Intelligent Data Analysis: Page Rank

Martin Russell

School of Computer Science, University of Birmingham

ATTENDANCE QUIZ ATTENDANCE CODE: PSKY94H4

March 12, 2020

Overview

1. Motivation
   - Not all documents are equal
2. Markov processes
   - What is a Markov model?
   - The state probability distribution $P^{(t)}$
   - Convergence of Markov processes
3. Probabilistic model of Page rank
   - Simplified page rank
   - Matrix formulation
   - The damping factor
   - Examples
   - Dangling pages

Not all documents are equal

- So far, whether or not a document d is retrieved in response to query q depends only on $sim(q, d)$
- But the number of documents is huge - there may be too many documents for which $sim(q, d)$ is large
- All documents assumed equal - relevance of a document for a query depends only on the similarity score
- This is not true (e.g., compare Wikipedia with my home page)
- The **Page rank** of a document is its **prior importance**
- Prior importance - measure of probability that the document is the one you want **before you formulate the query**

The *prior* probability of a document

- Assign a probability $P(d)$ to each document d in our corpus
- Think of $P(d)$ as the probability that d is relevant to q before the user creates query q
- $P(d)$ is the **prior** (or *a priori*) probability of d.
- Whether d is returned in response to query q depends on $sim(q, d)$ and $P(d)$
- Treat $P(d)$ as the **Page rank** of d

Ranking documents

- Retrieval based only on $sim(q, d)$ assumes that $P(d)$ is the same for all documents
- This case is called **equal priors**
- How can we estimate better priors?
- Assumption: the prior relevance of a document to any query is related to how often that document is accessed
- Compare with **citation index** of an academic paper
- **Self-evaluating groups** - each group member evaluates all other group members

The *authority* of a document

- For a document d on the web, we could defined the **authority** of d, denoted x_d, based on the number of documents that have a hyperlink to d
- Relies on the underlying democracy of the web users vote with their mouse buttons
- Now the ranking of a document d in response to q depends on both $sim(q, d)$ and x_d
- But not all links are equal

The authority of a document

Authority of a document d takes into account:

- Number of incoming links (citations)
- Authority of pages c that cite d
- Selective citations from c more valuable than uniform citations of a large number of documents
- Note that this definition of page authority is **self-referencing**
- A simple candidate mathematical model for page authority is is a **Markov model**

Definition of a Markov model

An N-state Markov model consists of:

- A set of N **states** $\{\sigma_1, \cdots, \sigma_N\}$
- An **initial state probability distribution**

$$P^{(0)} = \begin{bmatrix} P^{(0)}(1) \\ P^{(0)}(2) \\ \vdots \\ P^{(0)}(N) \end{bmatrix}$$

  where $P^{(0)}(n)$ is the probability that the model starts in state n, and

- An $N \times N$ **state transition probability matrix** A, where a_{ij} is the probability of a transition from state i to state j at time t

6 dice: d_4, d_5 weighted

$$p(n) = \begin{cases} 0.9 & \text{if } n=6 \\ 0.02 & \text{if } n \neq 6. \end{cases}$$

① Choose d_1, throw get X_1

② Choose d_{X_1}, throw get X_2

⋮

ⓣ Choose $d_{X_{t-1}}$, throw get X_t.

$$p(0) = \begin{bmatrix} 1/6 \\ \vdots \\ 1/6 \end{bmatrix}$$

$$A = \begin{array}{c} \\ \\ d_4 \\ d_5 \\ \\ \end{array} \begin{bmatrix} 1/6 & \cdots & 1/6 & 1/6 \\ \vdots & & & \vdots \\ 0.02 & \cdots & 0.02 & 0.9 \\ 0.02 & \cdots & 0.02 & 0.9 \\ 1/6 & \cdots & 1/6 & 1/6 \end{bmatrix}$$

N state Markov model

- An **initial state probability distribution**

$$P^{(0)} = \begin{bmatrix} P^{(0)}(1) \\ P^{(0)}(2) \\ \vdots \\ P^{(0)}(N) \end{bmatrix}, \sum_{n=1}^{N} P^{(0)}(n) = 1 \quad (1)$$

