CS276B

Text Information Retrieval, Mining, and Exploitation

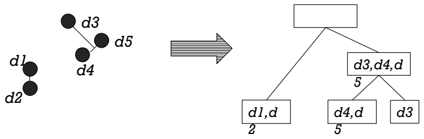
Lecture 3

Recap: Agglomerative clustering

- Given target number of clusters *k*.
- Initially, each doc viewed as a cluster
  - start with *n* clusters;
- Repeat:
  - **while** there are > *k* clusters, find the "closest pair" of clusters and merge them.

Recap: Hierarchical clustering

- As clusters *agglomerate*, docs likely to fall into a hieararchy of "topics" or concepts.



Recap: *k*-means basic iteration

- At the start of the iteration, we have *k* centroids.
- Each doc assigned to the nearest centroid.
- All docs assigned to the same centroid are averaged to compute a new centroid;
  - thus have *k* new centroids.

Recap: issues/applications

- Term vs. document space clustering
- Multi-lingual docs
- Feature selection
- Speeding up scoring
- Building navigation structures
  - "Automatic taxonomy induction"
- Labeling

Today's Topics

Clustering as dimensionality reduction

Evaluation of text clustering

Link-based clustering

Enumerative clustering/trawling

Clustering as dimensionality reduction

- Clustering can be viewed as a form of data compression
  - the given data is recast as consisting of a “small” number of clusters
  - each cluster typified by its representative “centroid”
- Recall LSI from CS276a
  - extracts “principal components” of data
    - attributes that best explain segmentation
  - ignores features of either
    - low statistical presence, or
    - low discriminating power

Simplistic example

- Clustering may suggest that a corpus consists of two clusters
  - one dominated by terms like *quark* and *energy*
  - the other by *disk* and stem-variants of *process*
- Dimensionality reduction likely to find linear combinations of these as principal axes
- See work by Azar et al.
  - (resources end of lecture)

Dimensionality reduction

- Clustering is not intrinsically linear algebraic
- Dimensionality reduction doesn’t have to be, either
  - which factors explain the data at hand?
  - probabilistic versions studied extensively
- Ongoing research area

Evaluation of clustering

- Perhaps the most substantive issue in data mining in general:
  - <u>how do you measure goodness?</u>
- Most measures focus on computational efficiency
  - ok when this is the goal - cosine scoring in search
  - presumption that search results close to those without clustering
    - in practice of course there are tradeoffs

Approaches to evaluating

- Anecdotal
- User inspection
- Ground “truth” comparison
  - Cluster retrieval
- Purely quantitative measures
  - Probability of generating clusters found
  - Average distance between cluster members
- Microeconomic

Anecdotal evaluation

- Probably the commonest (and surely the easiest)
  - “I wrote this clustering algorithm and look what it found!”
- No benchmarks of the form “Corpus plus the <u>useful</u> things that clustering should find”
- Almost always will pick up the easy stuff like partition by languages
- Generally, unclear scientific value.

User inspection

- Induce a set of clusters or a navigation tree
- Have subject matter experts evaluate the results and score them
  - some degree of subjectivity
- Often combined with search results clustering
- Not clear how reproducible across tests.

Ground "truth" comparison

- Take a union of docs from a taxonomy
  - Yahoo!, ODP, newspaper sections …
- Compare clustering results to prior
  - e.g., 80% of the clusters found map "cleanly" to taxonomy nodes
- But is it the "right" answer?
- For the docs given, the static prior taxonomy may be wrong in places
  - the clustering algorithm may have gotten right things not in the static taxonomy

"Subjective"

Ground truth comparison

- Divergent goals
- Static taxonomy designed to be the "right" navigation structure
  - somewhat independent of corpus at hand
- Clusters found have to do with vagaries of corpus
- Also, docs put in a taxonomy node may not be the most representative ones for that topic
  - cf Yahoo!

Microeconomic viewpoint

- Anything - including clustering - is only as good as the economic utility it provides
- For clustering: net economic gain produced by an approach (vs. another approach)
- Strive for a concrete optimization problem
  - will see later how this makes clean sense for clustering in recommendation systems

Microeconomic view

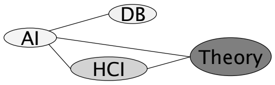
- This purist view can in some settings be simplified into concrete measurements, e.g.,
- Wall-clock time for users to satisfy specific information needs
  - people-intensive to perform significant studies
  - if every clustering paper were to do these …

