CS347

Lecture 11

May 16, 2001

Topics

Link-based clustering

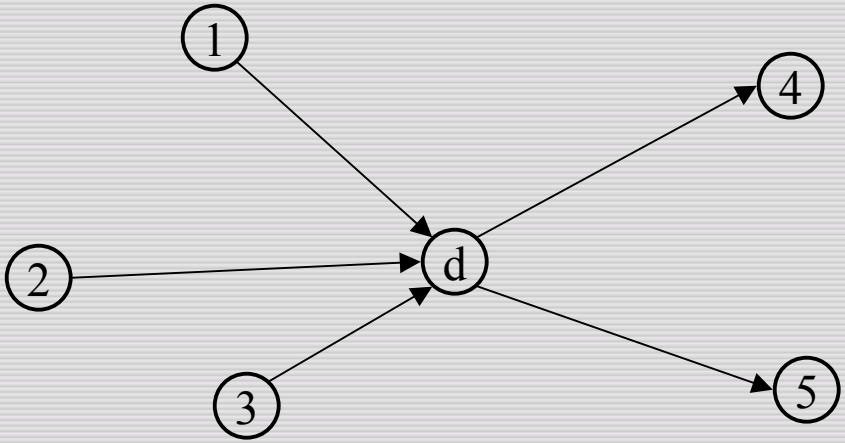
Enumerative clustering/trawling

Recommendation systems

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 as well as for in- and out-neighbors.

Example



Vector of terms in *d*	1 2 3 4 5 ….	1 2 3 4 5 ….
	1 1 1 0 0 ….	0 0 0 1 1 ….
	In-links	Out-links

Clustering

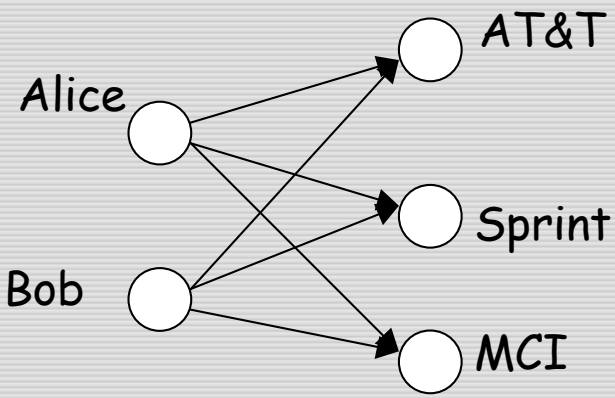
- Given vector space representation, run any of the clustering algorithms from lecture 8.
- Has been implemented on web search results.
- Other corpora: patents, citation structures.

Back up

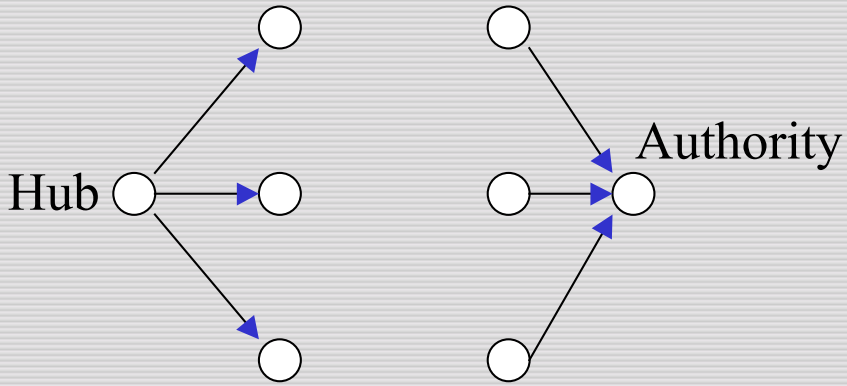
- In clustering, we partition input docs into clusters.
- In *trawling*, we'll enumerate subsets of the corpus that "look related"
  - will discard lots of docs
- Twist: will use purely link-based cues to decide whether docs are related

