Similarity

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

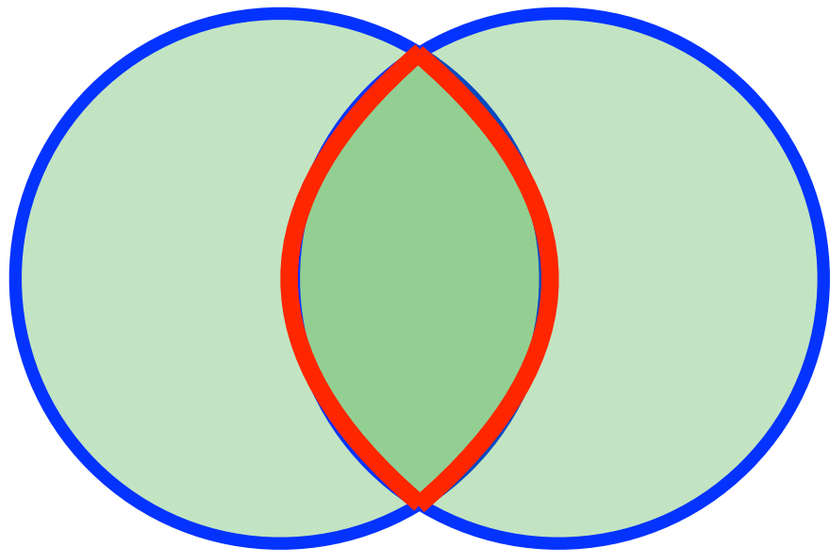
Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Sets

$$A = \{1, 2, 3, 4\} \qquad B = \{3, 4, 5, 6, 7\}$$

Union: items in either	$A \cup B$	{1, 2, 3, 4, 5, 6, 7}
Intersection: items in both	$A \cap B$	{3, 4}
Exclusion: items in A but not B	$A \setminus B$	{1, 2}
Symmetric difference: items in union but not intersection	$A \triangle B$	{1, 2, 5, 6, 7}

Sets



$$|A \cup B| = |A| + |B| - |A \cap B|$$

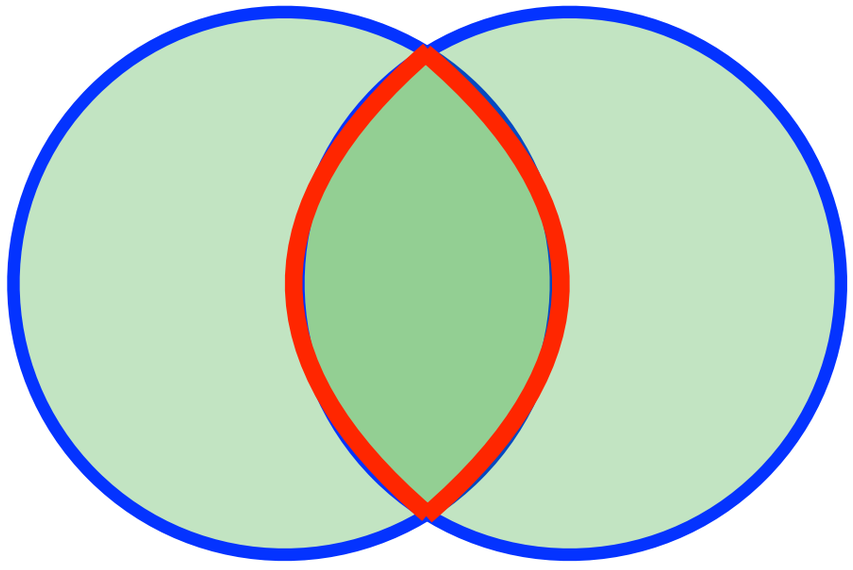
Inclusion–exclusion
or "double counting"
principle

$$|A \cap B| = |A| + |B| - |A \cup B|$$

Sets

To measure similarity of A & B, we're interested in the size of $A \cap B$ i.e. $|A \cap B|$

