Web retrieval: Link analysis for page ranking

Massimo Melucci
massimo.melucci@unipd.it

"Information Management Systems" Research Group
Department of Information Engineering
University of Padova, Italy

European Summer School on Information Retrieval
3rd September 2003, Aussois, France

Authority and relevance of Web pages

- the quality of conventional IR system document collection is homogeneous
- the Web is uncontrolled and quality is highly heterogeneous
- one aspect of quality is authority
- relevance is then crossed with authority
- however information about authority is not available
- directories and categories might be a means

The role of links to measure authority

- links might supply information about page authority
- if the author of page A thinks that page B is important, relevant or more generally related, he is likely to insert a link from A to B
- two assumptions
  - a link is likely to express authority of the target page
  - the more the links, the greater the authority

Note 3-1

The role of links to measure authority

- *the use of links to measure authority implies that the latter is conferred to a page by another page*
- *this is not necessary*
- *one might infer authority on the basis of "stand-alone" properties, e.g. typographical features or layout*
- *for example, the electronic version of a journal paper would be more authoritative than a "casual" HTML page*

Care when using links to measure authority

- it is uncertain that a link is likely to express the authority of the target page
  - a link might not point to an authority
  - a link might point to a non-authority
  - a page might be pointed-to w.r.t. one or more subjects
- the number of in-links might not be a measure of authority
  - a popular page is directly pointed to by many links
  - authoritative pages might be less pointed-to
- link analysis based methods might let authorities emerge because deal with large numbers

Note 4-1

Care when using links to measure authority

- *automatically generated links rarely point to authoritative pages*
- *if they did, there would exist an automatic method to detect authorities*
- *advertisement links are very often pointing to non-authorities*
- *the methods being illustrated in this lecture are unable to let young authorities emerge – they are little pointed-to by other pages*

Two link analysis approaches for Web page ranking

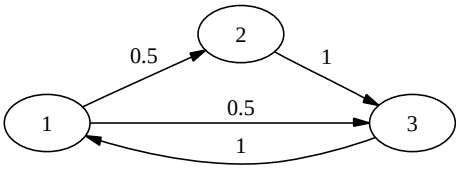
- Markov chains
  - model *navigation*
  - authorities link to authorities
  - rank by the probability that the page is reached
  - applied at indexing time
- mutual reinforcement relationship
  - model *authoring*
  - hubs link to authorities
  - rank by the degree to which the page is pointed-to by hubs that point to other authorities
  - applied at retrieval time

Related work

- bibliometrics and the measures of impact factors of scientific "units" (journals, papers, etc.)
- social networks and the measures of standing and social influence
- hypertext information retrieval
- hypertext structure analysis

Markov chains

- a set of states and a set of transitions between states
- p_{ij} is the transition probability that state j is reached from i
- depicted as a weighted and directed graph, where nodes are states and edge weights are probabilities of transition



the sequence of states $(1, 3, 1, 2)$ has probability $p_1 p_{13} p_{31} p_{12} = \frac{1}{3}\frac{1}{2}1\frac{1}{2} = \frac{1}{12}$

Note 7-1

Markov chains

- *a Markov chain is defined as follows*
- S *is a discrete and finite state space* $\{1, 2, \ldots, m\}$ *(but see below)*
- *the initial probability of state* i *is* p_i*, such that* $\sum_i p_i = 1$
- *each page has at least one out-link, i.e. there are not "sink" states*
- *the probability of transition from* i *to* j *is* p_{ij}*, given* i
- $\{p_{i1}, \ldots p_{im}\}$ *is a probability distribution*

$$p_{ij} \geq 0 \qquad \sum_j p_{ij} = 1$$

- *the probability of the sequence of states* $i_0, i_1, i_2, \ldots, i_{n-1}, i_n$ *is defined by* $p_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}$

Markov chains

- *we are considering time homogeneous or invariant Markov chains, which are a special case of Markov chains*
- *the transition probabilities of the more general case are defined as*

$$p_{ij}(t) = \Pr(j \text{ is reached at time } t \mid i \text{ is reached at time } t-1)$$

- *the invariant Markov chains have the property that their transition probabilities are independent of time* t
- *Markov chains are a special case of stochastic processes whose transition probabilities depend on states that are reached before the previous one*

Stochastic processes

- S *is the discrete state space and* T *is the discrete parameter space (in general,* S *or* T *might be continuous)*
- X_t *is a random variable depending on* $t \in T$ *and taking values in* S
- $(t_0, t_1, ..., t_n, t)$ *is finite or countably infinite and* $t_i < t_{i+1}, t_n < t$
- *the stochastic process* $\{X_t, t \in T\}$ *has probability function*

$$\Pr(X_t = x \mid X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_0} = x_0)$$

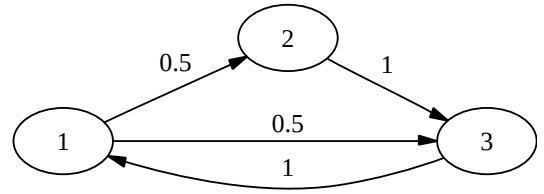
- *with Markov chains*

$$p_{x_n,x}(t) = \Pr(X_t = x \mid X_{t_n} = x_n)$$

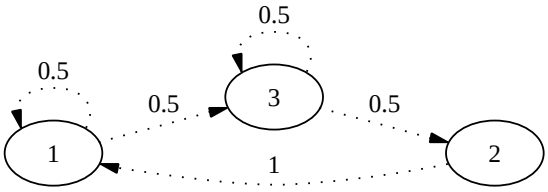
- *with invariant Markov chains*

$$p_{x_n,x} = \Pr(X_t = x \mid X_{t_n} = x_n) = \Pr(X_{t_1} = x \mid X_{t_0} = x_n)$$

n-step transition probability



one-step transition



two-step transition

$$
\begin{aligned}
p_{11}^{(2)} &= p_{11}p_{11} + p_{12}p_{21} + p_{13}p_{31} \\
&= 0 + 0.5 \cdot 0 + 0.5 \cdot 1 \\
&= \tfrac{1}{2} \\
p_{32}^{(3)} &= p_{31}^{(2)}p_{12} + p_{32}^{(2)}p_{22} + p_{33}^{(2)}p_{32} \\
&= 0 \cdot \tfrac{1}{2} + 0.5 \cdot 0 + 0.5 \cdot 0 \\
&= 0
\end{aligned}
$$

n-step transition probability

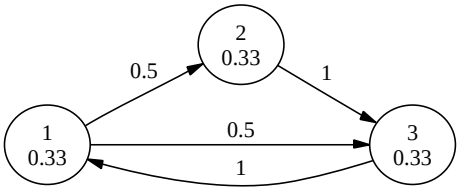
- by definition
$$p_{ij}^{(1)} = p_{ij}$$
is the one-step transition probability from i to j
- the two-step transition probability is
$$p_{ij}^{(2)} = p_{i1}p_{1j} + p_{i2}p_{2j} + \ldots + p_{im}p_{mj} = \sum_k p_{ik}p_{kj}$$
- in general,
$$p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)}\ p_{kj} \qquad n = 1, 2, \ldots$$
is the n-step transition probability, where $p_{ik}^{(0)} = 1$ if $i = k$, 0 otherwise

Chapman-Kolmogorov equation

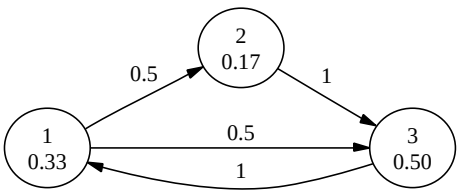
- *for time homogeneous, discrete and finite Markov chains*
- *for any* r *such that* $0 < r < n$,

$$p_{ij}^{(n)} = \sum_{k \in S} p_{ik}^{(r)} p_{kj}^{(r,n)} \tag{1}$$

State probability



$$p_1^{(0)} = p_2^{(0)} = p_3^{(0)} = 0.33$$



$$p_i^{(1)} = p_1^{(0)} p_{1i} + p_2^{(0)} p_{2i} + p_3^{(0)} p_{3i}$$

n-step state probability

- initial probability distribution

$$p_i^{(0)} \qquad i = 1, 2, \ldots m$$

- the probability of state i after one step is

$$p_i^{(1)} = p_1^{(0)} p_{1i} + p_2^{(0)} p_{2i} + \ldots + p_m^{(0)} p_{mi} = \sum_k p_k^{(0)} p_{ki}$$

