

Hubs and Authorities in a Hyperlinked Environment

1 Searching the World Wide Web

Because diverse users each modify the link structure of the WWW within a relatively small scope (by creating web-pages on their own sites that link to other web-pages), the global structure of the WWW is unpredictable and must be analyzed a posteriori. To make effective use of the wealth of information that the WWW provides, we need a way of extracting meaningful information with respect to a global search query that we make. We focus on *broad-topic* queries, i.e., “Tell me about automobile manufacturers.” A naive search strategy might be to return any page containing the phrase “automobile manufacturers”. Unfortunately, there are literally hundreds of thousands of webpages meeting this criteria, far too many for a person to look through alone. Furthermore, the websites of major automobile manufacturers are unlikely to even contain the phrase “automobile manufacturers”. So simply returning the pages with the most occurrences of the query string will obviously not suffice (we call these pages the “text-based results”). Moreover, we maintain that any real measure of the quality of a search result is inherently subjective, as it and has much to do with the context and desires of the individual making the query.

We want to identify the authoritative, informative web pages for a particular query. In the above example, the Toyota home page and important reviews might be considered authorities, while the personal home page of someone who makes his own personal cars might not. But without a formal definition of what makes a web-page an authority on a subject, we can’t codify an algorithm to identify them. Suppose you are provided the text-based search results, and have the ability to determine which pages link to which others. So what is a good heuristic for making this identification? A moment’s reflection will reveal that it is not at all obvious how to proceed. It has been suggested that authoritative pages may be identified by their popularity, as measured by the number of pages that link to it. That would imply that any very popular page was authoritative for any string it contains, e.g., Yahoo would be an authoritative page on Jessica Simpson. Clearly a more sophisticated approach is needed. In section 3 we describe the seminal work of J.M. Kleinberg [1] on using spectral analysis to find “hubs” and “authorities” for a broad-topic query over the WWW, an approach that has stood up well in practice. In section 4 we will discuss some generalizations of this technique and its relation to others, such as Google’s PageRank.

2 Linear Algebra Preliminaries

From a mathematical point of view, the underlying structure of the WWW at any given time is a directed graph. We employ spectral analysis techniques on graphs, meaning we study the properties of graphs using linear algebra. We assume the reader is familiar with elementary graph theory and linear algebra. In particular, we assume a knowledge of the basics of matrix algebra and the notions of eigenvectors, eigenvalues, and eigenspaces, as well as more advanced topics like diagonalization, positive (semi)-definite matrices, and the Spectral Theorem. For a refresher, we refer to the reader to [3]. The way to turn graphs into linear algebra is to consider the adjacency matrix, which represents the connectivity of the graph, and is formally defined as follows.

Definition 2.1 Let $G = (V, E)$ be an undirected (resp. directed) graph and let $|V| = n$. The *adjacency matrix* associated to G is the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ where $A_{i,j} = 1$ if $\{v_i, v_j\} \in E$ (resp. $(v_i, v_j) \in E$) and 0 otherwise.

Spectral analysis uses the spectra (the eigenvalues) of A to obtain information about the graph. We collect here some specialized results from linear algebra that we will use.

Lemma 2.2 Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix. Then $\mathbf{A}\mathbf{A}^T$ is a semi-positive definite matrix (so in particular, it is also square and symmetric).

Proof: Let λ be an eigenvalue of $\mathbf{A}\mathbf{A}^T$ and \mathbf{q} be the eigenvector corresponding to λ .

$$\begin{aligned} (\mathbf{A}\mathbf{A}^T)\mathbf{q} &= \lambda\mathbf{q} \\ \mathbf{q}^T(\mathbf{A}\mathbf{A}^T\mathbf{q}) &= \mathbf{q}^T(\lambda\mathbf{q}) \\ &= \lambda(\mathbf{q}^T\mathbf{q}) \\ \frac{\mathbf{q}^T\mathbf{A}\mathbf{A}^T\mathbf{q}}{\mathbf{q}^T\mathbf{q}} &= \lambda \end{aligned}$$

$\lambda = \frac{\mathbf{z}^T\mathbf{z}}{\mathbf{q}^T\mathbf{q}}$, where $\mathbf{z} = \mathbf{A}^T\mathbf{q}$. Note that for any vector $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v}^T\mathbf{v} > 0$. $\mathbf{q}^T\mathbf{q} > 0$ and $\mathbf{z}^T\mathbf{z} \geq 0$, therefore $\lambda \geq 0$. ■

The Spectral Theorem tells us that any positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be decomposed uniquely as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ for some orthonormal matrix \mathbf{Q} of eigenvectors forming an orthonormal basis for \mathbb{R}^n , and diagonal matrix $\mathbf{\Lambda}$ of corresponding eigenvalues. We stress that we have $\mathbf{Q}^T = \mathbf{Q}^{-1}$ by orthonormality. This gives us the following lemma.