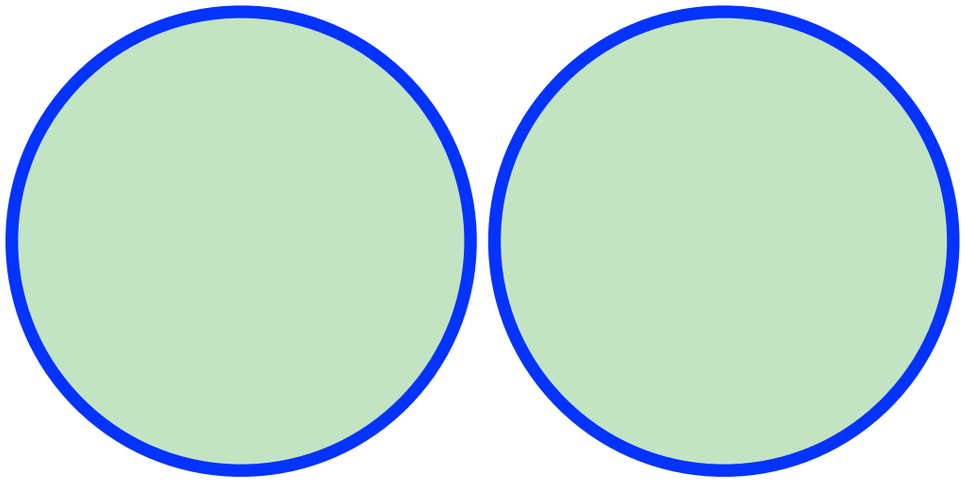
This is best understood relative to the sizes of the sets themselves, so let's normalize by $|A \cup B|$



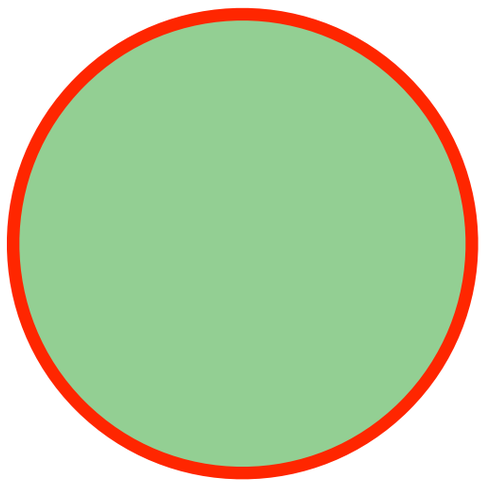
$$\frac{|A \cap B|}{|A \cup B|} = J$$

J is the *Jaccard coefficient*

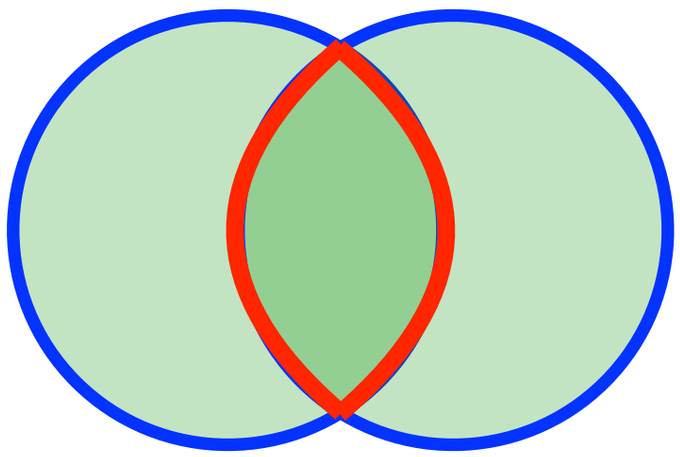
Sets



$$\frac{|A \cap B|}{|A \cup B|} = 0$$



$$\frac{|A \cap B|}{|A \cup B|} = 1$$



$$0 < \frac{|A \cap B|}{|A \cup B|} < 1$$

Sets

$$J = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{|A \cap B|}{|A \cap B| + |A \triangle B|}$$

Symmetric difference

Helps isolate what's happening in the denominator

$$= \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

"Double counting" to eliminate union

$$= \frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

"Double counting" to eliminate intersection

Sets

$$A = \{1, 2, 3, 4\} \qquad B = \{3, 4, 5, 6, 7\}$$

$$J = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{3,4\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{2}{7}$$

Cardinality

Say we find the 8 minimum hashes (bottom-8) for items in set A, and repeat for items in set B

