Clustering (Search Engine Results)

CSE 454

Motivation

Document 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- dead links and out of date pages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
(relevancy of doc ~ 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

www.cs.washington.edu/research/clustering

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Shared Phrases and Sample Document Titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>&quot;Clinton's personal rape - 1972&quot;, &quot;Hillary Clinton investigation&quot;</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>&quot;Bill Clinton's impeachment&quot;, &quot;Clinton impeachment&quot;</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>&quot;Bush's 2004 campaign&quot;, &quot;Bush's foreign policy&quot;</td>
</tr>
</tbody>
</table>
Desiderata

- Coherent cluster
- Speed
- Browsable clusters
  - Naming

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?
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)

Basic Concepts - 1

- Hierarchical vs. Flat

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

Basic Concepts - 3

- Distance / similarity function (for documents)
  - dot product of vectors
  - number of common terms
  - co-citations
  - access statistics
  - share common phrases

Basic Concepts - 4

- What is “right” number of clusters?
  - a priori 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

Hierarchical Clustering

- Agglomerative
  - bottom-up

Initialize: each item a cluster
Iterate: select two most similar clusters
  - merge them
Halt: when have required # of clusters
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

HAC Similarity Measures

- **Single link**
- **Complete link**
- **Group average**
- **Ward’s method**

### Single Link

- cluster similarity = similarity of two most similar members

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

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

### Complete Link

- cluster similarity = similarity of two least similar members

- 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
**Group Average**

- cluster similarity
  = average similarity of all pairs

**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

**Example: Clusters of Varied Sizes**

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

- single-link: chaining,
  but succeeds on this example

**Example - Outliers**

HAC:

**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

**STC Characteristics**

- Coherent
  - phrase-based
  - overlapping clusters
- 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
  - use a **suffix tree**
- Merge base clusters as needed

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

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

Example - suffix tree of the strings:

1) "cats eat cheese"
2) "mice eat cheese too" and
3) "cats eat mice too"

Suffix Tree

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

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

**Average Precision - WSR-DOCS**

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

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)

**Evaluation - Log Analysis**

- Number of clusters followed
- Number of documents followed
Northern Light

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

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