- in general, the probability of state i at step n is

$$p_i^{(n)} = \sum_k p_k^{(n-1)} p_{ki}$$

Matrix representation

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{p}^{(0)} = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}$$

one-step transition probability matrix

initial state probability vector

Matrix representation

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & & \vdots \\ p_{m1} & \cdots & p_{mm} \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix}$$

$$\mathbf{p}^{(n)} = \mathbf{P}'\mathbf{p}^{(n-1)} = \begin{bmatrix} \sum_i p_{i1} p_i^{(n-1)} \\ \vdots \\ \sum_i p_{im} p_i^{(n-1)} \end{bmatrix}$$

Matrix representation of n-step transition probability

$$\begin{bmatrix} 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{P}^2 = \mathbf{P} \qquad \mathbf{P}$$

$$\begin{bmatrix} 0.406 & 0.188 & 0.406 \\ 0.438 & 0.187 & 0.375 \\ 0.375 & 0.219 & 0.406 \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \cdots \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{P}^{10} = \underbrace{\mathbf{P} \quad \cdots \quad \mathbf{P}}_{10 \text{ times}}$$

Matrix representation of n-step transition probability

$$\begin{aligned}\mathbf{P}^n &= \underbrace{\mathbf{P}\cdots\mathbf{P}}_{n \text{ times}} \\ &= \begin{bmatrix} p_{11}^{(n)} & \cdots & p_{1m}^{(n)} \\ \vdots & & \vdots \\ p_{m1}^{(n)} & \cdots & p_{mm}^{(n)} \end{bmatrix}\end{aligned}$$

Matrix representation of n-step state probability

$$\begin{bmatrix} 0.333 \\ 0.167 \\ 0.500 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.333 \\ 0.333 \end{bmatrix}$$

$$\mathbf{p}^{(1)} = \mathbf{P}' \mathbf{p}^{(0)}$$

$$\begin{bmatrix} 0.417 \\ 0.166 \\ 0.417 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.250 \\ 0.417 \end{bmatrix}$$

$$\mathbf{p}^{(4)} = \mathbf{P}' \mathbf{p}^{(3)}$$

Transition probability and state probability at step n

$$
\begin{array}{ccccc}
\begin{bmatrix} 0.417 \\ 0.166 \\ 0.417 \end{bmatrix} & = & \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} & & \begin{bmatrix} 0.333 \\ 0.250 \\ 0.417 \end{bmatrix} \\
\mathbf{p}^{(4)} & = & \mathbf{P}' & & \mathbf{p}^{(3)} \\
 & = & \begin{bmatrix} 0.25 & 0.50 & 0.50 \\ 0.25 & 0 & 0.25 \\ 0.50 & 0.50 & 0.25 \end{bmatrix} & & \begin{bmatrix} 0.333 \\ 0.333 \\ 0.333 \end{bmatrix} \\
 & = & \mathbf{P}^{4\prime} & & \mathbf{p}^{(0)}
\end{array}
$$

Note 15-1

Transition probability and state probability at step n

- *n-step state probability*

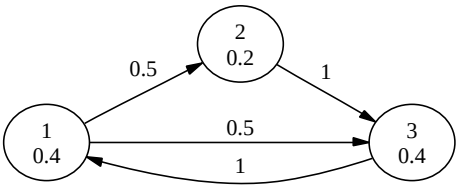
$$\mathbf{p}^{(n)} = \mathbf{P}'\mathbf{p}^{(n-1)}$$

- *relationship with n-step transition probability*

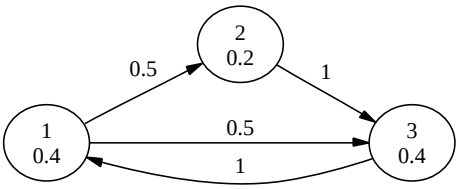
$$
\begin{aligned}
\mathbf{p}^{(n)} &= \mathbf{P}'\mathbf{p}^{(n-1)} \\
&= \mathbf{P}'(\mathbf{P}'\mathbf{p}^{(n-2)}) \\
&= \mathbf{P}'(\mathbf{P}' \dots (\mathbf{P}'\mathbf{p}^{(0)})) \\
&= \mathbf{P}^{n\prime}\mathbf{p}^{(0)}
\end{aligned}
$$

Stationarity

$$\begin{array}{ccc} \mathbf{p}^{(n)} & = & \mathbf{P}' \quad\quad \mathbf{p}^{(n-1)} \\ \begin{bmatrix} 0.4 \\ 0.2 \\ 0.4 \end{bmatrix} & = & \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.2 \\ 0.4 \end{bmatrix} \end{array}$$



chain at step 0



chain at step $n > 0$ *– note that the state probability is stationary*

Stationarity and matrices

- $\mathbf{p}$ *is stationary if*

$$p_j^{(n)} = p_j^{(n-1)} \qquad n = 1, 2, \ldots \qquad j = 1, 2, \ldots$$

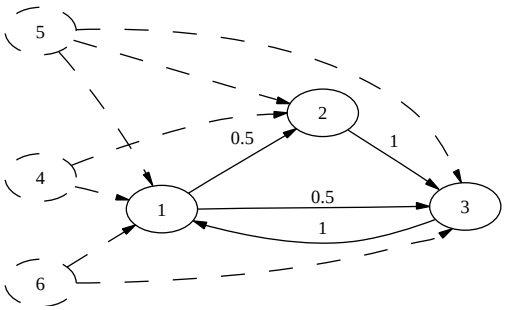
- *matrix form*

$$\mathbf{p} = \mathbf{P}'\mathbf{p}$$

where

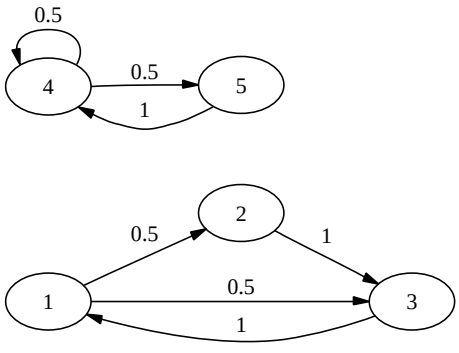
$$\mathbf{p} = \mathbf{p}^{(n)} \qquad n = 1, 2, \ldots$$

An irreducible chain



- let us consider the "solid" sub-graph with states $\{1, 2, 3\}$
- each state can be reached from any other state after $n \geq 0$ steps
- then every pair of states communicate between them within one chain
- the sub-chain is closed because outside states cannot be reached
- $\{1, 2, 3\}$ is irreducible because there are not closed subchains of it

A non-irreducible chain

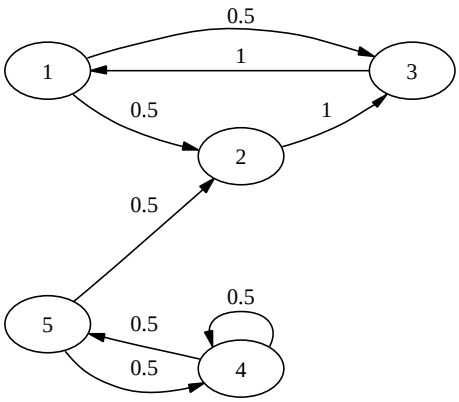


- some states cannot be reached from some states
- there are states that do not communicate between them within the chain
- the chain is not irreducible because contains two closed subchains
- which are irreducible

Irreducibility

- *state j is accessible from state i if j can be reached from i in a finite number of steps*
- *a chain is closed if no state outside is accessible from any state inside it*
- *states i and j are said to communicate if they are accessible to each other*
- *communication is an equivalence relationship and S can be partitioned into equivalence classes such that states belonging to different equivalence classes do not communicate with each other*
- *if there is one equivalence class, the chain is irreducible*

