

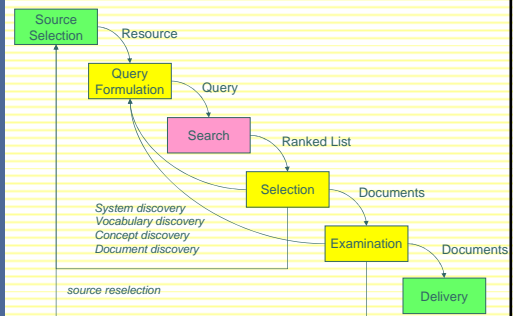
LBSC 796/INFM 718R: Week 10
Clustering, Classifying, and Filtering



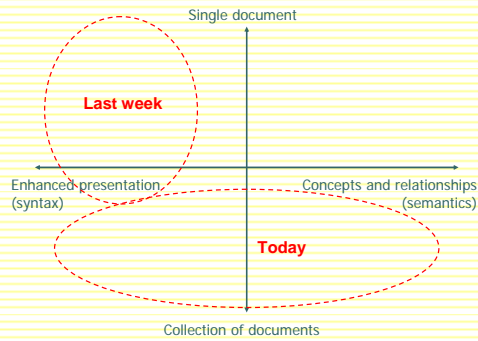
Jimmy Lin
College of Information Studies
University of Maryland

Monday, April 10, 2006

The Information Retrieval Cycle



Organizing Search Results



Today's Focus: Clustering

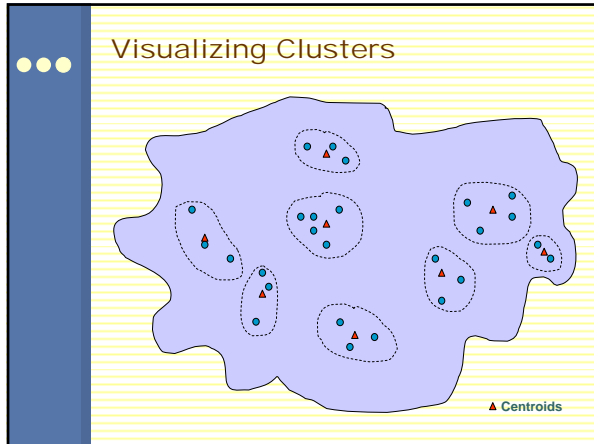
- o Can we do better than a ranked list?
- o How do we automatically group documents into clusters?
- o What are the issues to consider?

Related Topics

- o Using the tools in your toolbox to tackle related problems:
 - Classification: automatically assign labels to documents
 - Filtering: automatically decide if a document matches my information needs

Text Clustering

- o Automatically partition documents into clusters based on content
 - Documents within each cluster should be similar
 - Documents in different clusters should be different
- o Discover categories in an unsupervised manner
 - No sample category labels provided by humans



The Cluster Hypothesis

“Closely associated documents tend to be relevant to the same requests.”

van Rijsbergen 1979

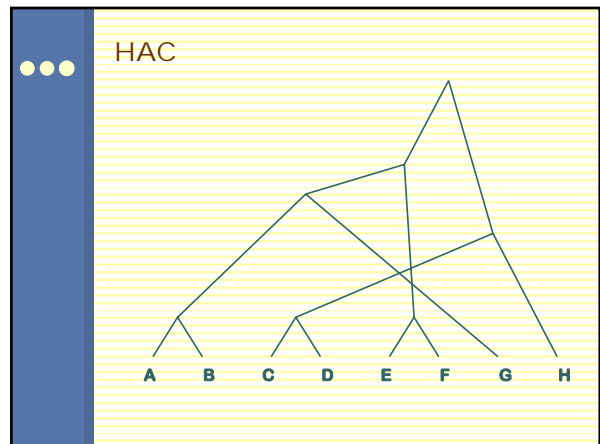
“... I would claim that document clustering can lead to more effective retrieval than linear search [which] ignores the relationships that exist between documents.”

van Rijsbergen 1979

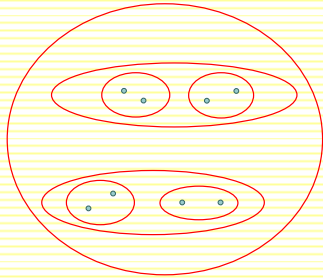
- ### Outline of Clustering
- How do you actually do it?
 - Why would you want to do it?
 - How can you build interfaces that support clustering?

