MS&E 235, Internet Commerce

Stanford University, Winter 2007-08

Instructor: Prof. Ashish Goel, Notes Scribed by Shrikrishna Shrin

Lecture 10: eigenvalue Methods

PageRank Continued

Let π_v be the PageRank of page(node) v. Also, $P_{uv} = 0$ if there is no link from u to v and $P_{uv} = 1/d_u$ otherwise (d_u is the number of out links from node u).

$$\pi_v = (\sum_u P_{uv}\pi_u)(1-\epsilon) + \frac{\epsilon}{N}$$

where N is the number of pages and ϵ is the reset probability.

$$\sum_v \pi_v = 1$$

In order to solve for the value of PageRank, think of π as a row vector

$$\pi = (\pi_1, \pi_2, \pi_3.....\pi_N)$$

Let I be the identity matrix.
$b = (\epsilon/N, \epsilon/N, \epsilon/N....., \epsilon/N)(1 \times N$ matrix $)$
In matrix form, the PageRank equation can be written as

$$\pi = (1-\epsilon)\pi P + b$$

Starting with an initial value of π, $\pi^{(0)} = (1/N, 1/N, ...1/N)$, we iteratively compute successive values of $\pi^{(i)}$ as:

$$\pi^{(i)} = (1-\epsilon)\pi^{(i-1)}P + b$$

This method converges to the actual value of PageRank.

Proof

Let $\pi^{(E)}$ be the true value of PageRank.

Let $\pi^{(i)} = \pi^{(E)} + \Delta^{(i)}$ where $\Delta^{(i)}$ is the vector by which the value of PageRank computed at the current iteration i differs from the equilibrium value $\pi^{(E)}$.

$$\pi^{(E)} + \Delta^{(i)} = (1-\epsilon)(\pi^{(E)} + \Delta^{(i-1)})P + b$$

by rearranging terms we get

$$\pi^{(E)} + \Delta^{(i)} = ((1-\epsilon)\pi^{(E)}P + b) + (1-\epsilon)\Delta^{(i)}P$$

since $\pi^{(E)}$ is a true PageRank vector, it is a solution to the equation $\pi^{(E)} = (1-\epsilon)\pi^{(E)}P + b$ and therefore we get

$$\Delta^{(i)} = (1-\epsilon)\Delta^{(i-1)}P$$

Since $\epsilon > 0$, $\Delta^{(i)}$ is multiplied by a constant factor (< 1) at each iteration and the value of $\Delta^{(i)}$ decreases with each iteration as a result. It can be shown that this equation corresponds to a contraction map and that the value of $\Delta^{(i)}$ converges to zero.

The idea behind a contraction map is that as we move from one iteration to the next, the value of $\Delta^{(i)}$ strictly decreases. This can be shown as follows:

$$\Delta_v^{(i)} = (1-\epsilon)\sum_u P_{uv}\Delta_u^{(i-1)}$$

Since in the above expression some terms in the expression $\Delta_u^{(i-1)}$ may be negative and $(1-\epsilon)$ is positive,

By triangle inequality we have

$$\mid \Delta_v^{(i)} \mid \leq (1-\epsilon)\sum_u P_{uv} \mid \Delta_u^{(i-1)} \mid$$

summing both sides over v we get,

$$\sum_v \mid \Delta_v^{(i)} \mid \leq (1-\epsilon)\sum_v\sum_u P_{uv} \mid \Delta_u^{(i-1)} \mid$$

by interchanging the order of summation we get,

$$\sum_v \mid \Delta_v^{(i)} \mid \leq (1-\epsilon)\sum_u \mid \Delta_u^{(i-1)} \mid \sum_v P_{uv}$$

Now, $\sum_v P_{uv} = 1$. This is because, as described in an earlier lecture, the "monkey" never leaves the system and therefore the probability that the "monkey" moves to some other page v from page u is 1.

Therefore the inequality becomes:

$$\sum_v \mid \Delta_v^{(i)} \mid \leq (1-\epsilon)\sum_u \mid \Delta_u^{(i-1)} \mid$$

Now, we can think of $\sum_v \mid \Delta_v^{(i)} \mid$ as size of error in iteration i. From the above inequality, this error is at most $(1-\epsilon)$ times the error in the previous iteration. The size of error therefore exponentially converges to 0. This is true because, the error in step zero ($\sum_u |\Delta_u^{(0)}|$ (where $\Delta_u^{(0)} = \pi_u^{(E)} - \pi_u^{(0)}$)) is at most 2 as both $\pi^{(E)}$ and $\pi^{(0)}$ are unit vectors.

