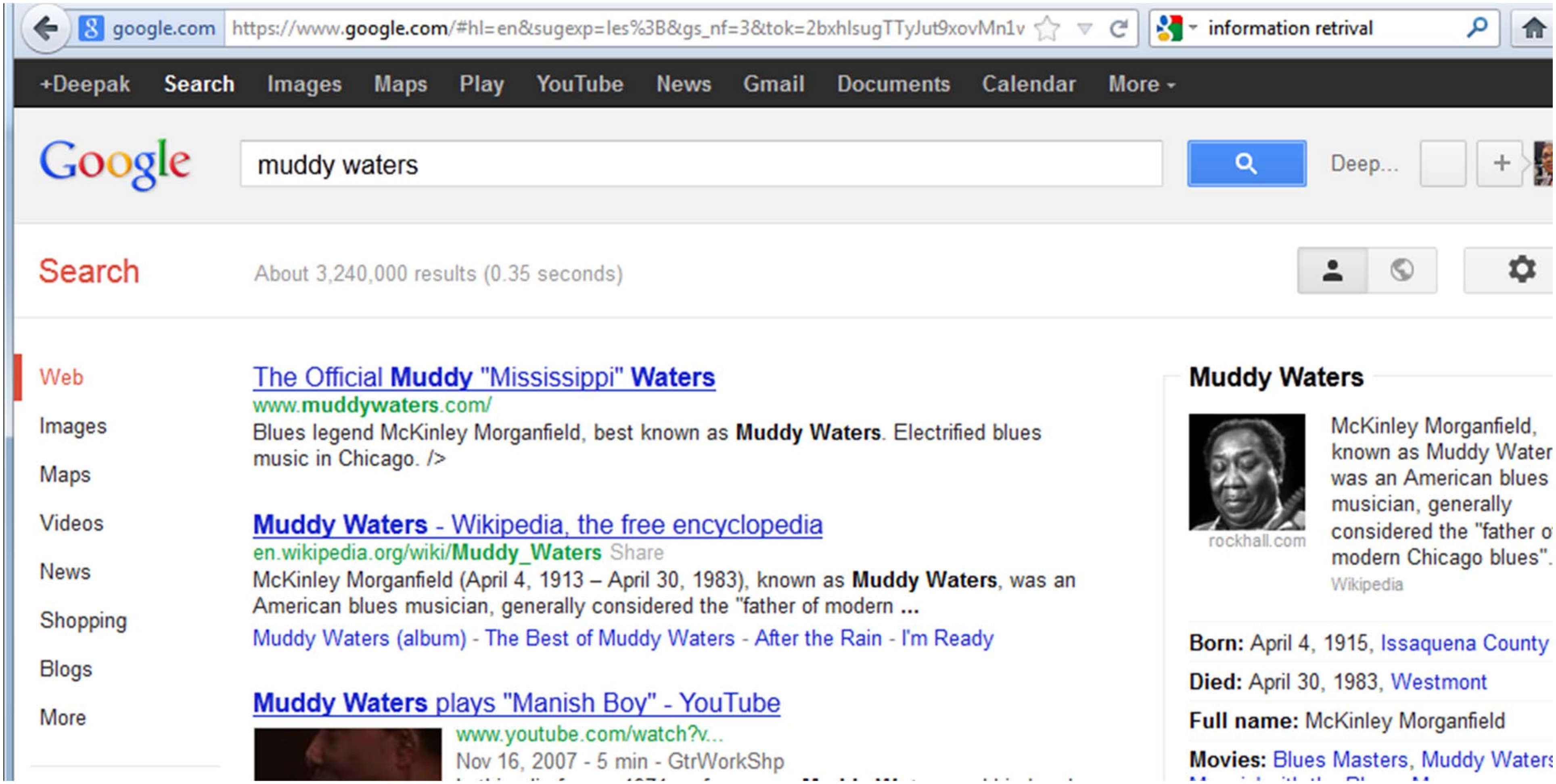
Information Retrieval

Deepak Kumar

Information Retrieval

Searching within a document collection for a particular information need.

Query



About 3,240,000 results (0.35 seconds)

The Official Muddy "Mississippi" Waters
www.**muddywaters**.com/
Blues legend McKinley Morganfield, best known as **Muddy Waters**. Electrified blues music in Chicago. />

Muddy Waters - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Muddy_Waters** Share
McKinley Morganfield (April 4, 1913 – April 30, 1983), known as **Muddy Waters**, was an American blues musician, generally considered the "father of modern **...**
Muddy Waters (album) - The Best of Muddy Waters - After the Rain - I'm Ready

Muddy Waters plays "Manish Boy" - YouTube
www.youtube.com/watch?v...
Nov 16, 2007 - 5 min - GtrWorkShp

Muddy Waters

McKinley Morganfield, known as Muddy Water was an American blues musician, generally considered the "father o modern Chicago blues".
Wikipedia

rockhall.com

Born: April 4, 1915, Issaquena County
Died: April 30, 1983, Westmont
Full name: McKinley Morganfield
Movies: Blues Masters, Muddy Waters