- ### Two Strategies
- Agglomerative (bottom-up) methods
 - Start with each document in its own cluster
 - Iteratively combine smaller clusters to form larger clusters
 - Divisive (partitional, top-down) methods
 - Directly separate documents into clusters

- ### HAC
- HAC = Hierarchical Agglomerative Clustering
 - Start with each document in its own cluster
 - Until there is only one cluster:
 - Among the current clusters, determine the two clusters c_i and c_j that are most similar
 - Replace c_i and c_j with a single cluster $c_i \cup c_j$
 - The history of merging forms the hierarchy



What's going on geometrically?



Cluster Similarity

- o Assume a similarity function that determines the similarity of two instances: $\text{sim}(x,y)$
 - What's appropriate for documents?
- o What's the similarity between two clusters?
 - Single Link: similarity of two most similar members
 - Complete Link: similarity of two least similar members
 - Group Average: average similarity between members

Different Similarity Functions

- o Single link:
 - Uses maximum similarity of pairs:
$$\text{sim}(c_i, c_j) = \max_{x \in c_i, y \in c_j} \text{sim}(x, y)$$
 - Can result in "straggly" (long and thin) clusters due to *chaining effect*
- o Complete link:
 - Use minimum similarity of pairs:
$$\text{sim}(c_i, c_j) = \min_{x \in c_i, y \in c_j} \text{sim}(x, y)$$
 - Makes more "tight," spherical clusters

Non-Hierarchical Clustering

- o Typically, must provide the number of desired clusters, k
- o Randomly choose k instances as seeds, one per cluster
- o Form initial clusters based on these seeds
- o Iterate, repeatedly reallocating instances to different clusters to improve the overall clustering
- o Stop when clustering converges or after a fixed number of iterations

K-Means

- o Clusters are determined by centroids (center of gravity) of documents in a cluster:
$$\bar{\mu}(c) = \frac{1}{|c|} \sum_{\bar{x} \in c} \bar{x}$$
- o Reassignment of documents to clusters is based on distance to the current cluster centroids

K-Means Algorithm

- o Let d be the distance measure between documents
- o Select k random instances $\{s_1, s_2, \dots, s_k\}$ as seeds.
- o Until clustering converges or other stopping criterion:
 - Assign each instance x_i to the cluster c_j such that $d(x_i, s_j)$ is minimal
 - Update the seeds to the centroid of each cluster
 - For each cluster c_j , $s_j = \mu(c_j)$

K-Means Clustering Example

Pick seeds
 Reassign clusters
 Compute centroids
 Reassign clusters
 Compute centroids
 Reassign clusters
 Converged!

K-Means: Discussion

- How do you select k ?
- Results can vary based on random seed selection
 - Some seeds can result in poor convergence rate, or convergence to sub-optimal clusters

Why cluster for IR?

- Cluster the collection
 - "Closely associated documents tend to be relevant to the same requests."
 - Retrieve clusters instead of documents
- Cluster the results
 - "... I would claim that document clustering can lead to more effective retrieval than linear search [which] ignores the relationships that exist between documents."

From Clusters to Centroids

▲ Centroids

Clustering the Collection

- Basic idea:
 - Cluster the document collection
 - Find the centroid of each cluster
 - Search only on the centroids, but retrieve clusters
- If the cluster hypothesis is true, then this should perform better
- Why would you want to do this?
- Why doesn't it work?