Set A

3	15
7	17
8	22
11	23

Set B

2	9
3	11
6	17
7	23

Cardinality

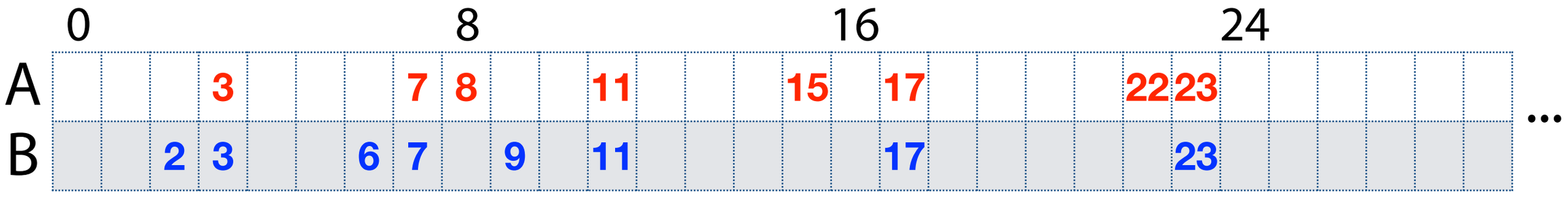
Say we find the 8 minimum hashes (bottom-8) for items in set A, and repeat for items in set B