Lemma 2.3 Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and \mathbf{q} be the eigenvector of $\mathbf{A}\mathbf{A}^T$ corresponding to λ . Then λ^k is an eigenvalue of $(\mathbf{A}\mathbf{A}^T)^k$ corresponding to the eigenvector \mathbf{q} .

Proof: By Lemma 2.2, $\mathbf{A}\mathbf{A}^T$ is positive semi-definite. So we write

$$\begin{aligned} (\mathbf{A}\mathbf{A}^T)^k &= (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T)^k \\ &= (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1})^k \\ &= \mathbf{Q}\mathbf{\Lambda}^k\mathbf{Q}^{-1}. \end{aligned}$$

Uniqueness of this decomposition proves the lemma. ■

Lemma 2.4 Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same multi-set of non-zero eigenvalues. Moreover, if \mathbf{q} is an eigenvector of $\mathbf{A}\mathbf{A}^T$ then $\mathbf{A}^T\mathbf{q}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$.

Proof: Let λ be a non-zero eigenvalue of $\mathbf{A}\mathbf{A}^T$ and \mathbf{q} be the eigenvector corresponding to λ .

$$\begin{aligned} (\mathbf{A}\mathbf{A}^T)\mathbf{q} &= \lambda\mathbf{q} \\ \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)\mathbf{q} &= \mathbf{A}^T(\lambda\mathbf{q}) \\ (\mathbf{A}^T\mathbf{A})(\mathbf{A}^T\mathbf{q}) &= \lambda(\mathbf{A}^T\mathbf{q}), \end{aligned}$$

which implies both statements. ■

3 Hubs, Authorities, and Focused Subgraphs

Automotive enthusiasts across the world have already compiled lists of useful, authoritative web pages regarding automobile manufacturers, and posted them online. We’ll refer to pages like these, which link to many authoritative sources, as “hubs”. We then somewhat circularly define an “authority” to be a page that is linked to by many “hubs”. An authoritative page p then has not only a large in-degree (cf. the naive approach to this problem outlined in the introduction), but among hubs must be a sort of consensus that p is worth linking to.

3.1 Focused Subgraph of a Query

Kleinberg’s algorithm takes as input a directed graph representing the link structure of a collection of web-pages. However, running the algorithm on the graph of the entire WWW would be impractical. To focus our computational effort, we want to find a subgraph of the WWW satisfying the following properties:

- It is small enough to be manipulated efficiently.
- It contains many pages relevant to the query.
- It contains many authorities.

We call such a subgraph the “focused subgraph” of a query. The first t results from a text-based search engine such as AltaVista should satisfy the first two requirements (the choice of search engines is arbitrary). As a heuristic, we suggest that the authoritative sites are quite likely have outgoing links to or incoming links from the pages in the first t text-based results. Taking these adjacent pages together with the original t text-based results gives us a subgraph which should have the properties we want. Note that a website may have hundreds of thousands of websites linking to it, so we need a parameter d specifying how many in-links to take.

We let our focused subgraph be the one induced by the set

$$\mathbf{S} = \mathbf{R} \cup \{\text{up to } d \text{ arbitrary pages linking to } \mathbf{R}\} \cup \{\text{pages linked to from } \mathbf{R}\}$$

where R is the set consisting of the first t text-based search results. In Kleinberg’s experiments he finds that reasonable values for t and d are around 200 and 50 respectively.