Clustering the Results

- Scatter/Gather (Hearst and Pedersen, 1996)
- Swish (Chen and Dumais, 2000)

Scatter/Gather

- o How it works
 - The system clusters documents into general “themes”
 - The system displays the contents of the clusters by showing topical terms and typical titles
 - User chooses a subset of the clusters
 - The system automatically re-clusters documents within selected cluster
 - The new clusters have different, more refined, “themes”
- o Originally used to give collection overview
- o Evidence suggests more appropriate for displaying retrieval results in context

Marti A. Hearst and Jan O. Pedersen. (1996) Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. Proceedings of SIGIR 1996.

Scatter/Gather Example

Query = “star” on encyclopedic text



Clustering and re-clustering is entirely automated

Cluster 1 Size: 8 key army war francis spangle banner air song scott word poem british

- o Star - Spangled Banner, The
- o Key, Francis Scott
- o Fort McHenry
- o Arnold, Henry Harley

Cluster 2 Size: 68 film play career win television role record award york popular stage p

- o Burstin, Ellen
- o Stanwyck, Barbara
- o Berle, Milton
- o Zukor, Adolph

Cluster 3 Size: 97 bright magnitude cluster constellation line type contain period spectr

- o star
- o Galaxy, The
- o extragalactic systems
- o interstellar matter

Cluster 4 Size: 67 astronomer observatory astronomy position measure celestial telescop

- o astronomy and astrophysics
- o astrometry
- o Agema
- o astronomical catalogs and atlases

Cluster 5 Size: 10 family specie flower animal arm plant shape leaf brittle tube foot hor

- o blazing star
- o brittle star
- o bishop's-cap
- o feather star

Cluster 1 Size: 14 player league hit game national set bat average season history baseba

- o Muniel, Stan
- o Bench, Johnny
- o Carew, Rod
- o Robertson, Oscar
- o Bellows, Jean
- o Casper, Billy
- o Chinese checkers
- o Best, George
- o Neuman, Bob

Cluster 2 Size: 47 role stage Broadway comedy performance actress production musical

- o Burstin, Ellen
- o Stanwyck, Barbara
- o Berle, Milton
- o Bankhead, Tallulah
- o Murphy, Eddie
- o Walsh, Basil
- o Martin, Mary
- o Zukor, Adolph
- o Cosby, Bill

Cluster 3 Size: 7 music country jazz folk pop paul cowboy leader william hampton boy

- o Williams, Hank
- o Crosby, Bing
- o Campbell, Glen
- o Belafonte, Harry
- o Shore, Dinah
- o Denver, John
- o Hampton, Lionel

Cluster 1 Size: 12 black white nuclear hole reaction helium neutron gravitational colla

- o stellar evolution
- o gravitational collapse
- o black hole
- o main sequence
- o carbon cycle
- o mass-luminosity relation

Cluster 2 Size: 49 galaxy type distance stellar variable spectral interstellar brightness ga

- o star
- o extragalactic systems
- o Galaxy, The
- o interstellar matter
- o cluster, star
- o population, stellar

Cluster 3 Size: 29 constellation northern hemisphere sky locate dipper celestial double

- o constellation (astronomy)
- o Auriga
- o Big Dipper
- o Cassiopeia
- o Cygnus
- o Taurus

Cluster 4 Size: 7 fraunhofer designate map joseph frown fur wollaston english von davis

- o Fraunhofer lines
- o Fraunhofer, Joseph von
- o Star Carr
- o Star of David
- o Star Chamber

Clustering Result Sets

- o Advantages:
 - Topically coherent sets of documents are presented to the user together
 - User gets a sense for the range of themes in the result set
 - Supports exploration and browsing of retrieved hits
- o Disadvantage:
 - Clusters might not “make sense”
 - May be difficult to understand the theme of a cluster based on summary terms
 - Additional computational processing required
- o Things to ponder:
 - What is the relationship between clusters and classification systems?
 - Why does this work?