Set A

3	15
7	17
8	22
11	23

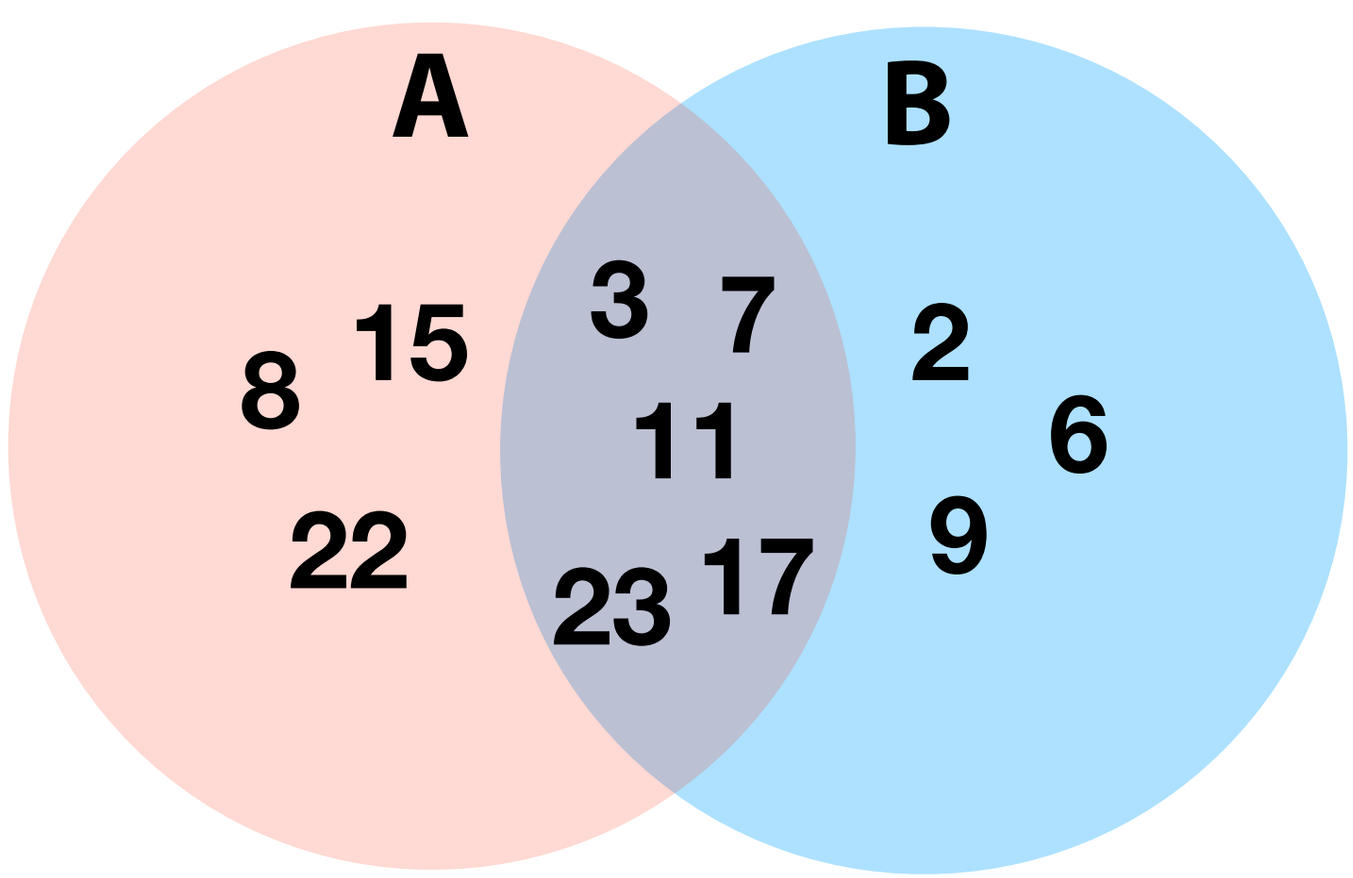
Set B

2	9
3	11
6	17
7	23



Cardinality

Can we localize hash values in a venn diagram? Assume no collisions; i.e. each hash value is a distinct item



