T-122.101 Modeling and Mining the Web

Link Analysis 1

Anne Patrikainen
Laboratory of Computer and Information Science
15.10.2003

Outline of the Presentation

- Link analysis is the study of the link (edge) structure of a graph
- Link analysis has been studied more than 50 years, for instance in areas such as
  - Social networks
  - Bibliometrics (study of citation structure of articles)
- After the development of the Web, link analysis reseach was revived, leading to creation of masses of new algorithms. The two most famous are
  - PageRank
  - HITS (Hubs & Authorities)

The need for link analysis

- Link analysis provides us with the means to rank the vertices of a graph
- In a social network, vertices represent individuals and edges represent relations between individuals (the edges might be directed and weighted)
- With link analysis, we can find a ranking for the individuals: who is the most important person in the society?
- Ranking based on relationships might give a better answer that ranking based on property, occupation, size of house...

Ranking based on in-degree

- Naturally, there are arbitrarily many ways to rank the vertices of a graph
- A simple idea is to rank the vertices based on their in-degrees
- The importance of vertex j: $r_j = \frac{d_j^{in}}{\sum_i d_i^{in}}$, where d_j^{in} is the in-degree of vertex i
- Is the in-degree a good measure of ranking? Consider a person who has only a few significant friends, as opposed to a person who has hundreds of insignificant acquaintances
- There are other ranking approaches, such as Katz path-counting or Hubbell equilibrium construction

Ranking based on path-counting

- Introduced by Katz in 1953
- For vertices i and j, let P_{ij}^r denote the number of paths of length r from i to j
- Idea: the more paths there are terminating at vertex j, the more important j is
- Problem: Summing over numbers of paths of all lengths does not converge
- Solution: Choose a damping factor b such that $0 < b < 1$.
- Define the weighted sum of paths from i to j as $Q_{ij} = \sum_{r=1}^{\infty} b^r P_{ij}^r$ (converges for small enough b)
- Define the importance of the vertex j as $\sum_i Q_{ij}$

The Hubbell equilibrium construction

- Proposed by Hubbell in 1965
- Let e_j denote the *a priori* estimate of the importance of vertex j
- Naturally, these estimates might be equal for all vertices
- Denote by A_{ij} the strenght of the endorsement (relationship) from i to j
- The importances (standings) of the vertices the are obtained as a solution to the system of equations $s_j = e_j + \sum_i s_i A_{ij}$, for all vertices j. Here s_j is the standing of vertex j
- The standings are thus defined to be a set of values for which the process of endorsement maintains an equilibrium

Impact factor

- Bibliometrics is the study of written documents and their citation structure
- Impact factor (Garfield, 1972) is a popular measure used to provide a numerical assesment of scientific journals
- The impact factor of a journal j in a given year is the average number of citations received by papers published in the previous two years in journal j
- Impact factor is essentially based on counting the in-degrees of vertices
- A more advanced approach is the Pinski-Naring construction

The Pinski-Narin construction 1/2

- The Pinski-Narin construction (1976) is based on the observation that not all citations are equally important
- A journal is influential if it is heavily cited by other influential journals
- Denote the number of journals by n
- A is nxn matrix with A_{ij} equal to the number of citations by the journal i to the journal j
- Normalize the rows of the matrix to one ($S = D^{-1}A$, where D is a diagonal matrix with the row sums of A on the diagonal) in order to find out the journalwise citation proportions

The Pinski-Narin construction 2/2

- The influence weight of the article j is defined as $w_j = \sum_i w_i S_{ij}$
- In a vector form, $w = S^T w$. Thus, the influence weight vector is an eigenvector of S^T
- The influence weights correspond to a stationary distribution of a random reader process (Geller, 1978)
- Beginning with an arbitrary journal j, the random reader chooses a random reference in j and moves to the journal specified in the reference
- Closely related to the Hubbell equilibrium construction

PageRank 1/5

- PageRank is one of the two famous Web link analysis algorithms
- PageRank was developed by Larry Page and Sergey Brin, also known as the creators of Google
- PageRank lies at the core of Google ranking capabilities
- Google began as a research project at Stanford University in spring 1995
- Back then, Page and Brin were PhD students working in the area of data miing
- Finding Google Inc., they put their PhD studies on hold and decided to make money instead

PageRank 2/5

- Consider a random surfer in the Web
- The surfer has an uniform probability to follow any of the links on the current page
- The surfer also has a small probability to teleport to a random Web page
- This prevents the surfer from getting stuck on pages with no out-links, or in disconnected components of the Web
- We thus have a Markov chain
- The stationary distribution of the Markov chain gives the PageRank scores for the Web pages

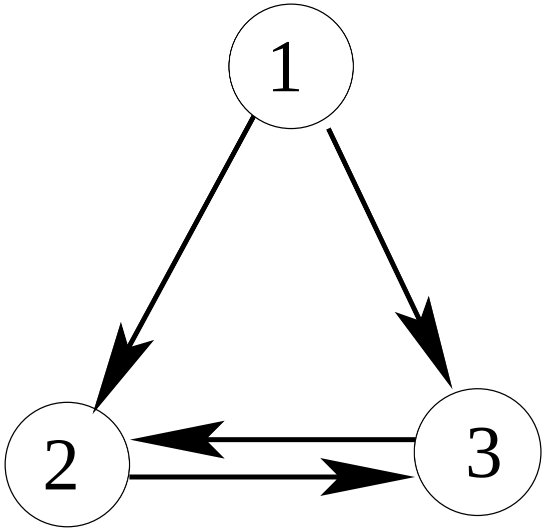
PageRank 3/5

- Consider the Web graph adjacency matrix A
- $A_{ij} = 1$ if there is a link from page i to page j
- We normalize the rows of A to one by $S = D^{-1}A$, where D is a diagonal matrix with the row sums of A on the diagonal
- Now S can be considered as the Markov chain transition matrix with S_{ij} as the probability of the surfer proceeding to page j from page i
- But we have not taken the random teleportation into account yet
- Let us introduce a stochastic escape matrix $E = D_{E'}^{-1}E'$, where $E' = \mathbf{1} - A$, where $\mathbf{1}$ is the matrix of all ones, and $D_{E'}$ is a diagonal matrix with row sums of E' on the diagonal

