Information Retrieval – Part 2

Deepak Kumar
Borges’ Library of Babel

“...each book contains four hundred ten pages; each page, forty lines; each line, approximately eighty black letters. There are also letters on the front cover of each book; these letters neither indicate nor prefigure what the pages inside will say.”

Q: How many books are in the library?
Q. How would you find what you’re looking for?
Elements of a Search Engine
Web Information Retrieval

• Search Engines
• Queries
  phrase queries
  structure queries (NEAR, intitle:, ...)
• Matching
• Inverted Index
  page number
  location
• Ranking & Relevance
• Metadata
Web Information Retrieval

• Search Engines
• Queries
  phrase queries
  structure queries
• Matching
• Inverted Index
  page number
  location
• Ranking & Relevance
• Metadata

Efficient matching is only one half the story.

The other grand challenge is how to **rank** the matching pages
Matching & Ranking

query

muddy waters

matching

matched pages

"hits"

ranking

ranked pages

1.
2.
3.
By far the most common cause of malaria is being bitten by an infected mosquito, but there are also other ways to contract the disease.

Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.

Nearness can resolve the ranking!
<table>
<thead>
<tr>
<th>Title</th>
<th>Text</th>
<th>Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>my cat</td>
<td>&lt;title&gt;my cat &lt;body&gt; the cat sat on the mat &lt;/body&gt;</td>
<td>a 3-10, cat 1-3 1-7 3-7, dog 2-3 2-7 3-11, mat 1-11 2-11, my 1-2 2-2 3-2, on 1-9 2-9, pets 3-3, sat 1-8 3-12, stood 2-8 3-8, the 1-6 1-10 2-6 2-10 3-6, while 3-9</td>
</tr>
<tr>
<td>my dog</td>
<td>&lt;title&gt;my dog &lt;body&gt; the dog stood on the mat &lt;/body&gt;</td>
<td></td>
</tr>
<tr>
<td>my pets</td>
<td>&lt;title&gt;my pets &lt;body&gt; the cat stood while a dog sat &lt;/body&gt;</td>
<td>&lt;body&gt; 1-5 2-5 3-5, &lt;/body&gt; 1-12 2-12 3-13, &lt;title&gt; 1-1 2-1 3-1, &lt;/title&gt; 1-4 2-4 3-4</td>
</tr>
</tbody>
</table>
Structure Queries

query
intitle: dog

<title>my dog
</title><body>
the dog stood on
the mat</body>
Exploiting Link Structure

• **PageRank** exploits the structure of the web:

  Use of Hyperlinks to
  - count # of incoming links
  - Identifying web authority

• Use the above in determining ranking & relevance.
The Garage

Garage at 232 Santa Margarita, Menlo Park, CA
Google 1.0 (1998)

2-proc Pentium II 300mhz, 512mb, five 9gb drives
2-proc Pentium II 300mhz, 512mb, four 9gb drives
4-proc PPC 604 333mhz, 512mb, eight 9gb drives
2-proc UltraSparc II 200mhz, 256mb, three 9gb drives, six 4gb drives
Disk expansion, eight 9gb drives
Disk expansion, ten 9gb drives

That's a total of:
**1792 megabytes** of memory
**366 gigabytes** of disk storage
**2933 megahertz** in **10** CPUs
The Disk Storage
Google 1.0 (1998)
**Ernie’s Scrambled Eggs Recipe**
Mix four eggs in a bowl with a little salt and pepper, ...

**Bert’s Scrambled Eggs Recipe**
First melt a tablespoon of butter, ...

Ernie’s recipe is good.

I really enjoyed Bert’s recipe.

Bert’s recipe is wonderful!

Bert’s recipe is fantastic!
Ernie’s Scrambled Eggs Recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Bert’s Scrambled Eggs Recipe
First melt a tablespoon of butter, ...

Ernie’s recipe is good.

I really enjoyed Bert’s recipe.

Bert’s recipe is wonderful!

Bert’s recipe is fantastic!

-ranked higher based on #incoming links-
Ernie’s Scrambled Eggs Recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Bert’s Scrambled Eggs Recipe
First melt a tablespoon of butter, ...

Ernie’s recipe is good.

I did not enjoy Bert’s recipe.

Bert’s recipe did not work.

Bert’s recipe is unhealthy!
Hyperlinks: Authority

Ernie’s Scrambled Eggs Recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Bert’s Scrambled Eggs Recipe
First melt a tablespoon of butter, ...

Deepak Kumar’s Home Page
I tried Ernie’s recipe once, and it’s not bad at all.

Alton Brown’s Home Page
Bert’s recipe is clearly one of the best.
Hyperlinks: Authority

Ernie’s Scrambled Eggs Recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Deepak Kumar’s Home Page
I tried Ernie’s recipe once, and it’s not bad at all.

Bert’s Scrambled Eggs Recipe
First melt a tablespoon of butter, ...

