Application of Kernels to Link Analysis

Takahiko Ito
takahi-i@is.naist.jp

Masashi Shimbo
shimbo@is.naist.jp

Taku Kudo*
taku@google.com

Yuji Matsumoto
matsu@is.naist.jp

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

ABSTRACT

The application of kernel methods to link analysis is explored. In particular, Kandola et al.'s Neumann kernels are shown to subsume not only the co-citation and bibliographic coupling relatedness but also Kleinberg's HITS importance. These popular measures of relatedness and importance correspond to the Neumann kernels at the extremes of their parameter range, and hence these kernels can be interpreted as defining a spectrum of link analysis measures intermediate between co-citation/bibliographic coupling and HITS. We also show that the kernels based on the graph Laplacian, including the regularized Laplacian and diffusion kernels, provide relatedness measures that overcome some limitations of co-citation relatedness. The property of these kernel-based link analysis measures is examined with a network of bibliographic citations. Practical issues in applying these methods to real data are discussed, and possible solutions are proposed.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Co-citation coupling, graph kernel, HITS, link analysis

1. INTRODUCTION

Link analysis aims to find useful information from the structure of graphs. In particular, much effort has been devoted in quantifying two types of information: *importance* of individual nodes in the graph, and *relatedness* between them. HITS [7] and PageRank [1] are the popular methods of evaluating importance of web pages. Co-citation [11] and bibliographic coupling [6] are two classic but still widely used measures of relatedness.

The objective of this paper is to show that the positive semidefinite kernels on graph nodes, proposed in the context of machine learning, provide a link analysis framework that enjoy many attractive properties.

As an instance of such a framework, we show that Kandola et al.'s Neumann kernels [5] subsume not only the co-citation and bibliographic coupling relatedness at an extreme of the parameter range, but also the HITS importance at the other extreme. This enables us to treat these popular link analysis measures uniformly in a parameterization scheme. Accordingly, the Neumann kernels can be interpreted as defining a spectrum of link analysis measures intermediate between co-citation/bibliographic coupling and HITS.

A formulation based on different kernels overcomes the limitations of co-citation (and bibliographic coupling) relatedness. The co-citation relatedness between documents is defined only on the basis of the number of joint citations made to them. It follows that co-citation coupling is not capable of computing relatedness if the documents are not jointly cited by any document. Moreover, it can be argued that the number of non-joint citations that are ignored by co-citation coupling is also a factor determining relatedness.

We show that the kernels based on the graph Laplacian [2, 8, 12] yield relatedness measures consistently over their parameter range, and these measures do not suffer from the above limitations. By introducing a new parameterization, we obtain link analysis measures that are intermediate between HITS importance and the relatedness measure given by the Laplacian-based kernels.

We also discuss the practical issues that may be encountered in the application of these kernels, including parameter tuning and approximation methods.

Due to lack of space, all the theorems in this paper are presented without proofs. These proofs can be found in [10] together with additional experimental results.

2. PRELIMINARIES

In this section, we review the link analysis measures of relatedness and importance that are relevant to the subsequent discussions. Throughout the paper, we denote matrices by capital letters, and column vectors by boldface letters. For a matrix A, $A(i,j)$ represents its (i,j)-element. Likewise, $\mathbf{v}(i)$ represents the i-th component of vector $\mathbf{v}$. Let $\rho(A)$ denote the spectral radius of A.

2.1 Relatedness

A general assumption underpinning link analysis is that in the target network structure such as a bibliographic citation graph and the web, an edge (a citation or hyperlink) between a pair of nodes (papers or web pages) signifies the nodes being in some sense re-

*Present address: Google Japan, Inc., Cerulean Tower 6F, 26-1 Sakuragaoka-cho, Sibuya-ku, Tokyo 150-8512, Japan.

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

lated. Hence the degree of relatedness can be inferred from the node proximity induced by the existence of edges.

Co-citation [11] and bibliographic coupling [6] are the standard methods of computing relatedness between documents in a citation network. Co-citation coupling defines relatedness between documents as the number of other documents citing them both. Bibliographic coupling defines relatedness between documents as the number of common references cited by them. Given an adjacency matrix A, the number of co-citations between nodes i and j is given by the (i,j)-element of the *co-citation matrix* $A^{\mathrm{T}}A$. Similarly, *bibliographic coupling matrix* AA^{T} gives the values of bibliographic coupling. These matrices are symmetric, so their graph counterparts, the *co-citation graph* and *bibliographic coupling graph*, are undirected. See Figure 1 for illustration.

2.2 Importance

Because of the difficulty in computing the importance of documents from their contents, citation counts have long been used as the index of document importance.

Kleinberg's HITS [7], along with PageRank, is a more recent and sophisticated method for evaluating document importance. HITS assigns two scores to each document (node), called the authority and hub scores. Let A be the adjacency matrix of a graph. HITS computes the following recursion over $n = 0, 1, \ldots$ starting with $\mathbf{a}_{(0)} = \mathbf{h}_{(0)} = \mathbf{1}$, the vector of all 1's.

$$\mathbf{a}_{(n+1)} = \frac{A^{\mathrm{T}}\mathbf{h}_{(n)}}{|A^{\mathrm{T}}\mathbf{h}_{(n)}|}, \quad \mathbf{h}_{(n+1)} = \frac{A\mathbf{a}_{(n+1)}}{|A\mathbf{a}_{(n+1)}|}. \tag{1}$$

The limit $\lim_{n\to\infty} \mathbf{a}_{(n)}$, if it exists, represents the *authority vector* whose i-th component represents the *authority score* of node i. Similarly, the *hub vector* $\lim_{n\to\infty} \mathbf{h}_{(n)}$ gives the *hub scores*. It is well known that when the dominant eigenvalue of $A^{\mathrm{T}}A$ (and AA^{T}) is simple, the authority and hub vectors exist and equal the dominant eigenvectors of $A^{\mathrm{T}}A$ and AA^{T}, respectively.