Two Queries: Two Clusterings

AUTO, CAR, ELECTRIC	AUTO, CAR, SAFETY
8 control drive accident ...	6 control inventory integrate ...
25 battery california technology ...	10 investigation washington ...
48 import j. rate honda toyota ...	12 study fuel death bag air ...
16 export international unit japan ...	61 sale domestic truck import ...
3 service employee automatic ...	11 japan export defect unite ...
...	...

The main differences are the clusters that are central to the query

The SWISH System


- Basic idea:
 - Use an existing hierarchical category structure to organize results of Web searches
 - Automatically classify Web pages into the relevant category
 - Present search results grouped according to categories
- Research questions:
 - How does a category interface compare with a list interface?
 - What features of a category interface would users find useful?

Hao Chen and Susan Dumais. (2000) Bringing Order to the Web: Automatically Categorizing Search Results. Proceedings of CHI 2000.


Organizing Search Results

Query: jaguar

Category Interface



List Interface



Category Structure

- Category hierarchy taken from LookSmart Web Directory (Spring, 1999)
 - 13 top-level categories
 - 150 second-level categories
- Top-level Categories:

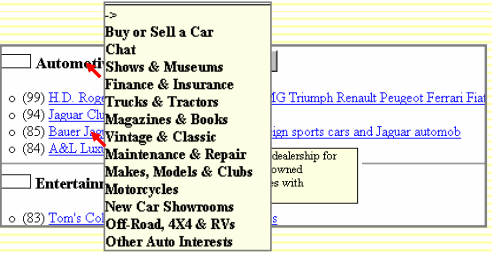
Automotive	People & Chat
Business & Finance	Reference & Education
Computers & Internet	Shopping & Services
Entertainment & Media	Society & Politics
Health & Fitness	Sports & Recreation
Hobbies & Interests	Travel & Vacations
Home & Family	

Interface Characteristics

- Problems
 - Large amount of information to display
 - Limited screen real estate
- Solutions
 - Information overlay ("mouseovers")
 - Expandable information display

Information Overlay

- Use "mouseovers" to show
 - Summaries of web pages
 - Category hierarchy



Expansion of Category Structure

The screenshot shows a hierarchical menu for 'Automotive' under 'MainCateg'. The categories are:

- Maintenance & Repair** (More (7))
 - (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
 - (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
- Buy or Sell a Car** (More (6))
 - (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
 - (84) [A&L Luxury Car Center - Jaguar Main Page](#)
- Vintage & Classic** (More (2))
 - (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
- Makes, Models & Clubs** (More (1))
 - (94) [Jaguar Club of Florida](#)

Expansion of Web Page List

The screenshot shows a list of 24 links under the 'Automotive' category, labeled 'SubCateg' and 'Less'. The links include:

- (99) [H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B](#)
- (94) [Jaguar Club of Florida](#)
- (85) [Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob](#)
- (84) [A&L Luxury Car Center - Jaguar Main Page](#)
- (78) [Exotic Car Rentals, Av& Rentals, BMW Rentals, Jaguar Rentals, Lexus Rentals,](#)
- (72) [Imperial Motors Jaguar dealer of Wilmette, Illinois. New autos, used cars, Sel](#)
- (71) [Coventry West - New, Rebuild, & Used Jaguar Parts](#)
- (56) [Ottawa Jaguar Club](#)
- (55) [Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us](#)
- (53) [Master Auto Service - Jaguar and Porsche Service](#)
- (51) [Jaguar Mercedes Cars in Melbourne, Florida](#)
- (49) [Audi BMW Jaguar Lexus Mercedes Benz Porsche Cars Automobiles Dealers Ohio Cleve](#)
- (47) [The Classic Car-Norton Classic, Jaguar Page](#)
- (38) [Autosport MG, Austin Healey, Jaguar, Volvo repair, maintenance, service](#)
- (33) [Jaguar History](#)
- (31) [Motorcars Ltd. - Jaguar Parts, Land Rover Parts and Service](#)
- (29) [Fischer Porsche Buick GMC, Jaguar, Subaru automobile dealer new and used cars](#)
- (24) [Auto Intelligence Luxury Cars Buyers Services BMW Jaguar Lexus Porsche Mercedes](#)