Persistency and transiency

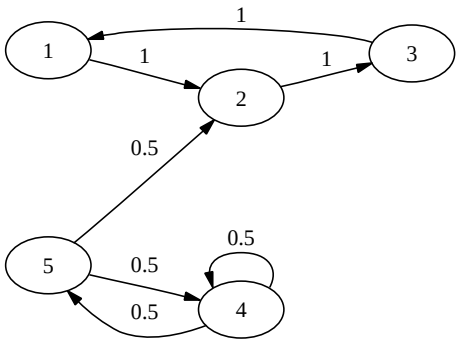


- $4, 5$ are transient – eventual return is uncertain
- $1, 2, 3$ are persistent – eventual return is certain

Persistency and transiency

- *a state i is persistent if and only if, starting from state i, eventual return of the chain to i is certain*
- *otherwise i is transient*
- *if a state i is an element of a equivalence class and i is persistent (transient), then all the other states of the same class are persistent (transient)*

A periodic chain



- $1, 2, 3$ are periodic

Periodicity

- *another example is given by*

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

- $p_{ii}^{(n)} = 0$ *if* n *is odd and* $p_{ii}^{(n)} > 0$ *if* n *is even, then state* i *is periodic and period is two*
- *state* i *has period is three if* $p_{ii}^{(n)} > 0$ *if* $n = 3k$, 0 *otherwise*
- *in general, the period of state* i *is the greatest common divisor of all integers* $n \geq 1$ *for which* $p_{ii}^{(n)} > 0$
- *if every state of a class has period one, then all the states are aperiodic and the class is aperiodic*
- *if* $p_{ij} > 0, i, j = 1, 2, \ldots, m$ *then the chain is irreducible, persisten and aperiodic*

Limit probabilities of an irreducible, a-periodic and persistent chain

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{P}^{30} = \begin{bmatrix} 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 0.4 \\ 0.2 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.4 & 0.4 \\ 0.2 & 0.2 & 0.2 \\ 0.4 & 0.4 & 0.4 \end{bmatrix} \begin{bmatrix} p_1^{(0)} \\ p_2^{(0)} \\ p_3^{(0)} \end{bmatrix}$$

$$\mathbf{p}^{(30)} = (\mathbf{P}^{30})' \mathbf{p}^{(0)}$$

Limit probabilities of two closed sub-chains

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

n	0	30	0	30
$p_1^{(n)}$	0.3	0.36	0.07	0.08
$p_2^{(n)}$	0.3	0.18	0.07	0.04
$p_3^{(n)}$	0.3	0.36	0.06	0.08
$p_4^{(n)}$	0.05	0.07	0.4	0.53
$p_5^{(n)}$	0.05	0.03	0.4	0.27

Limit probabilities of transient and persistent chains

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0.5 \end{bmatrix} \quad \mathbf{P}^{30} = \begin{bmatrix} 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.4 \\ 0.2 \\ 0.4 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1^{(0)} \\ p_2^{(0)} \\ p_3^{(0)} \\ p_4^{(0)} \\ p_5^{(0)} \end{bmatrix}$$

$$\mathbf{p}^{(30)} = (\mathbf{P}^{30})' \mathbf{p}^{(0)}$$

Limit probabilities of a periodic chain

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \end{bmatrix}$$

n	0	10	20	30	40	...
$p_1^{(n)}$	0.20	0.29	0.38	0.32	0.30	
$p_2^{(n)}$	0.20	0.31	0.30	0.38	0.32	
$p_3^{(n)}$	0.20	0.37	0.32	0.30	0.38	
$p_4^{(n)}$	0.20	0.02	0.00	0.00	0.00	
$p_5^{(n)}$	0.20	0.01	0.00	0.00	0.00	

Note 23-1

Limit transition probabilities of an irreducible chain

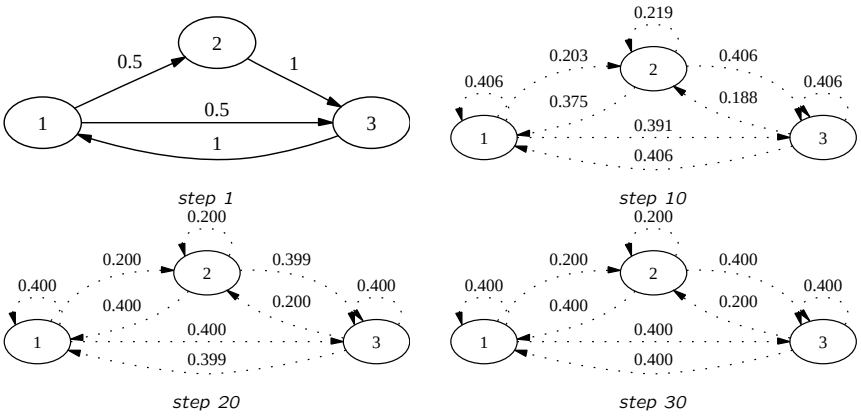
$\mathbf{P}^1$	1	2	3
1	0	0.50	0.50
2	0	0	1.00
3	1.00	0	0

$\mathbf{P}^5$	1	2	3
1	0.50	0.12	0.38
2	0.50	0.25	0.25
3	0.25	0.25	0.50

$\mathbf{P}^{10}$	1	2	3
1	0.41	0.19	0.41
2	0.44	0.19	0.37
3	0.37	0.22	0.41

$\mathbf{P}^{30}$	1	2	3
1	0.40	0.20	0.40
2	0.40	0.20	0.40
3	0.40	0.20	0.40

Limit transition probabilities of an irreducible chain



Limit transition probabilities of an non-irreducible chain

$\mathbf{P}^1$	1	2	3	4	5
1	0	0.50	0.50	0	0
2	0	0	1.00	0	0
3	1.00	0	0	0	0
4	0	0	0	0.50	0.50
5	0	0	0	1	0

$\mathbf{P}^5$	1	2	3	4	5
1	0.37	0.25	0.38	0	0
2	0.25	0.25	0.50	0	0
3	0.50	0.12	0.38	0	0
4	0	0	0	0.67	0.33
5	0	0	0	0.66	0.34

$\mathbf{P}^{10}$	1	2	3	4	5
1	0.41	0.20	0.39	0	0
2	0.37	0.22	0.41	0	0
3	0.40	0.19	0.41	0	0
4	0	0	0	0.67	0.33
5	0	0	0	0.67	0.33

$\mathbf{P}^{30}$	1	2	3	4	5
1	0.40	0.20	0.40	0	0
2	0.40	0.20	0.40	0	0
3	0.40	0.20	0.40	0	0
4	0	0	0	0.67	0.33
5	0	0	0	0.67	0.33

Limit transition probabilities of another non-irreducible chain

$\mathbf{P}^1$	1	2	3	4	5
1	0	0.50	0.50	0	0
2	0	0	1.00	0	0
3	1.00	0	0	0	0
4	0	0	0	0.50	0.50
5	0	0.50	0	0.50	0.50

$\mathbf{P}^5$	1	2	3	4	5
1	0.37	0.25	0.38	0	0
2	0.25	0.25	0.50	0	0
3	0.50	0.12	0.38	0	0
4	0.25	0.14	0.28	0.20	0.13
5	0.31	0.23	0.25	0.13	0.08

$\mathbf{P}^{10}$	1	2	3	4	5
1	0.41	0.20	0.39	0	0
2	0.38	0.22	0.39	0	0
3	0.41	0.19	0.40	0	0
4	0.39	0.19	0.37	0.07	0.04
5	0.39	0.19	0.36	0.04	0.03

$\mathbf{P}^{30}$	1	2	3	4	5
1	0.40	0.20	0.40	0	0
2	0.40	0.20	0.40	0	0
3	0.40	0.20	0.40	0	0
4	0.40	0.20	0.40	0	0
5	0.40	0.20	0.40	0	0

Limit transition probabilities of a periodic chain

$\mathbf{P}^1$	1	2	3	4	5
1	0	1.00	0	0	0
2	0	0	1.00	0	0
3	1.00	0	0	0	0
4	0	0	0	0.50	0.50
5	0	0.50	0	0.50	0

