Web Page Scoring Systems
for Horizontal and Vertical search

Michelangelo Diligenti, Marco Gori, Marco Maggini
{diligmic, marco, maggini}@dii.unisi.it

11-th World Wide Web Conference

5/9/2002
Introduction: Web surfing

- The goal of the paper is to model an user visiting the Web.

- The probability that the user is visiting a page, is proportional to the relevance of that page.
Summary

- Definition of the probabilistic model.
- Deriving Google’s PageRank and HITS from the model.
- Proposal of new models for vertical search engines.
- Experimental results.
Surfer model 1

Our surfer is allowed to perform the following basic operations:

- $j$ jump to a node of the graph;
- $l$ follow a hyperlink from the current page;
- $b$ follow a back-link (a hyperlink in the inverse direction);
- $s$ stay in the same node.
Surfer model 2

Surfer actions depend on the content of current page:

- \( x(l|q) \) the probability of following one hyperlink from page \( q \),
- \( x(b|q) \) the probability of following one back-link from page \( q \),
- \( x(j|q) \) the probability of jumping from page \( q \),
- \( x(s|q) \) the probability of remaining in page \( q \).
Surfer model 3

- $x(p|q, j)$ the probability of jumping from page $q$ to page $p$;

- $x(p|q, l)$ the probability of selecting a hyperlink from page $q$ to page $p$; $x(p|q, l) \neq 0 \iff p \in ch(q)$, being $ch(q)$ the set of the children of node $q$ in the graph $G$;

- $x(p|q, b)$ the probability of going back from page $q$ to page $p$; $x(p|q, b) \neq 0 \iff p \in pa(q)$, being $pa(q)$ the set of the parents of node $q$ in the graph $G$. 
Surfer model 4

The probability of being located at page \( p \) at time step \( t + 1 \) is

\[
x_p(t + 1) = \sum_{q \in G} x(p|q,j) \cdot x(j|q) \cdot x_q(t) + \\
+ \sum_{q \in pa(p)} x(p|q,l) \cdot x(l|q) \cdot x_q(t) + \\
+ \sum_{q \in ch(p)} x(p|q,b) \cdot x(b|q) \cdot x_q(t) + x(s|p) \cdot x_p(t)
\]

\( x(t) \) score vector at time \( t \). Starting from a given initial distribution \( x(0) \):

\[
x(t) = T^t \cdot x(0).
\]
Surfer model and Markov chains

Proposition 1

$T'$ is the state transition matrix of the Markov chain. $T'$ is stable, since $T'$ is a stochastic matrix having ($\lambda_{max} = 1$). If $\sum_{q \in G} x_q(0) = 1$, then $\sum_{q \in G} x_q(t) = 1$, $t = 1, 2, \ldots$.

By applying the results on Markov chains we can prove that:

Proposition 2

If $x(j|q) \neq 0 \land x(p|q,j) \neq 0$, $\forall p, q \in G$ then 1) $\lim_{t \to \infty} x(t) = x^*$ where $x^*$ does not depend on the initial state vector $x(0)$. 2) All pages get a score $\neq 0$, thus the resulting scoring system can be applied globally to the entire Web.
Google’s PageRank

- $x(b|p) = x(s|p) = 0$ for any page $p$.
- $x(j|p) = 1 - d$, $x(l|p) = d$ for any page $p$.
- $x(p|j) = 1/N$ for any page $p$, where $N$ is the number of pages on the Web Graph.
- $x(p|q, l) = 1/h_q$ where $h_q$ is the number of outlinks of page $q$.

For Proposition 2 PageRank converges to a vector independent from the starting distribution.
Note: setting $x(j|p) = 1$ and $x(l|p) = 0$ for any sink page $p$, the resulting model is still probabilistically coherent.
Focused Google’s PageRank

- PageRank: the random surfer follows each outlink of page $q$ with probability $1/\text{ch}(q)$;

- Focused PageRank (Domingos 2001): a surfer follows the links according to suggestions provided by a page classifier.

$$x(ch_i(q)|q,l) = \frac{s(ch_i(q))}{\sum_{j=0}^{h_q} s(ch_j(q))}$$
Double Focused Google’s PageRank

Surfer actions depend on content of current page:

- probability of following a link in page $p$ is proportional to classification score $s(p)$ of $p$

$$x(l|p) = d_1 \cdot \frac{s(p)}{\max_{q \in \mathcal{G}} s(q)}$$

- probability of jump to $p$ is proportional to $s(p)$

$$x(p|j) = \frac{s(p)}{\sum_{q \in \mathcal{G}} s(q)}$$

For Proposition 2 the resulting scoring system is stable and converges to a distribution independent from the initial conditions. All pages get a non-zero score (allowing global ranking).
Collaborative walks (Multi State models) 1

- A model based on a single variable may not capture relationships among pages (i.e. HITS scheme uses 2 variables).

- We define a multi-variable scheme by considering a pool of surfers each associated to a variable. A surfer can accept suggestions of surfer $i$, jumping to the page visited by $i$. 