- A $N \times N$ **state transition probability matrix**

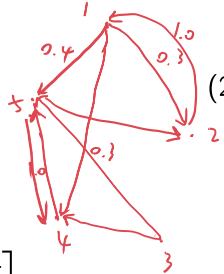
$0 \le P^t(N) \le 1$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{Nj} & \cdots & a_{NN} \end{bmatrix}, \sum_{n=1}^{N} a_{in} = 1, \forall i = 1, \cdots, N$$

Example

1. Number of states $N = 5$. Initial state probability vector

$$P^0 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.2 \\ 0.2 \end{bmatrix} \quad (2)$$



2. State transition probability matrix

$$A = \begin{bmatrix} 0 & 0.3 & 0 & 0.3 & 0.4 \\ 1.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0.5 & 0 & 0.5 & 0 \end{bmatrix} \quad (3)$$

Transition diagram

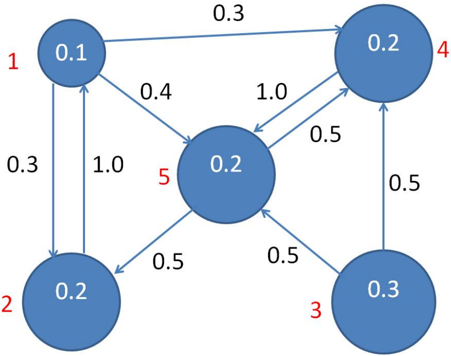


Figure: Simple state transition diagram ($N = 5$)

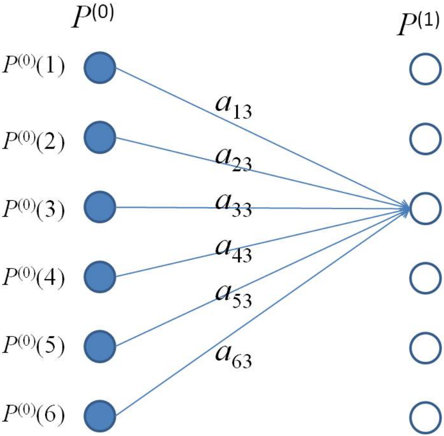
The state probability distribution at time t

- The state probability distribution at time $t = 0$ is $P^{(0)}$

$$P^{(0)} = \begin{bmatrix} P^{(0)}(1) \\ P^{(0)}(2) \\ \vdots \\ P^{(0)}(N) \end{bmatrix} \tag{4}$$

  is the state probability distribution at time $t = 0$
- What is the state probability distribution $P^{(1)}$ at time $t = 1$?
- What is the state probability distribution $P^{(t)}$ at a general time t?
- What happens to $P^{(t)}$ as $t \to \infty$?

Calculating the state probability distribution at time 1



$$P^{(1)}(3) = P^{(0)}(1)a_{13} + P^{(0)}(2)a_{23} + P^{(0)}(3)a_{33} + P^{(0)}(4)a_{43} + P^{(0)}(5)a_{53} + P^{(0)}(6)a_{63}$$

n-th column

A

$p^{(1)} = A^T[n] \cdot p^{(0)}$

Calculating general state probability distributions

- In matrix form this becomes

$$P^{(1)} = A^T P^{(0)} \tag{5}$$

and, more generally

$$P^{(t)} = A^T P^{(t-1)} = A^T A^T P^{(t-2)} = \cdots = (A^T)^t P^{(0)} \tag{6}$$

- Now suppose $P^{(t)} \to P$ as $t \to \infty$. Then

$$P = A^T P \tag{7}$$

$P^{(t+1)} = A^T P^{(t)}$

for $t+1$: only $\sim t$

stationary distribution

in other words, P is an **eigenvector** of A^T with **eigenvalue** 1

Example

Suppose $\mathcal{M}$ is a 3-state Markov model given by

$$P^0 = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.6 \end{bmatrix} \; A = \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ 0.5 & 0.2 & 0.3 \end{bmatrix} \tag{8}$$

and let $x = 113231$. Then

$$\begin{aligned} P(x|\mathcal{M}) &= P^0(1) \times a_{1,1} \times a_{1,3} \times a_{3,2} \times a_{2,3} \times a_{3,1} && (9) \\ &= 0.3 \times 0.2 \times 0.1 \times 0.2 \times 0.6 \times 0.5 && (10) \\ &= 0.00036 && (11) \end{aligned}$$