Alton Brown’s Home Page
Bert’s recipe is clearly one of the best.
Cycles
Computing Authority Scores

Diagram of authority scores:
- A points to B.
- C points to A.
- D points to A.
- E points to A.

Scores:
- A: 1
- B: 0
- C: 1
- D: 1
- E: 0
Computing Authority Scores
Computing Authority Scores

A 2
C 1
D 1

B 2
E

Diagram shows nodes A, B, C, D, and E connected with arrows indicating the direction of authority scores.
Computing Authority Scores

A → B
C → A
C → D
D → A
E → B
Computing Authority Scores

score for A is out of date
Computing Authority Scores

![Diagram showing the computing authority scores with nodes representing A, B, C, D, and E, and scores denoted as 1, 2, 4, and 2.]
Computing Authority Scores

A: 4
B: 4
C: 1
D: 1
E: 2
Computing Authority Scores
Computing Authority Scores

and so on...stuck in an infinite loop....
Sinks

dangling node
Sinks
The Random Surfer

restart probability = 15%
The Random Surfer

After 1000 page visits
The Random Surfer

after 1 million page visits
The Random Surfer

pages with many incoming links get high ranking
The Random Surfer

A

B

C

D

7% 10% 5% 2%

4% 4% 4% 2%

4% 5% 4% 2%

10% 2% 2% 2%

13% 14% 15%

authoritative score
The Random Surfer

#links+authoritative score
The Random Surfer

Ernie’s Scrambled Eggs Recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Bert’s Scrambled Eggs Recipe
First melt a tablespoon of butter, ...

Deepak Kumar’s Home Page
I tried Ernie’s recipe once, and its not bad at all.

Alton Brown’s Home Page
Bert’s recipe is clearly one of the best.
The Random Surfer

A (33%) → B (31%)

C (3%) → D (3%) → E (30%)
Formalizing PageRank

• Given a web page, $P_i$
• Set of pages pointing into $P_i$, $B_{P_i}$
• Number of outgoing links from page $P_j$, $|P_j|$
• PageRank of a page, $r(P_i)$

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$
Computing PageRank

• \( r(P_1) = r(P_3) \)

• But, \( r(P_3) \) is unknown

• To start, assume all pages have rank \( \frac{1}{n} \) (\( n = 6 \))

• \( \therefore r(P_1) = \frac{1}{6} \)
Computing PageRank

\[ r_0(P_1) = \frac{1}{6} \]
\[ r_0(P_2) = \frac{1}{6} \]
\[ r_0(P_3) = \frac{1}{6} \]
\[ r_0(P_4) = \frac{1}{6} \]
\[ r_0(P_5) = \frac{1}{6} \]
\[ r_0(P_6) = \frac{1}{6} \]

\[ r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \]
Computing PageRank

\[ r_1(P_1) = \frac{1}{18} \]
\[ r_1(P_2) = \frac{5}{36} \]
\[ r_1(P_3) = \frac{1}{12} \]
\[ r_1(P_4) = \frac{1}{4} \]
\[ r_1(P_5) = \frac{5}{36} \]
\[ r_1(P_6) = \frac{1}{6} \]

\[ r_{k+1}(P_i) = \sum_{P_j \in \mathcal{B}_P_i} \frac{r_k(P_j)}{|P_j|} \]
Computing PageRank

\[ r_2(P_1) = 1/36 \]
\[ r_2(P_2) = 1/18 \]
\[ r_2(P_3) = 1/36 \]
\[ r_2(P_4) = 17/72 \]
\[ r_2(P_5) = 11/72 \]
\[ r_2(P_6) = 14/72 \]

\[ r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \]
Computing PageRank

\[ r_2(P_1) = \frac{1}{36} \]
\[ r_2(P_2) = \frac{1}{18} \]
\[ r_2(P_3) = \frac{1}{36} \]
\[ r_2(P_4) = \frac{17}{72} \]
\[ r_2(P_5) = \frac{11}{72} \]
\[ r_2(P_6) = \frac{14}{72} \]

\[ r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|} \]
Matrix Representation

- Adjacency Matrix

\[
A = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 1 & 0 & 0 & 1 & 0 \\
4 & 0 & 0 & 0 & 0 & 1 & 1 \\
5 & 0 & 0 & 0 & 1 & 0 & 1 \\
6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{bmatrix}
\]
Matrix Representation

- Hyperlink Matrix, $H$

$$
\begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{bmatrix}
$$

- $\pi_{k+1}^T = \pi_k^T H$

where $\pi_k^T$ is the $k^{th}$ PageRank vector
The PageRank Equation

- \( \pi^T = \pi^T (\alpha S + (1 - \alpha)E) \)

where
- \( S \) is the stochastic \( H \) matrix
- \( E \) is the teleportation matrix
- \( \alpha \) is the scaling parameter

- Certain stochastic conditions apply!
Google Data Center
References