Interface Conditions

The image shows two side-by-side screenshots of a web browser. The left window is titled 'Category Interface' and shows a hierarchical menu structure with categories like 'Computers & Internet', 'Automotive', 'Travel & Vacations', and 'Business & Finance'. The right window is titled 'List Interface' and shows a list of links under the 'Automotive' category, similar to the one shown in the 'Expansion of Web Page List' slide.

User Study Interface

The screenshot shows a search results page for 'Jaguar' in a browser window. The search engine is 'Query Jaguar'. The results are organized into categories: 'Computers & Internet', 'Automotive', 'Entertainment & Media', 'Travel & Vacations', 'Business & Finance', and 'Shopping & Services'. A sidebar on the right shows 'Jaguar Clubs of North America' with a 'HOME' button and a 'SEARCH' button.

User Study

- Participants: 18 "intermediate" Web users
- Tasks
 - 30 search tasks, e.g., "Find home page for Seattle Art Museum"
 - Search terms are fixed for each task (cached Web pages)
- Experimental Design
 - Category/List – within subjects (15 search tasks with each interface)
 - Order (Category/List first) – counterbalanced between subjects
- Both Subjective and Objective Measures

Subjective Results

- 7-point rating scale (1=disagree; 7=agree)

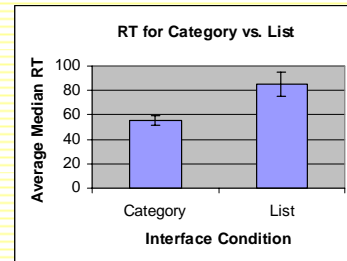
Question	Category	List	significance
It was easy to use this software.	6.4	3.9	p<.001
I liked using this software	6.7	4.3	p<.001
I prefer this to my usual Web Search engine	6.4	4.3	p<.001
It was easy to get a good sense of the range of alternatives	6.4	4.2	p<.001
I was confident that I could find information if it was there.	6.3	4.4	p<.001
The "More" button was useful	6.5	6.1	n.s.
The display of summaries was useful	6.5	6.4	n.s.

Use of Interface Features

- Average number of uses of feature per task:

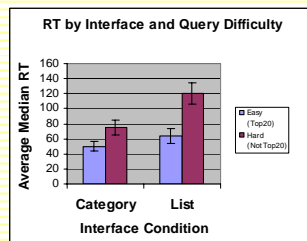
Interface Features	Category	List	significance
Expanding / Collapsing Structure	0.78	0.48	p<.003
Viewing Summaries in Tooltips	2.99	4.60	p<.001
Viewing Web Pages	1.23	1.41	p<.053

Search Time



Category: 56 sec.
List: 85 sec. ($p < .002$)
50% faster with category interface!

Search Time by Query Difficulty



Category interface is helpful for both easy and difficult queries!