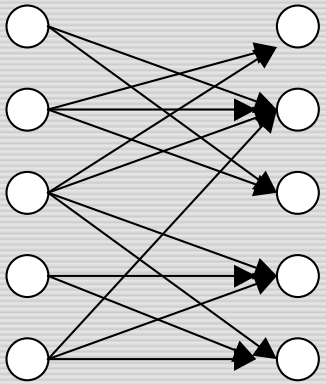
Trawling/enumerative clustering

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



Insights from hubs



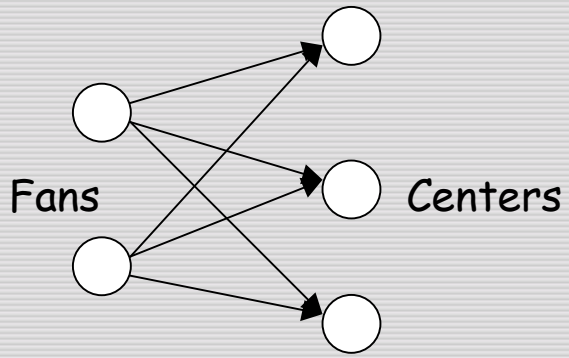


Link-based hypothesis:

Dense bipartite subgraph $\Rightarrow$ web community.

Communities from cores

- not easy, since web is huge
- what is a "dense subgraph"?
- define (*i*,*j*)-core: complete bipartite subgraph with *i* nodes all of which point to each of *j* others



(2,3) core

Random graphs inspiration

Every “large” enough “dense” bipartite graph “almost surely” has “non-trivial” core

e.g.,:

large = 3 by 10

dense = 50% edges

almost surely = 90% chance

non-trivial = 3 by 3

Approach

- Find all (i,j)-cores ($3 \leq i \leq 10$, $3 \leq j \leq 20$).
- Expand each core into its full community.

Finding cores

- “SQL” solution: find all triples of pages such that intersection of their outlinks is at least 3? Too expensive.
- Iterative pruning techniques actually work!

Initial data & preprocessing

- Crawl, then extract links
- Work with potential fans:

  nodes with $\geq j$ non-nepotistic links
- Eliminate mirrors
- Represent URLs by $2 \times 32 = 64$-bit hash
- Can sort URL's by either source or destination using disk-run sorting

Popular page elimination

- Don’t want “popular” communities (Yahoo!, Excite, DejaNews, webrings, …)
- Popular community has popular page(s)
- Define popular page: indegree ≥50

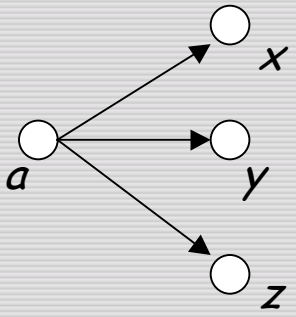
Main requirements

- Main memory conservation
- Few disk passes over data

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 <u>eliminate</u> a page (*a*) or <u>generate</u> a core

