A convenient way to normalize measurements is to define a "length" of $\mathbf{R} = [2, 2, 1]$ as the sum of absolute values of its components: $2 + 2 + 1 = 5$. The citation rank is then defined in terms of the normalized eigenvector $\mathbf{v} = \frac{1}{5}[2, 2, 1] = \left[\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right]$, normalized to have "length" equal to 1. The first document therefore gets citation rank $\frac{2}{5}$, the second also gets $\frac{2}{5}$, and the third gets $\frac{1}{5}$.

Example—For an extreme case, consider a collection A, B, C, X where document X is not cited by any other document. It follows that regardless of how many times X cites other documents, the citation rank of X within the collection is zero. This is intuitively the right result. One of several ways to prove this is to note that citation rank added to document X is then $(0)R_A + (0)R_B + (0)R_C + (0)R_X$. Consequently, the column space of **M** (and in particular the defining eigenvector) has no X component.

Existence of Eigenvectors for the Citation Rank

While it is non-trivial to prove rigorously (even in 3-D), one way to obtain the existence of a suitable eigenvector is to look to the geometry of the situation. See, for example, (Lay, 1992). Suppose that **R** is any initial rank vector $[R_A \geq 0, R_B \geq 0, R_C \geq 0]$. Whatever the citation pattern, the entries of the matrix **M** are the non-negative relative citation ratios. But, under matrix multiplication, the components of **MR** consist of sums of products formed from the entries of **M** and **R**. So, the components of **MR** also are all non-negative. In other words, **M** takes a vector **R** with non-negative entries to a vector **MR** again with non-negative entries. In 3-D, **M** therefore preserves the octant of non-negative triples $[R_A \geq 0, R_B \geq 0, R_C \geq 0]$. This "confinement" may stretch, contract or even switch directions within the non-negative octant. However, because multiplication by **M** leaves the octant as a whole invariant, this "confinement" of the octant necessarily will leave at least one direction invariant as well. Hence, there is at least one vector **R** with non-negative components such that $\mathbf{MR} = a\mathbf{R}$, and where a is necessarily also positive. If the eigenvalue were negative, the resulting vector $a\mathbf{R}$ would not be in the non-negative octant. The direction **R** with the largest non-negative eigenvalue then gives the eigenvector that solves the problem. That is, we obtain a major axis of the weight distribution determined by the citation pattern **M**. For a matrix of exclusively non-negative entries, the existence of at least one fixed direction in the non-negative octant is a classical result in mathematics that has been proved from several viewpoints. The topological-geometric approach just described can be made rigorous by using the Brower Fixed Point Theorem (Istratescu, 1981). Or, since the equation involves matrix multiplication in finite dimensions, one may instead appeal to the Perron-Frobenius theorem (Berman and Plemmons, 1994; Radjavi, 1999; Meyer, 2000).

Algorithm for Finding the Eigenvectors for Citation Rank

The calculations involved in trying to directly solve an eigenvalue equation for an actual system of say, one million equations, would be formidable. Brin and Page (1998) took another approach. Instead of requiring an exact result, they looked to obtaining an approximation. As it happens, one may use a classical power iteration technique to approximate the eigenvector solution. Again, the rigorous proof of this result is beyond the scope of this expository article. A key ingredient is the Hahn-Banach theorem (Istratescu, 1981).

To illustrate the technique, consider the example given by the 2×2 matrix $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2/3 \end{bmatrix}$. Notice that the largest positive eigenvalue is 1. Now let $\mathbf{v} = [x \geq 0, y \geq 0]$ be any seed vector with non-negative entries. Calculating $\mathbf{A}^n\mathbf{v}$, we have the sequence $[x, \left(2/3\right)^n y]$. This converges to the eigenvector $\mathbf{v} = [x, 0]$ corresponding to the largest positive eigenvalue $a = 1$. Note that in order for the resulting eigenvector $\mathbf{v} = [x, 0]$ to be a non-zero, it is crucial that the first component x of the seed vector $\mathbf{v} = [x \geq 0, y \geq 0]$ be non-zero. This is mentioned here because, in addition to its mathematical significance, this also plays a key role in the algorithm for the Google *PageRank*.

