Information Retrieval

Deepak Kumar
Information Retrieval

Searching within a document collection for a particular information need.
Search Engines...

Altavista  Ask  Baidu  Bing  Blekko  ChaCha  Dogpile  Daum  DuckDuckGo
Entireweb  Excite  Faroo  Info.com  Gigablast  Google  Go  Hakia  HotBot
Leapfish  Lycos  Monster Crawler  Naver  Omgili  Dmoz  Scrub The Web  Spezify  Stinky Teddy
Stumpdedia  Teoma  WebCrawler  Yahoo! Search  Yandex
Search Engine Market Share
March 2011

- Google: 65.7%
- Yahoo: 16.1%
- Bing: 13.6%
- Ask: 3.2%
- AOL: 1.7%
Matching & Ranking

query: muddy waters

matched pages

“hits”

ranked pages:
1.
2.
3.

matching

ranking
Index

Gregorian Calendar 242
Grey Poupon 38, 94

H
Hallway Cruiser 120
Hektor robot 261
Hertz (Hz) 169
hi-fidelity 170
HiLo game 154
Hoare, C. A. R. 227
Hogg, David 127
Hugs & Kisses 210
Human-robot interaction 262-63

I
iCat robot 263
IDLE 8, 22, 23, 29, 38
Idle, Eric 23
if-statement 100, 103, 118, 128, 270-71
image 182
Image 168, 176, 280
image processing 190
image understanding 195
Imitation Game 206
import 138-39, 275
in 92, 103, 270
Indecisive 117
init 13, 275

Jones, Crispin 259
Jones, Mick 107
JPEG 183
Julia Sets 178

K
Kasparov, Gary 209
Kismet robot 263
Kitaoka, Akiyoshi 181
Koch Snowflakes 178
Konane 209

L
Ladybug 107
Larson, Doug 227
LavenderBlush 159
Law, Jude 205
Leap Frog 260
leap year 241-243
LED 73
LEGO Mindstorms 4
len 92, 104, 270
Lenhi, Jurg 261
Light following 121-22
Line 160, 176, 280
linear time algorithms 252
List comprehensions 214, 224
lists 49, 91-93, 270
Inverted Index

• A mapping from content (words) to location.

• Example:

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

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

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

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

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

Every word in every web page is indexed!
Searching

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

query

<table>
<thead>
<tr>
<th>word</th>
<th>doc 1</th>
<th>doc 2</th>
<th>doc 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>1 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mat</td>
<td>1 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td>1 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stood</td>
<td>2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1 2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>while</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Searching

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

query

cat
Searching

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

query

cat
Searching

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

Query: dog

Hits:

- the dog stood on the mat
- the cat stood while a dog sat
Searching

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

query

cat dog
Searching

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

Query: cat dog

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

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

query

cat dog

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

hits

the cat stood while a dog sat
Searching

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

query

cat the sat

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

???
Phrase Queries

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

query
“cat sat”

hits

the cat sat on the mat
the cat stood while a dog sat
Phrase Queries

How to tell if two words occur next to each other?
Phrase Queries

How to tell if two words occur next to each other? EFFICIENTLY???
Inverted Index with Location

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

- 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
Inverted Index with Location

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

query
“cat sat”

<table>
<thead>
<tr>
<th>Word</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>1-2</td>
<td>3-2</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>2-2</td>
<td>3-6</td>
<td></td>
</tr>
<tr>
<td>mat</td>
<td>1-6</td>
<td>2-6</td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>1-4</td>
<td>2-4</td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td>1-3</td>
<td>3-7</td>
<td></td>
</tr>
<tr>
<td>stood</td>
<td>2-3</td>
<td>3-3</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>1-1</td>
<td>1-5</td>
<td>2-1</td>
</tr>
<tr>
<td>while</td>
<td></td>
<td>2-5</td>
<td>3-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3-4</td>
</tr>
</tbody>
</table>
**Inverted Index with Location**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 the cat sat on the mat</td>
<td>1-2 3-2</td>
</tr>
<tr>
<td>2 the dog stood on the mat</td>
<td>1-4 2-4</td>
</tr>
<tr>
<td>3 the cat stood while a dog sat</td>
<td>1-1 1-5 2-1 2-5 3-1</td>
</tr>
</tbody>
</table>

**Query: “cat sat”**

- **cat**: 1-2 3-2
- **sat**: 1-3 3-7
- **the**: 1-1 1-5 2-1 2-5 3-1
- **on**: 1-4 2-4
- ** stood**: 2-3 3-3
- **a**: 3-5
- **dog**: 2-2 3-6
- **mat**: 1-6 2-6
Inverted Index with Location

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

query
“cat sat”

query results:
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
**Inverted Index with Location**

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

**Query:**  "cat sat"

<table>
<thead>
<tr>
<th>Term</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3-5</td>
</tr>
<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>

**Hits:**

1. the cat sat on the mat
NEAR* Queries

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

query: cat NEAR dog

*NEAR: distance <= 5
NEAR* Queries

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

query

cat NEAR dog

hits

the cat stood while a dog sat

Useful in ranking!