Cluster retrieval

- Cluster docs in a corpus first
- For retrieval, find cluster nearest to query
  - retrieve only docs from it
- How do various clustering methods affect the quality of what's retrieved?
- Concrete measure of quality:
  - Precision as measured by user judgements for these queries
- Done with TREC queries
  - (see Shütze and Silverstein reference)

Topic segmentation

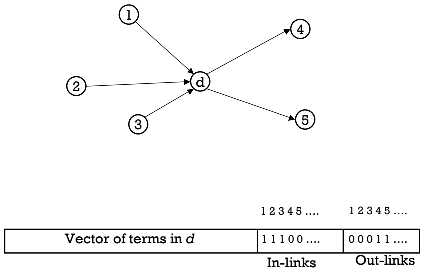
- P2P networks
  - content distributed amongst nodes
  - searches broadcast through neighbors
  - wasteful if you want high recall
- Cluster nodes with similar content
  - send queries only to germane “regions”
  - measure recall at a given level of traffic



Link-based clustering

- Given docs in hypertext, cluster into *k* groups.
- Back to vector spaces!
- Set up as a vector space, with axes for terms *and* for in- and out-neighbors.

Example



Clustering

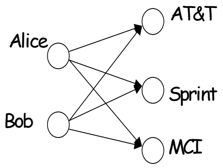
- Given vector space representation, run any of the previous clustering algorithms from.
- Studies done on web search results, patents, citation structures - some basic cues on which features help.

Back up

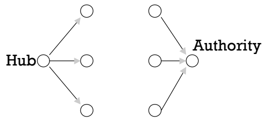
- In clustering, we partition input docs into clusters.
- In *trawling*, we’ll enumerate subsets of the corpus that “look related”
  - each subset a topically-focused community
  - will discard lots of docs
- Twist: will use purely link-based cues to decide whether docs are related.

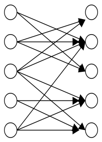
Trawling/enumerative clustering

- In hyperlinked corpora - here, the web
- Look for all occurrences of a linkage pattern
- Recall from hubs/authorities search algorithm in CS276a:



Insights from hubs

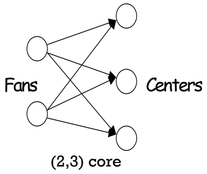




Link-based hypothesis:
Dense bipartite subgraph ⇔ Web community.

Communities from links

- Issues:
- Size of the web is huge - not the stuff clustering algorithms are made for
- What is a "dense subgraph"?
  - Define *(i,j)*-core: complete bipartite subgraph with *i* nodes all of which point to each of *j* others.



Random graphs inspiration

- Why cores rather than dense subgraphs?
  - hard to get your hands on dense subgraphs
- Every large enough dense bipartite graph almost surely has "non-trivial" core, e.g.,:
  - large: *i*=3 and *j*=10
  - dense: 50% edges
  - almost surely: 90% chance
  - non-trivial: *i*=3 and *j*=3.

Approach

- Find all (*i,j*)-cores
  - currently feasible ranges like 3 ≤ *i,j* ≤ 20.
- Expand each core into its full community.
- Main memory conservation
- Few disk passes over data

Finding cores

- "SQL" solution: find all triples of pages such that intersection of their outlinks is at least 3?
  - Too expensive.
- Iterative pruning techniques work in practice.

Initial data & preprocessing

- Eliminate mirrors
- Represent URLs by 2 ×32 = 64-bit hash
- Can sort URL's by either source or destination using disk-run sorting

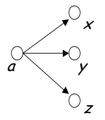
Pruning overview

- Simple iterative pruning
  - eliminates obvious non-participants
  - no cores output
- Elimination-generation pruning
  - eliminates some pages
  - generates some cores
- Finish off with “standard data mining” algorithms

Simple iterative pruning

- Discard all pages of
  - in-degree < *i* or
  - out-degree < *j*.
- Repeat Why?
- Reduces to a sequence of sorting operations on the edge list Why?

Elimination/generation pruning



a is part of a (*3, 3*) core if and only if the intersection of inlinks of *x, y,* and *z* is at least *3*

- pick a node *a* of degree 3
- for each *a* output neighbors *x*, *y*, *z*
- use an index on centers to output in-links of *x*, *y*, *z*
- intersect to decide if *a* is a fan
- at each step, either eliminate a page (*a*) or generate a core

Exercise

- Work through the details of maintaining the index on centers to speed up elimination-generation pruning.