Exercise

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

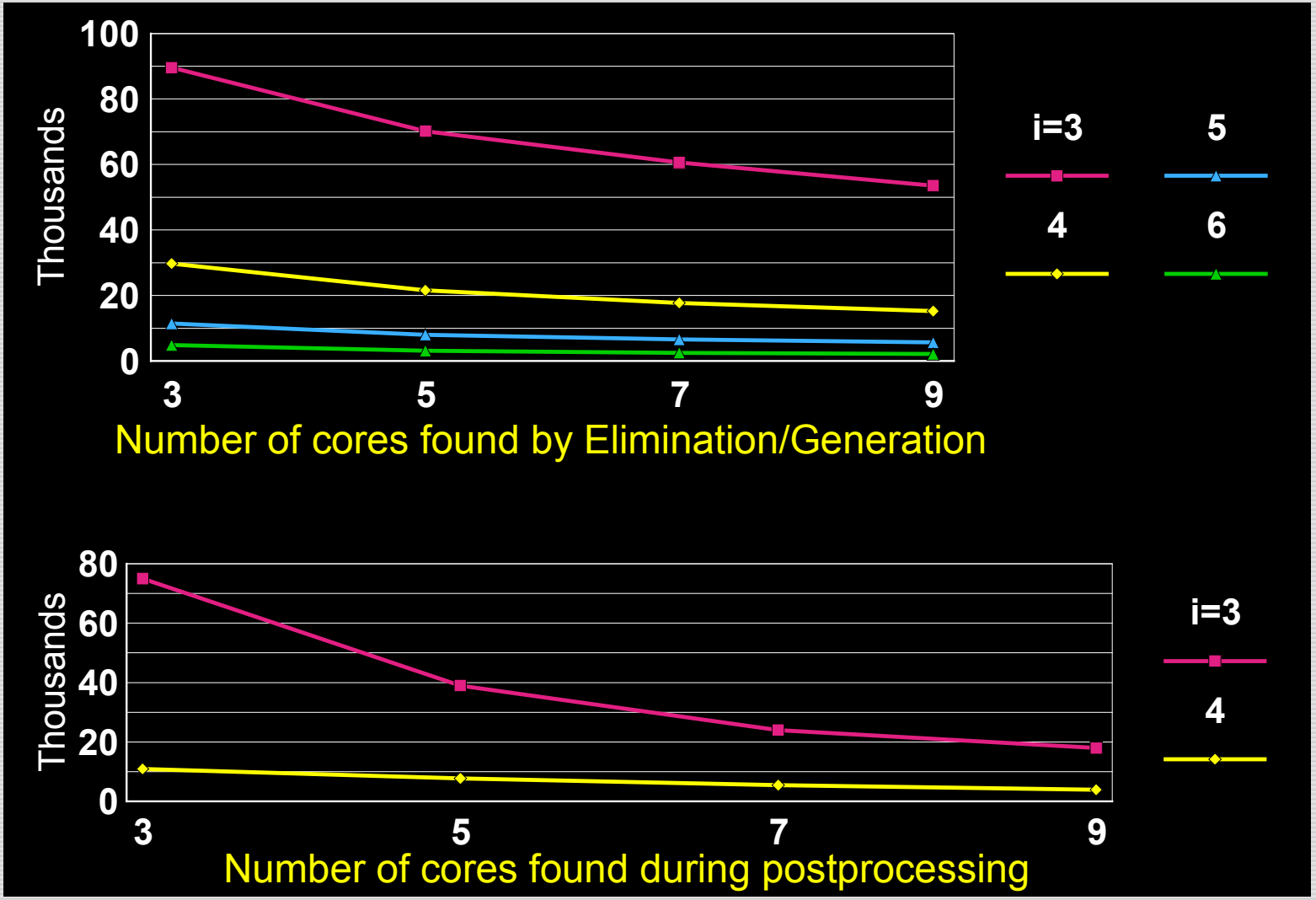
Results after pruning

- Elimination/generation pruning yields >100K non-overlapping cores for small *i,j*.
- 5M unpruned edges
  - small enough for post-processing by *a priori*
  - build (*i+1, j*) cores from (*i, j*) cores