$\mathbf{P}^5$	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0.13	0.20	0.34	0.20	0.13
5	0.56	0.17	0.06	0.13	0.08

$\mathbf{P}^{10}$	1	2	3	4	5
1	0	0	1	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0.24	0.42	0.22	0.07	0.04
5	0.20	0.11	0.62	0.04	0.03

$\mathbf{P}^{30}$	1	2	3	4	5
1	0	1	0	0	0
2	0	0	1	0	0
3	1	0	0	0	0
4	0.45	0.26	0.29	0	0
5	0.13	0.65	0.23	0	0

Limit transition probabilities

- *if a chain is irreducible, persistent (all states are persistent) and aperiodic*

$$\lim_{n\to\infty} p_{ij}^{(n)} = p_j \qquad j = 1, 2, \ldots$$

where $\{p_j\}$ *is stationary and* $p_j > 0$ *for every* j

- *if not persistent,* $p_j \geq 0$
- *matrix form*

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} p_1 & \cdots & p_m \\ \vdots & & \vdots \\ p_1 & \cdots & p_m \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ \vdots \\ \mathbf{p} \end{bmatrix} = \mathbf{p}$$

Limit state probabilities

- *since*

$$p_j^{(n)} = \sum_{i\to j} p_j^{(n-1)} p_{ij} = \sum_{i\to j} p_j^{(0)} p_{ij}^{(n)}$$

then

$$\begin{aligned} p_j &= \lim_{n\to\infty} p_j^{(n)} \\ &= \lim_{n\to\infty} \textstyle\sum_{i\to j} p_j^{(0)} p_{ij}^{(n)} \\ &= \textstyle\sum_{i\to j} p_j^{(0)} \lim_{n\to\infty} p_{ij}^{(n)} \end{aligned}$$

Intuitive view of PageRank

- the Web "is a" Markov chain
- the PageRank of page j is the probability that the user reaches j through $i \rightarrow j$ given that he reached i
- all the links from i to j are counted once
- the PageRank of page j depends on those of the pages linking to it
- the more j is linked by pages with high PageRank, the higher its PageRank
- the formulation is recursive thus requiring an initial probability
- proposed by Brin and Page (1998)

PageRank and Markov chains

- the set of Web pages is the set of states
- initial probability

  $$p_j^{(0)}$$

  that the user is at page j at the beginning of navigation
- after n steps

  $$p_j^{(n)} = \sum_{i \rightarrow j} p_j^{(n-1)} p_{ij} \text{ where } p_{ij} = \frac{1}{o_i}$$

- the PageRank of j is

  $$\lim_{n \rightarrow \infty} p_j^{(n)}$$

PageRank and Markov chains

- *S is the set of Web pages, and is finite yet very large*
- *the initial probability does not depend on the time at which the user starts navigation*
- *also the transition probabilities do not*
- *PageRank is then modeled by time homogeneous (invariant) chains*
- *the arrival at state j depends on the last step only and the states at which the user has arrived before are ignored - this is the Markov property*
- *transition probability is independent of visit time*
- *chains are time discrete and state discrete*

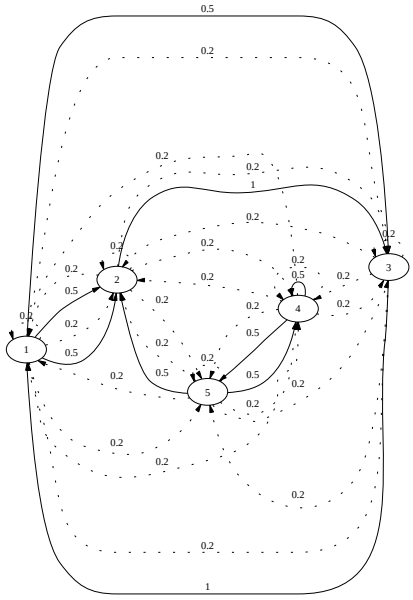
Requirements for PageRank formulation

- the chain must be a Markov chain, then $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$, therefore $o_i > 0$ for all i
- the chain must be *a-periodic* otherwise PageRank does not converge
- the chain must be *persistent* and *irreducible*, otherwise:
  - there are more than one irreducible and persistent disjoint subchains
    - the PageRank depends on the initial probability
  - there are transient pages and persistent pages
    - transient pages have null PageRank and are indistiguishable, while persistent pages absorb all the PageRank distribution

Decomposition of a chain

- *the states of a Markov chain can be divided, in a unique manner, into disjoint sets* $T, C_1, C_2, \ldots$ *such that*
- T *consists of all transient states*
- *if* $i \in C_k$ *then every* $j \in C_k$ *can be reached from* i*, whereas every* $j \in C_h,\ h \neq k$ *cannot be reached from* i
- *this implies that* C_k *is irreducible and contains only persistent states*

Converging to PageRank



- a page can be reached through actual links (solid edge) with probability $1 - d$ or other ways (dashed edge) with probability d
- other ways are URL typing, search engines, bookmarks, etc.
- a solid edge is weighted by the probability p_{ij} that $i \to j$ is followed
- a dashed edge is weighted by the probability q_{ij} that j is reached from i in another way
- if there were i such that $o_i = 0$, then $d = 1$

Note 28-1

Converging to PageRank

- *from the original formulation, PageRank of page i is the limit probability that a random surfer is at i when navigating*
- *if links are the only means the surfer easily get into a loop (periodicity) or leaves pages for ever (transiency)*
- *to extend it, note that surfers exploit alternative ways of access - search engines, "back" button, URL typing box - thus every page is potentially accessible*
- *if surfers fall into a "sink" page, then damping to another page is mandatory – this is why $d = 1$ for that page*

Converging to PageRank (cont.)

	1	2	3	4	5
1	0	0.5	0.5	0	0
2	0	0	1	0	0
3	1	0	0	0	0
4	0	0	0	0.5	0.5
5	0	0	0	1	0

$\mathbf{P}$

	1	2	3	4	5
1	0.2	0.2	0.2	0.2	0.2
2	0.2	0.2	0.2	0.2	0.2
3	0.2	0.2	0.2	0.2	0.2
4	0.2	0.2	0.2	0.2	0.2
5	0.2	0.2	0.2	0.2	0.2

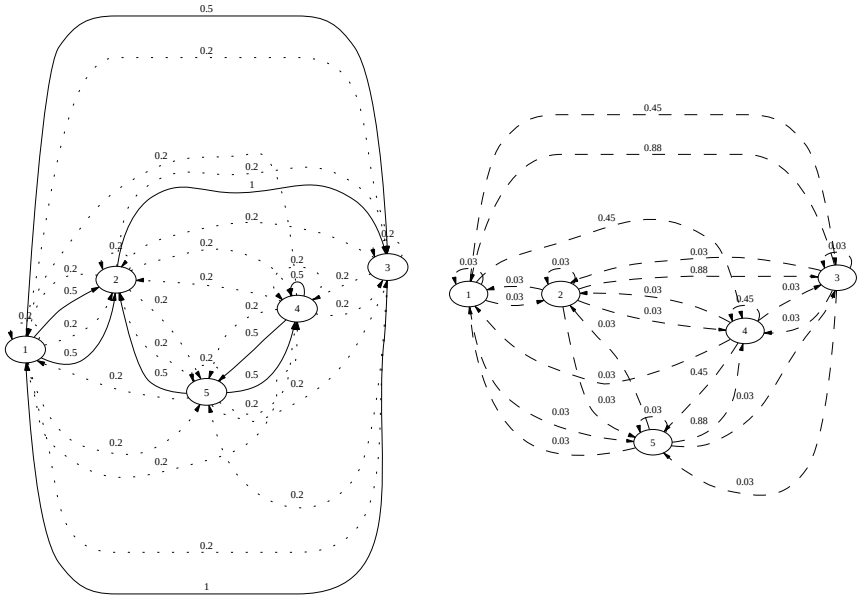
$\mathbf{Q} = (q_{ij}), q_{ij} = \frac{1}{m}$

$\mathbf{T} = (1-d)\mathbf{P}+d\mathbf{Q} =$

	1	2	3	4	5
1	0.03	0.45	0.45	0.03	0.03
2	0.03	0.03	0.88	0.03	0.03
3	0.88	0.03	0.03	0.03	0.03
4	0.03	0.03	0.03	0.45	0.45
5	0.03	0.03	0.03	0.88	0.03

provided $d = 0.15$

Converging to PageRank (cont.)