Now consider the example $\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$. Using the above technique directly, the sequence $\mathbf{A}^n\mathbf{v} = [3^n x, 2^n y]$ diverges. A main purpose of the technique, though, is to identify an eigenvector direction. To deal with the divergent sequence, one approach, therefore, is to keep track of the direction of each $\mathbf{A}^n\mathbf{v} = [3^n x, 2^n y]$, but to scale the terms so that they converge. The sequence has two divergent factors, the powers of 3 and the powers of 2. A natural choice, therefore, is to divide by the largest positive eigenvalue 3, and define $\mathbf{B} = \frac{1}{3}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2/3 \end{bmatrix}$. The original method now works for $\mathbf{B}$, with 1 as the largest positive eigenvalue. Once we have the solution for $\mathbf{Bv} = \mathbf{v}$, we can use $\mathbf{B} = \frac{1}{3}\mathbf{A}$ to get that $\mathbf{Av} = 3\mathbf{v}$. In other words, the eigenvector directions for $\mathbf{A}$ and $\mathbf{B}$ are the same.

For the typical link matrix $\mathbf{M}$, the number of rows is in the millions, and the eigenvalues are not known in advance. As it turns out, though, essentially the same effect as above may be obtained by normalizing one step at a time. This is accomplished by starting with any seed vector $\mathbf{u}$, with the only requirement being that all entries are strictly greater than zero. At each power iteration $\mathbf{M}^{(n)}\mathbf{u}$, divide the vector by its largest positive entry, p_n say. We then obtain a sequence $\frac{\mathbf{M}^{(n)}\mathbf{u}}{p_n}$ that converges to an eigenvector $\mathbf{v}$, for the largest positive eigenvalue. Just as in the example above, the limit vector $\mathbf{v}$ is normalized to 1, and the normalized components define the citation ranks for the collection of documents. In practice, one does not use the limit vector, but approximates the limit by the convergent terms $\frac{\mathbf{M}^{(n)}\mathbf{u}}{p_n}$. The number n of iterations used depends on the accuracy required and the convergence rate.

Google *PageRank*-Adjusted Citation Rank

The Brin and Page *PageRank* is a modification of citation rank. The modification is obtained by adding a damping term to the eigenvalue-eigenvector equation $\mathbf{MR} = a\mathbf{R}$. "The (extra and fixed) parameter d is a damping factor which can be set between 0 and 1" (Brin and Page, 1998). The *PageRank* is defined in the same way as citation rank (that is, by taking the components of a normalized eigenvector) except that a constant "rank source" is introduced into the defining equation through the function $\mathbf{E}(\mathbf{R}) = (1 - d)\mathbf{1} = (1 - d)[1, 1, \ldots, 1] = [(1 - d), (1 - d), \ldots, (1 - d)]$. Note that the constant function $\mathbf{E}(\mathbf{R}) = (1 - d)\mathbf{1}$ is not the same as the identity function $\mathbf{I}(\mathbf{R}) = \mathbf{R}$. Finally, the eigenvector used for the *PageRank* is taken to be a (normalized) solution of the non-linear equation given by the (convex) combination $[d\mathbf{M} + (1 - d)\mathbf{1}]\mathbf{R} = a\mathbf{R}$.

Brin and Page (1998) give an "Intuitive Justification" for the damping factor: "the d damping factor is the probability at each page the 'random surfer' will get bored and request another random page" (Brin and Page, 1998). On present showing, one of the mathematical roles of the damping term may be easily understood by noting that the Google algorithm is a power iteration. Because of the damping factor, even if some of the initial components of a seed vector are zero, the next term in the sequence will always be positive in all of its components. Consequently, the limit eigenvector is necessarily non-zero. Note also that the choice of damping factor will clearly affect the convergence rate. In practice, the damping factor d is taken to be approximately 0.85 (Brin and Page, 1998). The damping factor also helps resolve certain technical details when running the algorithm for graph structures that occur in actual collections. Going into further details, however, would go beyond the scope of this introductory article. Numerical examples are presented in Bryan and Leise (2006) and Pandurangan et al (2002).

Discussion and Natural Networks

As mentioned at the beginning of the article, a main purpose is to offer a generally accessible introduction to the mathematical basis of the Google search algorithm. For the science reader, a familiarity with matrix multiplication was assumed. The Google (relative) ranking is an adjusted citation rank based on the natural citation (graph) structure of the internet. We can formulate citation rank as a weight distribution, determined by the citation patterns of the documents. A ranking for the distribution can then be formulated relative to a principle rotational axis of the distribution. A principle rotational axis is obtained from an eigenvector $\mathbf{R}$ of the weight distribution matrix $\mathbf{M}$. The matrix $\mathbf{M}$ consists of non-negative entries and therefore preserves the non-negative octant (more generally non-negative cone). It follows that there is an eigenvector $\mathbf{R}$ with non-negative components and with largest positive eigenvalue $a > 0$. The components of the normalized eigenvector are the citation ranks. That this is possible can be seen in elementary diagrams. Known proofs, however, require 20th century mathematical theorems beyond the scope of this article. To implement the algorithm for actual networks of millions of documents, the Google algorithm uses the power-iteration technique to approximate the eigenvector. Note also that Google adds the damping term $(1 - d)\mathbf{1}$ to the original citation rank equation, thus giving the new eigenvector-eigenvalue equation $[d\mathbf{M} + (1 - d)\mathbf{1}]\mathbf{R} = a\mathbf{R}$. Since it is also assumed that $0 < d < 1$, the power iteration normally produces a non-zero eigenvector with positive eigenvalue $a > 0$. Exceptions are not treated in this introductory article.