Exercise

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

Results for cores



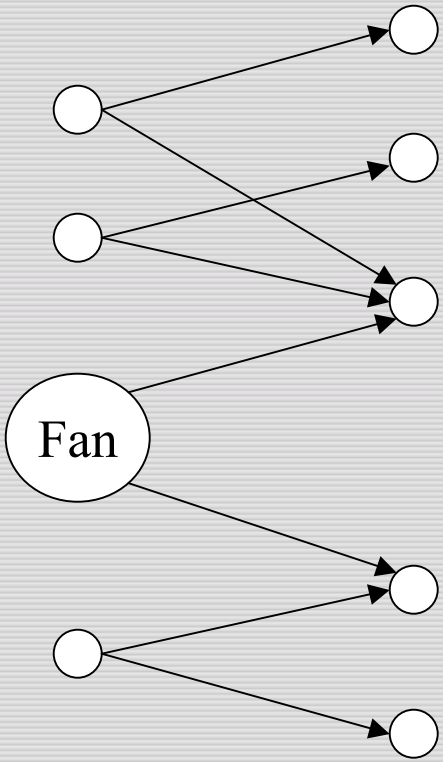
Sample cores

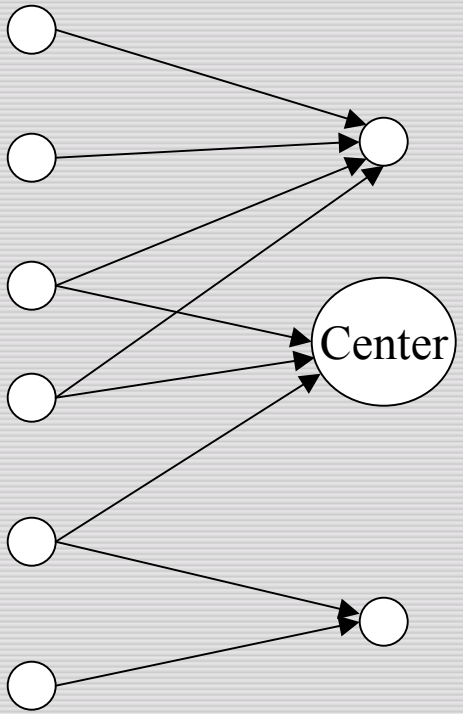
- hotels in Costa Rica
- clipart
- Turkish student associations
- oil spills off the coast of Japan
- Australian fire brigades
- aviation/aircraft vendors
- guitar manufacturers

From cores to communities

- Use hubs/authorities algorithm 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

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

Muslim student orgs.

- USC Muslim Students...ation Islamic Server
- The University of O...a Domain Name Change
- Caltech Muslim Students Home Page
- Islamic Society of Stanford University
- University of Texas...nformation Center...
- CSUN Muslim Students Association homepage
- HUDA
- Islamic Gateway
- Muslim Students' As...iversity of Michigan
- About Islam and Muslims
- Carnegie Mellon Uni...m Students Home Page
- Bookstore: The Onli...slamic Books, Isl...
- Islamic Texts and R... University at Bu...
- University of Warwick Islamic Society
- Muslim Students Ass...at Lehigh University
- MSA of CSU
- El Sagrado Corán
- Islamic Association... Palestine Home Page
- Kutkut - Islam
- Other MSAs and Organizations
- Other Resources rel...versity at Buffal...
- 777
- Huma's Mamalist of Islamic Links!
- Other MSAs
- ZUBAIR'S ISLAM PAGE
- MIDDLE EAST CONFLICTS
- Islamic Links at the Arabic Paper
- Middle East & Arab Hot Links
- MSA National: MSAs Home Page
- Islamic Page
- Info about Muslims (MSA @SUNY/Buffalo)
- Untitled: http://www...ev/mideast/islam.htm
- Aalim Fevens: Islam Home Page
- islam
- Links to MSAs
- THE ISLAM PAGE

Recommendation systems

Recommendation Systems

Recommend docs to user based on user's context (besides the docs' content).

Other applications:

- Re-rank search results.
- Locate experts.
- Targeted ads.

Input

Past transactions from users:

- which docs viewed
- which products purchased
- pages bookmarked….
- explicit ratings (movies, books….)

Current context:

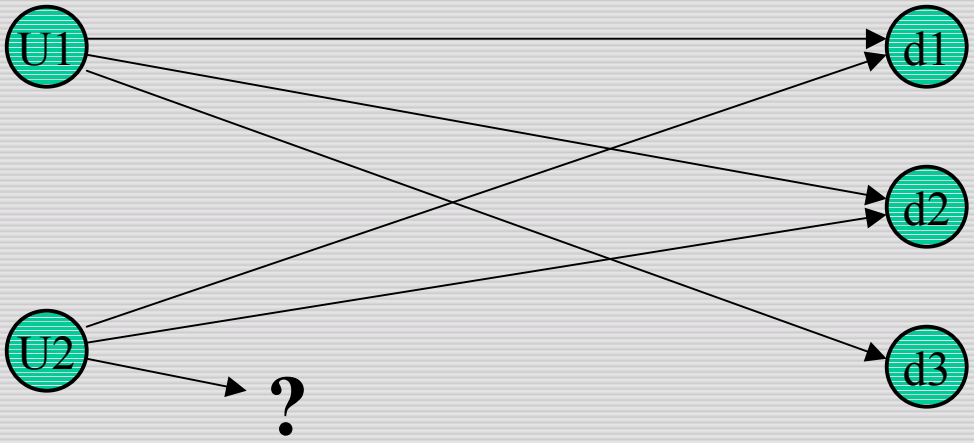
- browsing history
- search(es) issued

Explicit profile info:

