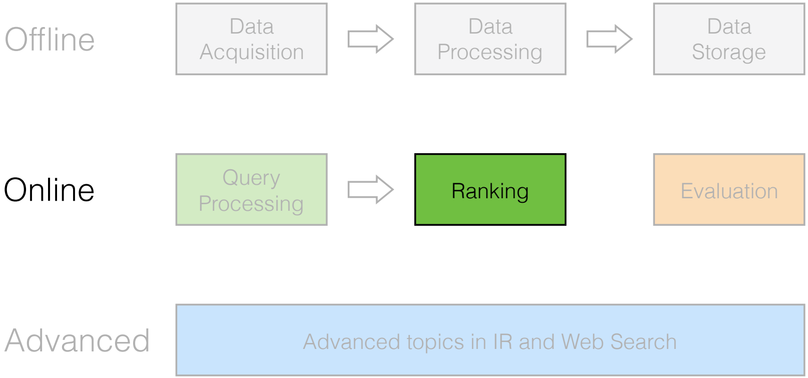
Information Retrieval

Link-based Retrieval

Ilya Markov
i.markov@uva.nl

University of Amsterdam

Ranking methods



Ranking methods

1. Content-based
    - Term-based
    - Semantic
2. **Link-based (web search)**
3. Learning to rank

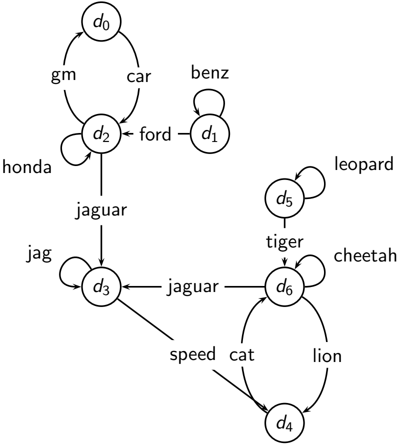
Linear algebra

- C – square $M \times M$ matrix
- $\vec{x}$ – M-dimensional vector
- $C\vec{x} = \lambda\vec{x}$
  - λ – eigenvalue
  - $\vec{x}$ – right eigenvector
- $\vec{y}^T C = \lambda\vec{y}^T$
  - $\vec{y}$ – left eigenvector
- Principal eigenvector – eigenvector corresponding to the largest eigenvalue
- There are many efficient algorithms to compute eigenvalues and eigenvectors

Outline

Outline

Web graph

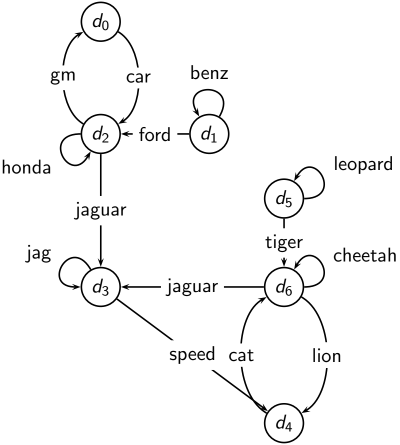


Manning et al., "Introduction to Information Retrieval"

Random walk

1. Start at a random page
2. Follow one of the outgoing links from this page
3. Repeat step 2

$$p(d_i) = \sum_{j:d_j \to d_i} \frac{p(d_j)}{|k : d_j \to d_k|}$$

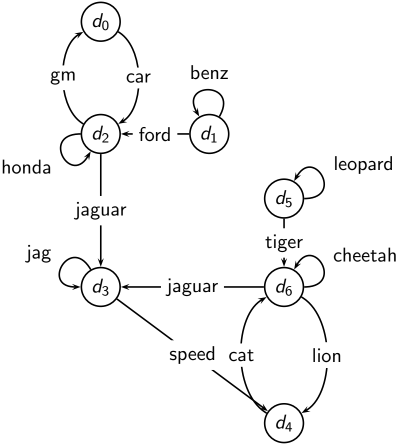


Manning et al., "Introduction to Information Retrieval"

Teleportation

- The surfer always teleports from a dead end to a random page
- At each step of a random walk the surfer teleports to a random page with probability α

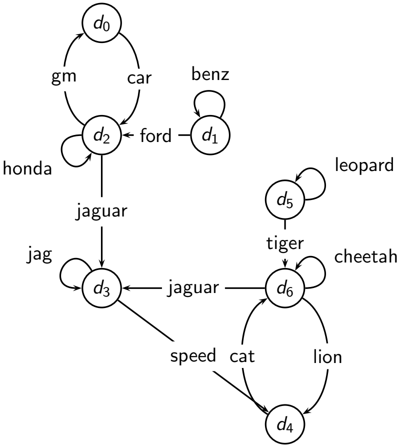
$$p(d_i) = \alpha \frac{1}{N}$$



Manning et al., "Introduction to Information Retrieval"

PageRank

- The more often a page is visited, the better the page
- In the steady state, each page has a long-term visit rate, called PageRank



$$p(d_i) = (1-\alpha) \sum_{j:d_j \rightarrow d_i} \frac{p(d_j)}{|k : d_j \rightarrow d_k|} + \alpha \frac{1}{N}$$

Markov chains

- N states
- P – transition probability matrix with dimensions $N \times N$
- P_{ij} – transition probability from i to j
- $\sum_{j=1}^{N} P_{ij} = 1$ for all i
- At each step, we are in exactly one state

Link matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Manning et al., "Introduction to Information Retrieval"