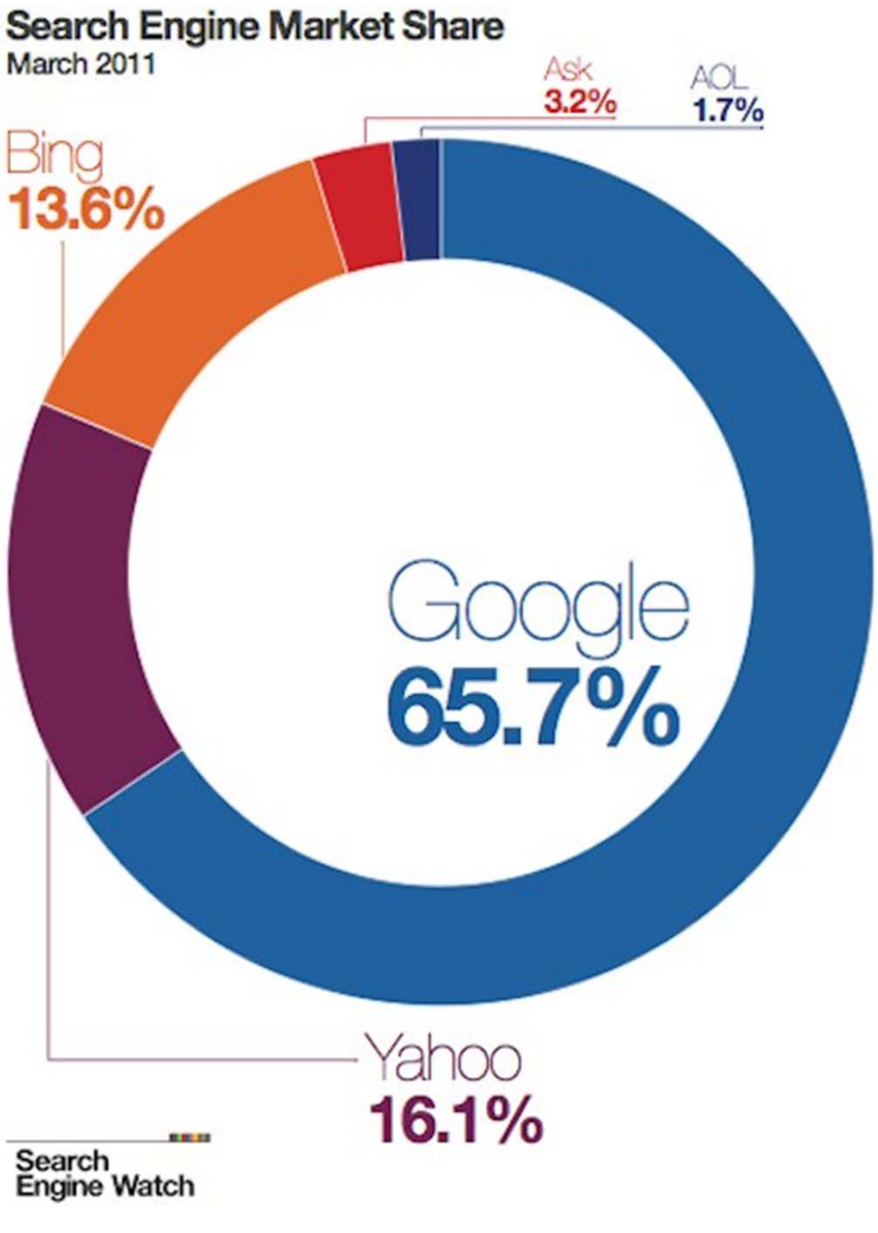
Search Engines…

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

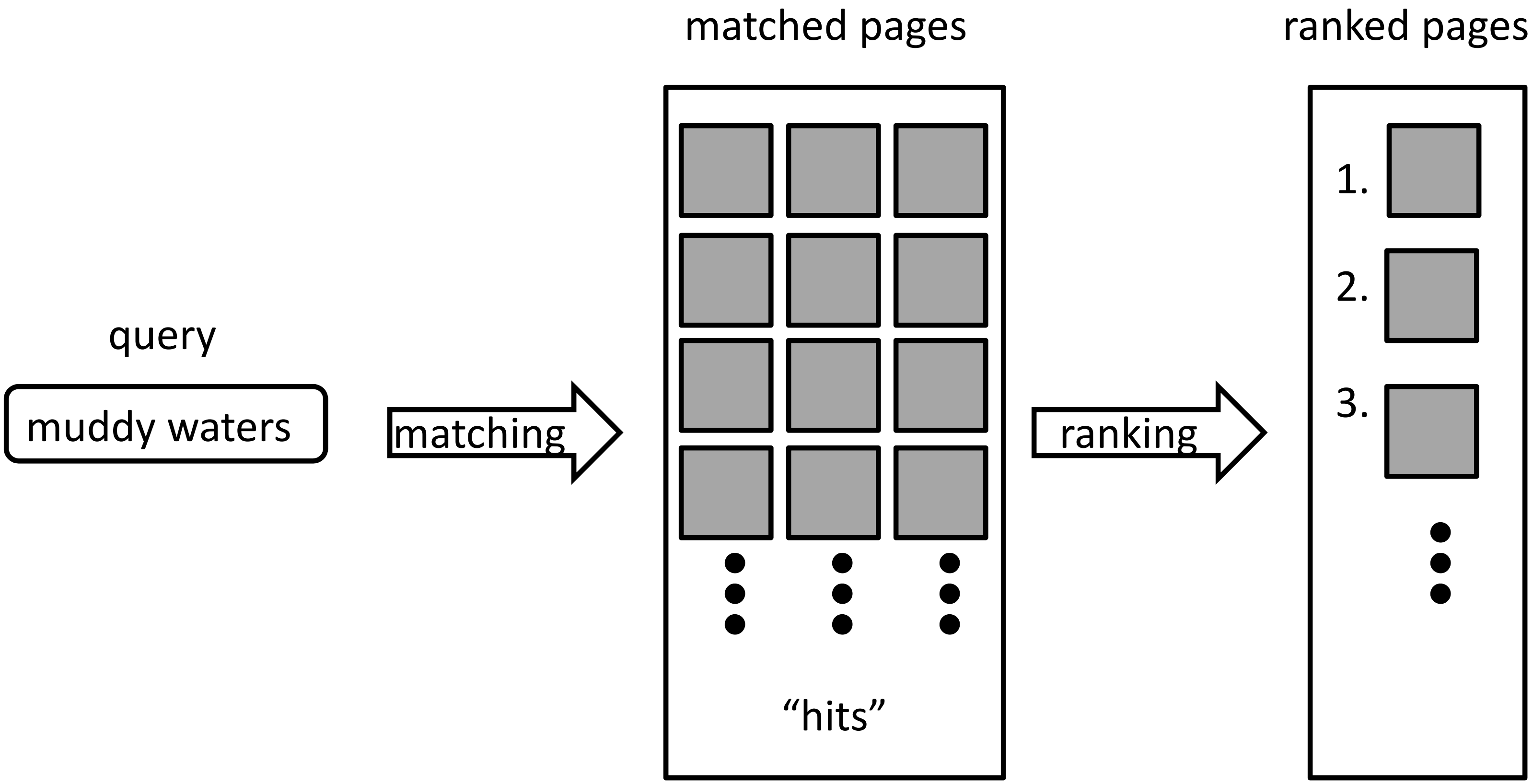
Stumpdedia
Teoma
WebCrawler
Yahoo! Search
Yandex



Search Engine Market Share
March 2011

Search
Engine Watch

Matching & Ranking



Index

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

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

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

cat

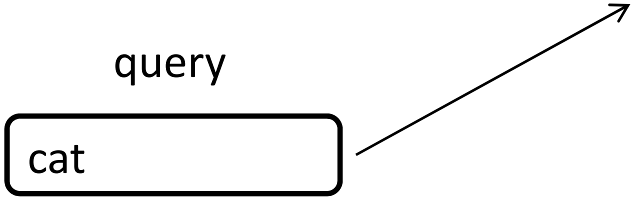
a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Searching

