Finding near-duplicate documents



A general paradigm:

- 1. Define function *f* capturing contents of each document in one number
 - "Hash function", "signature", "fingerprint"
- 2. Create $< f(doc_i)$, ID of doc_i> pairs
- 3. Sort the pairs
- 4. Recognize duplicate or near-duplicate documents as having the same *f* value or *f* values within a small threshold

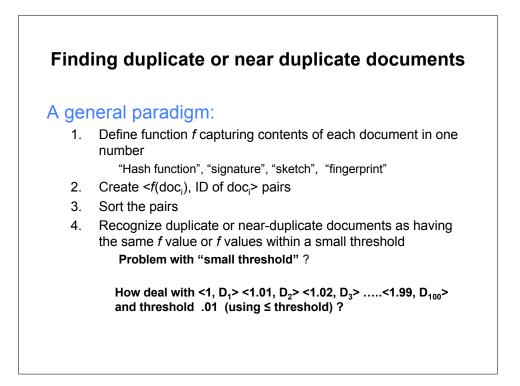
Compare: computing a similarity score on pairs of documents

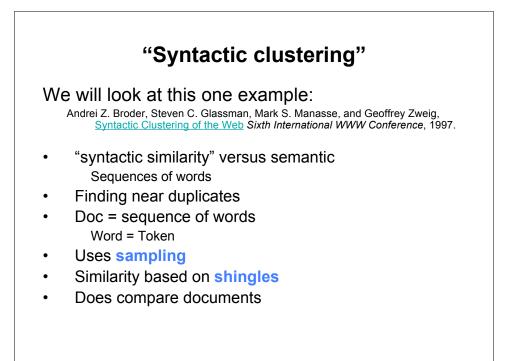
Finding duplicate or near duplicate documents

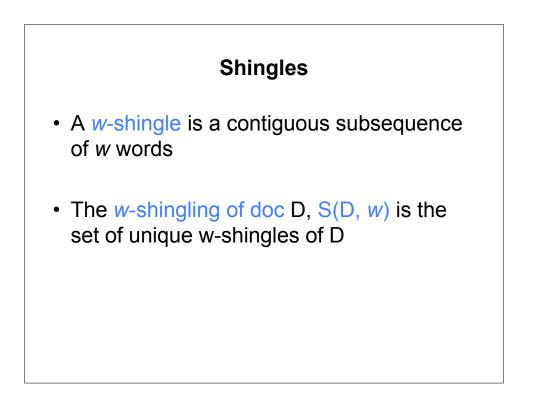
A general paradigm:

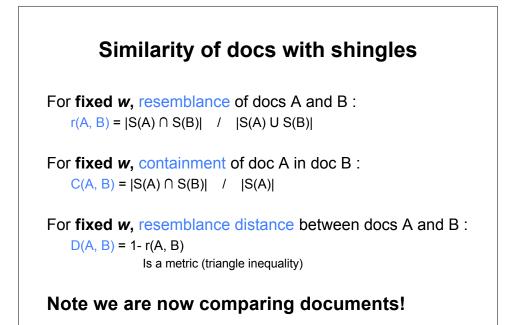
- 1. Define function *f* capturing contents of each document in one number
 - "Hash function", "signature", "sketch", "fingerprint"
- 2. Create $< f(doc_i)$, ID of doc_i> pairs
- 3. Sort the pairs
- 4. Recognize duplicate or near-duplicate documents as having the same *f* value or *f* values within a small threshold

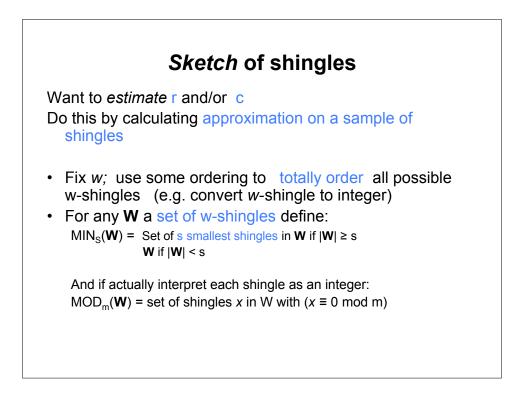
Problem with "small threshold" ?

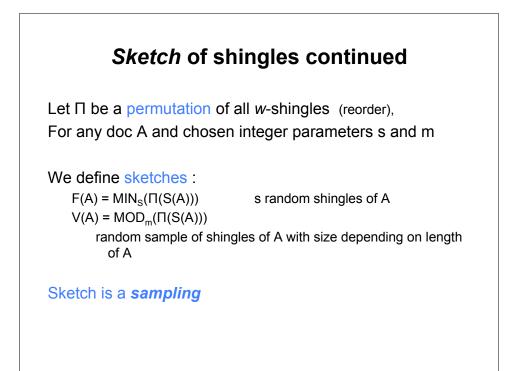


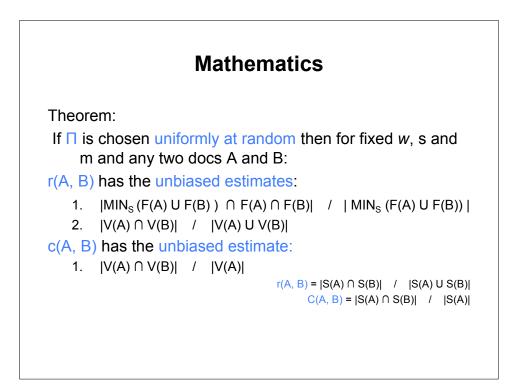


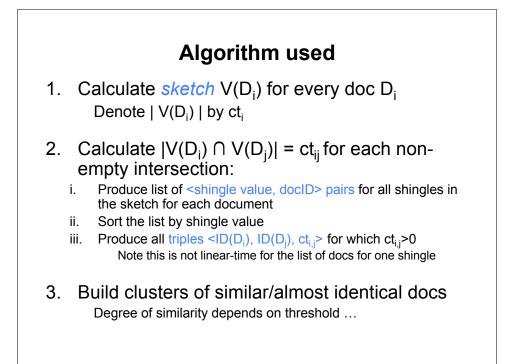


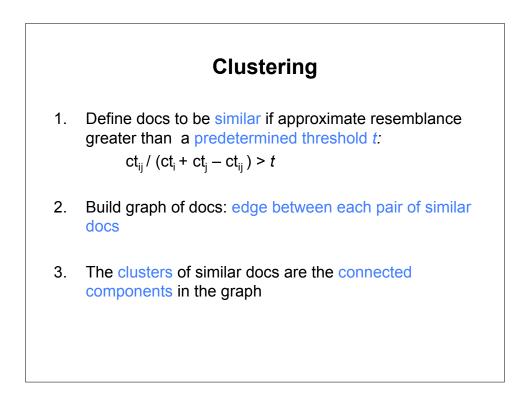












Paradigm?

- Does compare docs, so not same as paradigm we started with, but uses ideas
- Contents of doc captured by sketch a set of shingles (numbers)
- Similarity of docs scored by count of common shingles for docs
- Don't look at all doc pairs, look at all doc pairs that share a shingle
- Uses clustering by similarity threshold

