Lecture - 34

Thursday, 3 November 2016 (3:20 - 4:10)

Hubs and Authorities and their Spectral Analysis

In this lecture, we discussed hubs and authorities and their spectral analysis.

1 Hubs and Authorities

When we type "IIT Ropar" in google, the first result that it shows us is www.iitrpr.ac.in, which is the home page of IIT Ropar. Although there are a lot of webpages containing the substring IIT Ropar, yet how does Google "know" that this is the best answer? Search engines determine how to rank pages using automated methods. In this section, we look at a method using which google ranks its results.

The idea is that if a page receives many links from other *relevant* pages, then it receives a kind of collective endorsement. We can operationalize this idea by first collecting a large sample of pages that are relevant to the query (lets call them *lists*) and the let these pages *vote* through their links: whichever page receives the most number of in-links, is the most relevant to IIT Ropar. Another important point is that how do we know which of the pages in the *list* are more relevant than the others? For that, we assign some value to these lists as well. A page's value as a list is equal to the sum of the votes received by all pages for which it voted.

Repeated Improvement If we consider that the pages scoring well on the list are more sensible enough to provide a vote, then their votes should also carry higher weightage. So, we can repeat the voting process again, but this time, one vote will be equal to the value of the list assigning the vote, i.e, we will do weighted voting.

Another point is that, we can keep on repeating this process to get more refined votes on the right hand side. This is called *principle of repeated improvement.*

Hubs and Authorities We call the pages that we are originally seeking, i.e. the pages chosen to be the answers to the given query, as *authorities* of the query. The pages on the list which votee the authorities are called *hubs.*

1. For each page p, we try to find its hub score h(p) and authority score a(p), where h(p) initially starts with 1.

2. For each page p, we update a(p) to be the sum of the hub scores of all the pages that point to it.

3. For each page p, we update h(p) to be the sum of the authority scores of all the pages that it points to.

4. We keep repaeting this for some k steps.

5. In theb process, we may also normalize the hub (and authority) scores by dividing them by the sum of all hub (authority) values since we are only interested in their relative sizes. It turns out that these normalized values actually converge to limits.

1.1 Spectral Analysis of Hubs and Authorities

The use of eigenvalues and eigenvectors to study the structure of networks is referred to as *spectral analysis of graphs.*

We will view a set of n pages as a set of nodes in a directed graph represented by the adjacency matrix M. Now, for a node i, its hub score h_i is calculated as:

$$h_i = M_{i1}a_1 + M_{i2}a_2 + \cdots + M_{in}a_n$$

In matrix form, it may be written as:

$$h = Ma$$

where h and a are the vector of hub scores and authority scores respectively. Similarly,

$$a_i = M1ih_1 + M_{2i}h_2 + \cdots + M_{ni}h_n$$

In matrix form, it may be written as:

$$a = M^T h$$

The k-step Improvement What happens when we perform the k-step hub-authority computation for some k steps? Let a^k and h^k denote the authority and hub scores after k iterations respectively. Then,

$$a^1 = M^T h^0$$
$$h^1 = Ma^1 = MM^T h^0$$

Similarly,

$$a^2 = M^T h^1 = M^T MM^T h^0$$
$$h^2 = Ma^2 = MM^T MM^T h^0 = (MM^T)^2 h^0$$

And,

$$a^3 = M^T h^2 = MTMMTMM^T h^0 = (M^T M)^2 M^T h^0$$
$$h^3 = Ma^3 MM^T MM^T MM^T h^0 = (MM^T)^3 h^0$$

Proceeding in this way, we get,

$$a^k = (M^T M)^{k-1} M^T h^0$$

and

$$h^k = (MM^T)^k h^0$$

Now, our aim is to understand why this proces converges to stable values. Please note that the actula magnitudes of hubs and authorities will converge only if we normalize them at each step.

Hence, let us consider the constants c and d such that the sequences of vectors h^k/c^k and a^k/d^k converge to limits as k goes to infinity. Let us take the case of hubs first. We observe that if

$$\frac{h^k}{c^k} = \frac{(MM^T)^k h^0}{c^k}$$

is going to converge to a limit, say, h^*, it means that h^* will not change direction when further multiplied with MM^T and will only be multiplied by a value c, i.e.

$$(MM^T)h^* = ch^*$$

And we know that any vector h^* that satisfies this property is called an eigenvector and the scaling constant c is called its eigenvalue. So we should now show that the sequence of vectors h^k/c^k converges to an eigenvector of MM^T.

Fact : Any n*n symmetric matrix has n eigenvectors that are unit vectors and are mutually orthogonal, i.e. they form a basis of the space $\mathcal{R}^n$.

Now since MM^T is symmetric, we can apply the above fact to it. Let $z_1, z_2, \ldots, z_n$ be the orthogonal eigenvectors with $c_1, c_2, \ldots, c_n$ as the corresponding eigenvalues. Further, let $|c_1| > |c_2| > \cdots > |c_n|$. Now, for a given vector x, we can first write it as linear combination of the vectors $z_1, z_2, \ldots, z_n$ with $p_1, p_2, \ldots, p_n$ being the coefficients, i.e., $x = p_1 z_1 + p_2 z_2 + \cdots + p_n z_n$, we have,

$$\begin{aligned}
(MM^T)x &= (MM^T)(p_1 z_1 + p_2 z_2 + \cdots + p_n z_n) \\
(MM^T)x &= p_1 MM^T z_1 + p_2 MM^T z_2 + \cdots + p_n MM^T z_n) \\
(MM^T)x &= p_1 c_1 z_1 + p_2 c_2 z_2 + \cdots + p_n c_n z_n)
\end{aligned}$$

since each z_i is an eigenvector.

Further, we may observe that each multiplication of MM^T introduces an additional c_i in front of the ith term, i.e.

$$(MM^T)^k x = p_1 c_1^k z_1 + p_2 c_2^k z_2 + \cdots + p_n c_n^k z_n)$$