Visualization of Clusters

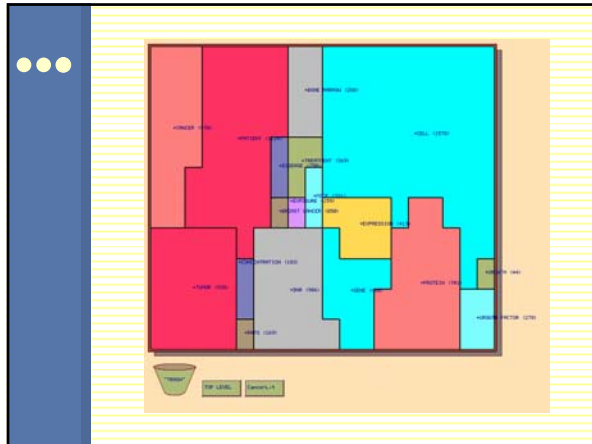
- Feature Maps
- Other 2D and 3D displays

Kohonen's Feature Maps

- AKA Self-Organizing Maps
- Expresses complex, non-linear relationships between high dimensional data on a 2D display
 - Geometric relationships on display preserve some relationships in original data set

Map Attributes

- Different areas correspond to different concepts in collection
- Size of area corresponds to its relative importance in set
- Neighboring regions share commonalities



Study of Kohonen Feature Maps

- Comparison: Kohonen Map and Yahoo
- Task:
 - "Window shop" for interesting home page
 - Repeat with other interface
- Results:
 - Starting with map could repeat in Yahoo (8/11)
 - Starting with Yahoo unable to repeat in map (2/14)

Hsinchun Chen, Andrea L. Houston, Robin R. Sewell, and Bruce R. Schatz. (1998) Journal of the American Society for Information Science, 49(7):582-603.

Feature Map Study (1)

- Participants liked:
 - Correspondence of region size to # documents
 - Overview (but also wanted zoom)
 - Ease of jumping from one topic to another
 - Multiple routes to topics
 - Use of category and subcategory labels

Feature Map Study (2)

- Participants wanted:
 - Hierarchical organization
 - Other ordering of concepts (alphabetical)
 - Integration of browsing and search
 - Correspondence of color to meaning
 - More meaningful labels
 - Labels at same level of abstraction
 - Fit more labels in the given space
 - Combined keyword and category search
 - Multiple category assignment (sports+entertainment)

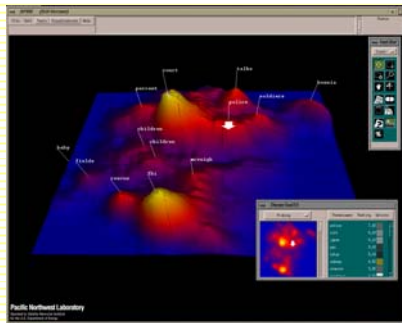
WEBSOM

Self-organizing map of Net newsgroups and postings

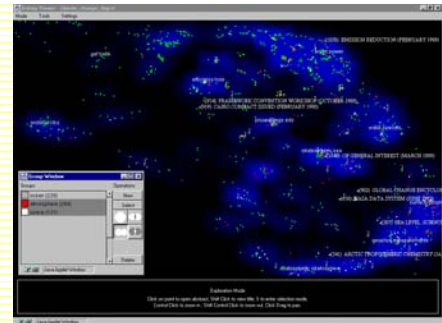
Galaxies

Galaxy view of 100,000 cancer literature abstracts

Themescape



WebTheme



TreeMaps

- o Demos
 - <http://www.smartmoney.com/marketmap/>
 - <http://www.cs.umd.edu/hcil/treemap/>

Deployment

- o Web Search engine that employ clustering:
 - <http://www.vivisimo.com>
 - <http://www.kartoo.com/>

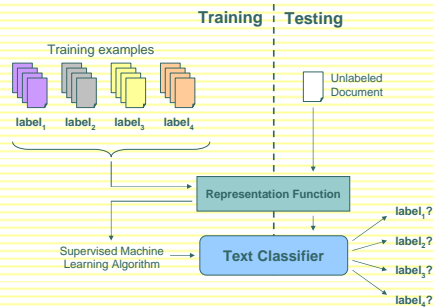
Summary: Clustering

- o Advantages:
 - Provides users with an overview of main themes in search results
 - Helps combat polysemy
 - Can improve retrieval effectiveness
- o Disadvantages:
 - Documents can be clustered in many ways
 - Not always easy to understand the theme of a cluster
 - What is the correct level of granularity?
 - More information to present; requires careful design of user interfaces