Example continued

$$P^{(1)} = A^T P^{(0)} = \begin{bmatrix} 0.39 \\ 0.34 \\ 0.27 \end{bmatrix} \tag{12}$$

$$P^{(2)} = A^T P^{(1)} = \begin{bmatrix} 0.315 \\ 0.361 \\ 0.324 \end{bmatrix} \tag{13}$$

$$P^{(3)} = A^T P^{(2)} = \begin{bmatrix} 0.333 \\ 0.321 \\ 0.345 \end{bmatrix} \tag{14}$$

$$\cdots$$

Example continued

From this example it appears that

$$P = lim_{t \to \infty} P^{(t)} = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}$$

Can confirm this using eigenvalue decomposition (MatLab [U,D] = eig(A');). This gives:

$$U = \begin{bmatrix} 0.58 & 0.13 - 0.39i & 0.13 + 0.39i \\ 0.58 & -0.65 & -0.65 \\ 0.58 & 0.52 + 0.39i & 0.52 - 0.39i \end{bmatrix}, \quad (15)$$

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.2 + 0.3i & 0 \\ 0 & 0 & -0.2 + 0.3i \end{bmatrix} \quad (16)$$

Example continued

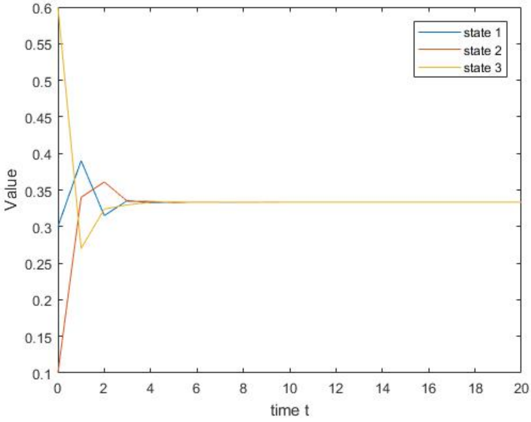
- We want an eigenvector with eigenvalue 1
- Eigenvalues are elements of D and first element is 1
- Corresponding eigenvector P icorresponds first column u_1 of U, i.e.

$$u_1 = \begin{bmatrix} 0.58 \\ 0.58 \\ 0.58 \end{bmatrix} \tag{17}$$

Hence, since P is a probability distribution:

$$P = \frac{u_1}{0.58 + 0.58 + 0.58} = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix} \tag{18}$$

MatLab simulation



doesn't correlate
to the initial
distribution

Another example

A company intranet consists of three pages W_1, W_2 and W_3. The way in which staff access and move between the pages in a browsing session is modelled as a 4-state Markov model $\mathcal{M}$, with initial state probability distribution P^0 and state transition probability matrix A given by:

$$P^{(0)} = \begin{bmatrix} 0.6 \\ 0.2 \\ 0.2 \\ 0 \end{bmatrix} A = \begin{bmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (19)$$

$\rightarrow$ exit state
control end of sequence

Pages W_1, W_2 and W_3 correspond to states 1, 2 and 3 of the Markov model

Analysis

1. We can think of state 4 as a *terminal state*. Once state 4 is entered the browsing session stops

$$P^{(2)} = A^T P^{(1)} = A^T A^T P^{(0)} \tag{20}$$

So,

$$P^{(1)} = \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0 \\ 0.3 & 0.2 & 0.3 & 0 \\ 0.3 & 0.2 & 0.2 & 0 \\ 0.1 & 0.4 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.2 \\ 0.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.28 \\ 0.26 \\ 0.2 \end{bmatrix} \tag{21}$$

Analysis (continued)