Note 31-1

Converging to PageRank

- *let d be the probability that the surfer gets page j through alternative ways of access independently of starting page (damping factor)*
- *$1-d$ is the probability that the surfer gets page j through in-links*
- *the transition probability is*

$$t_{ij} = \begin{cases} (1-d)\ p_{ij} + d\ q_j & \text{if } o_i > 0 \\ q_j & \text{if } o_i = 0 \end{cases}$$

- *state probability is defined as before*

$$p_j^{(n)} = \sum_{i \to j} p_i^{(n-1)} t_{ij} \qquad n = 1, 2, \ldots$$

- *this redefinition leads to a irreducible, persistent and aperiodic Markov chain – PageRank exists and is unique*
- *for nodes without out-links, d must be 1*

Converging to PageRank (cont.)

$$\begin{bmatrix} 0.23 \\ 0.13 \\ 0.24 \\ 0.26 \\ 0.14 \end{bmatrix} = \begin{bmatrix} 0.23 & 0.23 & 0.23 & 0.23 & 0.23 \\ 0.13 & 0.13 & 0.13 & 0.13 & 0.13 \\ 0.24 & 0.24 & 0.24 & 0.24 & 0.24 \\ 0.26 & 0.26 & 0.26 & 0.26 & 0.26 \\ 0.14 & 0.14 & 0.14 & 0.14 & 0.14 \end{bmatrix} \begin{bmatrix} p_1^{(0)} \\ p_2^{(0)} \\ p_3^{(0)} \\ p_4^{(0)} \\ p_5^{(0)} \end{bmatrix}$$

$$\mathbf{p} = (\lim_{n\to\infty} \mathbf{T}^n)' \mathbf{p}^{(0)}$$

Note 32-1

Converging to PageRank

- *let*

$$q_j = \frac{1}{m} \text{ and } \sum_j p_{ij} = 1$$

then

$$\sum_{j\in S} q_j = 1 \qquad \sum_{j\in S} t_{ij} = 1 \qquad i = 1, 2, \ldots, m$$

- *moreover* $t_{ij} > 0, i, j = 1, 2, \ldots, m$
- *then the chain is irreducible, persistent and aperiodic, and a unique PageRank exists*
- *note that this reformulation is sufficient yet not necessary to make PageRank unique*

Some extensions on PageRank

1. page-sensitive damping factor: damping factor is no longer uniform but changes according to the linked pages
2. topic-sensitive PageRank: damping factor is no longer uniform but changes according to the query topic
3. transition probabilities might be estimated in different ways

Page-sensitive damping factor

- a more general formulation of the transition probability of the PageRank chain would be

$$t_{ij} = (1 - d_i)\ p_{ij} + d_i\ q_j$$

  where the damping factor depends on page i
- the rationale is that damping is, for example, more likely if the current page is little useful

Page-sensitive damping factor

- *note that* $\{t_{ij}\}$ *is still a transition probability of a irreducible, persistent and aperiodic chain*

$$\begin{array}{ll} t_{ij} \geq 0 & \\ \sum_j t_{ij} & = (1-d_i)\sum_j p_{ij} + d_i \sum_j q_j \\ & = 1 - d_i + d_i \\ & = 1 \end{array}$$

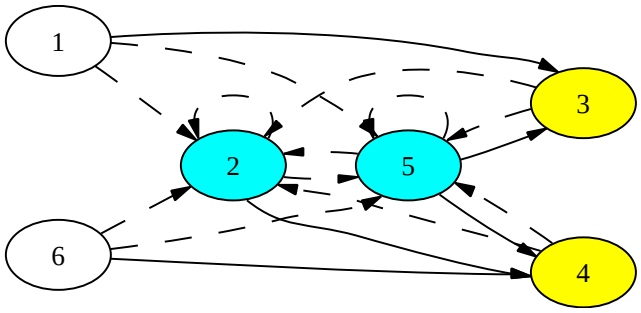
  provided that $\sum_j p_{ij} = 1$

- *an example is given by "sink pages" for which* $d_i = 1$

Topic-sensitive PageRank

- PageRank is computed once for each given Web graph
- it is independent of the query topic
- to make PageRank topic-sensitive, a set of predefined topics is selected
- for each topic a set of relevant pages is compiled
- PageRank is computed for each topic
- for each query the most probable topic is selected
- pages are ranked by the PageRank from the selected topic and the probability that the topic describes the query

Topic-sensitive PageRank



Note 36-1

Topic-sensitive PageRank

- *let us consider the 6-state Markov chain (t_{ij} is omitted for sake of simplicity)*
- *damping can occur to relevant pages only*
- $q_j = \frac{1}{2}$ *because there are 2 relevant pages*
- *to rank relevant pages, the subchain must be irreducible*
- *to make it irreducible, add all the pages, i.e.* $\{3, 4\}$ *being pointed-to by each relevant page*
- $\{2, 3, 4, 5\}$ *is the unique irreducible, persistent and a-periodic subchain − PageRank is unique*
- $\{1, 6\}$ *are transient − PageRank is null*

Topic-sensitive PageRank

- *let I be a subset of r relevant pages and*

$$q_j = \frac{1}{r} \qquad \textit{if } j \in I, \qquad q_j = 0 \textit{ otherwise}$$

- *C is the set of states that can be reached from I*
- *C is an irreducible, persistent and aperiodic class*
- *C is unique, PageRank exists positive and is unique for C whereas the pages outside I have null PageRank*
- *see Page et al. (1998), Haveliwala (2002), Pretto (2002)*

Alternative transition probability estimators

- if multiple links $i \to j$ are distinctly considered

$$p_{ij} = \frac{l_{ij}}{l_i}$$

  where l_{ij} is the number of distinct links $i \to j$ out of the $l_i = \sum_j l_{ij}$ total out-links from i

- using automatic hypertext generation methods

$$p_{ij} = \frac{\cos(\mathbf{v}_i, \mathbf{v}_j)}{\sum_j \cos(\mathbf{v}_i, \mathbf{v}_j)}$$

  means that p_{ij} is function of the cosine of the angle between the keyword vector representing i and the keyword vector representing j

Alternative transition probability estimators

- *the first estimator is based on the assumption that the distinct links are equivalently considered*
- *yet there might be some links being more likely to be followed, e.g. the one whose anchor is a bold text*
- *the second estimator might be replaced by the more "natural"*

$$p_{ij} = f(\Pr(\textit{relevance} \mid i, j))$$

that means that p_{ij} is function of the probability that j is relevant to the information need represented by i

- *note that transition probabilities are topic sensitive and must be either computed at retrieval time or pre-computed for pre-defined topics*

A couple of remarks on PageRank

1. in the seminal paper the PageRank formulation is slighly different
2. given a graph, page ranking depends on the damping factor

The original PageRank

- *in their seminal paper, Brin and Page wrote*

$$p'_j = (1-c) + c \sum_{i \to j} p'_i p_{ij}$$

- *note that*

$$\sum_j p'_j = \sum_j [(1-c) + c \sum_{i \to j} p'_i p_{ij}] = m(1-c) + c \sum_j \sum_{i \to j} p'_i p_{ij} = m$$