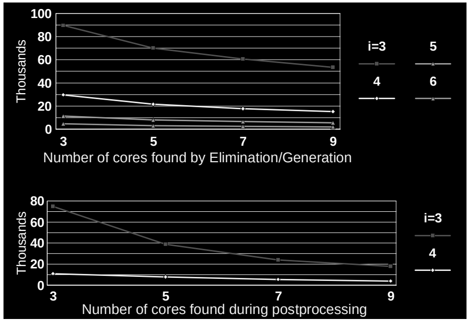
Results after pruning

- Typical numbers from late 1990’s web:
- Elimination/generation pruning yields >100K non-overlapping cores for *i,j* between 3 and 20.
- Left with a few (5-10) million unpruned edges What’s this?
  - small enough for postprocessing by *a priori* algorithm
  - build (*i+1*, *j*) cores from (*i*, *j*) cores.

Exercise

- Adapt the *a priori* algorithm to enumerating bipartite cores.

Results for cores



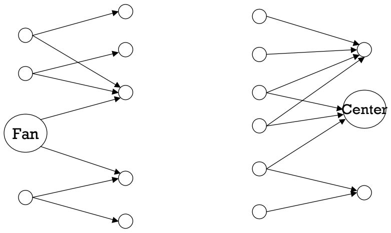
Sample cores

- hotels in Costa Rica
- clipart
- Turkish student associations
- oil spills off the coast of Japan
- Australian fire brigades
- aviation/aircraft vendors
- guitar manufacturers

From cores to communities

- Want to go from bipartite core to “dense bipartite graph” surrounding it
- Use hubs/authorities algorithm (CS276a) without text query - use fans/centers as samples
- Augment core with
  - all pages pointed to by any fan
    - all pages pointing into these
  - all pages pointing into any center
    - all pages pointed to by any of these
- Use induced graph as the base set in the hubs/authorities algorithm.

Using sample hubs/authorities



Costa Rican hotels and travel

- The Costa Rica Inte...ion on arts, busi...
- Informatica Interna...rvices in Costa Rica
- Cocos Island Research Center
- Aero Costa Rica
- Hotel Tilawa - Home Page
- COSTA RICA BY INTER@MERICA
- tamarindo.com
- Costa Rica
- New Page 5
- The Costa Rica Internet Directory.
- Costa Rica, Zarpe Travel and Casa Maria
- Si Como No Resort Hotels & Villas
- Apartotel El Sesteo... de San José, Cos...
- Spanish Abroad, Inc. Home Page
- Costa Rica's Pura V...ry - Reservation ...
- YELLOW\RESPALDO\HOTELES\Orquide1
- Costa Rica - Summary Profile
- COST RICA, MANUEL A...EPOS: VILLA
- Hotels and Travel in Costa Rica
- Nosara Hotels & Res...els &
- Restaurants...
- Costa Rica Travel, Tourism &
- Resorts
- Association Civica de Nosara
- Untitled: http://www...ca/hotels/mimos.html
- Costa Rica, Healthy...t Pura Vida
- Domestic & International Airline
- HOTELES / HOTELS - COSTA RICA
- tourgems
- Hotel Tilawa - Links
- Costa Rica Hotels T...On line
- Reservations
- Yellow pages Costa ...Rica Export
- INFOHUB Costa Rica Travel Guide
- Hotel Parador, Manuel Antonio, Costa Rica
- Destinations

Open research in clustering

- “Classic” open problems:
  - Feature selection
  - Efficiency
    - tradeoffs of efficiency for quality
  - Labeling clusters/hierarchies
- Newer open problems:
  - How do you measure clustering goodness?
  - How do you organize clusters into a navigation paradigm? Visualize?
  - What other ways are there of exploiting links?
  - How do you track temporal drift of cluster topics?

Resources

- A priori algorithm:
  - Mining Association Rules between Sets of Items in Large Databases: Agrawal, Imielinski, Swami. *http://citeseer.nj.nec.com/agrawal93mining.html*
  - R. Agrawal, R. Srikant. Fast algorithms for mining association rules.

http://citeseer.nj.nec.com/agrawal94fast.html

- Spectral Analysis of Data (2000): Y. Azar, A. Fiat, A. Karlin, F. McSherry, J. Saia. *http://citeseer.nj.nec.com/azar00spectral.html*
- Hypertext clustering: D.S. Modha, W.S. Spangler. Clustering hypertext with applications to web searching.

http://citeseer.nj.nec.com/272770.html

- Trawling: S. Ravi Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. Trawling emerging cyber-communities automatically.

http://citeseer.nj.nec.com/context/843212/0

- H. Schütze, C. Silverstein. Projections for Efficient Document Clustering (1997). http://citeseer.nj.nec.com/76529.html