Text Classification

- o Problem: automatically sort items into bins
- o Examples:
 - Spam vs. non-spam
 - Interesting vs. non-interesting
- o Machine learning approach
 - Obtain a training set with ground truth labels
 - Use a machine learning algorithm to "train" a classifier
 - kNN, Bayesian classifier, SVMs, decision trees, etc.
 - Apply classifier to new documents
 - System assigns labels according to patterns learned in the training set

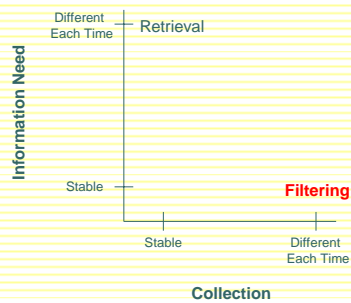
Machine Learning



kNN

- A simple text classification algorithm: k Nearest Neighbors
- Select k document that are similar to the test document
 - Have them vote on what the correct label should be
 - How can similarity be defined?

Information Access Problems



Information Filtering

- An abstract problem in which:
 - The information need is stable: characterized by a "profile"
 - A stream of documents is arriving: each must either be presented to the user or not
- Introduced by Luhn in 1958
 - As "Selective Dissemination of Information"
- Named "filtering" by Denning in 1975

A Simple Filtering Strategy

- Use any information retrieval system
 - Boolean, vector space, probabilistic, ...
- Have the user specify a "standing query"
 - This will be the profile
- Limit the standing query by date
 - For each use, show new documents since the last use

Social Filtering

- Exploit ratings from other users as features
 - Like personal recommendations, peer review, ...
- Reaches beyond topicality to:
 - Accuracy, coherence, depth, novelty, style, ...
- Applies equally well to other modalities
 - Movies, recorded music, ...
- Sometimes called "collaborative" filtering

Using Positive Information

	Small World	Space Mt.	Mad Tea Party	Dumbo	Speedway	Train Ride
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	C	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Source: Jon Herlocker, SIGIR 1999

Using Negative Information

	Small World	Space Mt.	Mad Tea Party	Dumbo	Speedway	Train Ride
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	C	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Source: Jon Herlocker, SIGIR 1999

- ### Some Things We (Sort of) Know
- Treating each genre separately can be useful
 - Separate predictions for separate tastes
 - Negative information can be useful
 - "I hate everything my parents like"
 - People like to know who provided ratings
 - Popularity provides a useful fallback
 - People don't like to provide ratings

- ### The Cold Start Problem
- Social filtering will not work in isolation
 - Without ratings, we get no recommendations
 - Without recommendations, we read nothing
 - Without reading, we get no ratings
 - An initial recommendation strategy is needed
 - Stereotypes
 - Content-based search

- ### Cold Start: Potential Solutions
- Provide motivation:
 - Self-interest
 - Altruism
 - Economic benefit
 - Implicit feedback

- ### Sample Observations
- User selects an article
 - Interpretation: Summary was interesting
 - User quickly prints the article
 - Interpretation: They want to read it
 - User selects a second article
 - Interpretation: another interesting summary
 - User scrolls around in the article
 - Interpretation: Parts with high dwell time and/or repeated revisits are interesting
 - User stops scrolling for an extended period
 - Interpretation: User was interrupted

Critical Issues

- Protecting privacy
 - What absolute assurances can we provide?
 - How can we make remaining risks understood?
- Scalable rating servers
 - Is a fully distributed architecture practical?
- Non-cooperative users
 - How can the effect of spamming be limited?

Recap

- Clustering
 - Automatically group documents into clusters
- Classification
 - Automatically assign labels to documents
- Filtering
 - Automatically decide if a document matches my information needs
- Many approaches to the same elephant