*NEAR: distance <= 5
Matching & Ranking

query
muddy waters

matched pages

"hits"

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.
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.
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.
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.

query
malaria cause

also 1-19
...
cause 1-6 2-2
...
malaria 1-8 2-19
...
whom 2-15

Nearness can resolve the ranking!
Using Metadata
Using Metadata

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>CS380: Science of Information (Course Page)</title>
</head>
<body>
P>
<CENTER>
<h3>Bryn Mawr College<br>Collections<br>CS 380: Recent Advances in Computer Science<br>Topic: Science of Information<br>Fall 2012<br>BMC Class Number: 1214<br>Course Materials</h3>
</CENTER>
...
my cat
the cat sat on the mat

my dog
the dog stood on the mat

my pets
the cat stood while a dog sat
my cat
the cat sat on the mat

my dog
the dog stood on the mat

my pets
the cat stood while a dog sat

<title>my cat</title> <body>the cat sat on the mat</body>

<title>my dog</title> <body>the dog stood on the mat</body>

<title>my pets</title> <body>the cat stood while a dog sat</body>
<title>my dog</title><body>the dog stood on the mat</body>

<title>my pets</title><body>the cat stood while a dog sat</body>
Structure Queries

query

intitle: dog

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
<body> 1-5 2-5 3-5
</body> 1-12 2-12 3-13
<title> 1-1 2-1 3-1
</title> 1-4 2-4 3-4
Structure Queries

intitle: dog

query

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
<body> 1-5 2-5 3-5
</body> 1-12 2-12 3-13
<title> 1-1 2-1 3-1
</title> 1-4 2-4 3-4
Structure Queries

<table>
<thead>
<tr>
<th>Word</th>
<th>Query Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3-10</td>
</tr>
<tr>
<td>cat</td>
<td>1-3 1-7 3-7</td>
</tr>
<tr>
<td>dog</td>
<td>2-3 2-7 3-11</td>
</tr>
<tr>
<td>mat</td>
<td>1-11 2-11</td>
</tr>
<tr>
<td>my</td>
<td>1-2 2-2 3-2</td>
</tr>
<tr>
<td>on</td>
<td>1-9 2-9</td>
</tr>
<tr>
<td>pets</td>
<td>3-3</td>
</tr>
<tr>
<td>sat</td>
<td>1-8 3-12</td>
</tr>
<tr>
<td>stood</td>
<td>2-8 3-8</td>
</tr>
<tr>
<td>the</td>
<td>1-6 1-10 2-6 2-10 3-6</td>
</tr>
<tr>
<td>while</td>
<td>3-9</td>
</tr>
<tr>
<td>&lt;body&gt;</td>
<td>1-5 2-5 3-5</td>
</tr>
<tr>
<td>&lt;/body&gt;</td>
<td>1-12 2-12 3-13</td>
</tr>
<tr>
<td>&lt;title&gt;</td>
<td>1-1 2-1 3-1</td>
</tr>
<tr>
<td>&lt;/title&gt;</td>
<td>1-4 2-4 3-4</td>
</tr>
</tbody>
</table>
Structure Queries

query

intitle: dog

- 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
- <body>: 1-5 2-5 3-5
- </body>: 1-12 2-12 3-13
- <title>: 1-1 2-1 3-1
- </title>: 1-4 2-4 3-4
Structure Queries

intitle: dog

query

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
<body> 1-5 2-5 3-5
</body> 1-12 2-12 3-13
<title> 1-1 2-1 3-1
</title> 1-4 2-4 3-4
Structure Queries

**query**

```
intitle: dog
```

```
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
<body> 1-5 2-5 3-5
</body> 1-12 2-12 3-13
<title> 1-1 2-1 3-1
</title> 1-4 2-4 3-4
```
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
References


• *Learning Computing with Robots*, Deepak Kumar, IPRE 2011.