Data Mining and Machine Learning

Page Rank

Peter Jančovič

Objectives

- To understand the basic idea of the PageRank of a document in a corpus
- To understand how to calculate PageRank
- To understand the Markovian model that underlies PageRank

Not all documents are equal

- So far, whether or not a document *d* is retrieved in response to a query *q* depends only on *sim*(*q,d*)
- Assumption is that all documents are equal - relevance of a document for a query depends only on the similarity score
- This is clearly not true (compare *Wikipedia* with my home page)
- Prior importance of a document is its Page rank
- Probabilistic interpretation of Page rank

The *prior* probability of a document

- Suppose that we could assign a probability $P(d)$ to each document d in our corpus
- Think of $P(d)$ as the probability that d is a relevant document before the user creates a query q
- $P(d)$ is the prior (or *a priori*) probability of d
- In this case, whether d is returned in response to a query q depends on $sim(q,d)$ and $P(d)$
- We will treat $P(d)$ as the Page rank of d

Retrieval using prior probabilities

- Retrieval based only on *sim*(*q,d*) assumes that *P*(*d*) is the same for all documents
- This case is called equal priors
- Intuitively we could do better if we could estimate more meaningful priors
- Assumption: the *prior* relevance of a document to any query is related to how often that document is accessed

Citation indices

- Similar idea used to measure quality of academic papers
- If a paper *p* contains important results or ideas, then lots of papers will refer to it
- The <u>citations index</u> *ci*(*p*) measures how many papers refer to a given paper *p*
- Citations index is a standard quality measure in research assessment
- But, quality of a paper depends not only on the <u>quantity</u> of papers that cite it but on their <u>quality</u> – their citation indices

Basics of Page Rank

- For a document, or a page, *d* on the web, we could defined the <u>Page Rank</u> *pr*(*d*) to be the number of documents that have a hyperlink to *d*
- This relies on the underlying democracy of the web – users 'vote with their mouse buttons'
- The ranking of a document *d* in response to *q* depends on both *sim*(*q,d*) <u>and</u> *pr*(*d*)
- But not all links are equal

The "Random Surfer Model"

- The solution is to allocate a <u>weight</u> of w_{de} to the hyperlink from document *d* to document *e*
- w_{de} can be thought of as the <u>probability</u> of following the link to page *e* if the user is on page *d*
- If $l(d)$ denotes the number of hyperlinks from *d*, setting $w_{de} = 1/l(d)$ corresponds to the <u>random surfer model</u>: on any page any of the available links are chosen with equal probability

UNIVERSITY OF BIRMINGHAM

The “Intentional Surfer Model”

- In reality all links on a page are not clicked with equal probability
- A better alternative is to estimate the w_{de}s using actual statistics of hyperlink use by surfers
- This is the intentional surfer model
- Organizations like Google collect and store this kind of information (I assume!)

Simplified Page Rank Calculation

- Once *pr*(*d*) is accepted as a measure of the importance of *d* there is a natural consequence
- In the calculation of *pr*(*d*), a hyperlink from a page d_1 to *d* should count for more than a hyperlink from page d_2 to *d* if $pr(d_1) > pr(d_2)$