1 the cat sat on the mat

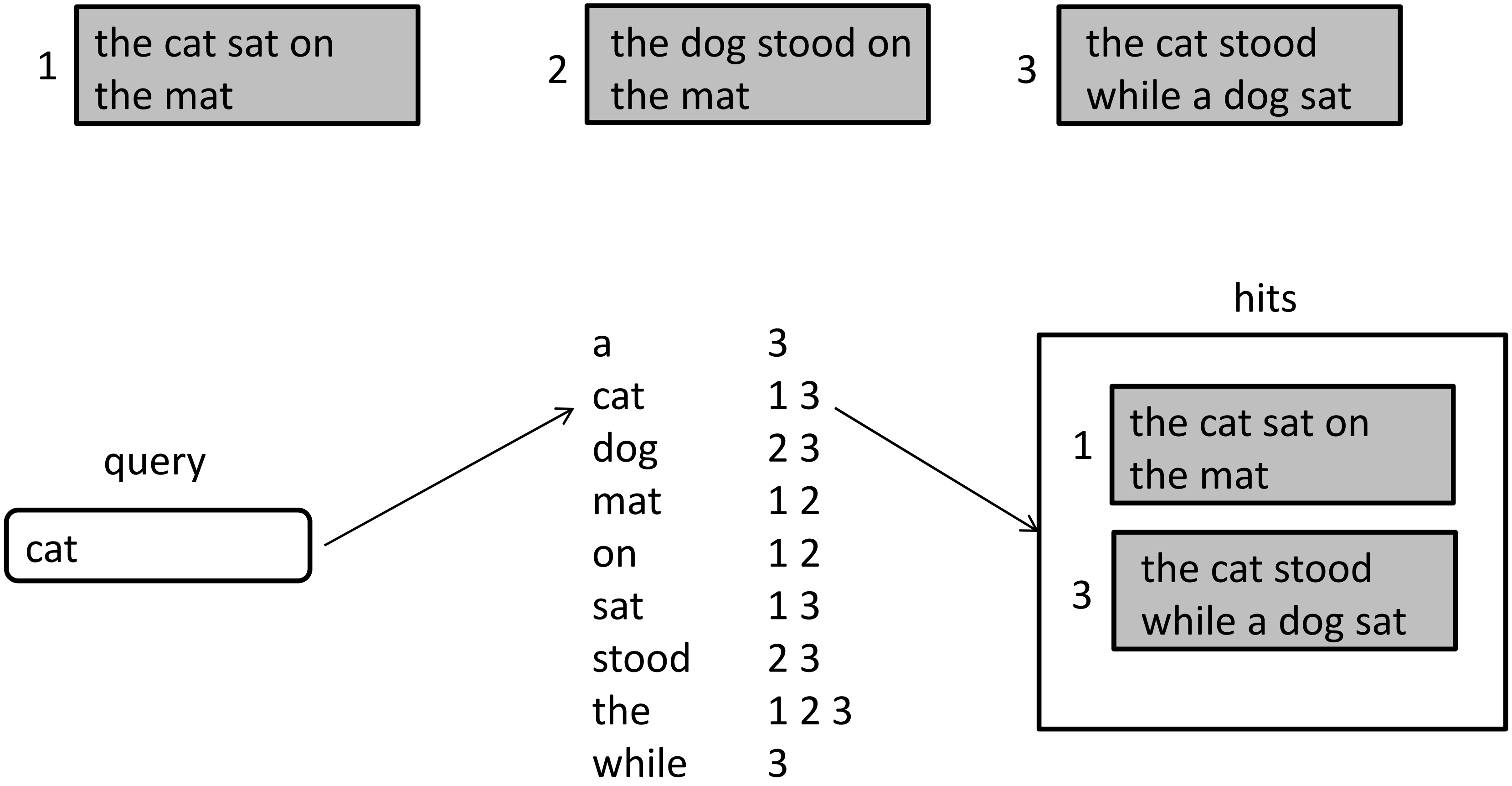
2 the dog stood on the mat

3 the cat stood while a dog sat



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

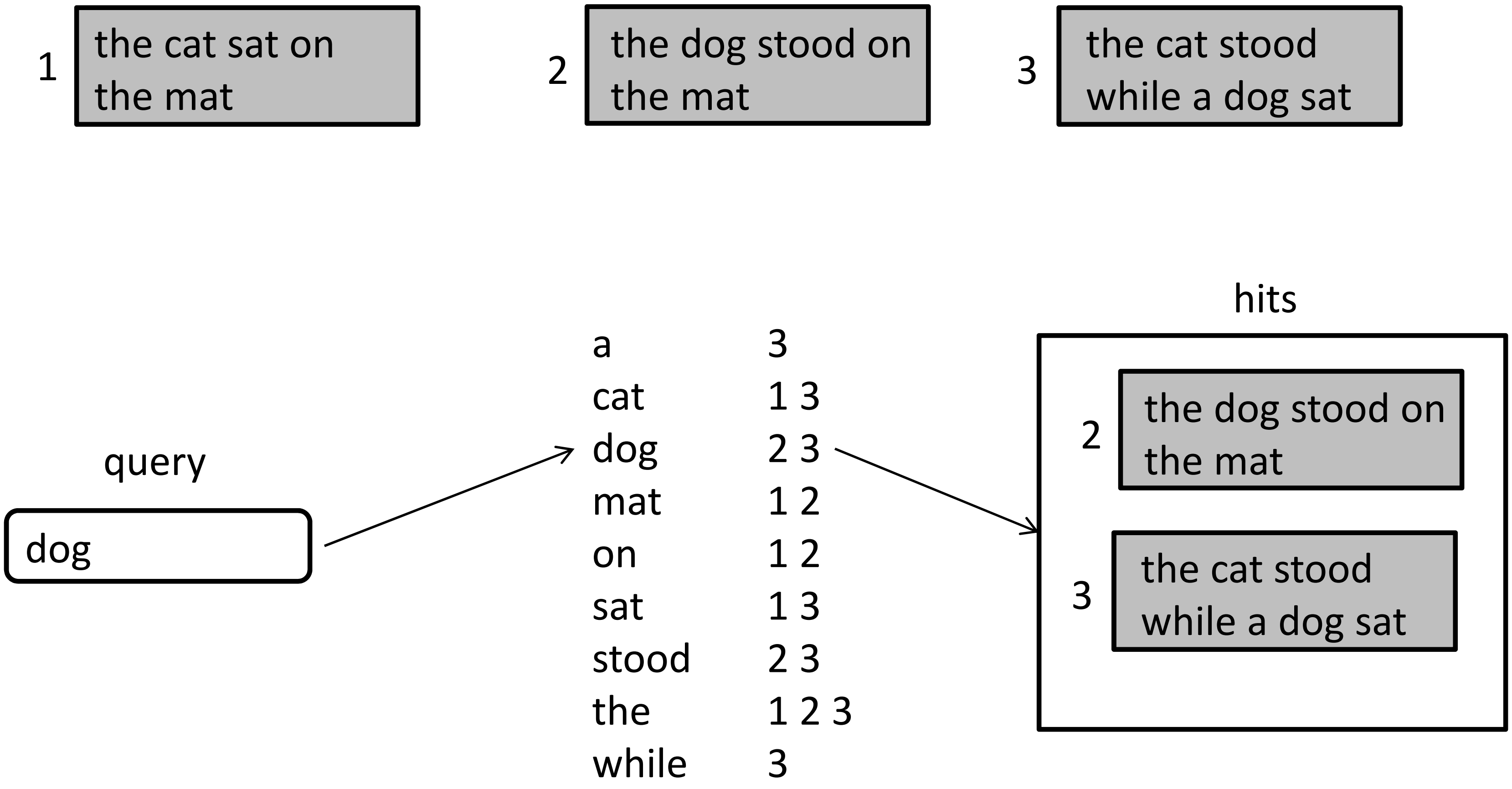
a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

hits

1 the cat sat on the mat

3 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

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

2 the dog stood on the mat

3 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

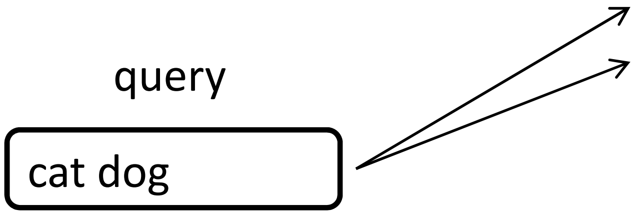
a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

