Clustering Outline

- Motivation
- Suffix Tree Clustering
- Offline evaluation
- Grouper I
- Grouper II
- Evaluation of deployed systems

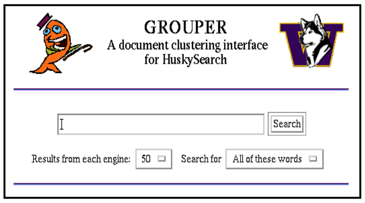
Low Quality of Web Searches

- System perspective:
  - small coverage of Web (<16%)
  - dead links and out of date pages
  - limited resources
- IR perspective (estimating relevancy of doc based on similarity to query):
  - very short queries
  - huge database
  - novice users

Document Clustering

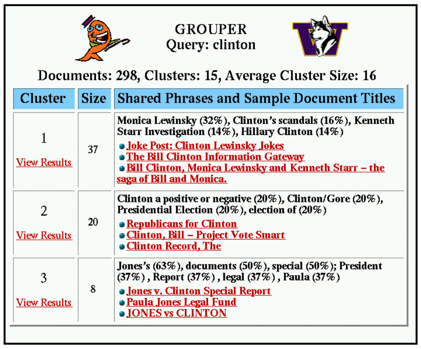
- User receives many (200 - 5000) documents from Web search engine
- Group documents in clusters - by topic
- Present clusters as interface

Grouper



GROUPER
A document clustering interface
for HuskySearch

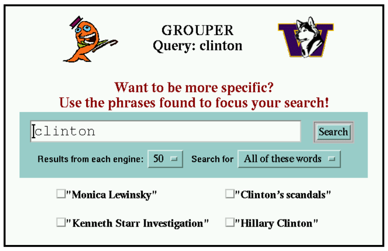
www.cs.washington.edu/research/clustering



GROUPER
Query: clinton

Documents: 298, Clusters: 15, Average Cluster Size: 16

Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results	37	Monica Lewinsky (32%), Clinton's scandals (16%), Kenneth Starr Investigation (14%), Hillary Clinton (14%) ● Joke Post: Clinton Lewinsky Jokes ● The Bill Clinton Information Gateway ● Bill Clinton, Monica Lewinsky and Kenneth Starr – the saga of Bill and Monica.
2 View Results	20	Clinton a positive or negative (20%), Clinton/Gore (20%), Presidential Election (20%), election of (20%) ● Republicans for Clinton ● Clinton, Bill – Project Vote Smart ● Clinton Record, The
3 View Results	8	Jones's (63%), documents (50%), special (50%); President (37%) , Report (37%) , legal (37%) , Paula (37%) ● Jones v. Clinton Special Report ● Paula Jones Legal Fund ● JONES vs CLINTON



GROUPER
Query: clinton

Want to be more specific?
Use the phrases found to focus your search!

clinton

Results from each engine: 50
Search for All of these words

□"Monica Lewinsky" □"Clinton's scandals"
□"Kenneth Starr Investigation" □"Hillary Clinton"

Desiderata

- Coherent clusters
- Speed
- Browsable clusters

Main Questions

- Is the automatic grouping of similar documents (document clustering) a feasible method of presenting the results of Web search engines?
- Will the use of phrases help in achieving high quality clusters? Can phrase-based clustering be done quickly?

Document Clustering Algorithms

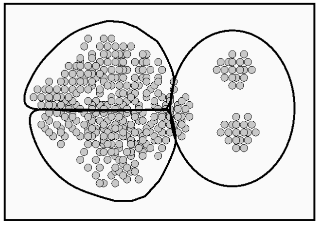
- Hierarchical Agglomerative Clustering:
  - O(n^2)
- Linear-time algorithms:
  - K-means (Rocchio, 66)
  - Single-Pass (Hill, 68)
  - Fractionation (Cutting et al, 92)
  - Buckshot (Cutting et al, 92)

Why Often Poor Results?

- model-based algorithms
- most work best for:
  - spherical clusters; equal size; few outliers
- text:
  - no model
  - not spherical; not equal size; overlap
- Web:
  - many outliers; lots of noise
- HAC often produce single large cluster

Example - Clusters of Varied Sizes

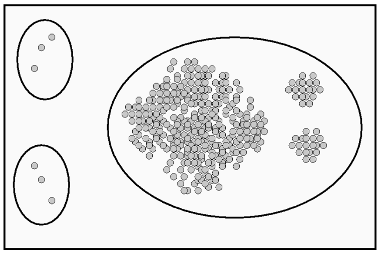
k-means; complete-link; group-average:



single-link: chaining, but succeed on example

Example - Outliers

HAC:



