uc3m | Universidad Carlos III de Madrid
Departamento de Matemáticas

Year 2024/2025

Further topics in numerical methods

Bachelor in Applied Mathematics and Computing

Search Engines Rate Page Quality [from *Numerical Analysis* by Timothy Sauer]

Web search engines such as `Google.com` distinguish themselves by the quality of their returns to search queries. We will discuss a rough approximation of Google's method for judging the quality of web pages by using knowledge of the network of links that exists on the web.

When a web search is initiated, there is a rather complex series of tasks that are carried out by the search engine. One obvious task is word-matching, to find pages that contain the query words, in the title or body of the page. Another key task is to rate the pages that are identified by the first task, to help the user wade through the possibly large set of choices. For very specific queries, there may be only a few text matches, all of which can be returned to the user. (In the early days of the web, there was a game to try to discover search queries that resulted in exactly one hit.) In the case of very specific queries, the quality of the returned pages is not so important, since no sorting may be necessary. The need for a quality ranking becomes apparent for more general queries. For example, the Google query "new automobile" returns several million pages, beginning with automobile buying services, a reasonably useful outcome. How is the ranking determined?

The answer to this question is that `Google.com` assigns a nonnegative real number, called the page rank, to each web page that it indexes. The page rank is computed by Google in what is one of the world's largest ongoing Power Iterations for determining eigenvectors. Consider a graph as in Figure 1, where each of n nodes represents a web page, and a directed edge from node i to node j means that page i contains a web link to page j.

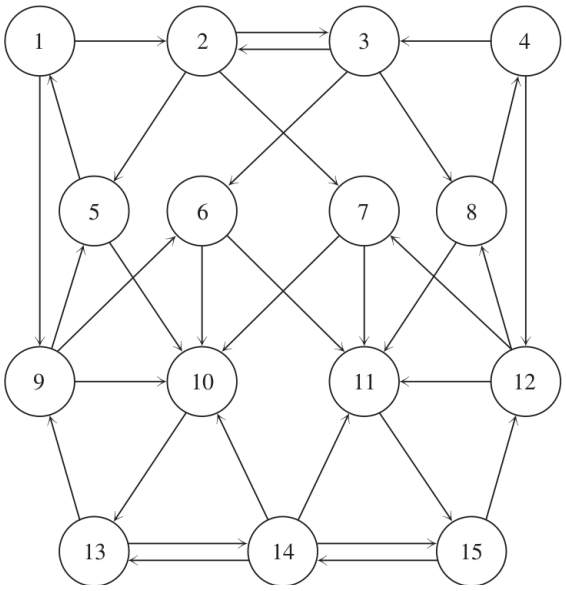

Figure 1: A network of web pages and links. Each directed edge from one page to another means that the first page contains at least one link to the second.

Let A denote the adjacency matrix, an $n \times n$ matrix whose ij-th entry is 1 if there is a link from node i to node j, and 0 otherwise. For the graph in Figure 1, the adjacency matrix is

$$A = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{bmatrix}.$$

The inventors of Google imagined a surfer on a network of n pages, who currently sits at page i with probability p_i. Next, the surfer either moves to a random page (with fixed probability q, often approximately 0.15) or, with probability $1 - q$, clicks randomly on a link from the current page i. The probability that the surfer moves from page i to page j after the click is $q/n + (1 - q)A_{ij}/n_i$, where A_{ij} is the entry of the adjacency matrix A and n_i is the sum of the i-th row of A (in effect, the number of links on page i). Since the time is arbitrary, the probability of being at node j is the sum of this expression over all i, and it is independent of time; that is,

$$p_j = \sum_i \left(\frac{qp_i}{n} + (1 - q) \frac{p_i}{n_i} A_{ij} \right)$$

which is equivalent in matrix terms to the eigenvalue equation

$$p = Gp,$$

where $p = (p_i)$ is the vector of n probabilities of being at the n pages, and G is the matrix

$$G_{ij} = \frac{q}{n} + \frac{(1 - q)\, A_{ji}}{n_j}.$$

We will call G the *google matrix*. Observe each column of the matrix G sums to one, reason why G is a *stochastic matrix*.[1] The eigenvector p corresponding to eigenvalue 1 is the set of steady-state probabilities of the pages, which are by definition the page ranks of the n pages. (This is the steady-state solution of the Markov process defined by G^T. The original idea to measure influence by steady-state probabilities goes back to Pinski and Narin [1976]. The jump probability q was added by Brin and Page [1998], the originators of Google.)

[1] Call a square matrix **stochastic** if the entries of each column add to one. A stochastic matrix (i) has an eigenvalue equal to one, and (ii) all eigenvalues are, at most, one in absolute value.

We will illustrate the definition of page rank with the example shown in Figure 1. Set $q = 0.15$. The principal eigenvector (corresponding to dominant eigenvalue 1) of the google matrix G is

$$p = \begin{bmatrix} 0.0268 \\ 0.0299 \\ 0.0299 \\ 0.0268 \\ 0.0396 \\ 0.0396 \\ 0.0396 \\ 0.0396 \\ 0.0746 \\ 0.1063 \\ 0.1063 \\ 0.0746 \\ 0.1251 \\ 0.1163 \\ 0.1251 \end{bmatrix}.$$

The eigenvector has been normalized, by dividing by the sum of all entries, to have sum equal to one, as probabilities should. The eigenvector with this normalization contains the page ranks. The page rank is highest for nodes 13 and 15, followed by node 14 and nodes 10 and 11. Note that the page rank does not simply depend on the"in-rank", or number of inward-pointing links to the page, but is more sophisticated at assigning ratings of importance. Although nodes 10 and 11 have the most inward-pointing links, the fact that they point to 13 and 15 transfers their authority down the line. This is the idea behind "google-bombing", the practice of artificially inflating the importance of a site by convincing high-traffic sites to link to it. Keep in mind that in defining page rank this way, we are using the word "importance", although no one really knows what that means. The page rank is a self-referential way of assigning importance that will probably suffice until a better method is found.

Questions

1. Prove that the google matrix G is a stochastic matrix.
2. Construct the matrix G for the network shown, and verify the given dominant eigenvector p.
3. Change the jump probability q to (a) 0 and (b) 0.5. Describe the resulting changes in the page rank. What is the purpose of the jump probability?
4. Suppose that Page 7 in the network wanted to improve its page rank, compared with its competitor Page 6–say, by persuading Pages 2 and 12 to more prominently display its links to Page 7. Model this by replacing $A_{2,7}$ and $A_{12,7}$ by 2 in the adjacency matrix. Does this strategy succeed? What other changes in relative page ranks do you see?
5. Study the effect of removing Page 10 from the network. (All links to and from Page 10 are deleted.) Which page ranks increase, and which decrease?
6. Design your own network, compute page ranks, and analyze according to the preceding questions.