3. A UNIFIED FRAMEWORK FOR IMPORTANCE AND RELATEDNESS

In this section and the next, we present some formulations of link analysis measures that are intermediate between importance of nodes and their relatedness. These formulations are based on the family of symmetric positive semidefinite kernels [9] defining an inner product of nodes in a graph.

Importance is a measure defined on individual nodes and is naturally represented as a vector, whereas relatedness is defined between nodes, and hence forms a matrix. To define intermediate measures between these two extremes, we use the symmetric matrix $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ instead of an importance score vector $\mathbf{v}$. When all the components of $\mathbf{v}$ is positive, every row (and column) vector in $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ yields the node ranking identical to that of $\mathbf{v}$.

3.1 Neumann kernels

Kandola et al. [5] proposed the *Neumann kernels* for computing document similarity from terms occurring in documents, in a spirit analogous to Latent Semantic Analysis. We discuss the interpretation of these kernels in the context of link analysis.

The Neumann kernel in its original form is defined in terms of the term-by-document matrix X whose (i,j)-element is the frequency of the i-th term occurring in document j. From X, document correlation matrix $K = X^{\mathrm{T}}X$ and term correlation matrix $M = XX^{\mathrm{T}}$ are first constructed.

Definition 3.1 Let X be a term-by-document matrix, and let $K = X^{\mathrm{T}}X$ and $M = XX^{\mathrm{T}}$. The *Neumann kernel* matrices with *diffusion factor* $\gamma\,(\geq 0)$, denoted by $\hat{K}_\gamma$ and $\hat{M}_\gamma$, are defined as the solution to the following system of equations.

$$\hat{K}_\gamma = \gamma X^{\mathrm{T}}\hat{M}_\gamma X + K, \quad \hat{M}_\gamma = \gamma X^{\mathrm{T}}\hat{K}_\gamma X + M. \tag{2}$$

The similarity between documents i and j is given by the (i,j)-element of $\hat{K}_\gamma$, and the term similarity is given by $\hat{M}_\gamma$. Eq. (2) implies an alternative representation based on the Neumann series.

$$\hat{K}_\gamma = K\sum_{n=0}^{\infty}\gamma^n K^n, \quad \hat{M}_\gamma = M\sum_{n=0}^{\infty}\gamma^n M^n. \tag{3}$$

Hence, when $\gamma < \rho(K)^{-1}\,(= \rho(M)^{-1})$, the solution exists and is given by $\hat{K}_\gamma = K(I-\gamma K)^{-1}$ and $\hat{M}_\gamma = M(I-\gamma M)^{-1}$.

3.2 Link analysis with Neumann kernels

The recurrence over $\hat{K}$ and $\hat{M}$ in eq. (2) implies that the Neumann kernels evaluate similarity between documents from term similarity, and vice versa. This complementary relation is reminiscent of the recursion (1) between the authorities and hubs in HITS. We apply the Neumann kernels to link analysis on the basis of this particular similarity to HITS. Specifically, we use the adjacency matrix A of a citation graph in place of the document-by-term matrix X. Thus we have $K = A^{\mathrm{T}}A$ and $M = AA^{\mathrm{T}}$, which coincide with the co-citation and bibliographic coupling matrices, respectively. Plugging them into eq. (3) yields the Neumann kernels based solely on citation information. For convenience, we introduce the shorthand

$$N_\gamma(B) = B\sum_{n=0}^{\infty}(\gamma B)^n, \tag{4}$$

and write

$$\hat{K}_\gamma = N_\gamma(A^{\mathrm{T}}A) = A^{\mathrm{T}}A\sum_{n=0}^{\infty}\gamma^n(A^{\mathrm{T}}A)^n. \tag{5}$$

Likewise, $\hat{M}_\gamma = N_\gamma(AA^{\mathrm{T}})$, but since $\hat{M}_\gamma$ can be obtained simply by transposing A in eq. (5), we focus on $\hat{K} = N_\gamma(A^{\mathrm{T}}A)$ below.

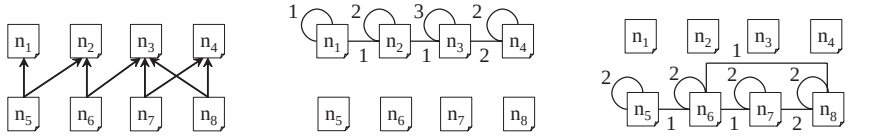
The Neumann kernels thus obtained from a citation graph possess much deeper relationship with HITS than just the superficial resemblance of their recursive forms. We will show that when viewed as a ranking method, the Neumann kernels subsume the HITS importance ranking as a special case.