PageRank 4/5

- The final transition matrix is $M = (1-\epsilon)S + \epsilon E$, where ϵ is the random teleportation probability
- The stationary distribution π is obtained from $M^T\pi = \pi$. In other words, π is the eigenvector of M^T corresponding to eigenvalue 1.
- Note that the eigenvector might be scaled in an arbitrary way; only the relative differences matter
- A common approach is to scale the eigenvector such that the maximum value is 1

PageRank 5/5 — an example



$$A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad M = 1/20 \begin{bmatrix} 2 & 9 & 9 \\ 1 & 1 & 18 \\ 1 & 18 & 1 \end{bmatrix}$$

Teleportation probability $\epsilon = 0.1$

The stationary distribution $\pi = [0.11, 1.00, 1.00]^T$

HITS (Hubs & Authorities) 1/5

- PageRank and all the other methods we have seen rank pages based on a single criterion
- In 1998, Jon Kleinberg came up with a two-level ranking scheme
- In the Web, there are some authoritative pages (for instance, the Microsoft home page and the IBM home page are authoritative among computer-related pages, unlike the local computer nerd's home page)
- But there are also hub pages, link lists pointing to many authoritative pages

HITS 2/5

- The authoritative pages seldom point to each other, so criteria based on the 'authoritative pages are pointed by other authoritative pages' principle fail
- Kleinberg proposed calculating two scores for each Web page: the hub score and the authority score
- Kleinberg's algorithm HITS stands for Hypertext Induced Topic Selection, and is also known as Hubs & Authorities

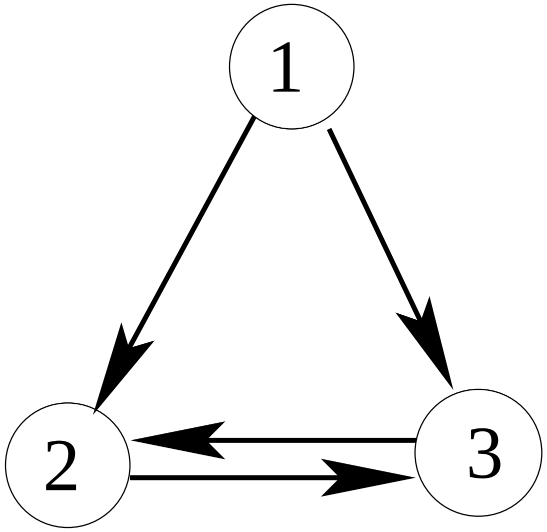
HITS 3/5

- Denote the n:th step hub score of page i by $h_i(n)$
- Denote the n:th step authority score of page i by $a_i(n)$
- A good hub points to good authorities
- The hub score can be iteratively computed as
  $h_i(n+1) = \sum_{j:(i \to j)} a_j(n)$
- A good authority is pointed by many good hubs
- For the authority score we obtain $a_j(n+1) = \sum_{i:(i \to j)} h_i(n+1)$
- The initial values may be chosen as the same constant
- Some normalization must be carried out

HITS 4/5

- The adjacency matrix of the Web graph is denoted by A, where $A_{ij} = 1$ if there is a link from page i to page j
- In vector form, we get for the hubs $h(n) = Aan - 1)$ and for the authorities $a(n) = A^T h(n)$
- Combining these, we notice that the hub scores are in the principal eigenvector of AA^T and the authority scores in the principal eigenvector of $A^T A$
- Note that it would make no sense to calculate general hub and authority scores for the whole Web
- As a preprocessing phase, a subgraph relevant to a given query is extracted (use text analysis here!), and only the subgraph is given to HITS

HITS 5/5 — an example



The largest eigenvalue of AA^T: 3

The corresponding eigenvector: $[1, 0.5, 0.5]^T$ (The hub score vector)

The largest eigenvalue of A^TA: 3

The corresponding eigenvector: $[0, 1, 1]^T$ (The authority score vector)

Homework problems for Link Analysis 1

In this homework assignment, you will try to rank vertices of a graph based on various relevance criteria. Use the sample of the Web you have crawled or any other suitable graph data set of your choosing (for instance, a scientific paper citation network or the graph you have generated yourself).

1. Rank the vertices of your graph (that is, Web pages, scientific papers, etc.) using in-degree as the measure of relevance. List the most important vertices.

2. Rank the vertices of the graph using PageRank. The choice of the random teleportation probability is yours to make. List the most important vertices. Does the choice of the teleportation probability have any effect on the results?

3. Rank the vertices of the graph using HITS (that is, Hubs & Authorities). Here, you'll obtain two scores for each web page. List the most important vertices with respect to each score. Compare the hub and authority scores for individual pages with the help of a plot.

4. Compare the results of the three rankings above. Are there any similarities? Why, why not?

5. Which ranking scheme would you choose for your data set? One of the above ones or some other? Which one you would not choose in any case? Justify your answers.

You will receive one exercise point by completing questions 1–3 and two points if you complete questions 4–5 as well.