Cardinality

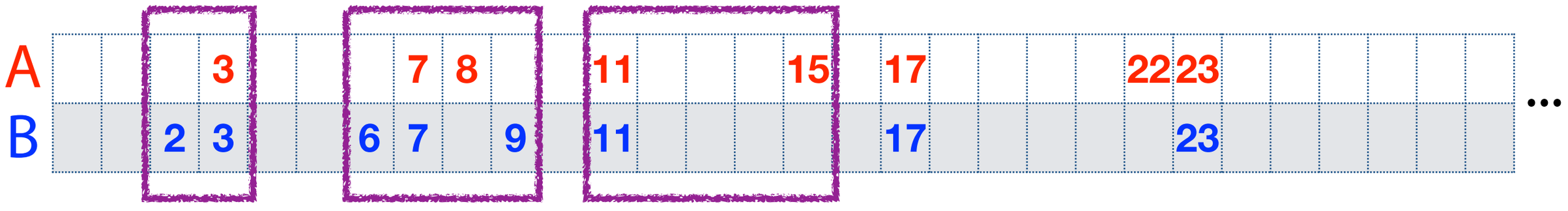
A	
3	15
7	17
8	22
11	23

∪

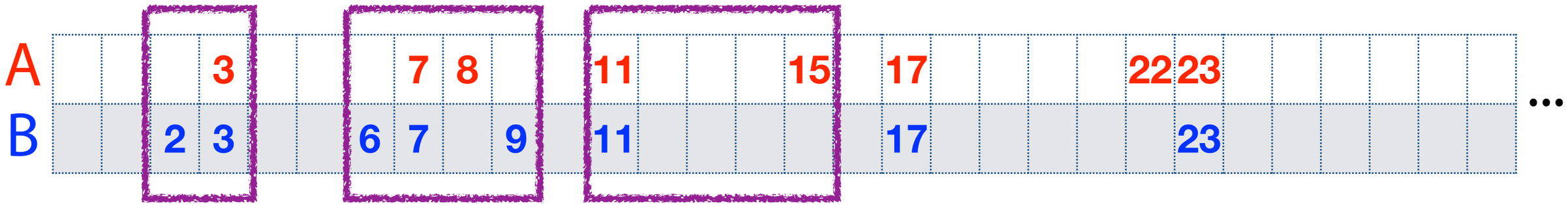
B	
2	9
3	11
6	17
7	23

=

A ∪ B	
2	8
3	9
6	11
7	15



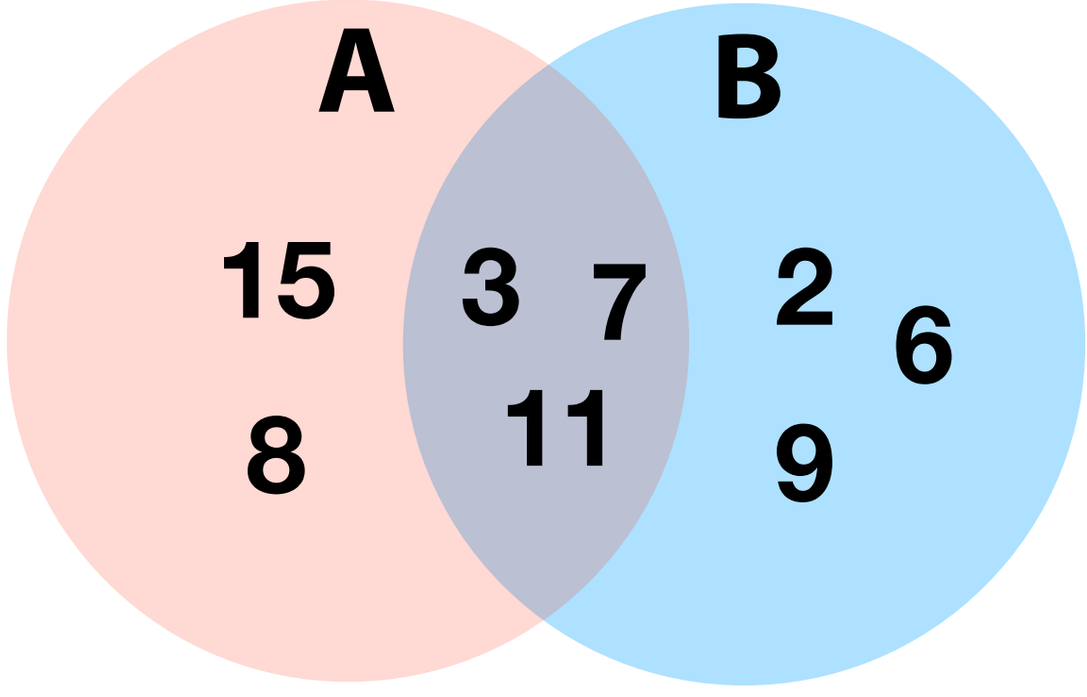
Cardinality



"Samples" from $|A \cup B|$

Fraction that are also in $|A \cap B|$ is an estimate for

$$J = \frac{|A \cap B|}{|A \cup B|}$$



Cardinality

To estimate Jaccard coefficient

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Find $\frac{|A \cap B|}{|A \cup B|}$ directly

Fraction of items in union sketch that are in both

Find $|A|, |B|, |A \cup B|$

$$J = \frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Direct:

Fraction of items in union sketch that are in both

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Direct:

Fraction of items in union sketch that are in both $= \frac{3}{8} = 0.375$

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Indirect:

Using KMV with $k = 8$, assuming hash range is integers in [0, 1000)

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Indirect:

Using KMV with $k = 8$, assuming hash range is integers in [0, 1000)

$$\frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

$$= \frac{8000/23 - 1 + 8000/23 - 1 - (8000/15 - 1)}{8000/15 - 1}$$

$$= \frac{347.82 + 347.82 - 533.33 - 1}{533.33 - 1}$$

$$= 0.30$$

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

Direct:

$$\frac{|A \cap B|}{|A \cup B|} = \frac{3}{8} = 0.375$$

Indirect:

$$\frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

$$= \frac{2 \cdot 8000/23 - 8000/15 - 1}{8000/15 - 1}$$

$$= 0.30$$

Cardinality

A

3	15
7	17
8	22
11	23

B

2	9
3	11
6	17
7	23

A ∪ B

2	8
3	9
6	11
7	15

All computation here is simple

- Hash functions
- Bottom k (heap / sorted list)
- k^{th} minimum value (lookup)
- Get union sketch (merge heaps / lists)
- Calculate Jaccard (during merge)

MinHash

On the resemblance and containment of documents

Andrei Z. Broder
DIGITAL Systems Research Center
130 Lytton Avenue, Palo Alto, CA 94301, USA
broder@pa.dec.com

Abstract

Given two documents A and B we define two mathematical notions: their *resemblance* $r(A, B)$ and their *containment* $c(A, B)$ that seem to capture well the informal notions of "roughly the same" and "roughly contained."

The basic idea is to reduce these issues to set intersection problems that can be easily evaluated by a process of random sampling that can be done independently for each document. Furthermore, the resemblance can be evaluated using a fixed size sample for each document.

This paper discusses the mathematical properties of these measures and the efficient implementation of the sampling process using Rabin fingerprints.