Therefore,

$$P^{(2)} = \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0 \\ 0.3 & 0.2 & 0.3 & 0 \\ 0.3 & 0.2 & 0.2 & 0 \\ 0.1 & 0.4 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 0.26 \\ 0.28 \\ 0.26 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.196 \\ 0.212 \\ 0.186 \\ 0.416 \end{bmatrix} \quad (22)$$

From this it is probably clear that as t gets bigger $P^{(t)}(4)$ gets bigger and that in the limit

$$P^{(t)} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (23)$$

$t \rightarrow \lim$

$\vec{p} = A^T \cdot \vec{p}$

$\boxed{A^T \cdot \vec{p} = \lambda \cdot \vec{p} = 1 \cdot \vec{p}}$

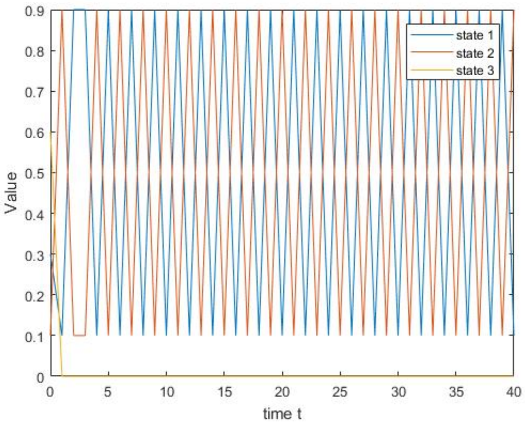
as $t \rightarrow \infty$. As before you can verify this with MatLab.

Convergence

- Does the state probability distribution $P^{(t)}$ always converge as $t \to \infty$?
- Consider: No.

$$P^0 = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.6 \end{bmatrix} A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (24)$$

MatLab simulation



Conditions for convergence of Markov processes

Sufficient conditions to ensure convergences are:

connected every state accessible after t.

1. The model must be *irreducible*: For every pair of states s_0 and s_1 there must be a time t and a state sequence $x_1, x_2, \cdots, x_t$ with $s_0 = x_1$, $s_1 = x_t$ and $P(x_1, x_2, \cdots, x_t) > 0$. In other words, it is possible to get from state s_0 to s_1 via a state sequence $x_1, x_2, \cdots, x_t$ with non-zero probability.
2. The model must be *aperiodic*: A state is aperiodic if the HCF of the set of *return times* for the state must be 1. A model is aperiodic if all of its states are aperiodic.

Simplified Page Rank

Returning to Page rank:

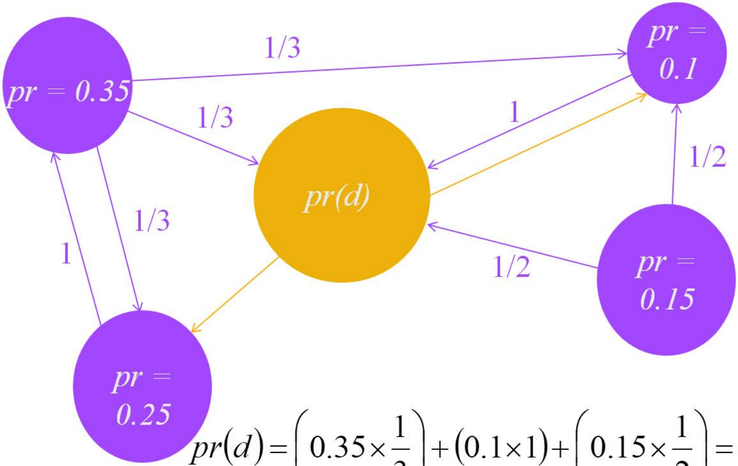
- Given a set of documents $D = \{d_1, \cdots d_N\}$, define:
- $pa(n)$ - the set of pages pointing to d_n
- h_n - the number of hyperlinks from d_n
- The **simplified Page rank** x_n for document d_n is given by:

$$x_n = \sum_{d_m \in pa(n)} \frac{x_m}{h_m} \tag{25}$$

- Because equation (26) is self-referencing, write

$$x_n^{(i+1)} = \sum_{d_m \in pa(n)} \frac{x_m^{(i)}}{h_m} \tag{26}$$

Simplified Page Rank



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

Simplified Page Rank - Matrix formulation

- Let W be the $N \times N$ matrix whose $(m, n)^{th}$ entry is given by

$$w_{m,n} = \begin{cases} \overset{p_{nm}}{\frac{1}{h_n}} \text{ if there is a hyperlink from } x_n \text{ to } x_m \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