3.3 Interpretation

Eq. (5) shows that the Neumann kernel matrix $N_\gamma(A^{\mathrm{T}}A)$ is a weighted sum of $(A^{\mathrm{T}}A)^n$ over every $n = 1, 2, \ldots$. Given that the (i,j)-element of the term $(A^{\mathrm{T}}A)^n$ represents the number of paths of length n between nodes i and j in the co-citation graph, we see that each element of the kernel matrix equals the sum of the number of paths between nodes weighted by a factor decaying exponentially with path length.

To grasp the meaning of the path counting and weight summation in the Neumann kernels, let us first examine what each term $(A^{\mathrm{T}}A)^n$, or the number of paths of length n, represents in terms of link analysis.

At $n = 1$, $(A^{\mathrm{T}}A)^n = A^{\mathrm{T}}A$ is the co-citation matrix giving a relatedness measure. It is not so obvious what $(A^{\mathrm{T}}A)^n$ represents when n is larger. With n sufficiently large, however, it can be shown that the number of paths of length n emanating from a node is an indicator of the importance of the node, as the following theorem asserts.



(a) citation graph (b) co-citation graph (c) bibliographic coupling graph

Figure 1: A citation graph, and the induced bibliographic coupling and co-citation graphs.

Theorem 3.1 *Let λ be the dominant eigenvalue of a nonnegative symmetric matrix $A^{\mathrm{T}}A$. If λ is a simple eigenvalue, there exists an eigenvector* $\mathbf{v}$ *corresponding to λ such that $(A^{\mathrm{T}}A/\lambda)^n \to \mathbf{v}\mathbf{v}^{\mathrm{T}}$ as $n \to \infty$.*

An implication of this theorem is that for every row (and column) vector of $(A^{\mathrm{T}}A)^n$, the node ranking induced by the magnitude of its components tends towards the HITS authority ranking (given by the dominant eigenvector $\mathbf{v}$), if the co-citation graph is connected.

Summing $(A^{\mathrm{T}}A)^n$ over $n = 1, 2, \ldots$ as in eq. (5) can thus be interpreted as the mixture of relatedness (when n is small) and importance (when n is large). As a special case, the Neumann kernels subsume co-citation and bibliographic coupling at $\gamma = 0$. On the other hand, at the ceiling of the parameter range, the rankings induced by the Neumann kernels are also identical to the HITS importance, as stated by the following theorem.

Theorem 3.2 *Let λ be the dominant eigenvalue of a nonnegative symmetric matrix $A^{\mathrm{T}}A$. If λ is a simple eigenvalue, there exists a unit eigenvector* $\mathbf{v}$ *corresponding to λ such that*

$$\left(\lambda^{-1} - \gamma\right) N_\gamma(A^{\mathrm{T}}A) \to \mathbf{v}\mathbf{v}^{\mathrm{T}} \qquad \text{as } \gamma \to \lambda^{-1} - 0.$$

4. LAPLACIAN KERNELS AS A RELATEDNESS MEASURE

4.1 Limitations of co-citation relatedness

In this section, we present different link analysis measures based on kernels, with the intention of overcoming the limitations of co-citation and bibliographic coupling relatedness. The two limitations we address are as follows.

Limitation 1 Co-citation coupling assigns a non-zero relatedness score to a pair of documents only if they are commonly referenced by a document.

In Figure 1(a), documents n_1 and n_3 are not jointly cited by any document, resulting in the absence of edge (n_1, n_3) in the co-citation graph (Figure 1(b)). Accordingly, they are not related to each other in terms of co-citation coupling. Because real-world networks are typically sparse, it is often desirable to even capture weak relationship between nodes such as n_1 and n_3 in this case. The relationship between these nodes might not be as strong as n_1 and n_2, or n_2 and n_3, but the fact that they are both co-cited with n_2 by other nodes still conveys a valuable piece of information.

Limitation 2 Co-citation coupling determines the relatedness between nodes i and j only on the basis of the number of nodes commonly citing the two. Nodes citing only one of i and j are neglected, and the number of citations from those nodes does not affect the relatedness between i and j in any way.

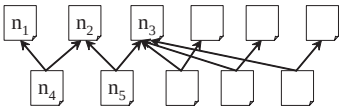


Figure 2: A citation graph illustrating the limitation of co-citation relatedness.

To see why Limitation 2 is an issue, consider the graph of Figure 2. In this graph, n_3 represents a frequently linked web page such as Google and Yahoo. Node n_1 is a much less popular page cited only by n_4. Our intuition dictates that n_2 is more related to n_1 than to n_3, as one would not conclude a page is related (or similar) to Google or Yahoo just because it is jointly cited with them. However, each of the pairs (n_1, n_2) and (n_2, n_3) has a co-citation count of one. In effect, n_1 and n_3 are estimated as equally related to n_2 in terms of co-citation relatedness.

The Neumann kernels do not give a solution to these limitations. Limitation 1 does not appear to be a problem for the Neumann kernels, as they count paths of any length if parameter $\gamma > 0$. However, an increase in γ, no matter how small, biases the induced measures towards importance at the same time, making them unsuitable for evaluating relatedness. This bias also incurs Limitation 2. When applied to the graph of Figure 2, the Neumann kernels with non-zero γ regards n_2 as more related to n_3 than to n_1, which contradicts our intuition even further than co-citation coupling.