3.2 Extracting Hubs and Authorities

We let h_p and a_p represent the “hub” and “authority” weights of a page p , respectively. We view pages with higher weights as “better” hubs and authorities. We begin by initializing $h_p = a_p = 1$ for each $p \in V$ (all pages are initially weighted equally). We will then proceed iteratively, updating the hub and authority weights for each page as follows:

$$\begin{aligned} a_p &\leftarrow \sum_{(q,p) \in E} h_q \\ h_p &\leftarrow \sum_{(p,q) \in E} a_q \end{aligned}$$

A page p ’s authority weight is proportional to the sum of the hub weights of the pages that link to p , while p ’s hub weight is proportional to the sum of the authority weights of the pages p links to. Each iteration therefore exploits the mutually reinforcing hub-authority relationship, and the “best” hubs and authorities should have their respective weights increased the most. After each update, we normalize the weights which we will see causes the algorithm to converge. But we stress here that convergence is natural and not the major result here. It is that dense bipartite graphs form in the internet, which we can identify efficiently. The internet was once thought to behave like a random graph!

Using matrix notation, where \mathbf{A} represents the adjacency matrix of the focused subgraph \mathbf{S} , a_i and h_i represent the set of weights $\{a_p\}$ and $\{h_p\}$ after the i -th iteration, one can easily verify that the above operations can also be expressed as:

$$\begin{aligned} a'_i &\leftarrow \mathbf{A}^T h_{i-1} \\ h'_i &\leftarrow \mathbf{A} a_{i-1}, \end{aligned}$$

Similarly, the normalization is expressed as (where $\|\cdot\|$ denotes the 2-norm of vectors in n -space):

$$\begin{aligned} a_i &\leftarrow \frac{a'_i}{\|a'_i\|} \\ h_i &\leftarrow \frac{h'_i}{\|h'_i\|} \end{aligned}$$

Note that in the notation of the above algorithm we have that

$$h_k = \frac{(\mathbf{A}\mathbf{A}^T)^k h_0}{\|(\mathbf{A}\mathbf{A}^T)^k h_0\|} \tag{1}$$

$$a_k = \frac{(\mathbf{A}^T\mathbf{A})^{k-1} \mathbf{A}^T h_0}{\|(\mathbf{A}^T\mathbf{A})^{k-1} \mathbf{A}^T h_0\|} \tag{2}$$

For our analysis, let λ_i and q_i , be the i -th eigenvalue and corresponding eigenvector, of $\mathbf{A}\mathbf{A}^T$. By Lemma 2.2, $\mathbf{A}\mathbf{A}^T = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ is positive definite. Reindexing if necessary, we assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$.

Theorem 3.1 The sequences h_1, h_2, h_3, \dots and a_1, a_2, a_3, \dots converge to limits q_1 and $\mathbf{A}^T q_1$, respectively. Moreover, both of these limits are positive (meaning have all positive entries).

Proof: It can be shown that q_1 (which is called the principal eigenvector) is positive, and since h_0 is the vector of all 1’s, we have that

$$q_1^T h_0 \neq 0 \tag{3}$$

For the first sequence, applying the Spectral Theorem and Lemma 2.3 to $\mathbf{A}\mathbf{A}^T$ gives us

$$(\mathbf{A}\mathbf{A}^T)^k h_0 = \lambda_1^k q_1 q_1^T h_0 + \lambda_2^k q_2 q_2^T h_0 + \dots + \lambda_n^k q_n q_n^T h_0,$$

where we have simply written out each term by the definition of matrix multiplication. Now using (1) we write

$$h_k = \frac{(\mathbf{A}\mathbf{A}^T)^k h_0}{\|(\mathbf{A}\mathbf{A}^T)^k h_0\|} \rightarrow q_1 \text{ as } k \rightarrow \infty$$

To see this last step, note that $q_i q_i^T h_0$ for any i is simply a scalar multiple of q_i , and that λ_i^k is an exponential in k , with λ_1^k having the largest base.