Searching

1 the cat sat on
the mat

2 the dog stood on
the mat

3 the cat stood
while a dog sat

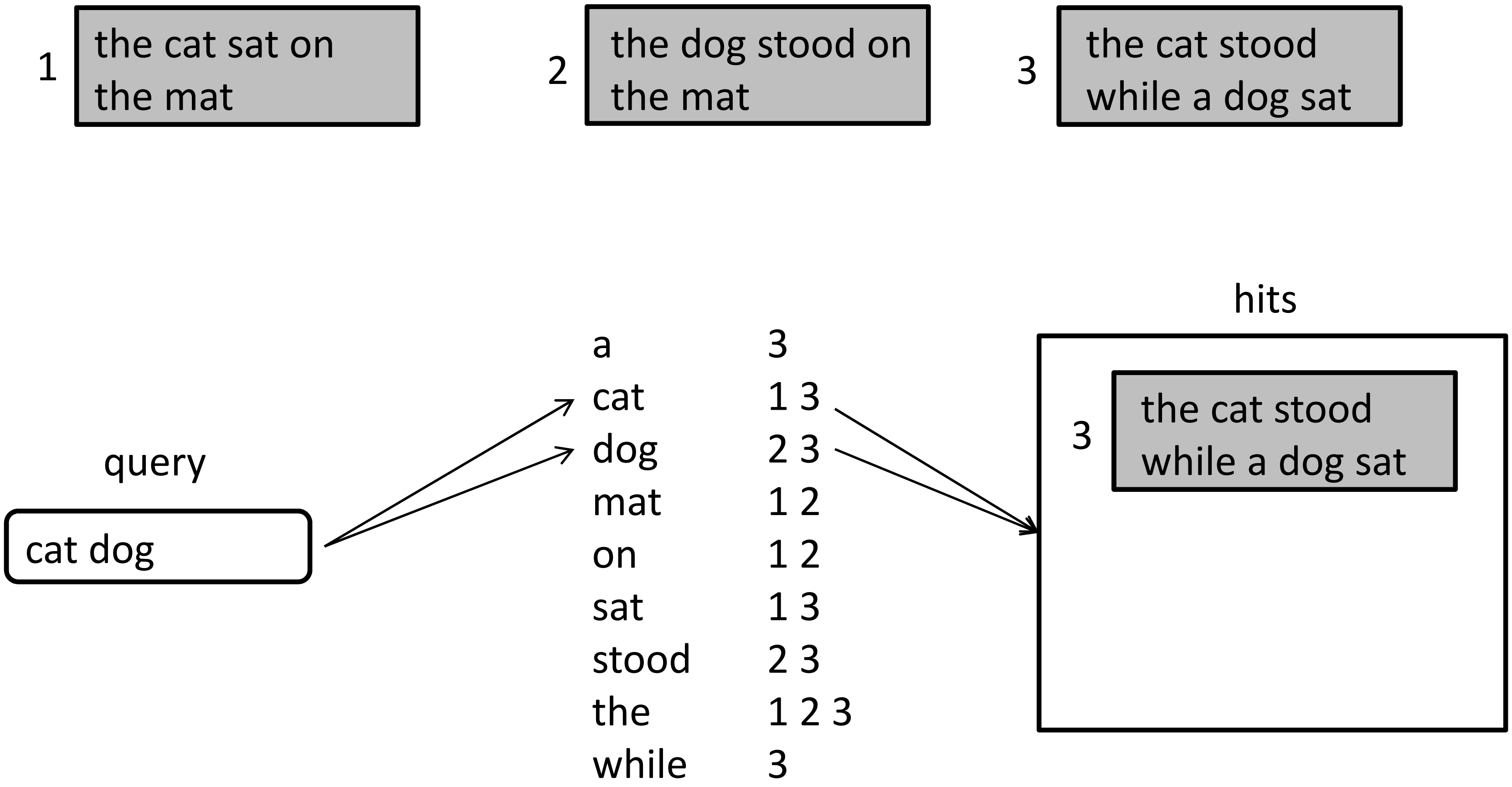


query

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

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

3 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

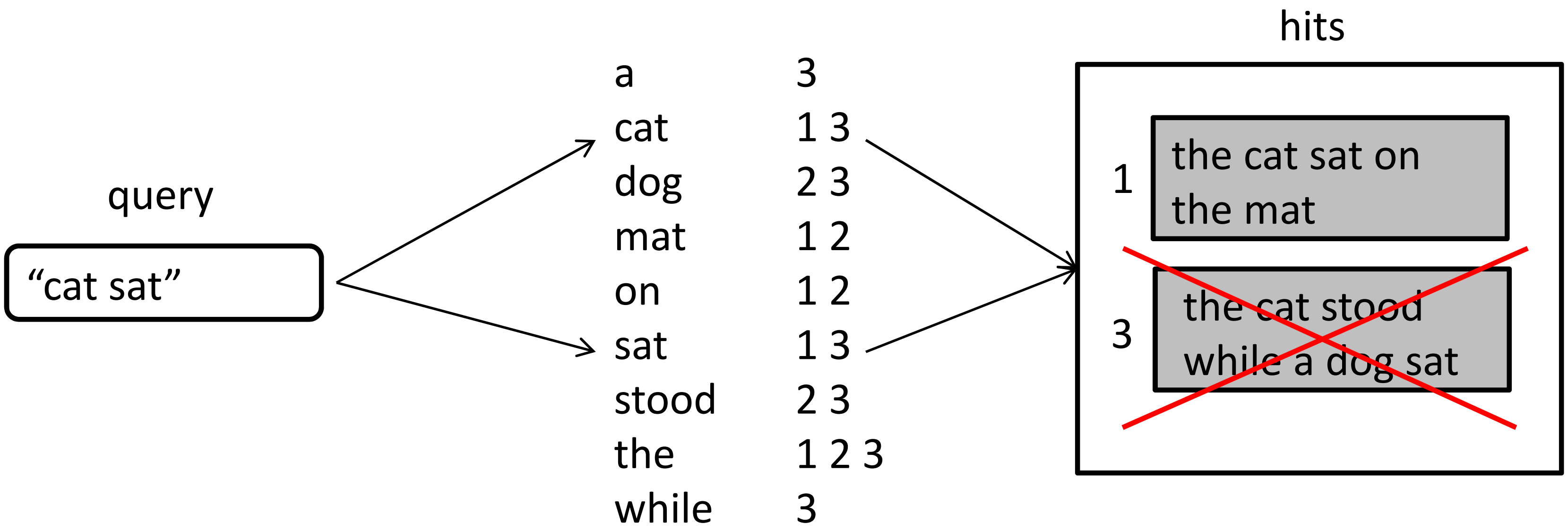
???

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"

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

1 the cat sat on the mat

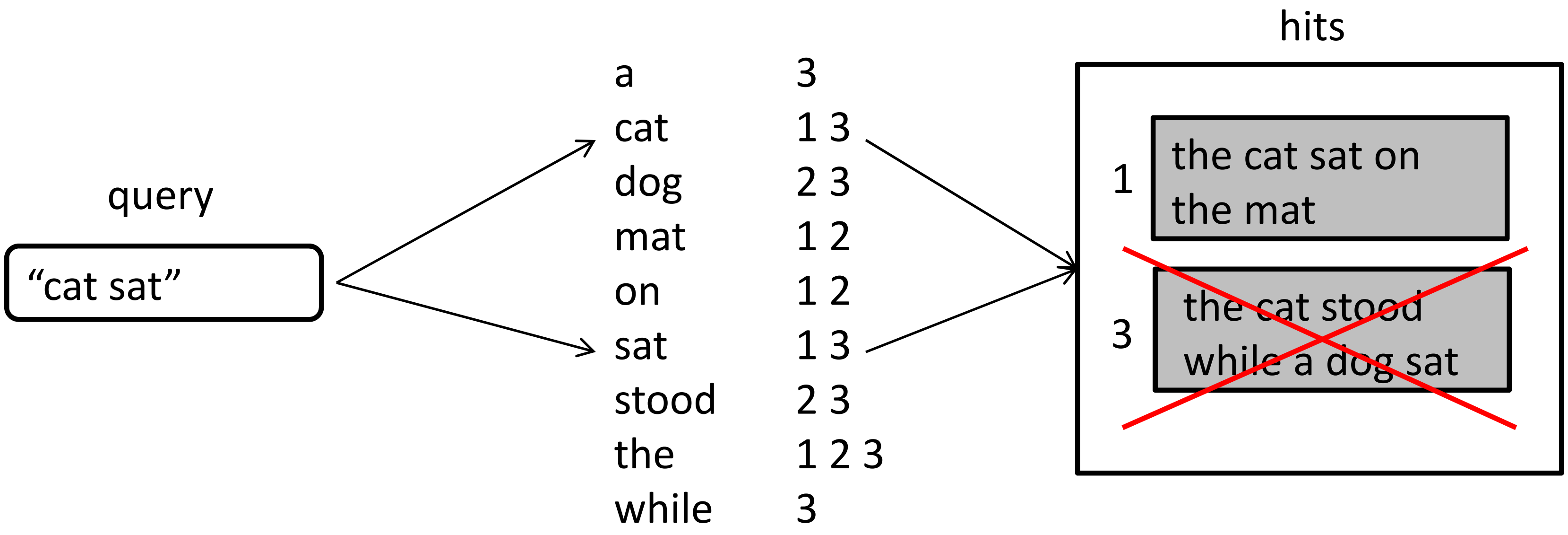
3 ~~the cat stood while a dog sat~~

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"

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

1 the cat sat on the mat

3 ~~the cat stood while a dog sat~~

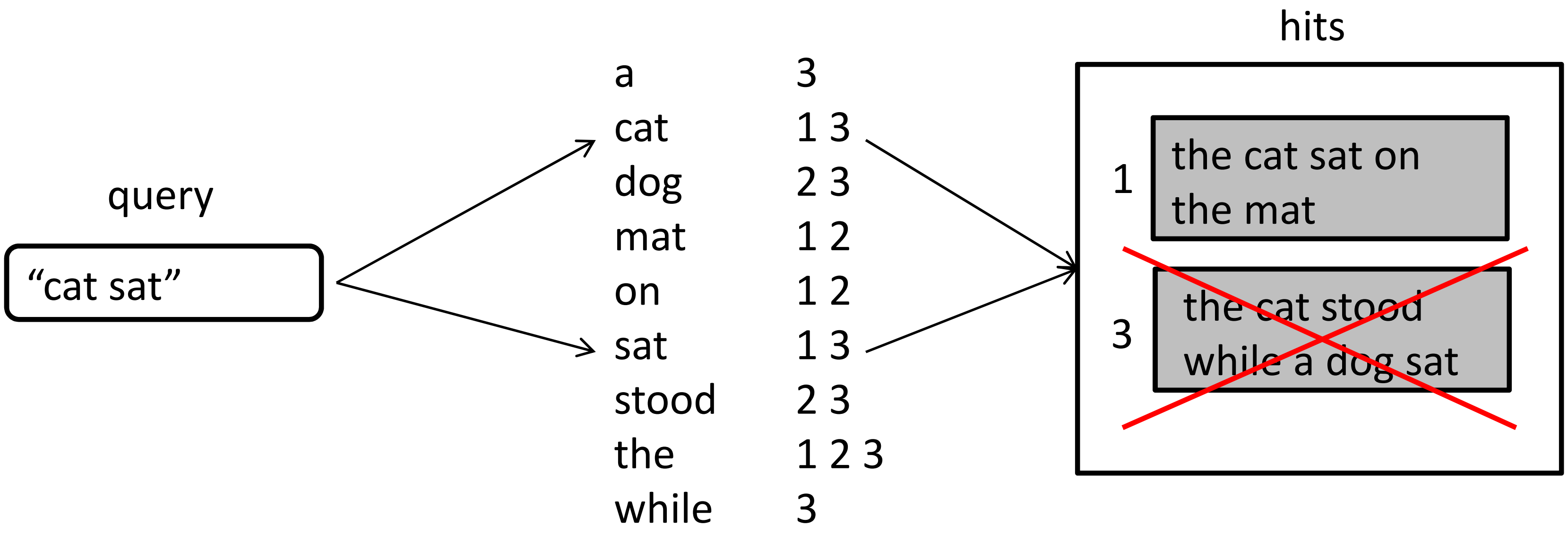
How to tell if two words occur next to each other?

