

Clustering (Search Engine Results)

CSE 454

Clustering Outline

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

© 2000-2005 Etzioni & Weld

Low Quality of Web Searches

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 - dead links and out of date pages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

© 2000-2005 Etzioni & Weld

Document Clustering

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

© 2000-2005 Etzioni & Weld

Grouper

GROUPER
A document clustering interface for HuskySearch

Search

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

www.cs.washington.edu/research/clustering

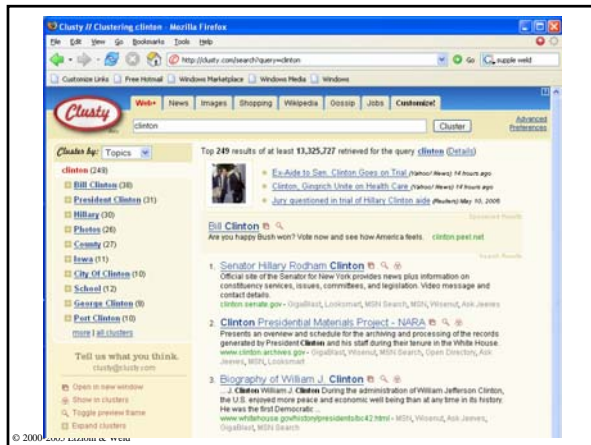
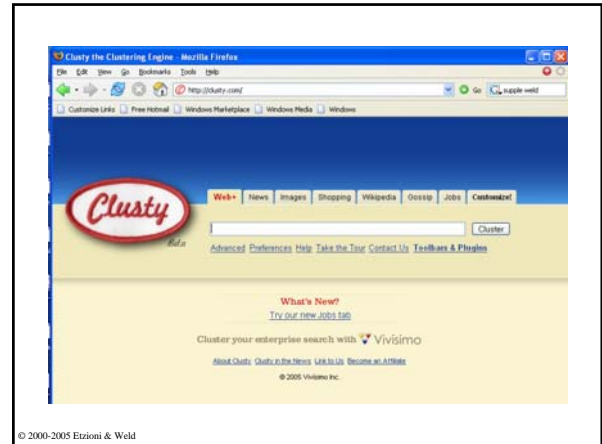
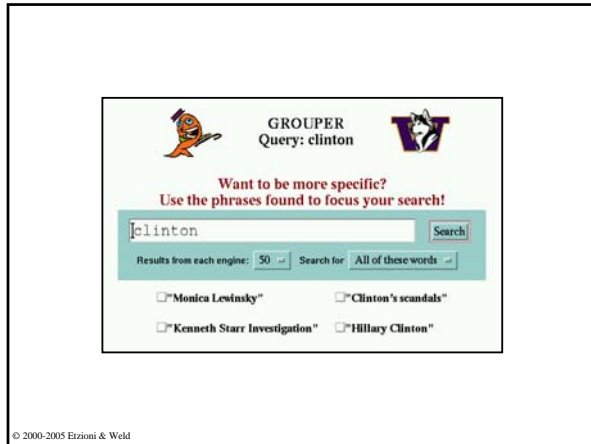
© 2000-2005 Etzioni & Weld

GROUPER
Query: clinton

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

Cluster	Size	Shared Phrases and Sample Document Titles
1	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	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	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

© 2000-2005 Etzioni & Weld



Desiderata

- Coherent cluster
- Speed
- Browsible clusters
 - Naming

© 2000-2005 Etzioni & Weld

Main Questions

- Is **document clustering** feasible for Web search engines?
- Will the use of **phrases** help in achieving high quality clusters?
- Can phrase-based clustering be done **quickly**?

© 2000-2005 Etzioni & Weld

1. Clustering

group together similar items
(words or documents)

The diagram shows three distinct groups of blue dots. Each group is enclosed in a green circle, representing a cluster. The dots are arranged in a way that suggests similarity within each group and dissimilarity between groups.

© 2000-2005 Etzioni & Weld

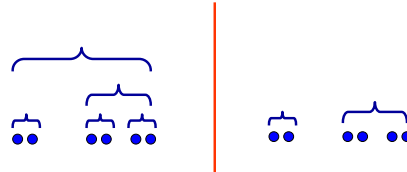
Clustering Algorithms

- 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)

© 2000-2005 Etzioni & Weld

Basic Concepts - 1

- Hierarchical vs. Flat



© 2000-2005 Etzioni & Weld

Basic Concepts - 2

- hard clustering:
 - each item in only one cluster
- soft clustering:
 - each item has a probability of membership in each cluster
- disjunctive / overlapping clustering:
 - an item can be in more than one cluster

