

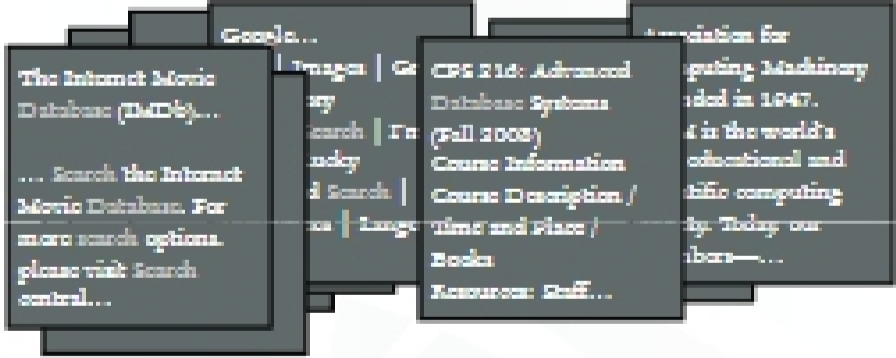
Web Searching & Indexing

CPS 116
Introduction to Database Systems

Announcements (December 2)

- ❖ Homework #4 sample solution available
- ❖ Course project demo period: December 8-13
 - Each project gets a 30-minute slot with me
 - Email me to schedule demo slots
- ❖ Final exam next Saturday, Dec. 13, 7-10pm
 - Early final next Wednesday, Dec. 10, 9am-12pm
 - Again, open book, open notes
 - Focus on the second half of the course
 - Sample final available
 - Sample final solution available Thursday

Keyword search



database AND search Search

What are the documents containing both "database" and "search"?

Keywords × documents

All keywords	All documents				
	Document 1	Document 2	Document 3	Document n	
"z"	1	1	1	...	1
"cat"	1	1	0	...	0
"database"	0	0	1	...	0
"dog"	0	1	0	...	1
"search"	0	0	1	...	0
...

1 means keyword appears in the document;
0 means otherwise

- ❖ Inverted lists: store the matrix by rows
- ❖ Signature files: store the matrix by columns

Inverted lists

- ❖ Store the matrix by rows
- ❖ For each keyword, store an inverted list
 - $\langle \text{keyword}, \text{doc-id-list} \rangle$
 - $\langle \text{"database"}, \{3, 7, 142, 857, \dots\} \rangle$
 - $\langle \text{"search"}, \{3, 9, 192, 512, \dots\} \rangle$
 - It helps to sort *doc-id-list* (why?)
- ❖ Vocabulary index on keywords
 - B⁺-tree or hash-based
- ❖ How large is an inverted list index?

Using inverted lists

- ❖ Documents containing "database"
 - Use the vocabulary index to find the inverted list for "database"
 - Return documents in the inverted list
- ❖ Documents containing "database" AND "search"
 -
- ❖ OR? NOT?

What are "all" the keywords?

- ❖ All sequences of letters (up to a given length)?
 - ... that actually appear in documents!
- ❖ All words in English?
- ❖ Plus all phrases?
 - Alternative: approximate phrase search by proximity
- ❖ Minus all stop words
 - They appear in nearly every document, e.g., a, of, the, it
 - Not useful in search
- ❖ Combine words with common stems
 - Example: database, databases
 - They can be treated as the same for the purpose of search

Frequency and proximity

- ❖ Frequency
 - $\langle \text{keyword}, \{ (\text{doc-id}, \text{number-of-occurrences}), (\text{doc-id}, \text{number-of-occurrences}), \dots \} \rangle$
- ❖ Proximity (and frequency)
 - $\langle \text{keyword}, \{ (\text{doc-id}, (\text{position-of-occurrence}_1, \text{position-of-occurrence}_2, \dots)), (\text{doc-id}, (\text{position-of-occurrence}_1, \dots)), \dots \} \rangle$
 - When doing AND, check for positions that are near

Signature files

- ❖ Store the matrix by columns and compress them
- ❖ For each document, store a w -bit signature
- ❖ Each word is hashed into a w -bit value, with only $s < w$ bits turned on
- ❖ Signature is computed by taking the bit-wise OR of the hash values of all words on the document

$\text{hash}(\text{"database"}) = 0110$ Docs d_{x_1} contain
 $\text{hash}(\text{"dog"}) = 1100$ d_{x_1} contains "database": 0110 "database"?
 $\text{hash}(\text{"cat"}) = 0010$ d_{x_2} contains "dog": 1100
 d_{x_3} contains "cat" and "dog": 1110

- ❖ Some false positives; no false negatives