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”

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

1 the cat sat on the mat

3 ~~the cat stood while a dog sat~~

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

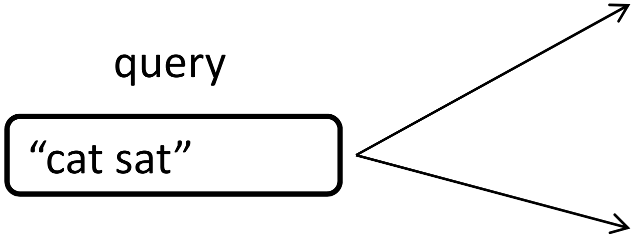
a	3-5
cat	1-2 3-2
dog	2-2 3-6
mat	1-6 2-6
on	1-4 2-4
sat	1-3 3-7
stood	2-3 3-3
the	1-1 1-5 2-1 2-5 3-1
while	3-4

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”

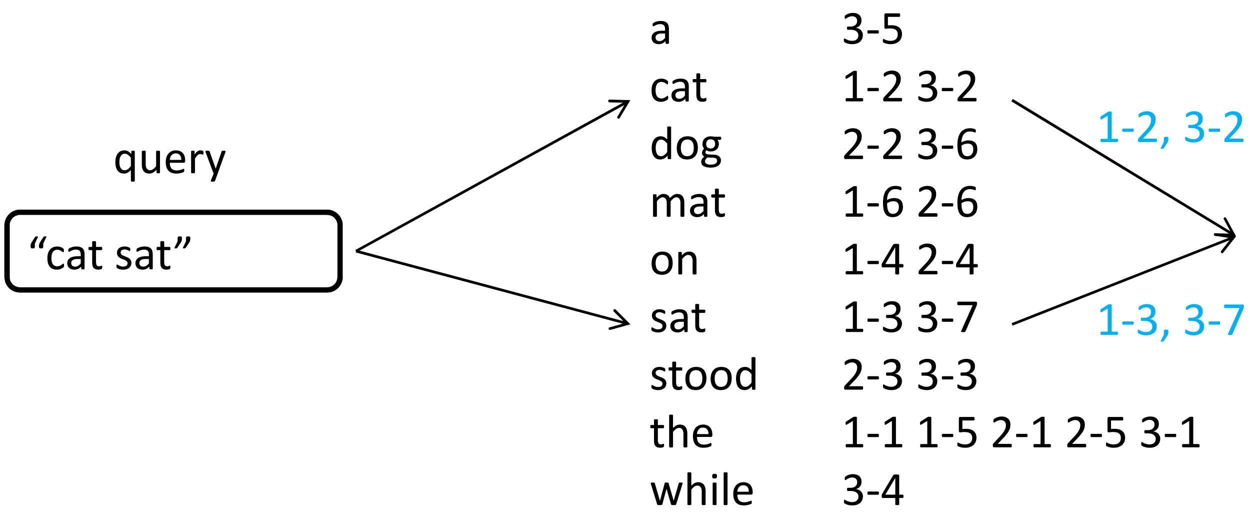
a	3-5
cat	1-2 3-2
dog	2-2 3-6
mat	1-6 2-6
on	1-4 2-4
sat	1-3 3-7
stood	2-3 3-3
the	1-1 1-5 2-1 2-5 3-1
while	3-4

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”

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

1-2, 3-2

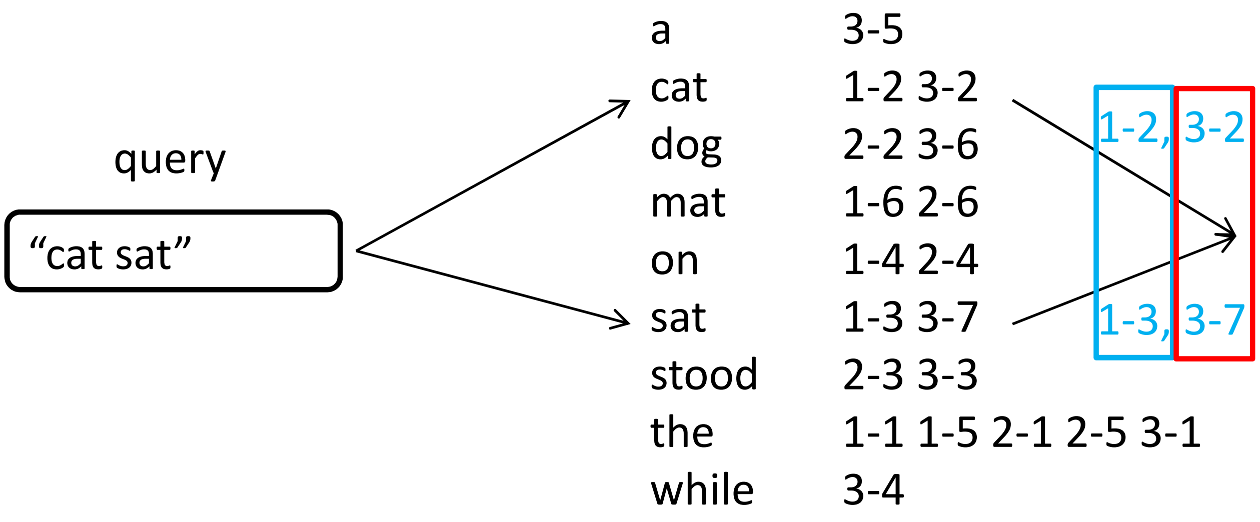
1-3, 3-7

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”

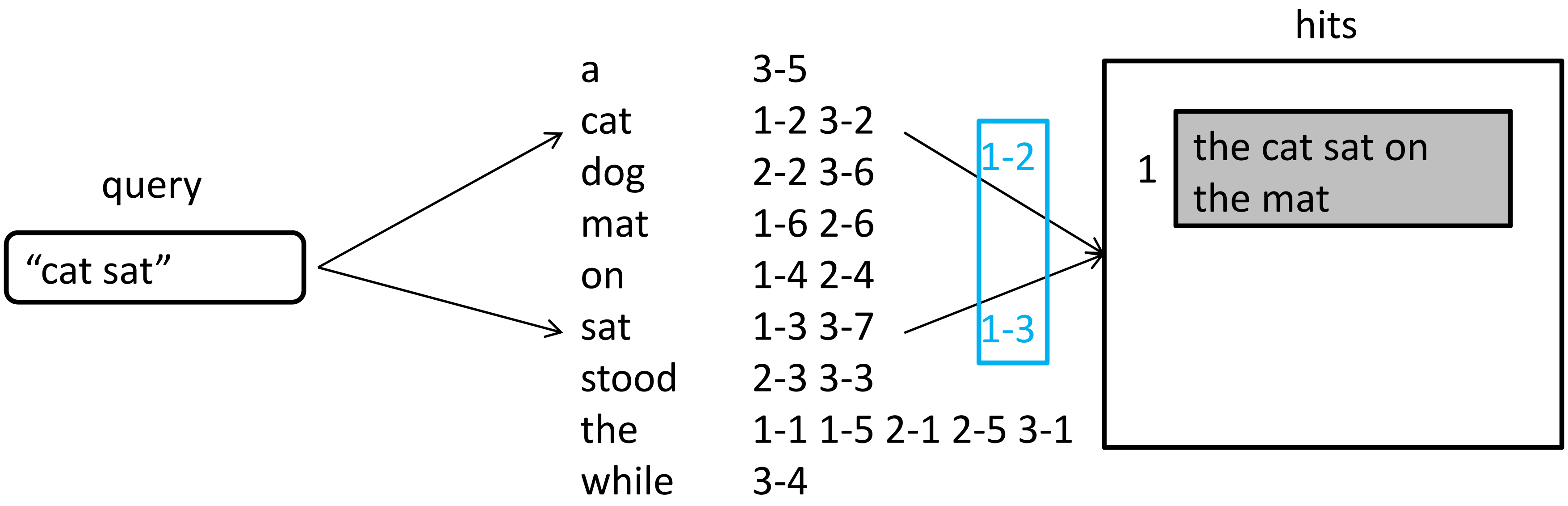
a	3-5
cat	1-2 3-2
dog	2-2 3-6
mat	1-6 2-6
on	1-4 2-4
sat	1-3 3-7
stood	2-3 3-3
the	1-1 1-5 2-1 2-5 3-1
while	3-4

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”