- *the question is whether this imprecision makes p' different from p*
- *one can show that*

$$p' = mp$$

thus the PageRank values change but ranking does not

Dependency of ranking on the damping factor

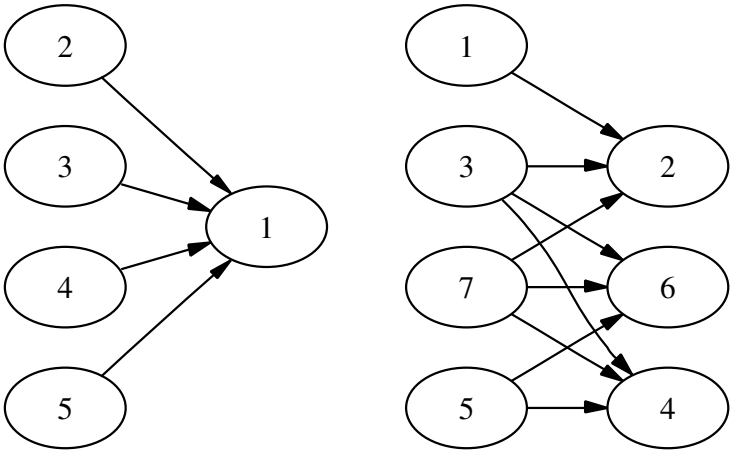
- *PageRank aims at ranking pages using links only*
- *the damping factor should be a parameter to make PageRank unique*
- *unfortunately, not only the PageRank values depend on the damping factor, but also page ranking does*
- *for example, the chain with transition probability matrix*

$$\begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

ranks pages with $d = 0.49$ differently if $d = 0.51$

- *see Pretto (2002)*

Mutual reinforcement relationship



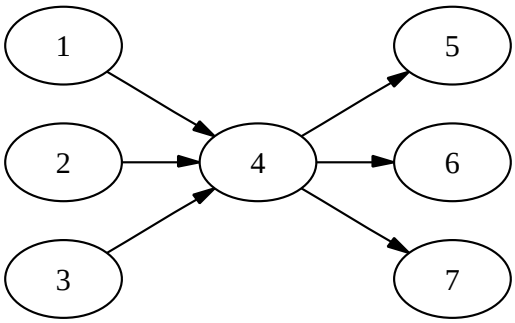
popular page authority and hubs

Note 39-1

Mutual reinforcement relationship

- *a popular page is directly pointed-to by many pages that do not frequently point to other pages*
- *an authoritative page is pointed-to by many pages, called "hub" that do frequently point to other (authoritative) pages*
- *authorities are pointed to by many hubs and hubs points to many authorities*
- *the more the page is pointed to by hubs, the more the page is authority*
- *the more the page point authorities, the more the page is hub*

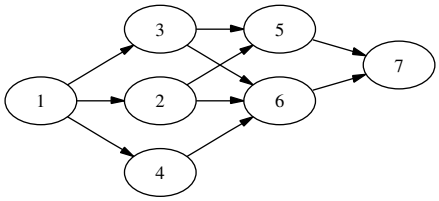
Computation of authority and page scores



$$a_4 = h_1 + h_2 + h_3$$

$$h_4 = a_5 + a_6 + a_7$$

An example of computation of authority and page scores



- after 10 steps, we have:

i	1	2	3	4	5	6	7
a_i	0	0.01	0.01	0.01	0.62	0.79	0
h_i	0.02	0.66	0.66	0.37	0	0	0

- note that 1 is a poor hub yet there are 3 out-links and
- that 7 is a poor authority yet there are 2 in-links

A formalization of the mutual reinforcement relationship

- *each page i is assigned an authority score a_i and a hub score h_i*
- *mutual reinforcement relationship*

$$a_i = \sum_{k \to i} h_k \qquad\qquad h_i = \sum_{i \to k} a_k$$

- *recursivity requires an iterative algorithm*
- *which score do we start computation from?*

An algorithm to measure mutual reinforcement

- each page i is assigned an authority score $a_i^{(n)}$ and a hub score $h_i^{(n)}$ at each step n

$$\begin{aligned} h_i^{(0)} &= 1 \\ a_i^{(1)} &= \textstyle\sum_{k \to i} h_k^{(0)} \\ h_i^{(1)} &= \textstyle\sum_{i \to k} a_k^{(1)} \\ a_i^{(2)} &= \textstyle\sum_{k \to i} h_k^{(1)} \\ &\vdots \end{aligned}$$

- when starting, hubs scores are set to constant values
- iteration continues until scores converge

An algorithm to measure mutual reinforcement

- let $h_i^{(0)} = 1$ for all $i = 1, 2, \ldots$
- let N be the number of iterations

$$a_i^{(n)} = \sum_{k \to i} h_k^{(n-1)} \qquad h_i^{(n)} = \sum_{i \to k} a_k^{(n)} \qquad n = 1, \ldots, N$$

- we will see that results change if $a_i^{(0)} = 1$ for all $i = 1, 2, \ldots$

$$h_i^{(n)} = \sum_{i \to k} a_k^{(n-1)} \qquad a_i^{(n)} = \sum_{k \to i} h_k^{(n)} \qquad n = 1, \ldots, N$$

An algorithm to measure mutual reinforcement: normalization

$h_i^{(0)} = 1, \ i = 1, 2, \ldots$	*initialize hub scores*
for $n = 1, 2, \ldots, N$	*at step* n
for $i = 1, \ldots, m$	*for each page* i
$a_i^{(n)} = \sum_{k \to i} h_k^{(n-1)}$	*update authority score*
$a_i^{(n)} = a_i^{(n)} / \sqrt{\sum_j (a_i^{(n)})^2}$	*normalize authority score*
$h_i^{(n)} = \sum_{i \to k} a_k^{(n)}$	*update hub score*
$h_i^{(n)} = h_i^{(n)} / \sqrt{\sum_j (h_i^{(n)})^2}$	*normalize hub score*
end for	
end for	

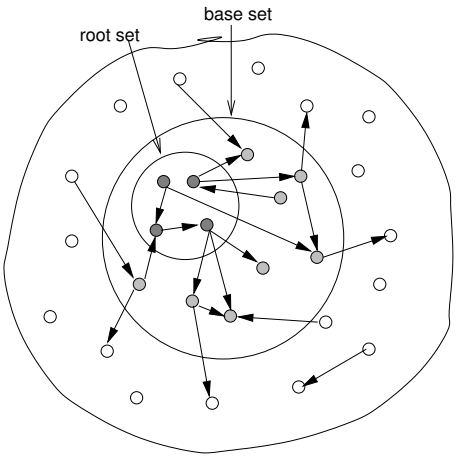
Remarks on the algorithm

- *the theory says that the number of iterations should be infinite*
- *in practice, the number of iterations must be finite yet "sufficiently" large – the answer to "how large?" depends on the instance*
- *two facts can be shown by a counter-example (see Pretto (2002)):*
  - *the algorithm is not symmetric, i.e. results change if computation starts after initializing authority scores instead of hub scores*
  - *results depend on the initial values given to the hub (authority) scores*
- *normalization changes scores but does not change page ranking*

Hyperlink Induced Topic Search

- an application of the mutual reinforcement algorithm
- target: broad topic queries
  - examples are "search engines", "java"
  - not only relevant pages but also authorities
- objective: discriminate authorities
- main ingredients:
  - a conventional search engine
  - some parameters
  - the algorithm based on mutual reinforcement relationship
- proposed by Kleinberg (1999)

Hyperlink Induced Topic Search (HITS)



given a query q:

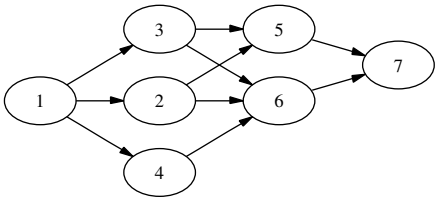
1. retrieve the root set (R_q)
2. expand R_q to the base set (B_q)
3. compute authorities and hubs in B_q
4. rank pages in B_q by authority or hub score

Note 44-1

Main steps of Hyperlink Induced Topic Search

- *let q be a query*
- *first a search engine retrieves the root set R_q matching q and selects the t top ranked pages*
  - *R_q is likely to contain many relevant pages yet they do not link each other*
- *then the base set B_q is built after adding all the pages that point to, or are pointed to by each page in R_q*
  - *B_q is still to contain many relevant pages but is likely to contain others that point to, or are pointed to by each page in R_q*
- *the algorithm is then performed on B_q*

Adjacency matrix



	1	2	3	4	5	6	7
1	0	1	1	1	0	0	0
2	0	0	0	0	1	1	0
3	0	0	0	0	1	1	0
4	0	0	0	0	0	1	0
5	0	0	0	0	0	0	1
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0