© 2000-2005 Etzioni & Weld

Basic Concepts - 3

distance / similarity function
(for documents)

- dot product of vectors
- number of common terms
- co-citations
- access statistics
- share common phrases

© 2000-2005 Etzioni & Weld

Basic Concepts - 4

- What is “right” number of clusters?
 - apriori knowledge
 - default value: “5”
 - clusters up to 20% of collection size
 - choose best based on external criteria
 - Minimum Description Length
 - Global Quality Function
- no good answer

© 2000-2005 Etzioni & Weld

Hierarchical Clustering

- Agglomerative
 - bottom-up

Initialize: - *each item a cluster*

Iterate: - *select two most similar clusters*
- *merge them*

Halt: *when have required # of clusters*

© 2000-2005 Etzioni & Weld

Hierarchical Clustering

- Divisive
 - top-bottom

Initialize: -all items one cluster

Iterate: - select a cluster (least *coherent*)
- divide it into two clusters

Halt: when have *required # of clusters*

© 2000-2005 Erzioni & Weld

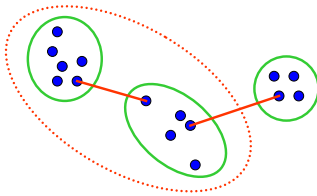
HAC Similarity Measures

- Single link
- Complete link
- Group average
- Ward's method

© 2000-2005 Erzioni & Weld

Single Link

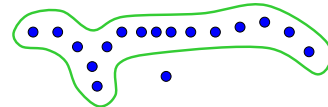
- cluster similarity = similarity of two *most* similar members



© 2000-2005 Erzioni & Weld

Single Link

- $O(n^2)$
- chaining:

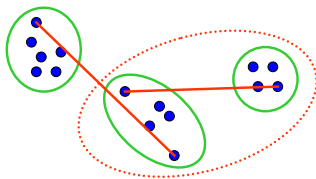


- bottom line:
 - simple, fast
 - often low quality

© 2000-2005 Erzioni & Weld

Complete Link

- cluster similarity = similarity of two *least* similar members



© 2000-2005 Erzioni & Weld

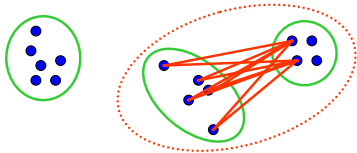
Complete Link

- worst case $O(n^3)$
- fast algo requires $O(n^2)$ space
- no chaining
- bottom line:
 - typically much faster than $O(n^3)$,
 - often good quality

© 2000-2005 Erzioni & Weld

Group Average

- cluster similarity
= average similarity of all pairs



© 2000-2005 Etzioni & Weld

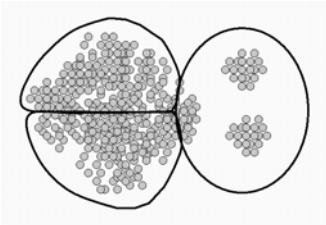
HAC Often Poor Results - Why?

- Often produces single large cluster
- Work best for:
 - spherical clusters; equal size; few outliers
- Text documents:
 - no model
 - not spherical; not equal size; overlap
- Web:
 - many outliers; lots of noise

© 2000-2005 Etzioni & Weld

Example: Clusters of Varied Sizes

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

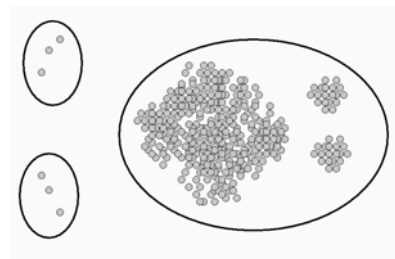


single-link: chaining,
but succeeds on this example

© 2000-2005 Etzioni & Weld

Example - Outliers

HAC:



© 2000-2005 Etzioni & Weld

Suffix Tree Clustering

(KDD'97; SIGIR'98)

- Most clustering algorithms aren't **specialized** for text:
Model document as **set** of words
- STC:
document = **sequence** of words

© 2000-2005 Etzioni & Weld

STC Characteristics

- Coherent
 - phrase-based
 - overlapping clusters
- Speed and Scalability
 - linear time; incremental
- Browsible clusters
 - phrase-based
 - simple cluster definition

© 2000-2005 Etzioni & Weld

STC - Central Idea

- Identify **base clusters**
 - a group of documents that share a phrase
 - use a **suffix tree**
- Merge base clusters as needed

© 2000-2005 Etzioni & Weld

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

© 2000-2005 Etzioni & Weld

Step 1 - Document “Cleaning”