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

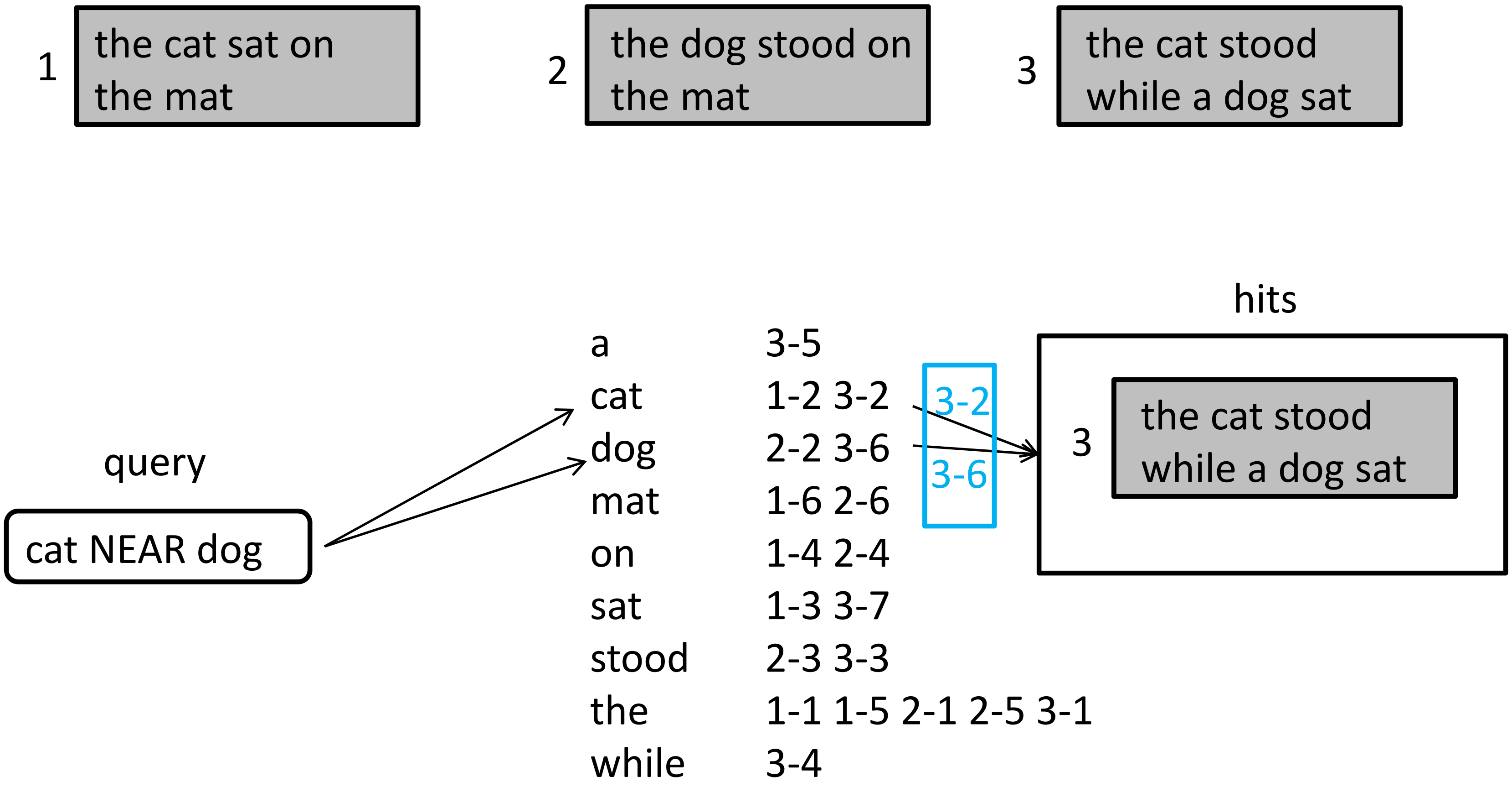
1-2

1-3

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

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

3-2
3-6

hits

3 the cat stood while a dog sat

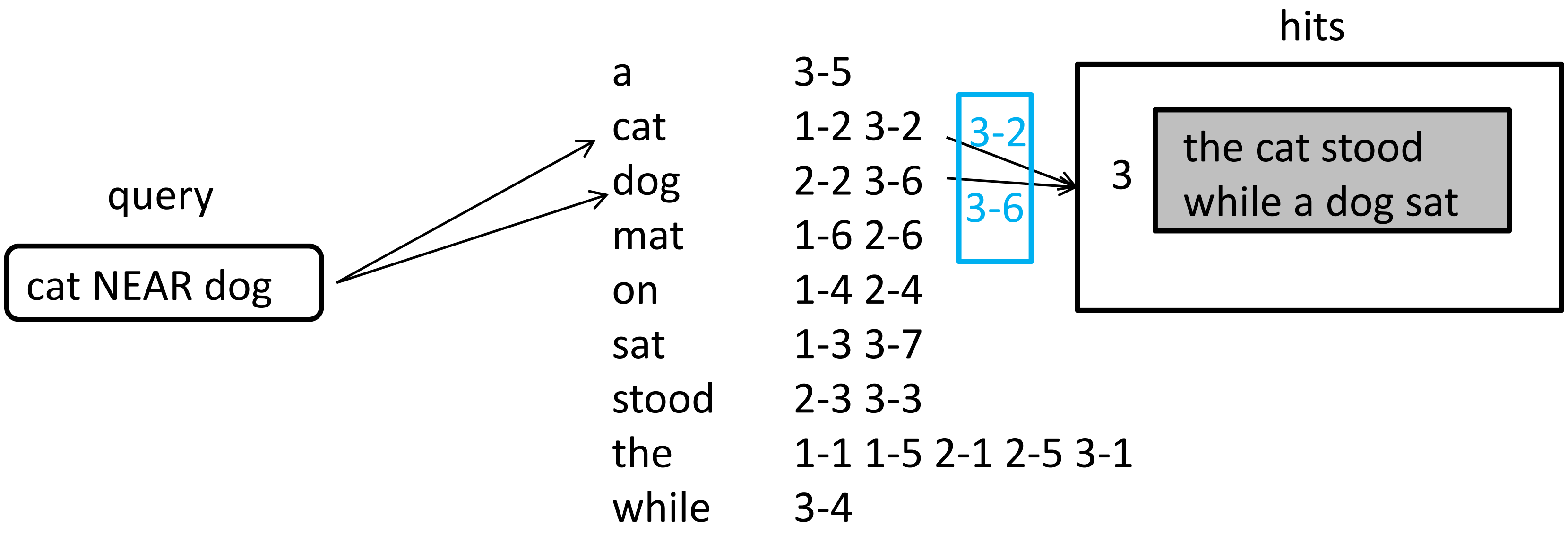
*NEAR: distance <= 5

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

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

3-2
3-6

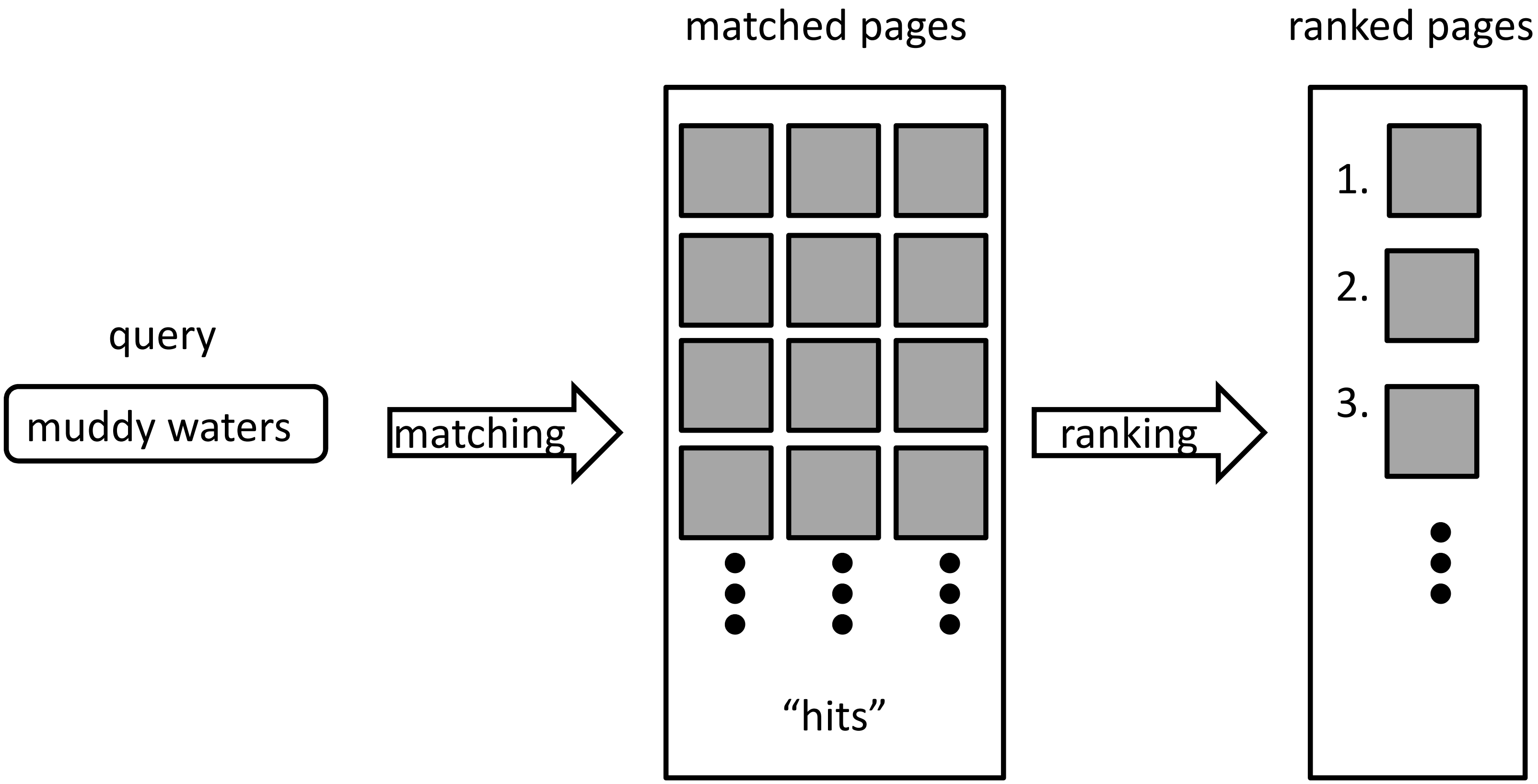
hits

3 the cat stood while a dog sat

Useful in ranking!

*NEAR: distance <= 5

Matching & Ranking



Ranking & Relevance

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.

Ranking & Relevance

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.

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

Ranking & Relevance