The submatrix of the Neumann kernel with $\gamma = 0.18 \simeq 0.9\lambda^{-1}$ for nodes n_1 through n_3 is shown below, where λ is the spectral radius of the co-citation coupling matrix.

$$N_{0.18} = \begin{pmatrix} 1.89 & 3.05 & 3.40 \\ 3.05 & 8.34 & 15.49 \\ 3.40 & 15.49 & 46.12 \end{pmatrix}.$$

Note that $N_{0.18}(2,1) = 3.05 < 15.49 = N_{0.18}(2,3)$. The same holds for smaller γ as well. At $\gamma = 0.02 \simeq 0.1\lambda^{-1}$, for instance, $N_{0.02}(2,1) = 1.06 < 1.13 = N_{0.02}(2,3)$.

4.2 Regularized Laplacian kernels

We show that the kernels based on the *graph Laplacian* [2] overcomes the limitations of Section 4.1. Let G be an undirected graph with positive edge weights, and B be its adjacency matrix. The *Laplacian* of G is defined as $L(B) = D(B) - B$, where $D(B)$ is a diagonal matrix with $D(B)(i,i) = \sum_j B(i,j)$. Smola and Kondor define the regularized Laplacian kernels [12]) as follows.

Definition 4.1 Let B be a nonnegative symmetric matrix, G be its induced undirected graph, and let $\gamma \geq 0$. The matrix

$$R_\gamma(B) = (I + \gamma L(B))^{-1} \tag{6}$$

is called the *regularized Laplacian kernel* on G with diffusion factor γ.

To ensure the symmetry of B, we take as B the co-citation matrix $A^{\mathrm{T}}A$ or the bibliographic coupling matrix AA^{T}.

Provided that $\gamma < 1/\rho(L(B))$, the right-hand side of (6) is the closed form solution to the series

$$R_\gamma(B) = \sum_{n=0}^{\infty} \gamma^n \left(-L(B)\right)^n . \tag{7}$$

It is also obtainable by using $-L(B)$ in place of the adjacency matrix B and dropping the first factor B from eq. (4). Note however that but $R_\gamma(B)$ in eq. (6) may exist even if $\gamma \geq 1/\rho(L(B))$, i.e., the infinite series in eq. (7) does not converge. In practice, restricting the parameter range to $\gamma < 1/\rho(L(B))$ has a merit in that an approximate computation method based on infinite series representation is applicable (see Section 5.1).

In addition, the infinite series representation allows the interpretation of kernel computation as path counting, parallelling the discussion of Section 3.3. In the case of the regularized Laplacian kernels, counting takes place not in co-citation or a bibliographic coupling graphs, but in the graph induced by taking the negative of their Laplacian as the adjacency matrix. The difference is that self-loop edges in the latter graph have negative weights.

4.3 Relatedness measure induced by the regularized Laplacian kernels

The regularized Laplacian kernels remain a relatedness measure even if diffusion factor γ is increased, by virtue of negative weights assigned to self-loop edges. During path counting, paths through these loops are also taken into account. As a result, authoritative nodes receive larger discounting, as the loops at authoritative nodes typically have a heavier weight; as seen from the definition of the Laplacian, the weight of a loop is the negated sum of the weight of the (non-loop) edges incident to the node.

Recall the graph depicted in Figure 2. As argued previously, it is more natural to regard n_2 as more related to n_1 than to n_3. The regularized Laplacian kernel matches this intuition and assigns a greater relatedness score to n_1 than to n_3 relative to n_2.

Below is the regularized Laplacian kernel with $\gamma = 0.18 \simeq 0.9\lambda^{-1}$, where λ is the spectral radius of the negative Laplacian of the co-citation matrix.

$$R_{0.18} = \begin{pmatrix} 0.87 & 0.12 & 0.01 \\ 0.12 & 0.76 & 0.08 \\ 0.01 & 0.08 & 0.62 \end{pmatrix} .$$

Again, only the submatrix of the kernel for nodes n_1 through n_3 is shown. Here, we have $R_{0.18}(2,1) > R_{0.18}(2,3)$.

If discounting high-degree nodes is all that is needed, one may argue that there should be simpler ways. Even though some of these straightforward discounting methods may work at a relatively small γ, as γ gets larger, they are either biased towards importance, or give a measure inconsistent with our intuition on relatedness. For instance, simply normalizing the Neumann kernel matrices by $\bar{N}(i,j) = N(i,j)/\sqrt{N(i,i)N(j,j)}$ gives $\bar{N}_{0.18}(2,1) = 0.77 < 0.79 = \bar{N}_{0.18}(2,3)$. Using the distance of nodes in the kernel-induced feature spaces does not work either.

Another possibility may be to use the column transition matrix $\bar{A}$, obtained by normalizing the adjacency matrix A so that its column sums equal to one, and apply the Neumann kernels to the matrix $\bar{A}^{\mathrm{T}}\bar{A}$. Again, this method works at relatively small γ, but increasing γ towards its ceiling yields a strange measure which does not appear to be either importance or relatedness.

At $\gamma = 0.68 \simeq 0.9\lambda^{-1}$ (λ is different from above due to reweighting), this method gives

$$N_{0.68}(\bar{A}^{\mathrm{T}}\bar{A}) = \begin{pmatrix} 9.25 & 5.67 & 0.87 \\ 5.67 & 3.80 & 0.81 \\ 0.87 & 0.81 & 1.31 \end{pmatrix} .$$

It evaluates n_2 as more related to n_1 than to n_3 as desired. Note however that $N_{0.68}(\bar{A}^{\mathrm{T}}\bar{A})(3,1) > N_{0.68}(\bar{A}^{\mathrm{T}}\bar{A})(3,2)$, which means that n_3 is more related to n_1 than to n_2. This is against intuition since the relatedness between n_3 and n_1 must be deduced from the paths from n_3 to n_1 in the co-citation graph, all of which pass through n_2 on the way. Indeed, Theorem 3.2 shows that $N_\gamma(\bar{A}^{\mathrm{T}}\bar{A})$ in the limit $\gamma \to 1/\lambda$ associates identical ranking $n_1 > n_2 > n_3$ to all nodes n_1, n_2, and n_3.