Note 45-1

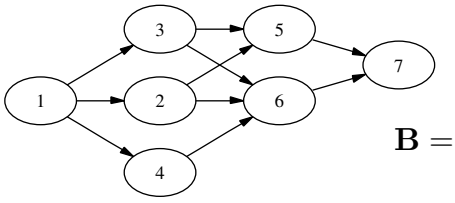
Adjacency matrix

- *a set of Web pages can be described by a graph such that a page is a node and a link is an edge*
- *an matrix, called "adjacency matrix", can be associated to a graph*
- *the adjacency matrix is*

$$\mathbf{D} = (d_{ij}) \qquad \textit{such that} \qquad d_{ij} = \begin{cases} 1 & \textit{if } i \rightarrow j \\ 0 & \textit{otherwise} \end{cases}$$

where $i, j = 1, 2, \ldots, m$ *and* m *is the number of pages*

Bibliographic coupling matrix



$$\mathbf{B} = \mathbf{D}\mathbf{D}' =$$

	1	2	3	4	5	6	7
1	3	0	0	0	0	0	0
2	0	2	2	1	0	0	0
3	0	2	2	1	0	0	0
4	0	1	1	1	0	0	0
5	0	0	0	0	1	1	0
6	0	0	0	0	1	1	0
7	0	0	0	0	0	0	0

Note 46-1

Bibliographic coupling matrix

- *both pages* i *and* k *cite* j *if* $i \rightarrow j$ *and* $k \rightarrow j$
- *note that*

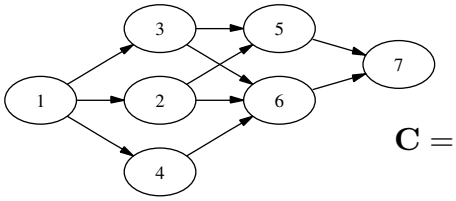
$$d_{ij}d_{kj} = \begin{cases} 1 & \textit{if and only if } i \rightarrow j \textit{ and } k \rightarrow j \\ 0 & \textit{otherwise} \end{cases}$$

- *the bibliographic coupling matrix is*

$$\mathbf{B} = \mathbf{D}\mathbf{D}' = (b_{ik}) \qquad \textit{such that} \qquad b_{ik} = \sum_{j=1}^{m} d_{ij}d_{kj}$$

- b_{ik} *is the number of pages that are cited by both* i *and* k
- *if* $\mathbf{D}$ *is* $r \times s$, $\mathbf{B}$ *is* $r \times r$

Co-citation matrix



$\mathbf{C} = \mathbf{D}'\mathbf{D} =$

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	1	1	1	0	0	0
3	0	1	1	1	0	0	0
4	0	1	1	1	0	0	0
5	0	0	0	0	2	2	0
6	0	0	0	0	2	3	0
7	0	0	0	0	0	0	2

Note 47-1

Co-citation matrix

- *both pages* j *and* k *are cited by* i *if* $i \to j$ *and* $i \to k$
- *note that*

$$d_{ij}d_{ik} = \begin{cases} 1 & \textit{if and only if } i \to j \textit{ and } i \to k \\ 0 & \textit{otherwise} \end{cases}$$

- *the co-citation matrix is*

$$\mathbf{C} = \mathbf{D}'\mathbf{D} = (c_{jk}) \qquad \textit{such that} \qquad c_{jk} = \sum_{i=1}^{m} d_{ij}d_{ik}$$

- c_{jk} *is the number of pages that cite both* k *and* j
- *if* $\mathbf{D}$ *is* $r \times s$, $\mathbf{C}$ *is* $s \times s$

Using the coupling and co-citation matrices to describe the algorithm

- let $\mathbf{h}^{(0)}$ be the initial hub scores, $\mathbf{C} = \mathbf{D}'\mathbf{D}$, $\mathbf{B} = \mathbf{D}\mathbf{D}'$
- the scores at step n are

$$\mathbf{h}^{(n)} = \mathbf{B}^n \mathbf{h}^{(0)}$$

and

$$\mathbf{a}^{(n)} = \mathbf{C}^{n-1}\mathbf{D}'\mathbf{h}^{(0)}$$

where $\mathbf{a}^{(1)} = \mathbf{D}'\mathbf{h}^{(0)}$

- $\mathbf{D}$ might be a non-square matrix, but $\mathbf{B}$ and $\mathbf{C}$ are always square matrices

Note 48-1

Using the coupling and co-citation matrices to describe the algorithm

$$\begin{array}{llll}
h_i^{(0)} & = 1 & & \mathbf{h}^{(0)} = [1,\ldots,1]' \\
a_i^{(1)} & = \sum_{k\to i} h_k^{(0)} & = \sum_k d_{ki} h_k^{(0)} & \mathbf{a}^{(1)} = \mathbf{D}'\mathbf{h}^{(0)} \\
h_i^{(1)} & = \sum_{i\to k} a_k^{(1)} & = \sum_k d_{ik} a_k^{(1)} & \mathbf{h}^{(1)} = \mathbf{D}\mathbf{a}^{(1)} \quad = (\mathbf{D}\mathbf{D}')\mathbf{h}^{(0)} \\
a_i^{(2)} & = \sum_{k\to i} h_k^{(1)} & = \sum_k d_{ki} h_k^{(1)} & \mathbf{a}^{(2)} = \mathbf{D}'\mathbf{h}^{(1)} \quad = (\mathbf{D}'\mathbf{D})\mathbf{D}'\mathbf{h}^{(0)} \\
\vdots & & & \vdots \\
 & & & \mathbf{h}^{(n)} \quad = (\mathbf{D}\mathbf{D}')^n\mathbf{h}^{(0)} \\
 & & & \mathbf{a}^{(n)} \quad = (\mathbf{D}'\mathbf{D})^{n-1}\mathbf{D}'\mathbf{h}^{(0)}
\end{array}$$

therefore

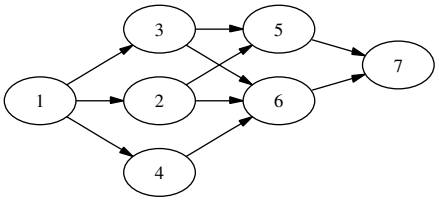
$$\mathbf{h}^{(n)} = \mathbf{B}^n\mathbf{h}^{(0)} \qquad \mathbf{a}^{(n)} = \mathbf{C}^{n-1}\mathbf{D}'\mathbf{h}^{(0)}$$

Uses and variations of the mutual reinforcement relationship

- **topic-based link weighting**: each link $i \to j$ is weighted with the degree to which the anchor is about the topic of j
- **statistical stemming**: the mutual reinforcement relationship is observed between stems and derivations and is applied to find the best word split
- **image retrieval**: links between pages and images are considered to find authority image, image hub/containers (pages pointing to/containing authority images)

Topic-based link weighting

- without weighting:



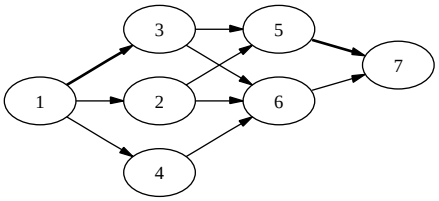
	1	2	3	4	5	6	7
1	0	1	1	1	0	0	0
2	0	0	0	0	1	1	0
3	0	0	0	0	1	1	0
4	0	0	0	0	0	1	0
5	0	0	0	0	0	0	1
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0

scores after 30 steps

i	1	2	3	4	5	6	7
a_i	0	0	0	0	0.62	0.79	0
h_i	0	0.66	0.66	0.37	0	0	0

Topic-based link weighting (cont.)

