CA FOSCARI UNIVERSITY of VENICE

INFORMATION RETRIEVAL AND WEB SEARCH [CM0473] -
PROF. S. ORLANDO

Page Rank / HITS Computation

[Compute efficient link analysis of algorithms]

Group Members:

1. **PATIL KHUSHBU MAHENDRA**

Matriculation number: **882697**

2. **KHALID VAFA**

Matriculation number: - **882677**

INTRODUCTION

- Now a day Web Mining is a data mining procedure which has become a key part of users. Users generally pay out a lot of time for the search queries and pull out the relevant information from web.
- We conclude that both page rank and HITS algorithm are different link analysis algorithms that employ different models to calculate web page rank.
- The Page Ranking algorithms which are an application of web mining play a vital role to easier navigation for users. As on today WWW is the largest information repository and set of all nodes which are interconnected by hypertext links. With the quick growth of the Web, users get easily vanished in the rich hyperlink structure. The main aim of website owners is to providing accurate data based on the user's requirement.
- So, discover the content of the Web pages and retrieving the users' interests from their actions has become gradually more important. Higher page rank of websites that means that website is more visited by users.
- Careful optimization of web sites by Search Engine Optimization that increase the websites visibility in the different search engine Google, Yahoo, Bing and many others. The results obtained by a search engines are a combination of large amount of appropriate and inappropriate information.
- Normally users visit only that website which is top of the lists. So various ranking algorithm such as PageRank, HITS are available that helps the users to navigate in the results.
- These ranking method uses by search engine that sort and displayed the result to users. So users can easily find the best result.

- The way in which the displaying of the web pages is done within a search is not a mystery.
- It involves applied math and good computer science knowledge for the right implementation. This relation involves vectors, matrixes and other mathematical notations.

PAGE RANK

- The PageRank vector needs to be calculated, that implies calculations for a stationary distribution, stochastic matrix.

- PageRank is a topic widely discussed by Search Engine Optimization (SEO) experts.

- Page rank algorithm was invented by Larry Page and Sergey Brin while they were graduate students at Stanford University and later it became a trademark of google in 1998.

- PageRank is an algorithm that drives Google that is world wide web (www)

- PageRank does not have an impact only in the programming industry, but also has an effect in the economic sector. It is also considered as a business goal of every firm to be ranked higher in the web page display, that is considered as a SEO strategy.

- Page rank algorithm was invented by Larry Page and Sergey Brin while they were graduate students at Stanford University and later it became a trademark of google in 1998.

- PageRank is an algorithm that drives Google that is world wide web (www)

- PageRank does not have an impact only in the programming industry, but also has an effect in the economic sector. It is also considered as a business goal of every firm to be ranked higher in the web page display, that is considered as a SEO strategy.

- Suppose we have web pages numbered 1*, . . . n*. The PageRank of webpage *i* based on its linking web pages (webpages *j* that link to*i*). But we don't just count the number of linking web pages, i.e., not all linking web pages are treated equally. Instead, we weight the links from different web pages.

This follows two ideas:

1) Webpages that link to *i*, and have high PageRank scores themselves, should be given more weight.
2) Webpages that link to *i*, but link to a lot of other webpages in general, should be given less weight.

Page Rank Definition

Let *Lij* = 1 if webpage *j* links to webpage *i* (written $j \rightarrow i$), and
Lij = 0 otherwise. Also let *mj* = P*n k*=1 *Lkj* , the total number of webpages that *j* links to.

- We're going to define something that's almost PageRank, but not quite, because it's broken. The BrokenRank *pi* of webpage .

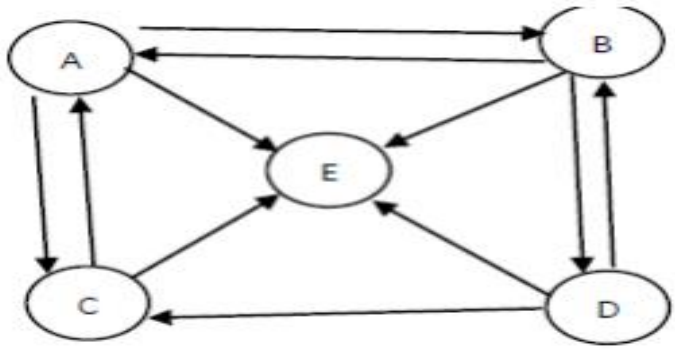
$$p_i = \sum_{j \to i} \frac{p_j}{m_j} = \sum_{j=1}^{n} \frac{L_{ij}}{m_j} p_j .$$

Calculating page rank with matrix method

- Today there are millions of websites and each holds tons of information. When a web surfer visits a web page, that point acts as the starting point for the listing.