For the reader who is familiar with Markov processes (Isaacson and Madsen, 1976), notice that entries of each column of a link matrix **M** are all non-negative and sum to 1. Also, given a matrix Markov process, one may investigate the citation ranks and *PageRanks* for the events. Markov processes are ubiquitous in contemporary studies, including: particle physics; probabilistic change of state chemical networks; cell and/or cell-virus compartment models (such as in cancer and HIV studies); population models; ecological systems; birth-death and catastrophe theory; and so on. In the case of internet websites, the matrix entries can be used to define probabilities that a reader move from website A to website B, etc. Having a higher citation rank then represents a website with a higher probability of a maintained reader, which would be appropriate for a website that is an "authority" in the field. In certain natural networks, citation ranks give relative probabilities of survival.

ACKNOWLEDGEMENTS

The author would like to thank the anonymous referees for their detailed and helpful comments.

LITERATURE CITED

BERMAN, A., AND PLEMMONS, R. J. 1994. Nonnegative matrices in the mathematical sciences. Soc. Indus. Appl. Math. Philadelphia.

BORNHOLDT, S., AND S. H. GEORGE (ed.). 2003. Handbook for graphs and networks: from the genome to the internet. Wiley-VCH Pub., Cambridge.

BRIN, S., AND L. PAGE. 1998. The anatomy of a large-scale hypertextual web search engine, Available via the Internet (*http://infolab.stanford.edu/~backrub/google.html*).

BRIN, S., L. PAGE, R. MOTWANI, AND T. WINOGRAD. 1998. The *PageRank* citation ranking: bringing order to the web. (White Paper), Stanford University.

BRYAN, K., AND T. LEISE. 2006. The $25,000,000,000 eigenvector—the linear algebra behind *Google*. SIAM Rev., 48:569–581.

CRAVEN, P. No date - Accessed November 2004. Google's *PageRank* explained and how to make the most of it. WebWorkshop. Available via the Internet (*http:www.webworkshop.net/pagerank.html*).

ISAACSON, D. L., AND R. W. MADSEN. 1976. Markov chains—theory and applications. New York, Wiley.

ISTRATESCU, V. I. 1981. Fixed Point Theory—an introduction. D. Reidel Pub. Co. Kluwer Group. Dordrecht, Holland.

KAMVAR, S. T., HAVELIWALA, T., AND G. GOLUB. 2004. Adaptive methods for the computation of *PageRank*. Linear Algebra Appl., 386:51–65.

KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. J. Assoc. Comp. Mach., 46(5): 604–632.

LANGVILLE, A. N., AND C. D. MEYER. 2005. A survey of eigenvector methods for web information retrieval. SIAM Rev., 47(1):135–161.

LAY, S. R. 1992. Convex sets and their applications. Robert E. Krieger Publishing, Malabar, Florida. Revised reprint of 1982 original.

MARION, J. B., AND S. T. THORNTON. 1995. Classical dynamics of particles and systems, 4th. ed. Saunders College Pub., Fort Worth, Texas.

MEYER, C. D. 2000. Matrix analysis and applied linear algebra. SIAM, Philadelphia.

PANDURANGAN, G., P. RAGHAVAN, AND E. UPFAL, Using PageRank to characterize web structure. 2002. Pp. 330–339 *in* Computing and combinatorics (lecture notes in computer science, Vol. 2387) (Oscar H. Ibarra and Louxin Zhang, eds.). Proc. 8th Ann. Inter. Conf., COCOON 2002, Singapore, August 15–17, 2002. Springer, Berlin.

RADJAVI, H. 1999. The Perron-Frobenius theorem revisited. Positivity, 3(4):317–332.

ROGERS, I. 2002. The Google Pagerank algorithm and how it works. (White Paper), IPR Computing Ltd. Available via the Internet (*http://www.iprcom.com/papers/pagerank*).

SANKARALINGAM, K., M. YALAMANCHI, S. SETHUMADHAVAN, AND J. C. BROWNE. 2003. *PageRank* computation and keyword search on distributed systems and P2P networks. J. Grid Comp., 1:291–307.