- with weighting:



	1	2	3	4	5	6	7
1	0	1	2	1	0	0	0
2	0	0	0	0	1	1	0
3	0	0	0	0	1	1	0
4	0	0	0	0	0	1	0
5	0	0	0	0	0	0	2
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0

scores after 30 steps

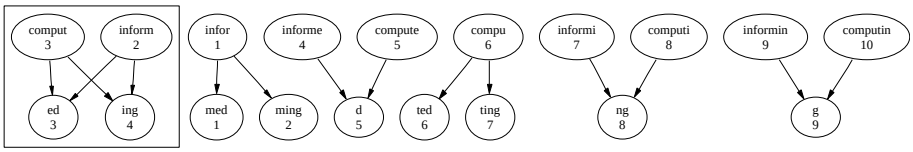
i	1	2	3	4	5	6	7
a_i	0	0.41	0.82	0.41	0	0	0
h_i	1	0	0	0	0	0	0

Note 51-1

Topic-based link weighting

- *the weight matrix* $\mathbf{W} = (w_{ij})$ *is used instead of* $\mathbf{D}$
- w_{ij} *is the measure of the degree to which page* i *confers authority to page* j *as regards to the topic*
- *needs to be computed at query time*
- *for example* $w_{ij} = 1 + f_{ij}$ *where* f_{ij} *is the number of topic terms occurring in the windows that are around the anchors*
- *in this way ranking is changed*

Link analysis-based stemming



- affix removal stemming – words are split into prefix and suffix, a stem is a prefix, a derivation is a suffix
- the key idea is mutual reinforcement among substrings:
  - stems are frequent prefixes that are followed by derivations
  - derivations are frequent suffixes that are preceded by stems

Link analysis-based stemming

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

- best authorities (derivations): "ed" (3) and "ing" (4) with score 0.71; the other scores are null
- best hubs (stems): "inform" (2) and "comput" (3) with score 0.71; the other scores are null

Link analysis-based stemming

- *let W be a set of words, X be the set of non-null prefixes and Y be the set of non-null suffixes*
- *a link $x \rightarrow y$ exists iff there exists $w \in W$ such that $w = xy$*
- *suffix/authority score and prefix/hub score*

$$s_y^{(k)} = \sum_{w \in W : w = xy} p_x^{(k-1)} \qquad\qquad p_x^{(k)} = \sum_{w \in W : w = xy} s_y^{(k)}$$

- *experimental results within CLEF are very similar to those obtained using the Porter's stemmers*

Link analysis-based image retrieval

- given a topic q, the set of pages matching q is retrieved
  - matching can be performed by any function
- the set of images contained in, or linked to by the retrieved pages is then collected
- mutual reinforcement is applied
  - pages are candidate hubs
  - images are candidate authorities

Link analysis-based image retrieval (cont.)

- let $\mathbf{D} = (d_{pi})$ be the page-image adjacency matrix such that

$$d_{pi} = \begin{cases} 1 & \text{if page } p \text{ contains or links to image } i \\ & \text{or to a page containing } i \\ 0 & \text{otherwise} \end{cases}$$

- authority image score is

$$a_i = \sum_p d_{pi} h_p = \sum_{p \to i} h_p$$

- image hub or image container score is

$$h_p = \sum_i d_{pi} a_i = \sum_{p \to i} a_i$$

Experimentation within the Web track

- it is one of the tracks of the Text Retrieval Conference (TREC)
- based on the test collection paradigm (test documents, test topics, relevance judgements)
- started on 1998
- main aims:
  - evaluate the effectiveness of link analysis-based methods
  - experiment other tasks than ad-hoc retrieval

Tasks

- ad-hoc: given a topic, retrieve relevant documents
- homepage finding: given a query string, find the homepage of the site described by the query
- topic distillation: given a topic, retrieve relevant *and* authority documents
- named page finding: given a query string, find the page described by the query

Test collections

- WT2g: 2GB and 250,000-document collection, used on 1999
- WT10g: 10GB and 1.69 million document collection, used on 2000 and 2001
- .GOV: 18GB and 1.25 million document collection, used on 2002 and 2003
  - less, but larger documents than WT10g
  - access to PDF and images available (67GB, binaries included)

Test topics

- the `title` field includes a real query
- very short queries
- sometimes misspelled queries, e.g. `angioplast7`

Main findings

- link-based methods are not beneficial for the ad-hoc task
  - it is not a necessary evidence
  - conventional yet advanced weighting schemes are necessary
- link-based methods might be useful for the homepage finding task
  - anchor text and URL are more effective
  - page structure is effective as well
- topic distillation and named page finding did not benefit from link structure
  - document structure and anchor text were more effective

Some final remarks

- link analysis-based algorithms for Web retrieval are in principle attractive
- they have shown their effectiveness in some experiments reported by single researchers, but failed within TREC
- one of the reasons is one of the assumptions, i.e. links represent authority assessment
- many links do not and their implementation does not incorporate any information about authority assessment
- automatic detection of link types/classes/labels would be a breakthrough

Some final remarks (cont.)

- these models have been successfully employed to perform other tasks (stemming, image retrieval)
- language can bias authority scores or PageRank values because links are likely to occur among pages written in the same language
- information on time is absent and young pages are less likely pointed to than older ones

Suggested bibliography

- **Markov chain theory**
  - W. Feller. *An introduction to probability theory and its applications*. Wiley, volume 1, 3rd edition, 1968.
- **PageRank**
  - S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the WWW-7 Conference* published in *Computer Networks and ISDN Systems*, volume 30, number 1–7, pages 107–117, 1998.
  - L. Page, S. Brin, R. Motwani and T. Winograd. The PageRank citation ranking: bringing order to the Web. Published at `http://www-db.stanford.edu/~backrub/pageranksub.ps` (Visited on August 2003).
  - T.H. Haveliwala. Topic-sensitive PageRank. *Proceedings of the World Wide Web Conference*, Honolulu, Hawaii, 2002.
- **Hyper-linked Induced Topic Search**
  - J. Kleinberg. Authorative sources in a hyperlinked environment. *Journal of the ACM*, volume 46, no. 5, pages 604–632, 1999.
  - S. Chakrabarti *et al.*. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Proceedings of the WWW-7 Conference* published in *Computer Networks and ISDN Systems*, volume 30, number 1–7, pages 65–74, 1998.
  - R. Lempel and S. Moran. SALSA: the Stochastic Approach for Link-Structure Analysis. *ACM Transactions on Information Systems*, volume 19, no. 2, pages 131-160, 2001.
  - R. Lempel and S. Soffer. PicASHOW: Pictorial authority search by hyperlinks on the Web. *ACM Transactions on Information Systems*, volume 20, no. 1, pages 1-24, 2002.
  - M. Bacchin, N. Ferro and M. Melucci. The Effectiveness of a Graph-based Algorithm for Stemming. *Proceedings of the Int.l Conf. on Asian Digital Libraries*, pages 117-128, Singapore, 2002.

- **Web track**
  - D. Hawking. Overview of the TREC-9 Web track. *Proceedings of TREC-9*, 2001.
  - D. Hawking and N. Craswell. Overview of the TREC-2001 Web track. *Proceedings of TREC-10*, 2002.
  - D. Hawking and N. Craswell. Overview of the TREC-2002 Web track. *Proceedings of TREC-11*, 2003.
  - I. Soboroff. Do TREC Web collections look like the Web? *SIGIR Forum*, volume 36, no 2, 2002.

- **Hypertext and information retrieval**
  - M. Agosti and A. Smeaton (eds). Information retrieval and hypertext. Kluwer Academic Press, 1996.
  - R. Botafogo *et al.*. Structural analysis of hypertext: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, volume 10, no. 2, pages 142-180, 1992.
  - M. Melucci. An evaluation of automatically constructed hypertexts for information retrieval. *Journal of Information Retrieval*, volume 1, numbers 1, pages 57-80, 1999.

- **Other**
  - L. Pretto. *A theoretical analysis to link analysis algorithms*. Ph. D. thesis, Department of Information Engineering, University of Padova, December 2002.
  - N.L. Geller. On the citation influence methodology of Pinski and Narin. *Information Processing and Management*, volume 14, pages 93–95, 1978.
  - P. Ingwersen. The calculation of Web impact factors. *Journal of Documentation*, volume 54, no. 2, pages 236–243, 1998.
  - E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, volume 178, pages 471-479, 1972.
  - L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, volume 18, pages 39-43, 1953.
  - C.H. Hubbell. An input-ouput approach to clique identification. *Sociometry*, volume 28, pages 377-399, 1965.