How Web Search Works

Dr. Ray Klump
Professor and Chair
Computer & Mathematical Sciences
Tuesday, September 8, 2015

CaMS Seminar Series - Fall 2015
Computer science is not just hardware and software.
Computer Science is best when it deals in ideas.
Great algorithms come from great ideas
Web Search
Web search is uses **indexes**.
History

Lycos & Infoseek (1994)
Altavista (1996)
Google (1998)
Steps

Indexing
Matching
 Ranking
Note: illustrations come from *Nine Algorithms that Changed the Future* (MacCormick)
An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.

<table>
<thead>
<tr>
<th>Word</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>cat</td>
<td>1 3</td>
</tr>
<tr>
<td>dog</td>
<td>2 3</td>
</tr>
<tr>
<td>mat</td>
<td>1 2</td>
</tr>
<tr>
<td>on</td>
<td>1 2</td>
</tr>
<tr>
<td>sat</td>
<td>1 3</td>
</tr>
<tr>
<td>stood</td>
<td>2 3</td>
</tr>
<tr>
<td>the</td>
<td>1 2 3</td>
</tr>
<tr>
<td>while</td>
<td>3</td>
</tr>
</tbody>
</table>
Matching

Search the “web” for the word “cat”

An imaginary World Wide Web that consists of only three pages, numbered 1, 2, and 3.
Phrases pose a challenge

Search the “web” for the phrase “cat sat”
Index words & locations

1. the cat sat on the mat
2. the dog stood on the mat
3. the cat stood while a dog sat

Index:

- a: 3-5
- cat: 1-2 3-2
- dog: 2-2 3-6
- mat: 1-6 2-6
- on: 1-4 2-4
- sat: 1-3 3-7
- stood: 2-3 3-3
- the: 1-1 1-5 2-1 2-5 3-1
- while: 3-4
Find “cat sat”
Near-To Queries

Find pages where cat & dog are within 5 words of each other

1. the cat sat on
   1 2 3 4
   the mat
   5 6

2. the dog stood
   1 2 3
   on the mat
   4 5 6

3. the cat stood
   1 2 3
   while a dog sat
   4 5 6 7

<table>
<thead>
<tr>
<th>a</th>
<th>3-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>1-2 3-2</td>
</tr>
<tr>
<td>dog</td>
<td>2-2 3-6</td>
</tr>
<tr>
<td>mat</td>
<td>1-6 2-6</td>
</tr>
<tr>
<td>on</td>
<td>1-4 2-4</td>
</tr>
<tr>
<td>sat</td>
<td>1-3 3-7</td>
</tr>
<tr>
<td>stood</td>
<td>2-3 3-3</td>
</tr>
<tr>
<td>the</td>
<td>1-1 1-5 2-1 2-5 3-1</td>
</tr>
<tr>
<td>while</td>
<td>3-4</td>
</tr>
</tbody>
</table>
Near-to queries help discern relevance.
Which page discusses the causes of malaria?

1. 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.

2. Our cause was not helped by the poor health of the troops, many of whom were suffering from malaria and other tropical diseases.
Word-location indexing maps **content** to **structure**
Web pages have parts

<html>
<head>
<link rel="icon" type="image/
<title>Not a Meta Tag, but required anyway </title>
<meta name="description" content="Awesome Description Here">
<meta http-equiv="content-type"
content="text/html; charset=UTF-8">
</head>

<body>

</body>
</html>
Searching with **structure**

1. `<titleStart> my cat <titleEnd> <bodyStart> the cat sat on the mat <bodyEnd>`
2. `<titleStart> my dog <titleEnd> <bodyStart> the dog stood on the mat <bodyEnd>`
3. `<titleStart> my pets <titleEnd> <bodyStart> the cat stood while a dog sat <bodyEnd>`
Find “dog” in the title

dog: 2-3 2-7 3-11
<titleStart>: 1-1 2-1 3-1
<titleEnd>: 1-4 2-4 3-4
But there’s more than *indexing* and *matching*.
PageRank

(Google’s MoneyMaker)
Strategies

Hyperlink Count

Weighted Hyperlink Count

Weighted Hyperlink Count with Randomness
Hyperlink Count

Ernie’s scrambled egg recipe
Mix four eggs in a bowl with a little salt and pepper, ...

Ernie’s recipe is good.

Bert’s scrambled egg recipe
First melt a tablespoon of butter, ...

I really enjoyed Bert’s recipe.

Bert’s recipe is wonderful!

Bert’s recipe is fantastic!
Weighted Hyperlink Count

- **Ernie's scrambled egg recipe**
  - Mix four eggs in a bowl with a little salt and pepper, ...
  - Weight: 2

- **Bert's scrambled egg recipe**
  - First melt a tablespoon of butter, ...
  - Weight: 100

- **John MacCormick's home page**
  - I tried Ernie's recipe once, and it's not bad at all.
  - Weight: 2

- **Alice Waters's home page**
  - Bert's recipe is clearly one of the best.
  - Weight: 100

- Multiple links to other pages, weighted with a total of 100 pages.
Cycles pose problems
Derive authority score using randomness
Weighted Hyperlink Count with Randomness
Pages can then be ranked even when there are cycles.
PageRank today includes up to 200 different factors.
Search =
Indexing + Matching + Ranking
Thank you.