- By making use of random selection of links, a website can be visited several times by the surfer.
- Graphs will provide a better visualization to make it more mathematically concrete. If a web page i has outlinks ≥ 1, then for each element from page i to another page j, the element in the row i and column j of matrix H is defined as = 1/ .
- If there is no link from page i to j then the value of matrix is 0. Thus, each non-zero element in a row sums to 1.
- A page with no outlinks is named as dangling node. All the elements in the row of a dangling node are set to 0. Consider the following graph with different nodes and links connecting them.

Lets take a example which shows Directed graph for an illustration to calculate PageRank



The matrix H for the directed graph with the vertex pointing from i to j is given as below.

$$H = \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Node	A	B	C	D	E
Number of Links	2	2	2	2	4

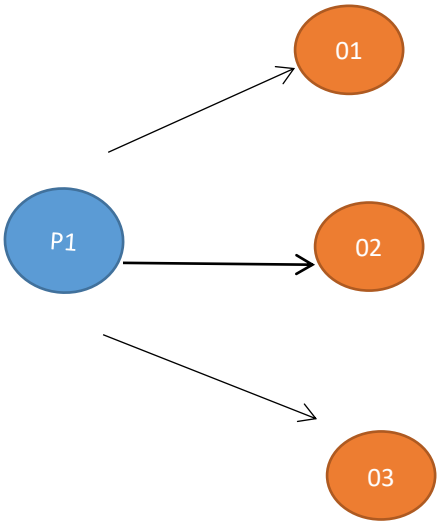
- Formula for counting page rank is $\mathrm{PR(I+1)(Pi)} = \sum_{pj} \mathrm{PR(i)(pj)\backslash C(Pj)}$ that is page rank of given site in next iteration is the page rank of given site in previous iteration divide by the numbers of outgoing links. SO the page rank will be

A	B	C	D	E
0.35	0.1925	0.235	0.1925	0.15

- So far, we have seen how Google calculates PageRank and how its matrix is a combination of stochastic matrices of link structure and behavior of web surfer.
- A strong property in the success of this algorithm is that it uses information not on the page content itself but on the linkage connection between pages.
- As the inspiration was from the ranking of scientific articles, PageRank itself can be thought as a link review.
- The formula is iteratively calculated and as the web pages keep growing in number and not in a connected manner, a damping constant is used to ease the duty.
- The mathematic that goes on with eigenvectors and eigenvalues makes it such that it converges

- Google has the most well known ranking algorithm called the Page Rank algorithm that has been claimed to supply top ranking pages that are relevant.
- The Page Rank algorithm was used and enhanced by Lawrence Page and Sergey Brin.
- Page Rank algorithm describes the popularity of web page or website. This Page Rank algorithm is depend on the link Analysis in which ranking of web page is decided based on outbound links and inbounds links.
- That means it's totally based on link of WWW and Google uses this algorithm for searching the web pages based on number of hyperlinks such as Inbound and outbound.

Inbound Links: Inbound links are those links that is comes from other site to your website, it is also known as “backlinks”. Google consider only relevant links point to your site but you cannot control which sites point to your site. If your website content is unique and rich then there are much chances those links will be “dofollow” otherwise links will be consider as “nofollow”



Outbound Links: Outbound links are those links that is pointing to other site from your website and you have more control over these links.

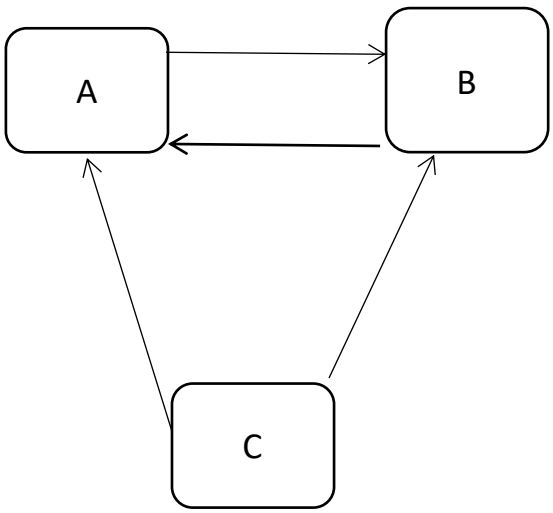
A page has high rank if the other pages with high rank
linked to it [7]. It is given by:-
PR (A) = (1-d) + d (PR (Ti)/C (Ti) + ... + PR (Tn)/C(Tn))
• Let A be the page and whose page rank is PR(A).
• Let PR (Ti) is the Pagerank of pages Ti which link to page A, C (Ti) is the number of outbound links going out from page Ti and d is a damping factor assume to be between 0 and 1 usually 0.85. Sometimes does not click on any links & jumps to another pages at random. It follows the direct links.