Suffix Tree Clustering
(KDD'97; SIGIR'98)

- Most clustering algorithms not unique for text:
  document as set of words
- STC:
  document as sequence of words

STC Characteristics

- Coherent
  - phrase-based
  - overlapping clusters
  - not model-based
- Speed and Scalability
  - linear time; incremental
- Browsable clusters
  - phrase-based
  - simple cluster definition

STC - Central Idea

- Identify ***base clusters*** - a group of documents that share a phrase - using a ***suffix tree***
- Merge base clusters to form clusters

STC - Outline

Three logical steps:

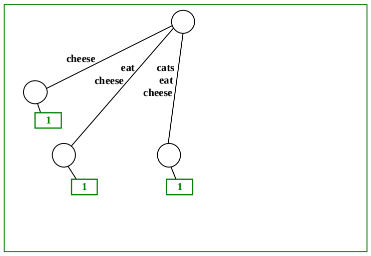
- "Clean" documents
- Use a ***suffix tree*** to identify ***base clusters*** - a group of documents that share a phrase
- Merge base clusters to form clusters

Step 1 - Document "Cleaning"

- Identify sentence boundaries
- Remove HTML tags, JavaScript, numbers, punctuation

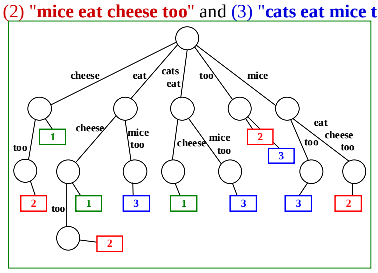
Suffix Tree
(Weiner, 73; Ukkonen, 95; Gusfield, 97)

Example - suffix tree of the string: (1) "**cats eat cheese**"



Suffix Tree (cont.)

Example - suffix tree of the strings: (1) "**cats eat cheese**", (2) "**mice eat cheese too**" and (3) "**cats eat mice too**"



Step 2 - Identifying Base Clusters via Suffix Tree

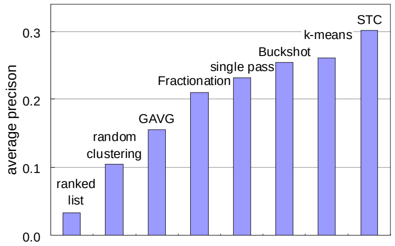
- Build one suffix tree from all sentences of all documents
- Suffix tree node = base cluster
- Score all nodes
- Traverse tree and collect top k (500) base clusters

Step 3 - Merging Base Clusters

- Motivation: similar documents share multiple phrases
- Merge base clusters based on the overlap of their document sets
- Example (query: "salsa")

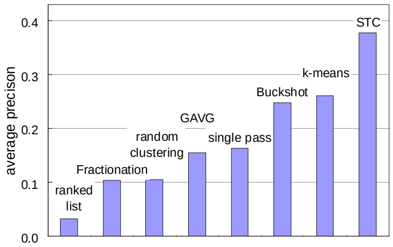
"*tabasco sauce*"	docs: **3,4,5,6**
"*hot pepper*"	docs: **1,3,5,6**
"*dance*"	docs: **1,2,7**
"*latin music*"	docs: **1,7,8**

Average Precision - WSR-SNIP



16% increase over k-means (not stat. sig.)

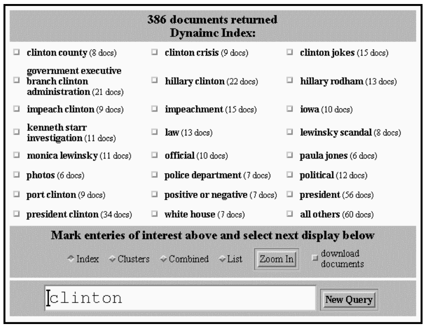
Average Precision - WSR-DOCS



45% increase over k-means (stat. sig.)

Grouper II

- Dynamic Index: non-merged based clusters
- Multiple interfaces: List, Clusters and Dynamic Index (key phrases)
- Hierarchical - interactive "Zoom In" feature (similar to Scatter/Gather)



386 documents returned
Dynaimc Index:

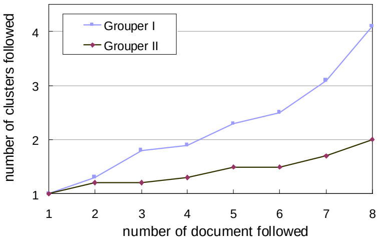
□ **clinton county** (8 docs)	□ **clinton crisis** (9 docs)	□ **clinton jokes** (15 docs)
□ **government executive branch clinton administration** (21 docs)	□ **hillary clinton** (22 docs)	□ **hillary rodham** (13 docs)
□ **impeach clinton** (9 docs)	□ **impeachment** (15 docs)	□ **iowa** (10 docs)
□ **kenneth starr investigation** (11 docs)	□ **law** (13 docs)	□ **lewinsky scandal** (8 docs)
□ **monica lewinsky** (11 docs)	□ **official** (10 docs)	□ **paula jones** (6 docs)
□ **photos** (6 docs)	□ **police department** (7 docs)	□ **political** (12 docs)
□ **port clinton** (9 docs)	□ **positive or negative** (7 docs)	□ **president** (56 docs)
□ **president clinton** (34 docs)	□ **white house** (7 docs)	□ **all others** (60 docs)

Mark enteries of interest above and select next display below

◇ Index ◇ Clusters ◇ Combined ◇ List [Zoom In] □ download documents

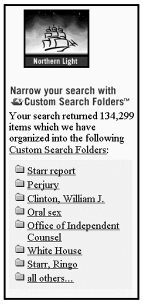
clinton [New Query]

Evaluation - Log Analysis



Northern Light

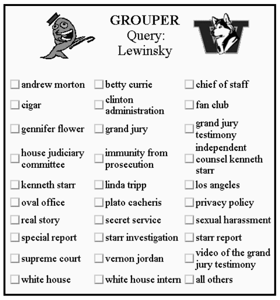
- "Custom Folders"
- 20000 predefined topics in a manually developed hierarchy
- Classify document into topics
- Display "dominant" topics in search results



Narrow your search with Custom Search Folders™

Your search returned 134,299 items which we have organized into the following Custom Search Folders:

- Starr report
- Perjury
- Clinton, William J.
- Oral sex
- Office of Independent Counsel
- White House
- Starr, Ringo
- all others...



GROUPER
Query:
Lewinsky

□ andrew morton	□ betty currie	□ chief of staff
□ cigar	□ clinton administration	□ fan club
□ gennifer flower	□ grand jury	□ grand jury testimony
□ house judiciary committee	□ immunity from prosecution	□ independent counsel kenneth starr
□ kenneth starr	□ linda tripp	□ los angeles
□ oval office	□ plato cacheris	□ privacy policy
□ real story	□ secret service	□ sexual harassment
□ special report	□ starr investigation	□ starr report
□ supreme court	□ vernon jordan	□ video of the grand jury testimony
□ white house	□ white house intern	□ all others

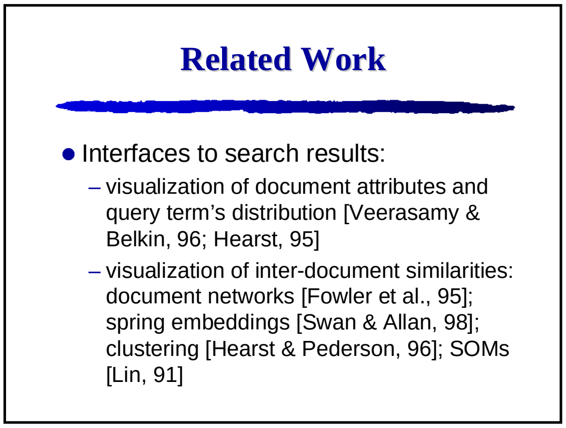
Summary

- Post-retrieval document clustering to address the low precision of Web searches
- STC - phrase-based; overlapping clusters; fast
- Offline evaluation - quality of STC, advantages of using phrases, compared to n-grams and FS
- Deployed two systems on the Web
- Log analysis: Promising initial results

www.cs.washington.edu/research/clustering

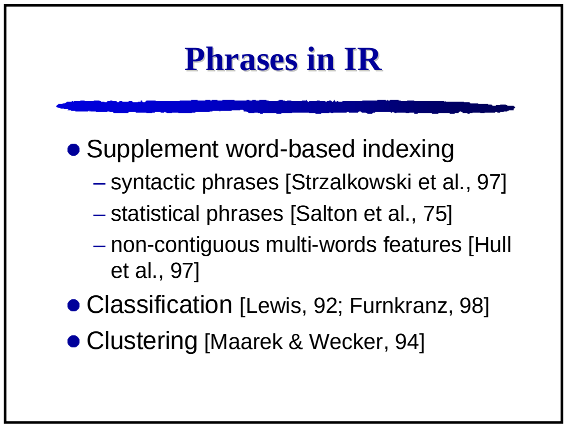
Related Work

- Increasing precision of Web searches
  - hyperlink structure: Google, Clever
  - popularity
- Helping users in low precision searches
  - sort by site, date
  - "Search Within": Infoseek, Lycos
  - "Similar Searches": IS, AV, Hotbot, Excite
  - "Find Similar": IS, Excite, Lycos
  - relate to predefined categories: Y!, IS, NL