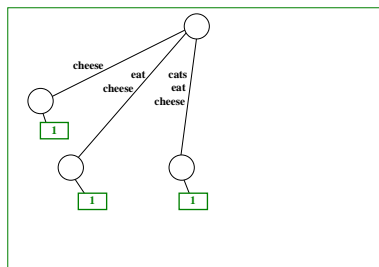
- Identify sentence boundaries
- Remove
 - HTML tags,
 - JavaScript,
 - Numbers,
 - Punctuation

© 2000-2005 Etzioni & Weld

Suffix Tree

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

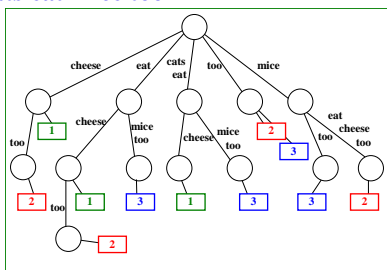
Example - suffix tree of the string: (1)
"cats eat cheese"



© 2000-2005 Etzioni & Weld

Example - suffix tree of the strings:

- (1) "cats eat cheese",
- (2) "mice eat cheese too" and
- (3) "cats eat mice too"



© 2000-2005 Etzioni & Weld

Step 2 - Identify 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

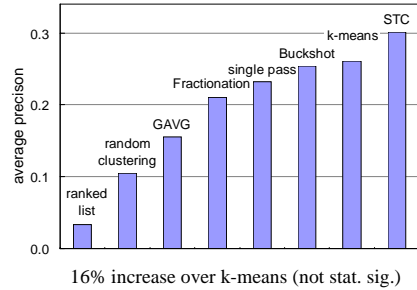
© 2000-2005 Etzioni & Weld

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

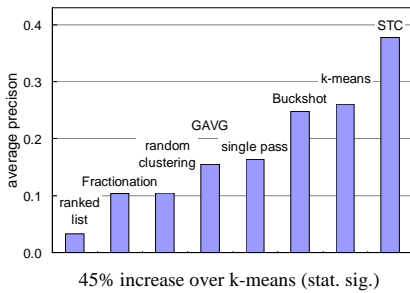
© 2000-2005 Etzioni & Weld

Average Precision - WSR-SNIP



© 2000-2005 Etzioni & Weld

Average Precision - WSR-DOCS



© 2000-2005 Etzioni & Weld

Grouper II

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

© 2000-2005 Etzioni & Weld

386 documents returned
Dynamic Index:

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

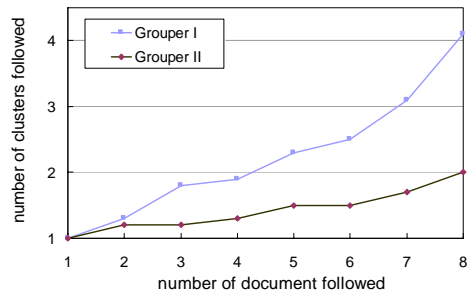
Mark entries of interest above and select next display below

Index
 Clusters
 Combined
 List

 download documents

© 2000-2005 Etzioni & Weld

Evaluation - Log Analysis

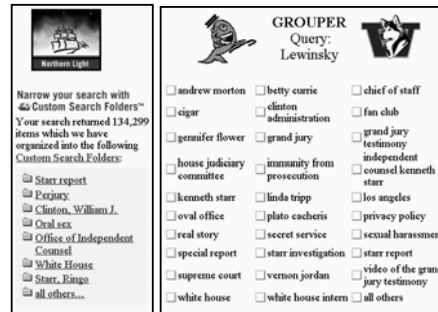


© 2000-2005 Etzioni & Weld

Northern Light

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

© 2000-2005 Etzioni & Weld



The screenshot shows the Northern Light search interface. On the left, there is a search box with the text "Northern Light" and a search button. Below the search box, it says "Narrow your search with 40 Custom Search Folders™" and "Your search returned 134,299 items which we have organized into the following Custom Search Folders:". A list of folders is shown, including "Starr report", "Perjury", "Clinton, William J.", "Oral sex", "Office of Independent Counsel", "White House", "Starr, Einge", and "all others...". On the right, there is a "GROUPER" section with the query "Lewinsky" and a list of related terms, each with a checkbox. The terms include "andrew morton", "betty currie", "chief of staff", "cigar", "clinton administration", "fm club", "genieifer flower", "grand jury", "grand jury testimony", "house judiciary committee", "immunity from prosecution", "independent counsel kenneth star", "kenneth star", "linda tripp", "los angeles", "oval office", "plato eacheris", "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", and "all others".

© 2000-2005 Etzioni & Weld

Summary

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

www.cs.washington.edu/research/clustering

© 2000-2005 Etzioni & Weld