The fact that $\sum_v P_{uv} = 1$ and $(1-\epsilon) < 1$ ensure that the above equations converge.

PageRank and Product Recommendations

Let us consider a product graph which is obtained by adding links between products such that if a link exists from product A to product B with weight w, it means that a person who has bought product A is likely to buy product B with probability w. We can define something like PageRank on this product graph starting with a node(product) A except that all resets bring the "random surfer" back to node A. What we obtain by such a method is the correlation of product A with other products.

Personalized PageRank

In the original PageRank algorithm, suppose instead of resetting to a random web page, resets were allowed only to web pages already viewed by the user, the PageRank values sort of get personalized depending on each user's browsing history. Also, if resets were allowed only to educational(.edu) web pages, for example, and resets occur after $1/\epsilon$ page views on average, the PageRank values get biased towards .edu web pages.

PageRank and eigenvalue Methods

If for a given matrix A and a vector v, $vA = \lambda v$ for some constant λ, then λ is known as the eigenvalue and v as an eigenvector of matrix A. Furthermore, if all rows of A sum to 1 (as is the case in all examples we consider), then $\lambda = 1$ is an eigenvalue and the vector v satisfying $vA = v$ is an eigenvector.

Consider the PageRank equation $\pi = (1 - \epsilon)P\pi + b$.

Define Q such that $Q = (1 - \epsilon)P$.

The PageRank equation becomes $\pi = \pi Q + b$. This equation can be written as:

$$\pi = \pi Q + b(Q - I)^{-1}(Q - I)$$

Rearranging terms we get,

$$\pi + bI(Q - I)^{-1} = \pi Q + b(Q - I)^{-1}Q$$

That is,

$$(\pi + b(Q - I)^{-1}) = (\pi + b(Q - I)^{-1})Q$$

Therefore, $\pi + b(Q - I)^{-1}$ is an eigenvector of the matrix Q. These methods are therefore known as eigenvalue methods.

HITS: Hypertext Induced Topics Search

HITS is another example of an eigenvalue method that was developed by Kleinberg around the same time as the PageRank algorithm. The primary goal of HITS was to help search engines rank web pages. When you search using a

keyword, you get a bunch of web pages that match the keyword. The HITS algorithm defines two kinds of web pages. A web page is called a hub if it links to many pages with information relevant to the query. A web page is called an authority if it contains relevant information to the query. If the search query was for a "shoe", examples of hubs would be Yahoo directory pages on shoes or a Top 10 Shoes list and examples of authorities would be web pages of Nike and Reebok.

A hub is considered good if it links to many good authorities and an authority is considered good if it is linked to from many good hubs. The iterative definition of reputation in this case hints that HITS is an eigenvalue method.

Let h_u be the hub score of page u. $L_{uv} = 1$ if there is a link from u to v, else, it is equal to 0. Let a_v be the authority score for page v.

$$h_u = \sum_v L_{uv} a_v$$

$$a_v = \sum_u L_{uv} h_u$$

Therefore, in matrix notation, $h = La$ or $h = LL^T h$ and $a = L^T h$ or $a = L^T La$.

However, the above equations do not converge. This is because the HITS algorithm rewards hubs for pointing to many good authorities and so $\sum_v L_{uv}$ is not equal to 1 (similarly $\sum_u L_{uv}$ is not equal to 1 for authorities). As mentioned earlier, such an equality is necessary for the eigenvalue method to converge.

In order to make the equations converge, the values for h_u and a_v are normalized after each iteration such that their L_2 norm ($\sum_u (h_u^{(i)})^2$ and $\sum_v (a_v^{(i)})^2$) is equal to 1 .

HITS and Product Recommendations

The HITS algorithm can also be used to generate product recommendations in the following fashion. Users are modeled as hubs and the products they buy as authorities. A link between a user and a product indicates that the user has bought the product. We can then determine product recommendations as follows:

1. Begin with a user (a hub)
2. Determine the set of products that he has bought (authorities liked to by that hub)
3. Determine the set of users that have bought the same products (hubs that link to the same authorities)
4. Determine the products that these users have bought (authorities linked to by these hubs)

This set of products can then be used to generate recommendations for the user. This type of analysis can also be carried out in a nested fashion .