Marco often discovers good links, I want to move to the page where he is located.
Collaborative walks (Multi State models) 2

The set of $M$ interacting surfers can be described as a set of matrix equations as follows

$$
\begin{align*}
    x^{(1)}(t + 1) &= T^{(1)} \cdot X(t) \cdot A^{(1)} \\
    \vdots \\
    x^{(M)}(t + 1) &= T^{(M)} \cdot X(t) \cdot A^{(M)}
\end{align*}
$$

where the j-th element of vector $A^{(i)}$ indicates the probability that surfer $i$ will relocate to the actual position of surfer $j$. 
The HITS algorithm assigns an *authority* and *hubness* score to each page $p$. It is modeled by a collaborative walk of 2 surfers:

- Surfer 1 associated to the page hubness.
- Surfer 2 associated to the page authority.

- $x^{(1)}(l|p) = 0$, $x^{(1)}(b|p) = 1$ for each page $p$.
- $x^{(2)}(l|p) = 1$, $x^{(2)}(b|p) = 0$ for each page $p$. 
Hubs/Authority

- $x^{(1)}(p|q, b) = 1$ for each page $q$ and $p \in pa(q)$.
- $x^{(2)}(p|q, l) = 1$ for each page $q$ and $p \in ch(q)$.

Surfer interaction: $A^{(1)} = (0, 1)'$, $A^{(2)} = (1, 0)'$

HITS does not respect the probabilistic model:

$$\sum_{p \in ch(q)} x^{(1)}(p|q, b) = |ch(q)| > 1$$
$$\sum_{p \in pa(q)} x^{(2)}(p|q, l) = |pa(q)| > 1$$

HITS can be modified to respect the probabilistic model and the conditions stated on Proposition 2 (more details on the paper).
Experimental results

2 focus crawling sessions for the topic “Linux” (50,000 pages) and “cooking recipes” (300,000 pages). We report the rank values of pages (sorted by the rank value).
<table>
<thead>
<tr>
<th>PageRank</th>
<th>HITS</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.zdnet.com">www.zdnet.com</a></td>
<td><a href="http://www.openbsdapps.com/?page=category&amp;">www.openbsdapps.com/?page=category&amp;</a>...</td>
</tr>
<tr>
<td><a href="http://www.google.com">www.google.com</a></td>
<td><a href="http://www.openbsdapps.com/?page=category&amp;">www.openbsdapps.com/?page=category&amp;</a>...</td>
</tr>
<tr>
<td>search.internet.com/power_search</td>
<td><a href="http://www.openbsdapps.com/?page=category&amp;">www.openbsdapps.com/?page=category&amp;</a>...</td>
</tr>
<tr>
<td><a href="http://www.yahoo.com">www.yahoo.com</a></td>
<td><a href="http://www.openbsdapps.com/?page=category&amp;">www.openbsdapps.com/?page=category&amp;</a>...</td>
</tr>
<tr>
<td>java.sun.com</td>
<td><a href="http://www.openbsdapps.com/?page=newupdate">www.openbsdapps.com/?page=newupdate</a>...</td>
</tr>
<tr>
<td><a href="http://www.osdn.com">www.osdn.com</a></td>
<td><a href="http://www.openbsdapps.com/?page=linkus">www.openbsdapps.com/?page=linkus</a></td>
</tr>
</tbody>
</table>

8 top “Linux” score pages, using either the PageRank surfer, or a HITS surfer pool (considering the authority value).
Qualitative results 2

<table>
<thead>
<tr>
<th>Focused PageRank</th>
<th>Double Focused PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.slackware.com">www.slackware.com</a></td>
<td><a href="http://www.slackware.com">www.slackware.com</a></td>
</tr>
<tr>
<td><a href="http://www.linux.org">www.linux.org</a></td>
<td><a href="http://www.li.org">www.li.org</a></td>
</tr>
<tr>
<td><a href="http://www.zdnet.com">www.zdnet.com</a></td>
<td><a href="http://www.linux.org">www.linux.org</a></td>
</tr>
<tr>
<td>jobs.osdn.com</td>
<td><a href="http://www.linuxhq.com">www.linuxhq.com</a></td>
</tr>
<tr>
<td><a href="http://www.yahoo.com">www.yahoo.com</a></td>
<td><a href="http://www.slackware.org">www.slackware.org</a></td>
</tr>
<tr>
<td><a href="http://www.python.org">www.python.org</a></td>
<td><a href="http://www.linuxusers.org">www.linuxusers.org</a></td>
</tr>
</tbody>
</table>

8 top “Linux” score pages, using the proposed focused versions of the PageRank surfer.
Expert judgments

(a) Percentage of authoritative pages among the $N$ pages with highest score. 10 experts labelled the pages as “authoritative” or not “authoritative” for the topic.
Conclusions

- We defined a probabilistic model from which many popular scoring algorithms can be derived.

- Properties of a scoring system based on our model:
  1. stable (at each iteration the sum of scores is equal to 1);
  2. converges to a solution independent from initial condition;
  3. non-zero score to each page (allowing global ranking).

- We proposed new scoring algorithms for vertical and horizontal search. Experts judgments confirm that proposed algorithms provide better results than other scoring systems.