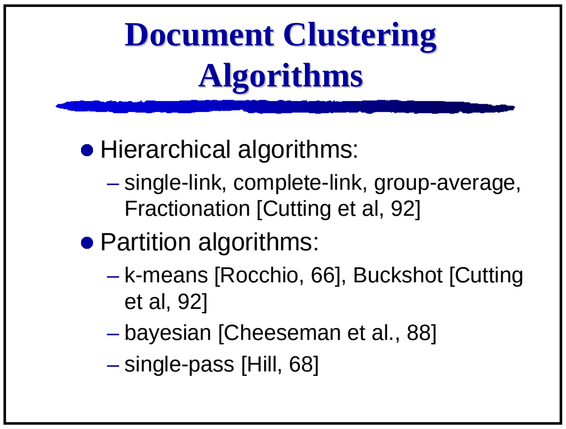
Related Work

- Interfaces to search results:
  - visualization of document attributes and query term’s distribution [Veerasamy & Belkin, 96; Hearst, 95]
  - visualization of inter-document similarities: document networks [Fowler et al., 95]; spring embeddings [Swan & Allan, 98]; clustering [Hearst & Pederson, 96]; SOMs [Lin, 91]



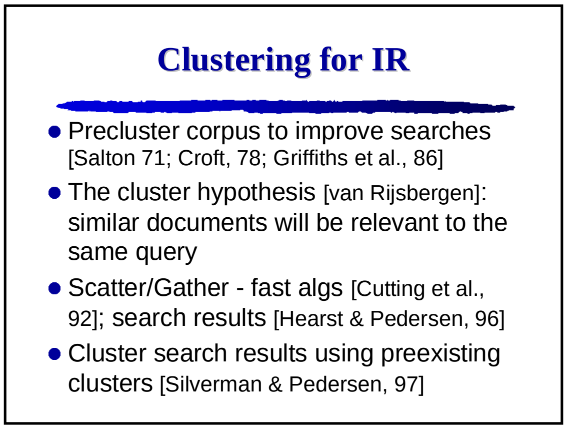
Phrases in IR

- Supplement word-based indexing
  - syntactic phrases [Strzalkowski et al., 97]
  - statistical phrases [Salton et al., 75]
  - non-contiguous multi-words features [Hull et al., 97]
- Classification [Lewis, 92; Furnkranz, 98]
- Clustering [Maarek & Wecker, 94]



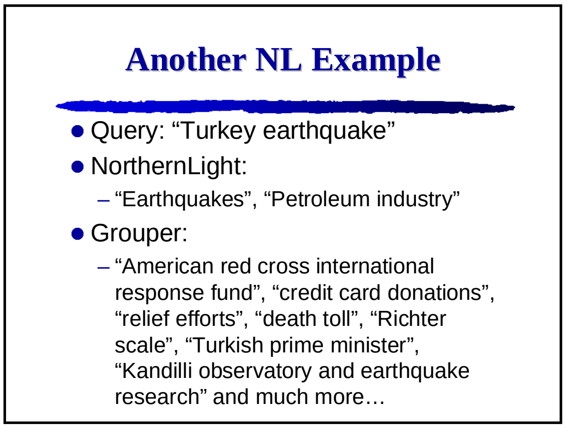
Document Clustering Algorithms

- Hierarchical algorithms:
  - single-link, complete-link, group-average, Fractionation [Cutting et al, 92]
- Partition algorithms:
  - k-means [Rocchio, 66], Buckshot [Cutting et al, 92]
  - bayesian [Cheeseman et al., 88]
  - single-pass [Hill, 68]



Clustering for IR

- Precluster corpus to improve searches [Salton 71; Croft, 78; Griffiths et al., 86]
- The cluster hypothesis [van Rijsbergen]: similar documents will be relevant to the same query
- Scatter/Gather - fast algs [Cutting et al., 92]; search results [Hearst & Pedersen, 96]
- Cluster search results using preexisting clusters [Silverman & Pedersen, 97]



Another NL Example

- Query: “Turkey earthquake”
- NorthernLight:
  - “Earthquakes”, “Petroleum industry”
- Grouper:
  - “American red cross international response fund”, “credit card donations”, “relief efforts”, “death toll”, “Richter scale”, “Turkish prime minister”, “Kandilli observatory and earthquake research” and much more…