For the second sequence, it can be shown that \mathbf{A}^T is not orthogonal to q_1 , so that $A^T q_1$ is positive. Convergence follows by identical argument to the above. The second statement of the theorem follows from (3) and positive semi-definiteness. ■

For those familiar with the singular value decomposition (a generalization of the decomposition we used in Lemma 2.4), we remark that the limits $q_1, \mathbf{A}^T q_1$ are the left and right singular vectors of the matrix \mathbf{A} , respectively. Since these limits are positive vectors, our *FindHubsAndAuthorities* algorithm really does return the “best” hubs and authorities, in the sense that they have the highest weights in absolute value. One might wonder why we cannot just calculate these limits directly using their symbolic forms. The answer is that computing the eigenvalues and eigenvectors of large matrices is highly unreliable (basically meaningless) due to numerical roundoff. However, in Kleinberg’s experiments, convergence of the vectors using the iterative algorithm was found to be quite rapid – 15 to 20 iterations usually suffices. Though we focus on the theory here, the experiments also show that our technique is extremely effective. We urge the reader to look at Kleinberg’s experiments, in particular the ones that contrast the results of naive approaches with the approach that we develop here. We also emphasize that our approach, most remarkably, is based solely on hyperlink-based analysis of the text-based results. Moreover, though our algorithm performs *global* analysis on the WWW, it does not need space to index all web pages; only local analysis on the text-based results is performed.

4 Some Generalizations and Related Techniques

4.1 Multiple Hubs and Authorities

The beauty of Lemma 2.4 is that it tells us *all* the eigenvectors and eigenvalues of AA^T and $A^T A$ have exactly the same mutually reinforcing relationship exploited by our algorithm. In particular, the non-principal eigenvectors provide a way to get further information about the hub-authorities structure of the focused subgraph. This is especially useful for queries such as “jaguar,” which have multiple meanings. In the experiments, Kleinberg found that the highest weighted pages according to the different eigenvectors give clear separations between the dense bipartite subgraphs corresponding to the different meanings. Note that, unlike for the principal eigenvectors, these eigenvectors have both positive and negative entries. Usually taking the highest weighted positive ones suffices, but this also produces a nice separation in some cases. In particular, for highly divisive issues such as “abortion,” the positive and negative eigenvalues are seen to correspond to websites representing different sides. The algorithms for identifying the non-principle eigenvectors are not as simple, and we omit them here.

4.2 Page Ranking

Many early techniques for searching the WWW based on structural properties were based on some method of counting the number of in-links and out-links. More recent techniques, have centered on the goal of assigning each web page a global “page rank,” a number that somehow tells how important the page is. This notion of importance is propagated via hyperlinks and is analyzed using spectral analysis. This analysis is based on a model in which a user takes a kind of random-walk around the WWW: with probability p the user follows one of the out-links of the current page, and with probability $1 - p$ the user jumps to a completely random page. The probability that this walk converges to page i is the page rank of i .

More concretely, we consider the Brin-Page model [2]. Let us that the WWW has n pages, let A be the adjacency matrix of the (directed) WWW, and let $d - i$ denote the out-degree of page i . In the Brin-Page model, the probability of a transition from page i to page j is given by

$$\mathbf{A}'_{ij} = pn^{-1} + (1 - p)d_i^{-1}\mathbf{A}_{ij}$$

The page rank vector r (i.e., i th component of r is the page rank of page i) is then the positive solution to $(\mathbf{A}')^T r = r$, i.e., the principal eigenvector of $(\mathbf{A}')^T$. In this model, the walk is likely to converge to pages that are somehow “highly connected” to others, and are therefore most important. The random jump to any page is due to the fact that some sequences of pages hit a “dead end” in their conferral of importance.

We highlight some similarities and differences between this type of model and the “hubs-and-authorities approach.” There are obvious similarities – both are based on spectral analysis, and view hyperlinks as conferring some notion of authority or importance (which are similar concepts). However, a key difference is that the approach presented here uncovers further structure in the structure of the WWW using the concept of “hubs.” Moreover, the page ranking technique ranks pages with respect to the whole WWW (which are then used to reorder the results of a specific query later on), whereas the technique presented here computes the hubs and authorities “on-the-fly” for each query; that is to say, hubs and authorities are *local* to each query and do not need to be updated with time.

4.3 Conclusion

We note in passing that there are several other related points one can make here. One is that similar-page queries can be handled by an entirely analogous approach, the only difference being that instead of text-based results one uses all the pages with out-links to the page under consideration. Also, in cases where the query is not sufficiently broad, some further text-based analysis can be helpful in finding desired results. The reader is encouraged to consult [1] for further details.

References

- [1] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [3] Gilbert Strang. *Linear Algebra and its Applications*. Brooks Cole, 1988.