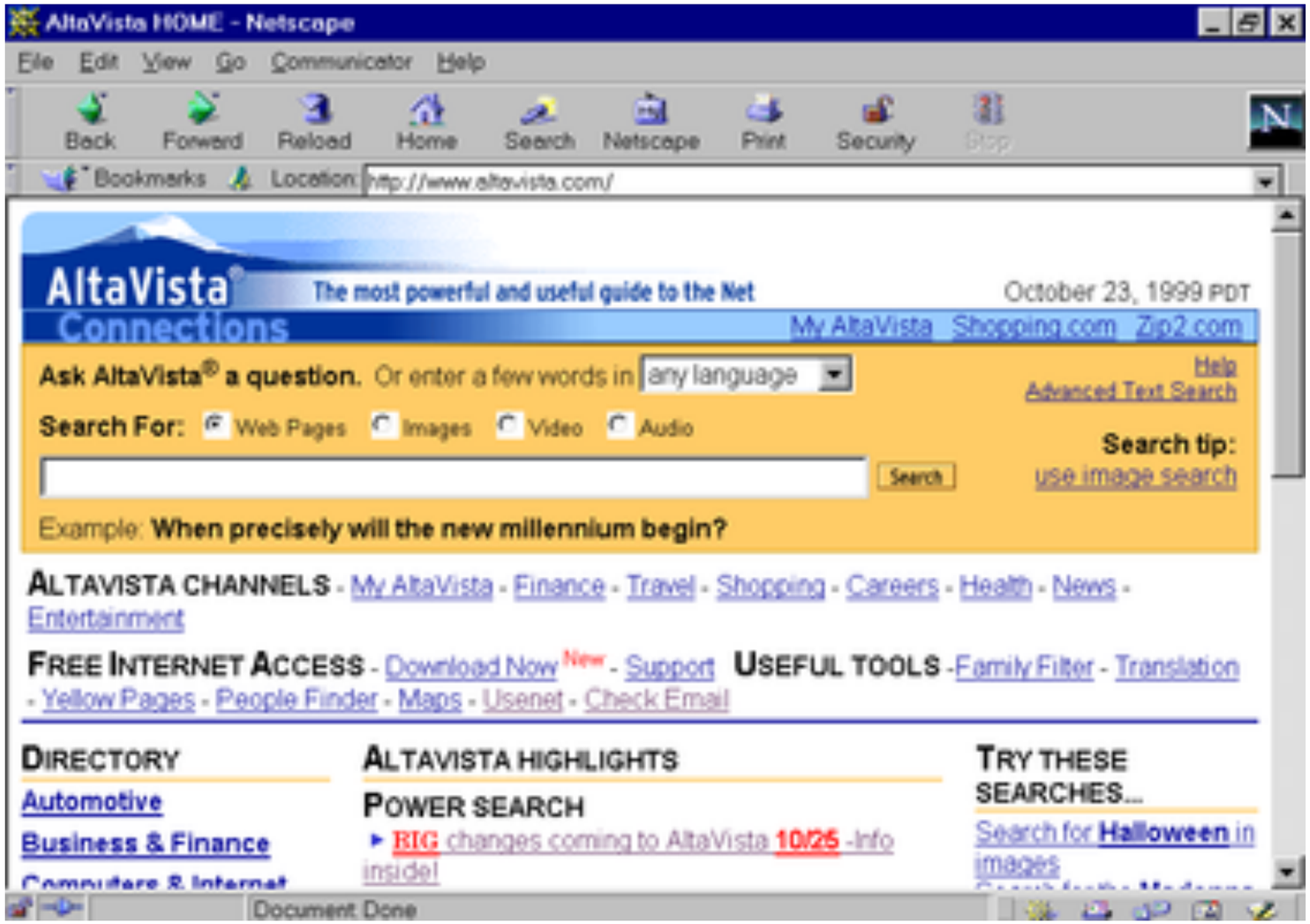
Originally developed for web search. Anyone heard of AltaVista?

"resemblance" = Jaccard

MinHash approach: sketch A & B and estimate Jaccard from union sketch

Broder, Andrei Z. "On the resemblance and containment of documents." *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997.

AltaVista



AltaVista in 1999, as seen in Netscape web browser

Cardinality

A	
3	15
7	17
8	22
11	23

B	
2	9
3	11
6	17
7	23

A ∩ B

?			?
?			?

Cardinality

A	
3	15
7	17
8	22
11	23

B	
2	9
3	11
6	17
7	23

$A \cap B$

? ?

3	23
7	?
11	?
17	?

? ?

Cannot generally construct intersection sketch from individual sketches

Problem most pronounced when J (similarity) is small

Cardinality

A	
3	15
5	19
8	22
12	23

B	
2	9
3	11
6	17
7	23

$A \cap B$

? ?

? ?

3	?
23	?
?	?
?	?

Cannot generally construct intersection sketch from individual sketches

Problem most pronounced when J (similarity) is small

MinHash

Sometimes explained this way but not so efficient

k hashes

Original MinHash proposal

Bottom k

AKA 1-permutation hashing / 1-permutation MinHash

k partitions