By contrast, such anomaly is not present in the regularized Laplacian kernels in the limit of γ. The following theorem states that these kernels in the limit assign a uniform score to all the nodes in the same connected component of the graph induced by B.

Theorem 4.1 *Let $B \in \mathbb{R}^{m\times m}$ be a nonnegative symmetric irreducible matrix. The regularized Laplacian kernel $R_\gamma(B)$ converges to $(1/m)\mathbf{1}\mathbf{1}^{\mathrm{T}}$ as $\gamma \to \infty$.*

A question at this point is whether the regularized Laplacian kernels are nonnegative so that one can use the vectors in the kernel matrices as a score vector, in a manner similar to the Neumann kernels. The proof does not seem so straightforward as the latter because of the negative elements in the Laplacian, but the proof for the case where $\gamma < 1/\rho(B)$ can be obtained as a corollary to Theorem 4.2 we present in the next section.

4.4 Controlling bias

In the regularized Laplacian kernels, γ cannot be used for controlling the bias between importance and relatedness, as they remain a relatedness measure regardless of the value of γ. To control bias in these kernels, we introduce a new parameterization scheme.

Definition 4.2 Let G be an undirected graph with positive weights, and B be its adjacency matrix, and let $0 \leq \alpha \leq 1$. We define the *modified Laplacian* $L_\alpha(B)$ of G as $L_\alpha(B) = \alpha D(B) - B$, where $D(B)$ is a diagonal matrix with $D(B)(i,i) = \sum_j B(i,j)$ as before.

Definition 4.3 Let B be a nonnegative symmetric matrix, and G be its induced graph. For $\gamma \geq 0$ and $0 \leq \alpha \leq 1$, if the series

$$R_{\gamma,\alpha}(B) = \sum_{n=0}^{\infty} \gamma^n \left(-L_\alpha(B)\right)^n . \tag{8}$$

is convergent, we call $R_{\gamma,\alpha}(B)$ the *modified regularized Laplacian kernel* on G.

At $\alpha = 1$, $R_{\gamma,\alpha}(B)$ reduces to the (original) regularized Laplacian kernel $R_\gamma(B)$ representing relatedness between nodes. As α decreases towards 0, each row (column) vector of the kernel matrix bears more and more the character of an importance vector, provided that γ is sufficiently large. In particular, at $\alpha = 0$, $R_{\gamma,\alpha}(B)$ reduces to $I + \gamma N_\gamma(B)$, where $N_\gamma(B)$ is given by eq. (4). The following theorem states the property of the modified regularized Laplacian kernels.

Theorem 4.2 *For any nonnegative symmetric matrix B, the modified regularized Laplacian kernel $R_{\gamma,\alpha}(B)$, if it converges, is doubly nonnegative, i.e., (element-wise) nonnegative and symmetric positive semidefinite.*

Nonnegativity means that the vectors in the kernel matrices can be interpreted as score vectors. Positive semidefiniteness implies that the kernels define an inner product in some feature space and hence they are compatible with Support Vector Machines and other state-of-the-art kernel-based machine learning tools.

4.5 Diffusion kernels

All the kernels presented above are based on the Neumann series, but other series can be used. Using the matrix exponential in place of the Neumann series yields the so-called *diffusion (heat) kernels*, originally developed in the context of spectral graph theory [2]. It was first introduced to machine learning community by Kondor and Lafferty [8].

Definition 4.4 Let G be an undirected graph with positive weights, and B be its adjacency matrix. The *diffusion kernel* matrix H_γ on G with diffusion factor $\gamma \geq 0$ is given by

$$H_\gamma(B) = \exp(-\gamma L(B)) = \sum_{n=0}^{\infty} \frac{\gamma^n \left(-L(B)\right)^n}{n!}. \qquad (9)$$

It can be shown that $H_\gamma(B)$ also converges to a uniform matrix as $\gamma \to \infty$.

The modified Laplacian $L_\alpha(B)$ can be used in place of the Laplacian $L(B)$ with diffusion kernels as well. The resulting kernel $H_{\gamma,\alpha}(B) = \exp(-\gamma L_\alpha(B))$ is positive semidefinite, and allows for controlling the bias between relatedness and importance just like the modified regularized Laplacian kernels.

In parallel to Theorem 3.2, it can be shown that when $\mathbf{v}$ is the HITS authority vector of a graph whose adjacency matrix is A and λ is the dominant eigenvector of $A^{\mathrm{T}}A$, $H_{0,\gamma}(A^{\mathrm{T}}A)/\exp(\gamma\lambda)$ converges to $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ as $\gamma \to \infty$.

5. PRACTICAL ISSUES

In this section, we discuss some issues that may be encountered in the practical application of the kernel-based link analysis. Empirical results demonstrating the effectiveness of the methods proposed below are presented in the companion technical report [10].