- Role in an enterprise
- Demographic info
- Interest profiles

Example

Users | Docs viewed

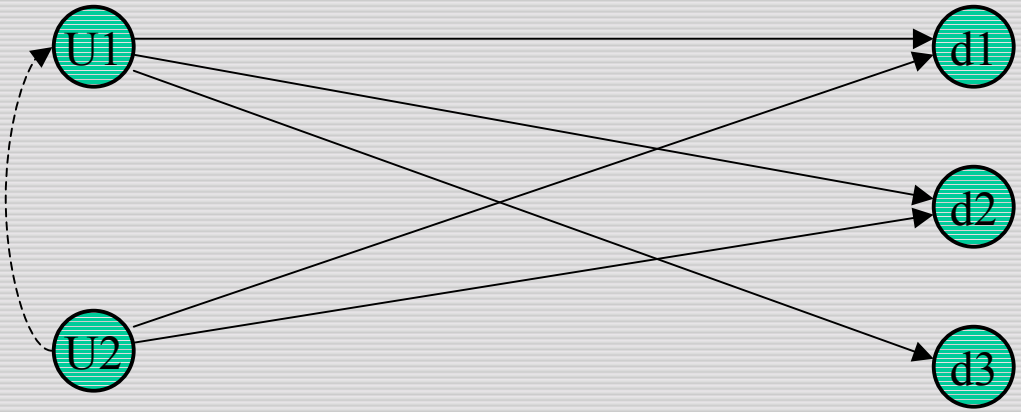


U1 viewed d1, d2, d3.

U2 views d1, d2.

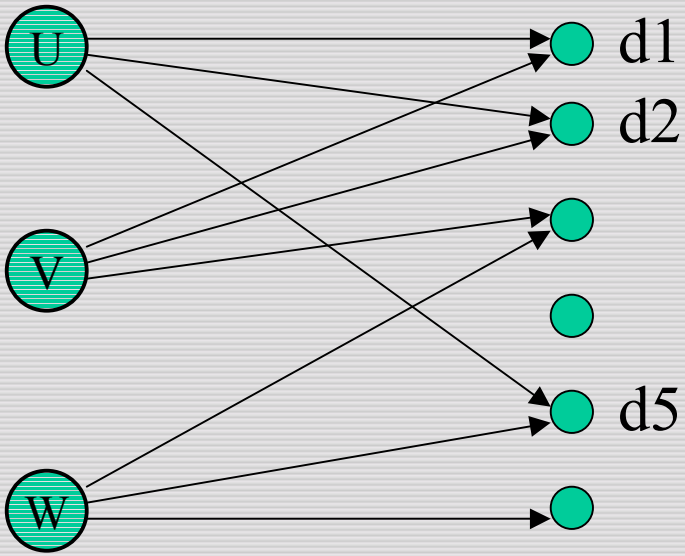
Recommend d3 to U2.

Expert finding



In an enterprise setting, recommend U1 to U2 as an expert.

Simple Algorithm

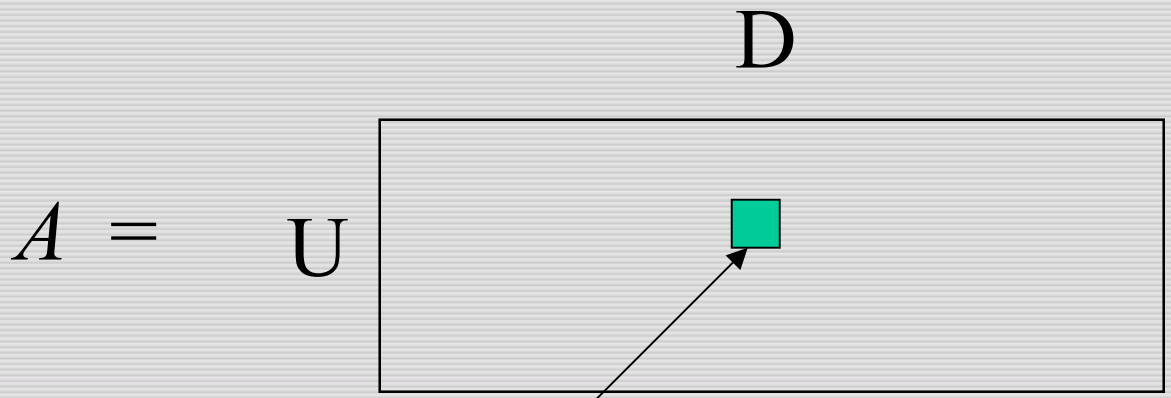


U viewed d1, d2, d5.

Look at who else viewed d1, d2 or d5.

Recommend to U the doc(s) most "popular" among these users.