Transition probability matrix P

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Manning et al., "Introduction to Information Retrieval"

Random walk revisited

- $\vec{x}_t = [p_t(d_1), \ldots, p_t(d_N)]$ – vector of probabilities at time t of a random walk
- $\vec{x}_{t+1} = \vec{x}_t P = x_0 P^{t+1}$

Ergodic Markov chains

- A Markov chain is ergodic iff it is irreducible and aperiodic
  - **Irreducibility.** Roughly: there is a path from any page to any other page
  - **Aperiodicity.** Roughly: the pages cannot be partitioned such that the random walker visits the partitions sequentially
- **Theorem.** For any ergodic Markov chain, there is a unique long-term visit rate for each state
- A random walk with teleportation is an ergodic Markov chain $\Longrightarrow$ there is a unique PageRank value for each page

PageRank revisited

- $\vec{\pi} = [PR(d_1), \ldots, PR(d_N)]$ – vector of stationary probabilities
- $1\vec{\pi} = \vec{\pi}P$
- $\lambda = 1$ – the largest eigenvalue
- $\vec{\pi}$ – principal eigenvector

Computing PageRank using power iteration

- For any initial distribution vector $\vec{x}$
- For large t
- $\vec{x}P^t$ is very similar to $\vec{x}P^{t+1}$
- $\vec{\pi} \approx \vec{x}P^t$

Example

$$P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

$\vec{x}_0$	1	0	0
$\vec{x}_1$	1/6	2/3	1/6
$\vec{x}_2$	1/3	1/3	1/3
$\vec{x}_3$	1/4	1/2	1/4
$\vec{x}_4$	7/24	5/12	7/24
...	...	...	...
$\vec{x}$	5/18	4/9	5/18

Manning et al., "Introduction to Information Retrieval"

PageRank summary

- PageRank is a query-independent indicator of the page quality
- PageRank is a stationary state of a random walk with teleportation
- A random walk with teleportation is an ergodic Markov chain $\Longrightarrow$ there is a unique PageRank value for each page
- PageRank is a principal eigenvector of the transition matrix P $\Longrightarrow$ it can be computed using any algorithm for finding eigenvectors

Outline

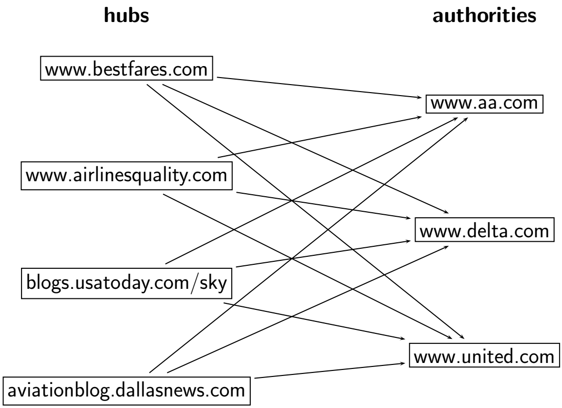
1. PageRank
2. HITS
3. Summary

Intuition

- **Hub** – a page with a good list of links to pages answering the information need
- **Authority** – a page with an answer to the information need

- A good hub for a topic *links to* many authorities for that topic
- A good authority for a topic *is linked to* by many hubs for that topic

Example

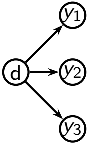


Manning et al., "Introduction to Information Retrieval"

Computing hub and authority scores

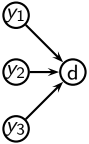
- Hub score

$$h(d) \leftarrow \sum_{y:d \to y} a(y)$$



- Authority score

$$a(d) \leftarrow \sum_{y:y \to d} h(y)$$



Manning et al., "Introduction to Information Retrieval"

Computing hub and authority scores

- A – incidence matrix
- Vectorized form of the hub and authority scores

$$\vec{h} \leftarrow A\vec{a}$$
$$\vec{a} \leftarrow A^T\vec{h}$$

- Can be rewritten as

$$\vec{h} \leftarrow AA^T\vec{h}$$
$$\vec{a} \leftarrow A^TA\vec{a}$$

- $\vec{h}$ and $\vec{a}$ are the eigenvectors of AA^T and A^TA respectively

Hypertext-induced topic search (HITS)

1. Assemble the target query-dependent subset of web pages
2. Form the graph, induced by their hyperlinks
3. Compute AA^T and A^TA
4. Compute the principal eigenvectors of AA^T and A^TA
5. Form the vector of hub scores $\vec{h}$ and authority scores $\vec{a}$
6. Output the top-scoring hubs and the top-scoring authorities

Selecting pages for HITS

1. Do a regular web search
   - The obtained search results form the *root set*
2. Find pages that are linked from or link to pages in the root set
   - These pages form the *base set*
3. Compute hubs and authorities for the base set

HITS summary

- HITS is a query- and link-dependent indicator of the page quality
- Can be computed using any algorithm for finding eigenvectors
- Usually, too expensive to be applied at a query time
- In practice, usually a good hub is also a good authority
- Therefore, the actual difference between PageRank ranking and HITS ranking is not large

Outline

Link-based retrieval summary

- PageRank
  - Query-independent
  - Can be precomputed
- HITS
  - Query-dependent
  - Cannot be precomputed
  - In practice, could be similar to PageRank

Materials

- Manning et al., Chapters 21.2–21.3
- Croft et al., Chapter 4.5

Ranking methods

1. Content-based
   - Term-based
   - Semantic
2. Link-based (web search)
3. **Learning to rank**