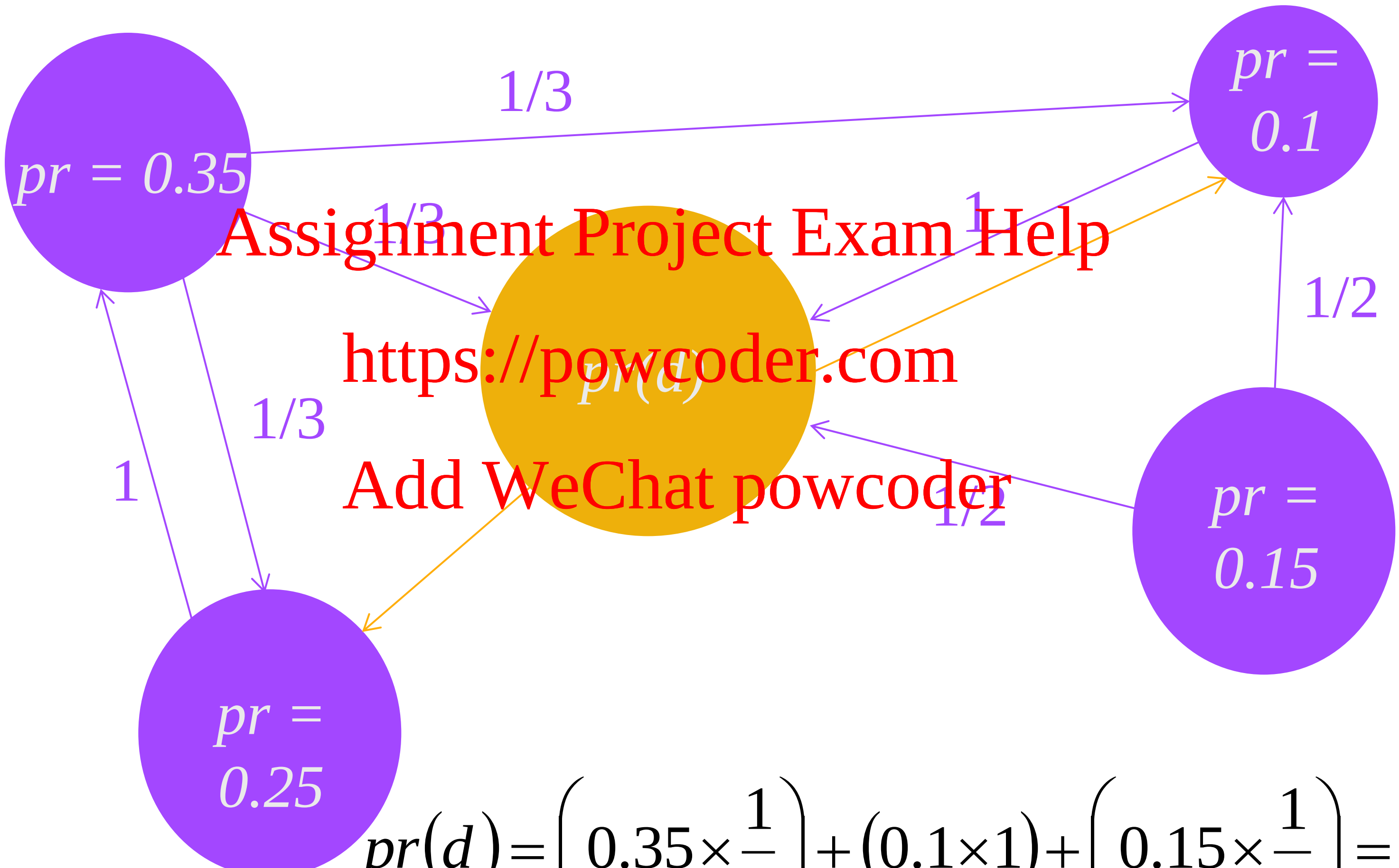
- This motivates:

$$pr(d) = \sum_{e \in L(d)} pr(e) w_{ed}$$

where *L*(*d*) is the set of pages which link to page *d*

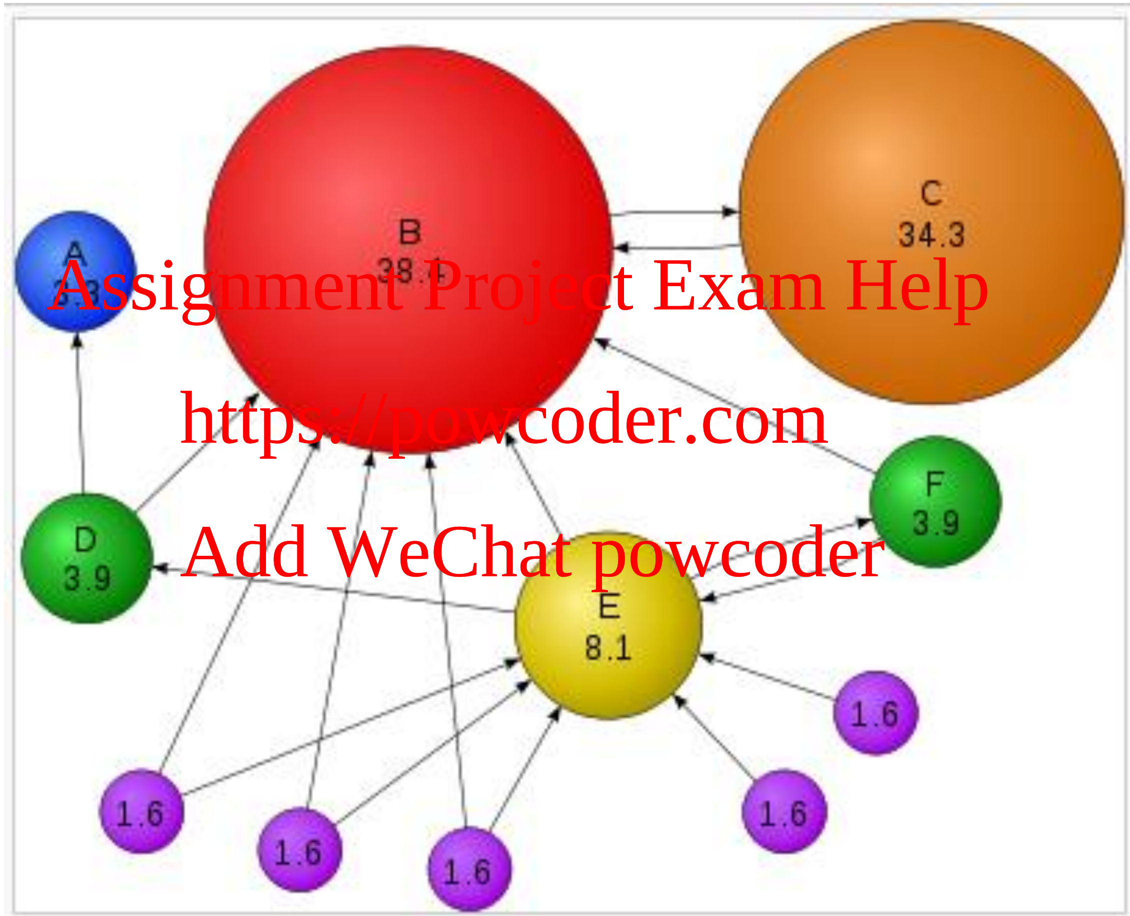
- This is the simplified Page rank calculation