- Let $x^{(i)}$ be the $N \times 1$ column vector whose n^{th} entry is $x_n^{(i)}$
- Then

$$x^{(i+1)} = Wx^{(i)} \tag{28}$$

The matrix W

$$W\vec{x}^{(i)} = \begin{bmatrix} \frac{1}{h_1} & \frac{1}{h_2} & 0 & \cdots & \frac{1}{h_n} & \cdots & \frac{1}{h_N} \\ 0 & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & \frac{1}{h_n} & \cdots & \frac{1}{h_N} \\ \frac{1}{h_1} & 0 & \frac{1}{h_3} & \cdots & \frac{1}{h_n} & \cdots & \frac{1}{h_N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{h_1} & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & \frac{1}{h_n} & \cdots & \frac{1}{h_N} \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix} \tag{29}$$

Simplified Page rank - Markov model interpretation

This is a Markov model of simplified Page rank, where

$$P^{(0)} = x^{(0)} \tag{30}$$

$$A = W^T \tag{31}$$

In other words:

- $P^{(0)}$ is the initial estimate of Page rank
- W is the **transpose** of the state transition probability matrix A

$A = W^T$ means $P^{(t)} = WP^{(t-1)}$ (instead of $W^T P^{(t-1)}$)

The "damping factor"

- Until now, all the authority, x_n of a page d_n comes from the pages that have hyperlinks to it, from equation (28).
- A page with no incoming hyperlinks will have authority 0 (c.f. irreducibility)
- A solution is the **damping factor** $d \in \mathbb{R}, 0 < d < 1$
- d is the proportion of authority that a page gets by default
- The simple Page rank equation (28) becomes

$$x^{(i+i)} = (1-d)Wx^{(i)} + \frac{(d)}{N} 1_N \quad \Rightarrow + \frac{d}{N} \qquad (32)$$

uniformly distributed

where $1_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ is a $N \times 1$ column vector of 1s

Convergence

- The expression

$$x = (1-d)Wx + \frac{d}{N}1_N \tag{33}$$

  is a system of N equations in N unknowns.
- Considered as a dynamical system

$$x^{(t+1)} = (1-d)Wx^{(t)} + \frac{d}{N}1_N \tag{34}$$

  converges for any initial condition $x^{(0)}$ to a **unique** fixed point x^* such that

$$x^* = (1-d)Wx^* + \frac{d}{N}1_N \tag{35}$$

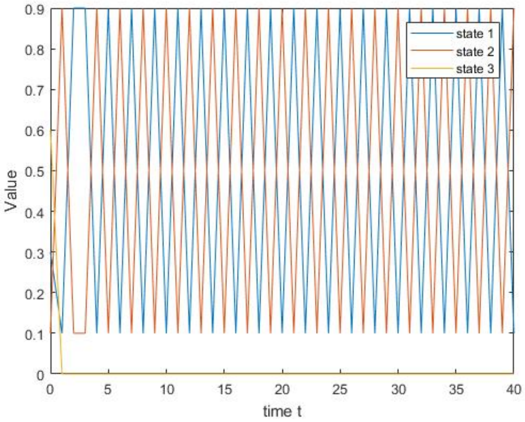
Convergence - damping factor

- Recall earlier example of **non-convergent** process:

$$P^0 = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.6 \end{bmatrix} A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{36}$$