5.1 Computational issues

Computing the entire kernel matrix requires matrix inversion or exponentiation, and hence its computational complexity is roughly $O(|V|^3)$ where $|V|$ is the number of nodes in the graph, and this may be a computational burden with large graphs. However, the standard techniques for matrix computation [4, §11.2] allow approximating kernel computation with the sum of the first k terms of the infinite series in eqs. (5), (6), and (8). The approximation error is bounded by $(|V|/k!)\,((\gamma\lambda)^{-1} - 1)^{-1/2}$, where λ is the spectral radius of the co-citation matrix.

Furthermore, if one is concerned with the importance of nodes relative to a single node i rather than the entire kernel matrix, or if the entire kernel matrix cannot be kept on memory, we can reduce the space requirement by summing $\left(\gamma A^{\mathrm{T}}A\right)^n \mathbf{u}_i$ over $n = 1, \ldots, k$, where $\mathbf{u}_i$ is a unit vector with only 1 at the i-th component; the computation now reduces to that of vector sums and the matrix-vector multiplication similar to HITS.

5.2 Parameter tuning

All the kernels in the previous sections are parameterized. In the Neumann kernels, the parameter γ controls the tradeoff between relatedness and importance. Setting the right parameter value hence emerges as an issue in practical application of these kernels. Unfortunately, the optimality condition according to which the parameter must be tuned seems highly dependent on individual tasks. Consider a paper recommendation system, for example. Given a small list of the 'root' papers the user considered interesting, the system should recommend other papers that may be of interest to the user. The degree of the user's acquaintance with the field of the root papers should affect how much the system should bias (through parameter tuning) its decision towards authoritative papers in the field, but such knowledge on the user is often outside the scope of link analysis.

If parameter setting requires external knowledge (e.g., user modeling), a practical alternative should be to present the user (or the external user-modeling module) kernel matrices with various parameter settings, and let them choose the most suitable one. This approach requires a way to choose sample points efficiently; as we will see in Section 6, the character of the link analysis measures induced by these kernels is far from linear to the parameters, making sampling at uniform intervals a non-viable option.

We point out that the derivatives of kernel matrices with respect to the bias parameter can be used to efficiently determine sample points. For some kernels, derivatives can be analytically computed from kernel matrices at a given point γ as follows.

$$\frac{\partial N_\gamma(B)}{\partial\gamma} = \left(N_\gamma(B)\right)^2, \qquad (10)$$
$$\frac{\partial R_\gamma(B)}{\partial\gamma} = -L(B)\left(R_\gamma(B)\right)^2,$$
$$\frac{\partial H_\gamma(B)}{\partial\gamma} = -L(B)\,H_\gamma(B).$$

Let us take the Neumann kernel $N_\gamma(B)$ as an example. Suppose we have $N_\gamma(B)$ for some γ at hand. We can compute the first order approximation $\tilde{N}_{\gamma+\Delta\gamma}(B)$ of the matrix $N_{\gamma+\Delta\gamma}(B)$ as

$$\tilde{N}_{\gamma+\Delta\gamma}(B) = N_\gamma(B) + \Delta\gamma\frac{\partial N_\gamma(B)}{\partial\gamma}, \qquad (11)$$

where $\partial N_\gamma(B)/\partial\gamma$ is given by eq. (10). By comparing the rankings induced by $\tilde{N}_{\gamma+\Delta\gamma}(B)$ and $N_\gamma(B)$, we can estimate how likely the change may occur in a given range $[\gamma, \gamma+\Delta\gamma]$. The cost of this estimation is that of matrix multiplication and summation in eqs. (10) and (11); there is no need to compute $N_{\gamma+\Delta\gamma}$ every time from scratch, until a suitable sampling interval $\Delta\gamma$ is determined.

6. EXPERIMENTAL EVALUATION

To evaluate the characteristics of the kernel-based link analysis measures introduced in the previous sections, we applied them to a co-citation graph of papers on natural language processing, which is a connected graph consisting of 2280 nodes (papers).

Each kernel matrix was treated as a ranking method by taking the i-th row vector of the matrix as the score vector for the i-th node. Given the ranking induced by the i-th row vector, we call the i-th node as the *root node* of this ranking.

Following [13], we use the minimizing Kendall (K-min) distance [3] between the top-10 items to evaluate the (dis)similarity of rankings. A small K-min distance means the two top-10 rankings are similar. It is equal to 0 if all top-10 items are identical, and takes the maximum value of 100 if there are no common items in the top-10 lists.

6.1 Neumann kernels

Table 1 shows the K-min distance between the top-10 lists induced by the Neumann kernels and HITS, averaged over all 2280 root nodes. The diffusion factor γ for the kernels is shown as a

Table 1: K-min distance between HITS and the Neumann kernels.

$\gamma\lambda$	0.01	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999
K-min	87.4	87.3	86.3	72.0	26.4	5.5	1.1	0.0

Table 2: K-min distance between HITS and the regularized Laplacian kernels.

$\gamma\lambda$	0.01	0.1	0.5	0.999	10	100	1000	HITS
0.01	0.0	0.1	0.1	0.5	6.2	17.1	24.3	95.7
0.1		0.0	0.1	0.4	6.2	17.1	23.2	95.7
0.5			0.0	0.4	6.1	17.0	22.5	95.8
0.999				0.0	5.8	16.8	22.3	96.1
10					0.0	11.4	19.5	99.1
100						0.0	8.4	99.7
1000							0.0	99.9

normalized factor relative to $1/\lambda$, where λ is the spectral radius of the co-citation matrix; thus the admissible parameter range is $0 \leq \gamma\lambda < 1$. Table 1 indicates that the rankings induced by the Neumann kernels are biased towards the HITS ranking as γ is increased.