(1-d) is the probability of jumping off to some random pages; every page has a minimum page rank of (1-d). It follows the non-direct links. To calculate the Page Rank of any Page We required to know the Page Rank of each page that point to it and number of the outbound links from each of those pages.

Example Illustrating functioning of Page Rank

Let us consider a simple example of three web page A,B and C shown in figure.

1. Page A contains 1 outbound link that is pointing to Page B.
2. Page B contains 2 outbound links that is pointing to Page A and Page C.
3. And Page contains 1 outbound link that is pointing to Page A
4. The initial page Rank of each page is considered to be 1.



The Page Rank of each page is computed by following equation

PR (A) = 0.2 + 0.4PR (B) + 0.8PR (C)

PR (B) = 0.2 + 0.8PR (A)

PR (C) = 0.2 + 0.4PR (B)

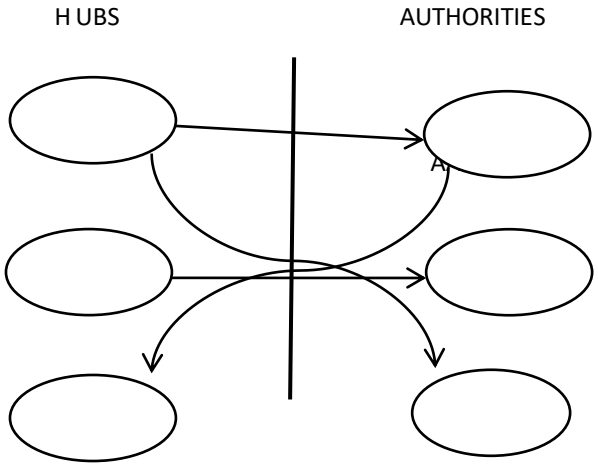
The result of above equation is given

PR (A) = 1.2

PR (B) = 1.0 PR PR (C) = 0.66

HITS

- Hypertext Induced Topic Search (HITS) or hubs and authorities is a link analysis algorithm developed by Jon Kleinberg in 1998 to rate Web pages.
- A precursor to PageRank, HITS is a search query dependent algorithm that ranks the web page by processing its entire in links and out links. Thus, ranking of the web page is decided by analyzing its textual contents against a given query.
- When the user issues a search query, HITS first expands the list of relevant pages returned by a search engine and then produces two rankings of the expanded set of pages, authority ranking and hub ranking.
- In this algorithm a web page is named as authority if the web page is pointed by
- many hyper links and a web page is named as HUB if the page point to various hyperlinks.
- HITS is applied on a sub graph after a search is done on the complete graph.
- Uses hubs and authorities to define a recursive relationship between web pages.
- An authority is a page that many hubs link to.
- A hub is a page that links to many authorities.



- The scores for authority nodes x can be determined from the hub scores x = AT y and similarly the hub scores from the authority scores y = Ax I Substituting into the equations
  we get

x = AT Ax

y = AAT y

- The algorithm produces two types of pages :-

Authority: pages that provide an important, trustworthy information on a given topic
Hub: Pages that contain links to authorities

Advantages of HITS

1. HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.
2. The ranking may also be combined with other information retrieval based rankings.
3. HITS is sensitive to user query (as compared to PageRank)
4. Important pages are obtained on basis of calculated authority and hubs value.
5. HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.
6. HITS induces Web graph by finding set of pages with a search on a given query string.
7. Results demonstrates that HITS calculates authority nodes and hubness correctly

Comparision

CRITERIA	PAGE RANK	HITS
Basic Criteria	Link analysis algorithm based on random surfer model	Link analysis algorithm
Technique	Web Structure Mining	Web Structure Mining, Web content Mining
Efficiency	Computes a single measure of quality for a page at crawl time	Htts invokes traditional search Engines to retrieve set of pages relevant to it
Mutual Reinforcement	Page rank does not attempt to capture the distinction between hubs and authority.	Htts emphasize mutual reinforcement between auhorities and hub webpages

Summary

- Now a day Web Mining is a data mining procedure which has become a key part of users. Users generally pay out a lot of time for the search queries and pull out the
- relevant information from web. We conclude that both page rank and HITS algorithm are different link analysis algorithms that employ different models to calculate web
- page rank.

- The Page Ranking algorithms which are an application of web mining play a vital role to easier navigation for users.

- HITS is applied on a subgraph after a search is done on the complete graph.
- HITS defined hubs and authorities recursively.
- PageRank is used for ranking all the nodes of the complete graph and then applying a search.
- PageRank is based on the 'random surfer' idea and the web is seen as a Markov Chain Power Iteration an efficient way to calculate with sparse Matrices