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.

query

malaria cause

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

Ranking & Relevance

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.

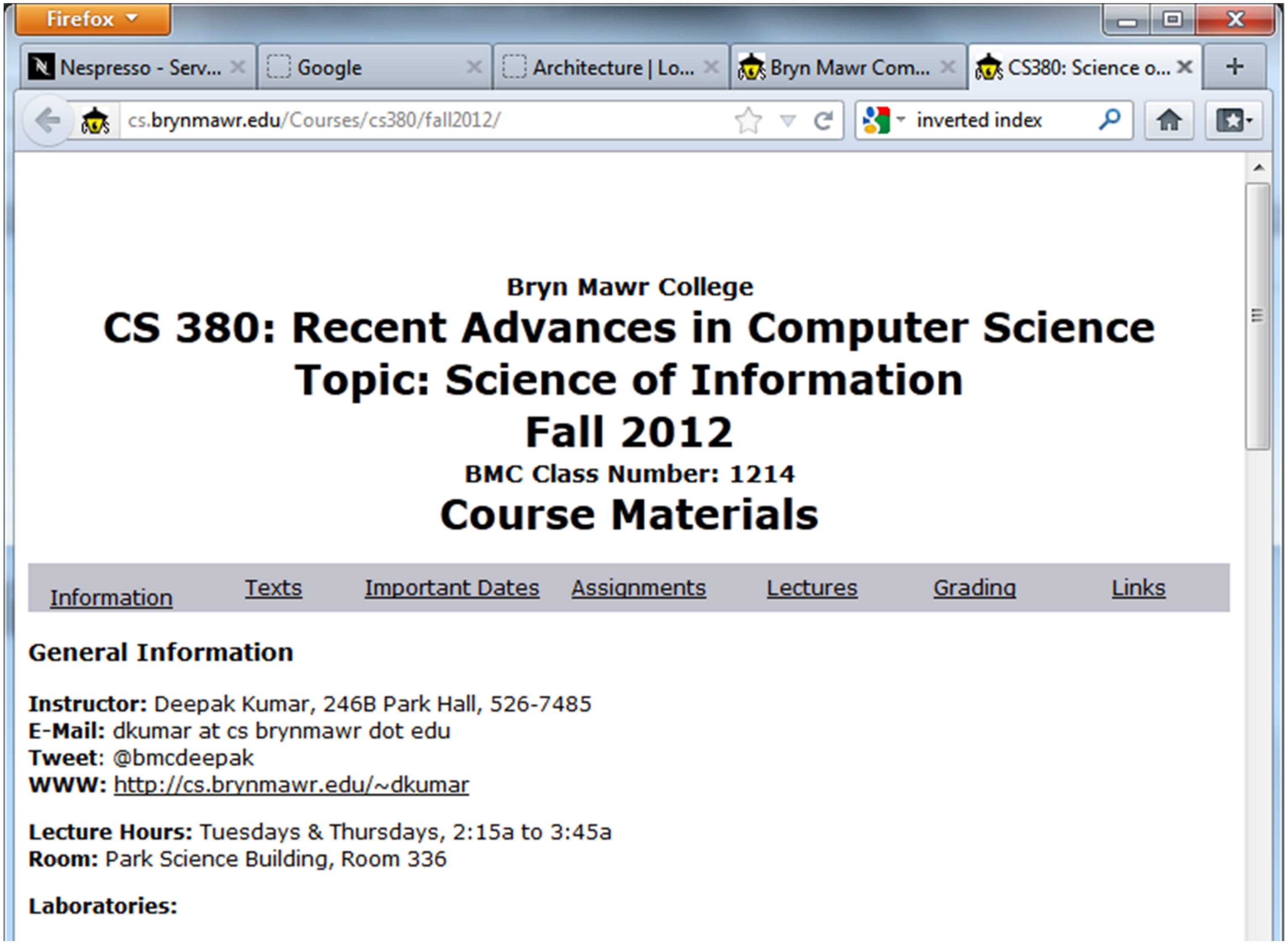
query

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



Bryn Mawr College

CS 380: Recent Advances in Computer Science

Topic: Science of Information

Fall 2012

BMC Class Number: 1214

Course Materials

Information Texts Important Dates Assignments Lectures Grading Links

General Information

Instructor: Deepak Kumar, 246B Park Hall, 526-7485
E-Mail: dkumar at cs brynmawr dot edu
Tweet: @bmcdeepak
WWW: http://cs.brynmawr.edu/~dkumar

Lecture Hours: Tuesdays & Thursdays, 2:15a to 3:45a
Room: Park Science Building, Room 336

Laboratories:

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 CLEAR="ALL">
<B><FONT SIZE="+2">CS 380: Recent Advances in Computer Science<br>
 Topic: Science of Information
 </FONT></B><BR CLEAR="ALL">
 <B><FONT SIZE="+2">Fall 2012</FONT></B><br>
 BMC Class Number: 1214<BR CLEAR="ALL">
 <B><FONT SIZE="+2">Course Materials</FONT></B>
 </h3>
 </CENTER>
…
```

Metadata

1 **my cat**
the cat sat on the mat

2 **my dog**
the dog stood on the mat

3 **my pets**
the cat stood while a dog sat

Metadata

1 **my cat**
the cat sat on the mat

2 **my dog**
the dog stood on the mat

3 **my pets**
the cat stood while a dog sat

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

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

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

Metadata

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

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

3 <title>my pets
</title><body>th
e cat stood while
a dog sat

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

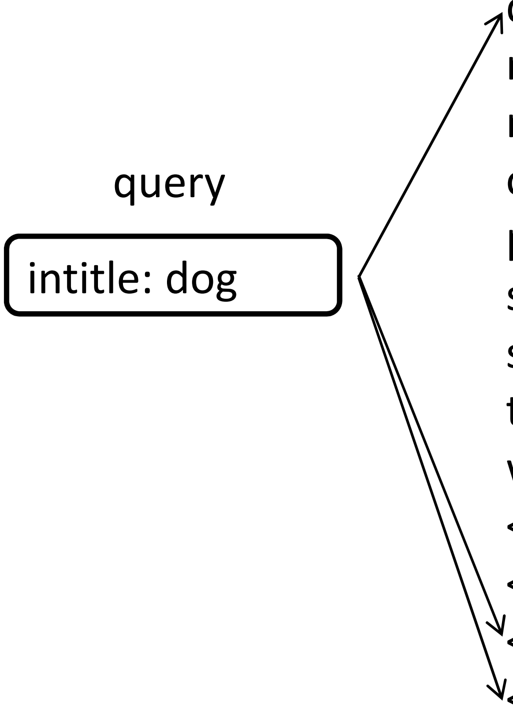
Structure Queries

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

Structure Queries

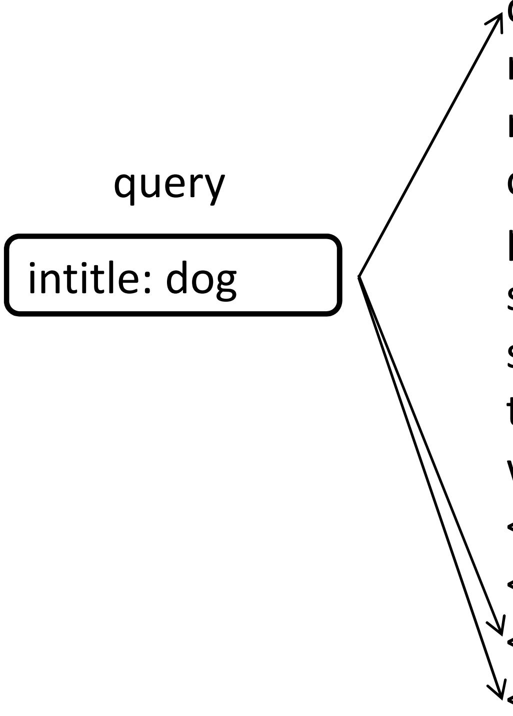


query

intitle: dog

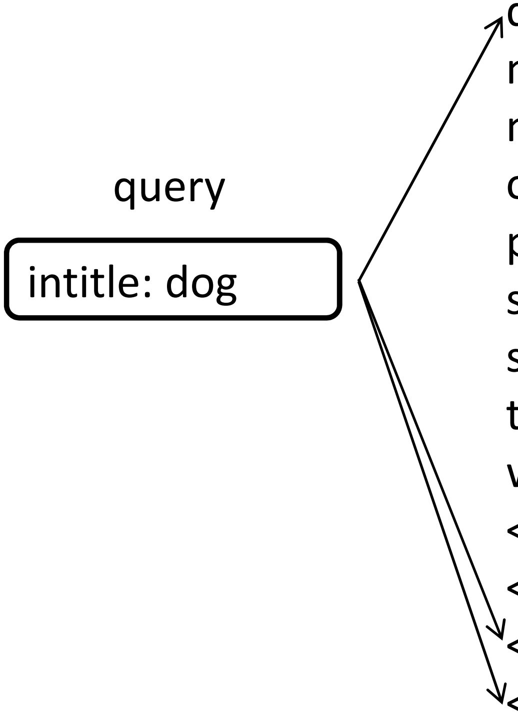
a	3-10
cat	1-3 1-7 3-7
dog	2-3 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<body>	1-5 2-5 3-5
</body>	1-12 2-12 3-13
<title>	1-1 2-1 3-1
</title>	1-4 2-4 3-4

Structure Queries



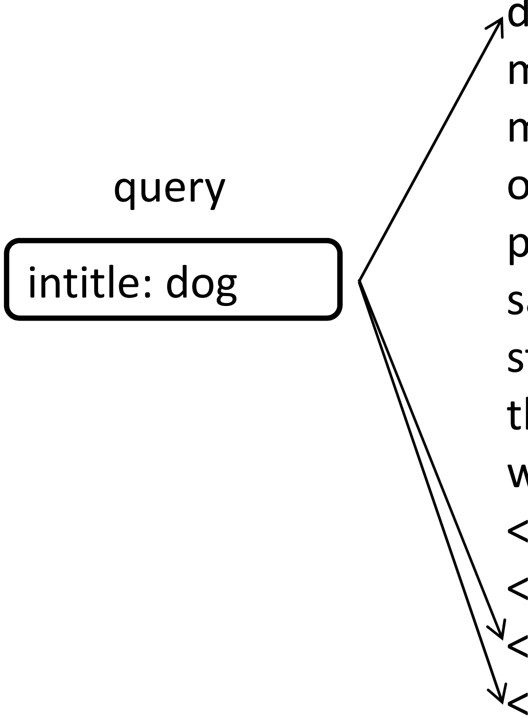
a	3-10
cat	1-3 1-7 3-7
dog	**2-3** 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<body>	1-5 2-5 3-5
</body>	1-12 2-12 3-13
<title>	1-1 **2-1** 3-1
</title>	1-4 **2-4** 3-4

Structure Queries



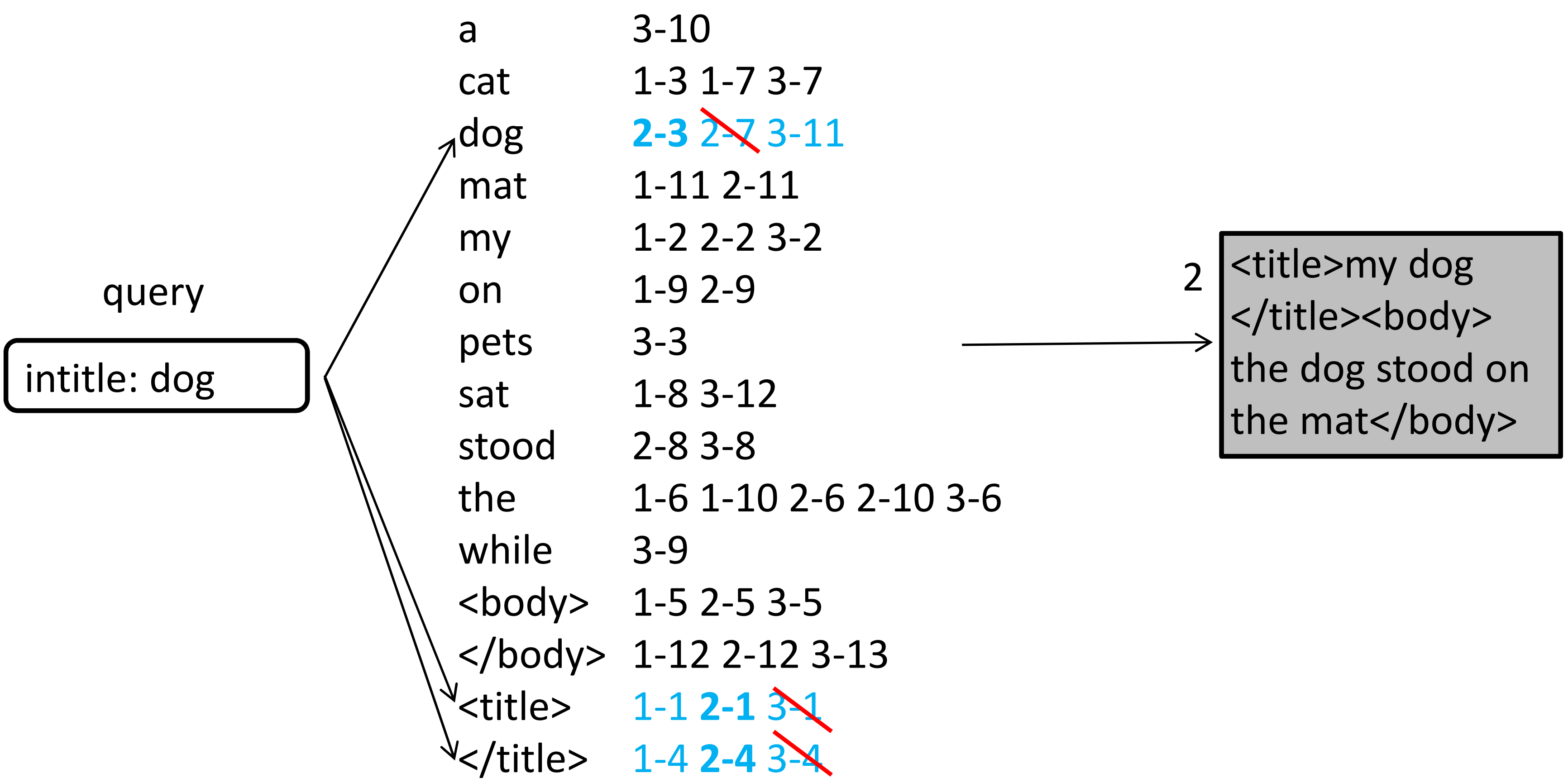
a	3-10
cat	1-3 1-7 3-7
dog	**2-3** ~~2-7~~ 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<body>	1-5 2-5 3-5
</body>	1-12 2-12 3-13
<title>	1-1 **2-1** 3-1
</title>	1-4 **2-4** 3-4

Structure Queries



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~~

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

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

- *Google’s PageRank and Beyond*, Amy N. Langville and Carl D. Meyer, Princeton University Press, 2006.
- *Nine Algorithms That Changed The Future*, John MacCormick, Princeton University Press, 2012.
- *Learning Computing with Robots*, Deepak Kumar, IPRE 2011.