6.2 Regularized Laplacian kernels

Table 2 lists the average K-min distance among the rankings of HITS and the regularized Laplacian kernels with various γ, with the average taken over 2280 root nodes.

All over the parameter range shown in the table, the K-min distance consistently exceeds 95, meaning that the induced rankings do not resemble that of HITS. Recall that the distance between the rankings of HITS and the Neumann kernel at $\gamma = 0.01$ was 87.4 (see Table 1), which is also large, but not as large as that of the regularized Laplacian kernels. This result suggests that bias towards importance persists in the Neumann kernels even with a small γ.

The regularized Laplacian kernels are extremely stable over the parameter range of $0.01 \leq \gamma < 1$; the difference in K-min distance between the rankings at $\gamma = 0.01\lambda^{-1}$ and $0.999\lambda^{-1}$ is less than 1. The increase in the distance of rankings between $\gamma = 0.01\lambda^{-1}$ and $\gamma \geq 10\lambda^{-1}$ is not because the measure is inclined towards importance as γ is increased, but because increased γ makes the kernel more and more 'uniform,' as asserted in Theorem 4.1.

To see if the regularized Laplacian kernels indeed give a relatedness measure, we need to measure the correlation between these kernels and co-citation coupling. However, this time we cannot use the K-min distance as the index of dissimilarity of their ranking lists, because the co-citation ranking often includes a number of ties which cannot be handled by the K-min distance. Instead, we have verified that for every root paper in the dataset, all the papers that are co-cited with the root paper are ranked topmost by the regularized Laplacian kernels with $\gamma = 0.1\lambda^{-1}$ and $0.01\lambda^{-1}$.

Although omitted for the lack of space, the rankings of diffusion kernels show tendencies similar to the regularized Laplacian kernels.

6.3 Modified regularized Laplacian kernels

To verify parameter α controlling the tradeoff between relatedness and importance in the modified regularized Laplacian kernels, we compared the rankings induced by these kernels with those of HITS and the regularized Laplacian kernel with $\gamma = 0.1\lambda^{-1}$. The latter two are used as the benchmark of the importance and relatedness measures, respectively. Again, we did not use co-citation coupling as the baseline for relatedness because of ties in its ranking. The regularized Laplacian kernel was used instead, on the basis of the result presented in Section 6.2, which showed that this kernel is a legitimate alternative to co-citation. The diffusion factor γ is set to $0.99999\lambda^{-1}$, where λ is the spectral radius of the Laplacian.

Table 3: K-min distance between the modified regularized Laplacian kernels and other link analysis measures: HITS (as a baseline measure of importance), and the unmodified regularized Laplacian kernel with $\gamma = 0.1\lambda^{-1}$ (as a measure of relatedness).

α	0.01	0.05	0.1	0.15	0.2	0.3	0.5	0.75
HITS	0.0	12.9	35.0	51.8	89.6	94.7	96.0	96.1
Unmodified RLK	95.0	93.3	95.3	94.3	33.5	17.3	8.2	4.1

The result is shown in Table 3. The modified regularized Laplacian kernels tend to be more similar to HITS as α is decreased.

6.4 An illustrating example

We conclude this section with an example illustrating the characters of the kernels. Table 4 shows the part of the ranking lists for the root paper 'Empirical studies in discourse' by M. A. Walker and J. D. Moore, *Computational Linguistics* 23(1):1–12, 1997. The table lists all the 22 papers that are ranked as top 10 in at least one of the 9 ranking lists shown in the right-hand side of the table.

An interesting observation with this root paper is that it makes a real-world example of the toy graph of Figure 2; the root paper is n_2, and the most authoritative paper (Penn Treebank) is n_3. All the other co-cited papers concern discourse just like the root paper, and presumably, they are more related to the root paper than the Penn Treebank paper is. As a result, each of these co-cited papers on discourse corresponds to n_1.

Compare the ranking lists of the Neumann kernel (NK) and the regularized Laplacian (RLK) with $\gamma\lambda = 0.1$. In the two lists, the set of the top seven papers (including the root paper itself) is identical, with all the seven papers being those with non-zero co-citation (CC) scores.

Looking inside the rankings of these seven papers, we find that both kernels place the root paper at the top of the ranking, but for the other six papers, the rankings they produce are the inverse of each other. The Neumann kernel's ranking of these six matches that of HITS, while that of the regularized Laplacian kernel is in the opposite order of their HITS rankings.

7. RELATED WORK

7.1 Relative importance

Relative importance is a new link analysis measure recently proposed by White and Smyth [13]. This measure is defined as the 'importance of nodes in a graph relative to one or more root nodes.' In this view, HITS and PageRank are 'global' importance algorithms. White and Smyth made a convincing argument that simply applying global importance algorithms to the subgraph surrounding the root nodes does not yield a precise estimate of relative importance, because the root nodes are not given any special preference during importance computation.

Our kernel-based link analysis measures fit naturally as relative importance, and as a bonus clarify the relationship between relative importance and relatedness (namely, the co-citation and bibliographic coupling relatedness), an issue not addressed previously.