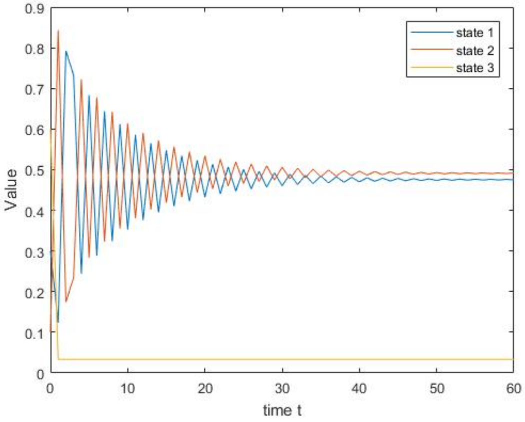
MatLab simulation

No damping factor ($d = 0$)



MatLab simulation (with $d = 0.1$)

With damping factor $d = 0.1$



Dangling pages

- A page which contains no hyperlinks (out) is called a **dangling page**
- If d_n is a dangling page then the n^{th} column of W consists entirely of zeros
- Hence the n^{th} column sums to 0 and W is no longer a (column) stochastic matrix
- In this case some parts of the above analysis no longer holds

Dangling pages - solution

- Introduce a new "dummy" page d_{N+1}
- Add a hyperlink to d_{N+1} from every dangling page
- Extend the transition matrix W to get a new $(N+1) \times (N+1)$ matrix $\overline{W}$
  - introduce a **dangling page indicator** $\vec{r} = [r_1, r_2, \cdots, r_N]$, where

$$r_n = \begin{cases} 1 \text{ if } d_n \text{ is a dangling page} \\ 0 \text{ otherwise} \end{cases} \quad (37)$$

  - Add $\vec{r}$ to the bottom of W
  - Add an additional column of N 0s followed by a single 1

The extended matrix $\overline{W}$

Suppose that page n is a dangling page:

$$\overline{W} = \begin{bmatrix} \frac{1}{h_1} & \frac{1}{h_2} & 0 & \cdots & 0 & \cdots & \frac{1}{h_N} & 0 \\ 0 & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & 0 & \cdots & \frac{1}{h_N} & 0 \\ \frac{1}{h_1} & 0 & \frac{1}{h_3} & \cdots & 0 & \cdots & \frac{1}{h_N} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 \\ \frac{1}{h_1} & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & 0 & \cdots & \frac{1}{h_N} & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 1 \end{bmatrix} \tag{38}$$

- Each dangling page has a hyperlink to d_{N+1}
- d_{N+1} only has a link to itself.

An alternative solution to dangling pages

- Create links from dangling pages to **all** pages
- Construct $N \times N$ matrix V with N equal rows $\frac{\vec{r}}{N}$
- The $N \times N$ matrix $(W + V)$ has the form:

$$V + W = \begin{bmatrix} \frac{1}{h_1} & \frac{1}{h_2} & 0 & \cdots & \frac{1}{N} & \cdots & \frac{1}{h_N} \\ 0 & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & \frac{1}{N} & \cdots & \frac{1}{h_N} \\ \frac{1}{h_1} & 0 & \frac{1}{h_3} & \cdots & \frac{1}{N} & \cdots & \frac{1}{h_N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{h_1} & \frac{1}{h_2} & \frac{1}{h_3} & \cdots & \frac{1}{N} & \cdots & \frac{1}{h_N} \end{bmatrix} \tag{39}$$

Modified Page rank equation becomes:

$$x = (1 - d)(W + V)x + \frac{d}{N} 1_N \tag{40}$$

Probabilistic interpretation of Page rank

- Think of each page as a state of a Markov model (or node in a graph)
- Nodes connected by hyperlink structure
- Connection between node p and q is weighted by the **probability** of its usage
- Weights depend only on current node, not how we got there (Markov property)
- Surfer never stops surfing
- At eany time t the surfer becomes bored with probability $1 - d$ and jumps to any web page with equal probability $\frac{1}{N}$

Summary

- Motivation - not all pages are equal
- Markov prcesses
- Simplified Page rank $x^{(i+1)} = Wx^{(i)}$
- Page rank with damping factor $x^{(i+1)} = (d-1)Wx^{(i)} + \frac{d}{N}1_N$
- Dealing with dangling pages
- Probabilistic interpretation of Page rank