Simplified Page Rank Calculation



$$pr(d) = \left(0.35 \times \frac{1}{3}\right) + (0.1 \times 1) + \left(0.15 \times \frac{1}{2}\right) = 0.292$$

Example



Taken from wikipedia: see http://en.wikipedia.org/wiki/PageRank

Simplified Page Rank Calculation

- Of course, changing *pr*(*d*) will change the Page Ranks of the other pages, which in turn will change *pr*(*d*)....
- Hence the definition of Page Rank is recursive, and *pr*(*d*) is calculated iteratively:

$$pr_{n+1}(d) = \sum_{e \in L(d)} pr_n(e) w_{ed}$$

Markov Chain interpretation

- Let

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1D} \\ w_{21} & w_{22} & \cdots & w_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ w_{d1} & w_{d2} & \cdots & w_{dD} \\ \vdots & \vdots & \vdots & \vdots \\ w_{D1} & w_{D2} & \cdots & w_{DD} \end{bmatrix}$$

where w_{ij} is the probability of a user following a hyperlink between the i^{th} and j^{th} pages and D is the number of pages – this is the <u>page transition probability matrix</u>

- Notice that each row of W sums to 1

Markov Chain interpretation

- Let $pr_n^T = [pr_n(1), pr_n(2), \ldots, pr_n(D)]$ - $pr_n(i)$ is the Page Rank of the i^{th} page after n iterations
- Then $pr_{(n+1)} = W^T pr_n$, or $pr_n = (W^T)^n pr_0$
- In Markov Chain terminology, w_{de} is the transition probability from page d to page e
- Can think of w_{de} as the probability of page e at time $t+1$ given page d at time t: $P(e\ @\ t+1\ |\ d\ @\ t)$
- pr_n is an estimate of the probability distribution over all of the pages after the n^{th} iteration
- In this case $\sum_d pr_n(d) = 1$

Markov chain interpretation

$$\begin{bmatrix} pr_{n+1}(1) \\ pr_{n+1}(2) \\ . \\ pr_{n+1}(d) \\ . \\ . \\ pr_{n+1}(D) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{D1} \\ w_{12} & w_{22} & \cdots & w_{D2} \\ . & . & . & . \\ w_{1d} & w_{2d} & \cdots & w_{Dd} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ w_{1D} & w_{2D} & \cdots & w_{DD} \end{bmatrix} \begin{bmatrix} pr_n(1) \\ pr_n(2) \\ . \\ pr_n(d) \\ . \\ . \\ pr_n(D) \end{bmatrix}$$

UNIVERSITY OF BIRMINGHAM

Markov Chain interpretation

- If this system converges, then

$$\begin{bmatrix} pr(1) \\ pr(2) \\ \cdot \\ pr(d) \\ \cdot \\ \cdot \\ pr(D) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{D1} \\ w_{12} & w_{22} & \cdots & w_{D2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1d} & w_{2d} & \cdots & w_{Dd} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1D} & w_{2D} & \cdots & w_{DD} \end{bmatrix} \begin{bmatrix} pr(1) \\ pr(2) \\ \cdot \\ pr(d) \\ \cdot \\ \cdot \\ pr(D) \end{bmatrix}$$

- $pr = W^T pr$
- In other words *pr* is an eigenvector of W^T with eigenvalue 1

Damping Factor

- The model we have used to develop Page Rank is a "random surfer" model with 'proper' hyperlink probabilities
- The random surfer will eventually stop clicking
- The probability that the random surfer continues clicking when he arrives at a page is called the <u>damping factor</u> and denoted by δ
- A typical value of δ is 0.85

Page Rank

- Taking into account the damping factor,

$$pr(d) = \left(\frac{1-\delta}{N} \right) + \delta \left(\sum_{e \in L(d)} pr(e) \times w_{ed} \right)$$

where *N* is the number of documents

Notes

- Assuming that *p*(*e*) is the probability of the page d, then this formula preserves $\sum_{d} pr(d) = 1$
- The formula assigns a 'floor' value of $\frac{1-\delta}{N}$ to a page that has no incoming hyperlinks (so that it has non-zero page rank)
- In addition, the damping factor reduces the effect of past estimates of PageRank on the present estimate

Notes

- This lecture presents a probabilistic approach to Page rank
- “PageRank” is a trademark of Google
- It was developed by Larry Page between 1995 and 1998
- Larry Page is one of the founders of Google Inc.
- A high PageRank is a valuable asset for a www page, for example to attract advertising
- Hence the precise details of the Google PageRank algorithm are secret!