7.2 Kernels and link analysis

Smola and Kondor [12] pointed out the connection between graph kernels and importance computation methods including HITS, in a formulation different from ours. There are two differences between

Table 4: Rankings relative to the root paper 'Empirical studies in discourse': HITS (H), co-citation (CC), Neumann kernels (NK) and regularized Laplacian (RLK) kernels. The column 'Topic' shows the topic category of each paper: discourse (D), parsing (P), standard data set (S), machine translation (T), and word-sense disambiguation (W). A '–' in column CC indicates that the paper was not co-cited with the root paper.

				NK ($\gamma\lambda$)			RLK ($\gamma\lambda$)			
Paper title	Topic	H	CC	0.1	0.9	0.99	0.1	0.999	10	100
Building a large annotated corpus of English: the Penn Treebank	C	**1**	**2**	**2**	**1**	**1**	**7**	**7**	**7**	15
A stochastic parts program and noun phrase parser for unrestricted text	P	**2**	–	12	**7**	**2**	12	23	184	407
Statistical decision-tree models for parsing	P	**3**	–	**8**	**5**	**3**	**8**	**8**	86	323
A new statistical parser based on bigram lexical dependencies	P	**4**	–	**9**	**6**	**4**	**10**	**10**	90	334
Unsupervised word sense disambiguation rivaling supervised methods	W	**5**	–	62	15	**5**	75	134	328	523
Word-sense disambiguation using statistical models of Roget's categories trained	W	**6**	–	365	12	**6**	32	54	207	380
The mathematics of statistical machine translation	T	**7**	–	374	26	**9**	551	553	745	986
Three generative, lexicalised models for statistical parsing	P	**8**	–	11	11	**7**	11	14	113	365
Transformation-based error-driven learning and natural language processing	P	**9**	–	21	14	**8**	24	47	242	550
Integrating multiple knowledge sources to disambiguate word sense	W	**10**	–	63	18	**10**	74	113	264	449
Attention, intentions, and the structure of discourse	D	50	**2**	**3**	**3**	25	**6**	**6**	**6**	**8**
Assessing agreement on classification tasks: the kappa statistic	D	76	**2**	**4**	**4**	41	**5**	**5**	**5**	**5**
Centering: a framework for modeling the local coherence of discourse	D	96	–	**10**	33	90	**9**	**9**	**10**	65
A prosodic analysis of discourse segments in direction-giving monologue	D	198	–	15	96	182	15	11	**8**	14
The reliability of a dialogue structure coding scheme	D	201	**2**	**5**	**8**	94	**4**	**4**	**4**	**4**
Message Understanding Conference tests of discourse processing	D	604	**2**	**6**	**9**	156	**3**	**3**	**3**	**3**
Discourse segmentation by human and automated means	D	685	–	62	336	691	52	31	**9**	**9**
Empirical studies in discourse	D	771	**1**	**1**	**2**	95	**1**	**1**	**1**	**1**
Human-machine problem solving using spoken language systems	D	1026	–	203	786	1171	146	133	40	**10**
Experiments in evaluating interactive spoken language systems	D	1046	–	204	805	1199	144	131	24	**7**
Preventing false inferences	D	1054	–	205	812	1213	145	132	25	**6**
Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue	D	1061	**2**	**7**	**10**	205	**2**	**2**	**2**	**2**

their formulation and ours. (1) They state that for a given node in a regular graph, its HITS score is given by the length of its corresponding vector in the feature space induced by Laplacian-based kernels. By contrast, we interpret the elements of kernel matrices, not just the diagonals, as indicating the score of importance (or relatedness, or relative importance). (2) Our formulation obtains the HITS importance scores as an extremum of the Neumann kernels, which does not rely on the Laplacian. In addition, their argument is solely concerned with importance, and relatedness and intermediates between these are left out of discussion.

8. CONCLUSIONS AND FUTURE WORK

This paper has investigated the properties of graph kernels viewed as link analysis measures. The Neumann kernels provide a unified framework that accounts for the co-citation/bibliographic coupling relatedness and the HITS importance. Laplacian-based kernels define relatedness measures that are free from the limitations of co-citation and bibliographic coupling. We have also proposed an alternative parameterization for the Laplacian-based kernels to control the tradeoff between the Laplacian-based relatedness and HITS importance.

In future work, we plan to investigate the property of the kernels based on the normalized version of the Laplacian [2] as link analysis measures.

9. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Network and ISDN Systems*, 30(1–7):107–117, 1998.

[2] F. R. K. Chung. *Spectral Graph Theory*. Am. Math. Soc., Providence, RI, USA, 1997.

[3] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computation*. Johns Hopkins Univ. Press, 3rd edition, 1996.

[5] J. Kandola, J. Shawe-Taylor, and N. Cristianini. Learning semantic similarity. In *NIPS 15*, pages 673–680, 2003.

[6] M. M. Kessler. Bibliographic coupling between scientific papers. *Am. Documentation*, 14(1):10–25, 1963.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[8] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proc. 18th ICML*, pages 21–24, 2001.

[9] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, USA, 2002.

[10] M. Shimbo and T. Ito. Application of kernels to link analysis: proofs and additional experimental results. Technical report, Grad. School of Inform. Science, Nara Institute of Science and Technology, 2005. In preparation.

[11] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Am. Soc. for Inform. Science*, 24:265–269, 1973.

[12] A. J. Smola and R. Kondor. Kernels and regularization of graphs. In *Proc. COLT'03*, pages 144–158, 2003.

[13] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proc. KDD'03*, pages 266–275, 2003.