More formally



$A_{ij} = 1$ if user *i* viewed doc *j*.

AA^t : Entries give # of docs viewed by pairs of users.

Voting Algorithm

- Row i of AA^t : Vector whose j^{th} entry is the # of docs viewed by both i and j.
- Call this row r_i, e.g., (0, 7, 1, 13, 0, 2, ….)
- Then $r_i \circ A$ is a vector whose k^{th} entry gives a vote count to doc k
  - emphasizes users who have high weights in r_i .
- Output doc(s) with highest vote counts.

What's on the diagonal of AA^t?

Voting Algorithm - implementation issues

- Wouldn't implement using matrix operations
  - use weight-propagation on data structures.
- Need to log and maintain "user views doc" relationship.
  - typically, log into database
  - update vote-propagating structures periodically.
- For efficiency, discard all but the heaviest weights in each r_i.

What good was the matrix formulation?

AA^t entries give us a similarity measure between users.

r_i has proximities from user *i* to the rest.

$r_i \circ A$ gives proximities from user *i* to the docs.

Need a more general formulation

- If a user is close to two docs d1 and d2, are the docs d1 and d2 close to each other?
- How do we combine different sources of content and context?
  - terms in docs
  - links between docs
  - users' access patterns
  - users' info.

Vector spaces again

Turn every entity into a vector.

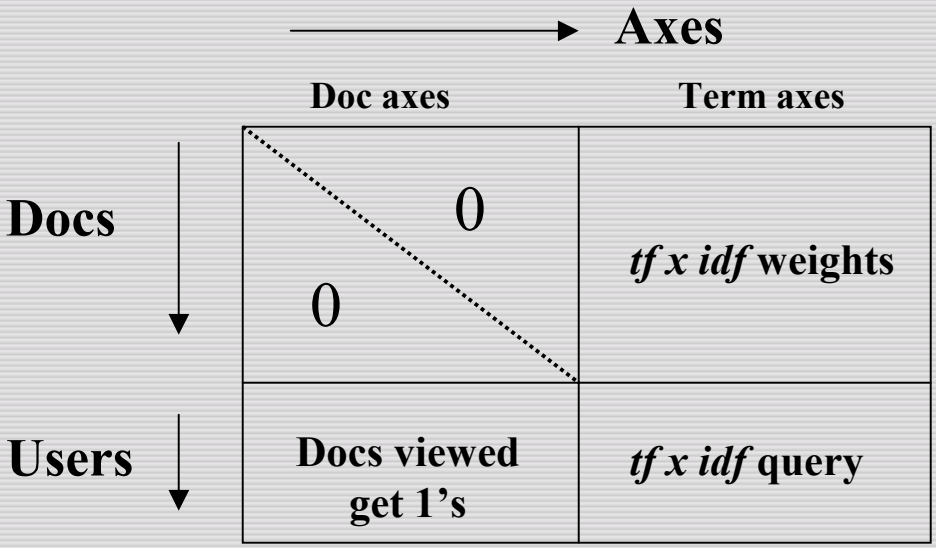
Axes are terms, docs, user info …

e.g.,

- Some axes for terms
- One axis for each doc.
- Additional axes for user attributes like gender, enterprise role, etc.

Vector Space

Each doc represented by *tf* × *idf* weights for terms, plus a 1 entry for its own axis, and 0's elsewhere.



Users represented by 1's for docs viewed, 0's elsewhere.

User posing a query: *tf* × *idf* weights for terms.

Context with content

- Docs' content captured in term axes.
- Other attributes (user behavior, current query etc.) captured in other axes.
- A probe consists of

  1 : a vector *v* (say, a user vector plus a query)

  2 : a type of vector to be retrieved (say, a doc)
- Result = vectors of chosen type closest to *v*

Implementation details

- Don’t really want to maintain this gigantic (and sparse) vector space
- Dimension reduction
- Fast near neighbors (of vectors from a given type)
- Incremental versions needed

Resources

- Hypertext clustering: D.S. Modha, W.S. Spangler. Clustering hypertext with applications to web searching.

  http://citeseer.nj.nec.com/272770.html
- Duplicate detection: A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the Web.
  http://citeseer.nj.nec.com/context/109312/0
- *a priori* algorithm: R. Agrawal, R. Srikant. Fast algorithms for mining association rules.

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