Information Filtering

LBSC 796/INFM 718R Douglas W. Oard Session 10, April 13, 2011

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

Information Filtering



Standing Queries

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
 Each use, show what arrived since the last use

What's Wrong With That?

• Unnecessary indexing overhead

- Indexing only speeds up <u>retrospective</u> searches

- Every profile is treated separately
 - The same work might be done repeatedly
- Forming effective queries by hand is hard
 The computer might be able to help
- It is OK for text, but what about audio, video, ...
 Are words the only possible basis for filtering?

Stream Search: "Fast Data Finder"

- Boolean filtering using custom hardware
 Up to 10,000 documents per <u>second</u> (in 1996!)
- Words pass through a pipeline architecture
 Each element looks for one word



Profile Indexing (SIFT)

- Build an inverted file of profiles

 Postings are profiles that contain each term
- RAM can hold 5 million profiles/GB
 And several machines could run in parallel
- Both Boolean and vector space matching
 - User-selected threshold for each ranked profile
 - Hand-tuned on a web page using today's news

Profile Indexing Limitations

• Privacy

- Central profile registry, associated with known users

- Usability
 - Manual profile creation is time consuming
 - May not be kept up to date
 - Threshold values vary by topic and lack "meaning"

Vector space example: query "canine" (1)



Similarity of docs to query "canine"



User feedback: Select relevant documents



Results after relevance feedback





 $\vec{\mu}_{R}$: centroid of relevant documents



 $\vec{\mu}_R$ does not separate relevant / nonrelevant.



 $\vec{\mu}_{NR}$: centroid of nonrelevant documents.



 $\vec{\mu}_R \quad \vec{\mu}_{NR}$: difference vector



Add difference vector to $\vec{\mu}_R$...



... to ge \vec{q}_{opt}





 \vec{q}_{opt} separates relevant / nonrelevant perfectly.



 \vec{q}_{opt} separates relevant / nonrelevant perfectly.

Adaptive Content-Based Filtering



Latent Semantic Indexing



Content-Based Filtering Challenges

- IDF estimation
 - Unseen profile terms would have infinite IDF!
 - Incremental updates, side collection
- Interaction design
 - Score threshold, batch updates
- Evaluation
 - Residual measures

Machine Learning for User Modeling

- All learning systems share two problems
 - They need some basis for making predictions
 - This is called an "inductive bias"
 - They must balance adaptation with generalization

Machine Learning Techniques

- Hill climbing (Rocchio)
- Instance-based learning (kNN)
- Rule induction
- Statistical classification
- Regression
- Neural networks
- Genetic algorithms

Rule Induction

- Automatically derived Boolean profiles
 (Hopefully) effective <u>and</u> easily explained
- <u>Specificity</u> from the "perfect query"
 AND terms in a document, OR the documents
- <u>Generality</u> from a bias favoring short profiles

 e.g., penalize rules with more Boolean operators
 Balanced by rewards for precision, recall, …

Statistical Classification

• Represent documents as vectors

– Usual approach based on TF, IDF, Length

- Build a statistical models of rel and non-rel – e.g., (mixture of) Gaussian distributions
- Find a surface separating the distributions – e.g., a hyperplane
- Rank documents by distance from that surface

Linear Separators

• Which of the linear separators is optimal?



Maximum Margin Classification

• Implies that only support vectors matter; other training examples are ignorable.



Soft Margin SVM



Original from Ray Mooney

Non-linear SVMs



Original from Ray Mooney

Training Strategies

- Overtraining can hurt performance

 Performance on training data rises and plateaus
 Performance on new data rises, then <u>falls</u>
- One strategy is to learn less each time
 But it is hard to guess the right learning rate
- Usual approach: Split the training set
 Training, DevTest for finding "new data" peak

NetFlix Challenge

near window Swiss Family the Christ Star Wars: Episode V: Sleeping Beauty: Singin' in Robinson Φ Special Edition The Empire Strikes Back the Rain Some Like The Maltese Seven Samurai The Phantom of the The Incredibles Psycho It Hot Silverado Falcon **Opera: Special Edition** Lord of the Rings: **Finding Nemo** Indiana Jones and the Ran £ The Wizard of Oz: Lawrence of (Widescreen) The Two Towers Citizen Kane Last Crusade The Fox and Chicago **Collector's Edition** Arabia Harold and Lilo and the Hound The Big Rocky IV The Matrix JFK: Special Maude Stitch **Red Dawn** One Flew Over the The Princess Easy Dune Edition Tarzan Shrek (Full-screen) Tombstone Cuckoo's Nest Taxi Driver Bride Black Hawk Spider-Man Rocky The Ta -Batman Malcolm X Gandhi Amadeus Anastasia Down **Driving Miss** Ice Age Veronica Guerin Mr. F Starman Pirates of the Caribbean: The Curse Apocalypse Now Signs Daisy American History Braveheart Silkwood Life Is Gladiator The Sopranos: of the Black Pearl Diner X **Beautifu** Season 1 The Village Philadelphia Schindler's List Wall Street Secondhand Lions The Godfather The Hours The Mask Findi Planet of Dragonheart **Boys Don't** Beetlejuice The Sixth the Apes Forrest Gump of Zorro Sliding Doors Miracle Identity Cry Colors The Silence of Saw Sense **Dead Poets** The Joy Slin Apollo 13 the Lambs Luck Club Love Actually Society Seabiscuit Grease The Terminator **Basic Instinct** The Devil's Say Anything The World Is Poltergeist Father of Advocate Not Enough Whale **Practical Magic** Pleasantville K-Pax The Firm the Bride Life as Fallen A Time Saved! Mad Max **Reality Bites** A Nightmare on a House Top Gun *lountain* Ra to Kill Pretty Woman Beaches My Big Fat Elm Street The Bl Crimson Tide Dogma Hearts in Greek Wedding Clerks Kung Fu What Lies Soul Food **Runaway Jury Big Fish** The Birdcage Witch Pre Atlantis Hustle Ransom Beneath **Erin Brockovich** Ray Under the Collateral Air Force Snatch The Day Sweet Home **Pulp Fiction** The Bone American Beauty **Tuscan Sun** One After Tomorrow **Dangerous Minds** Men of Alabama Collector **Matchstick** Men Blade Gothika John Q **Fight Club** Honor **Fools Rush** Miss Congeniality Van Helsing **Darkness Falls** The Rock House on Zoolander Man on Blade 2 Swordfish Confessions of a Haunted Hill Alien vs. The Last National Treasure Fire Dangerous Mind Predator The Truman Dodgeball: A True Old School Samurai The Patriot The Big Lost in The Terminal Ronin Show sion to The Fast and Underdog Story Drugstore Cowboy Translation Lebowski Austin Powers: International Man Sahara Mars the Furious National Lampoon's American Pie of Mystery XXX: Special Vanilla Sky aider Van Wilder The Mummy Edition About Sc The Nightmare Bruce Almighty Starsky & Ace Ventura: Wayne's World Ferris Bueller's Napoleon Dynamite Miss Congeniality 2: Armed Meet the **Before Christmas** Deep Impact **Jurassic Park** Pet Detective Ghost World, Hutch Day Off and Fabulous Parents Heathers Best in Monty Python and the Maid in Show Anger Management Mr. Deeds Independence Day Are We Holy Grail Encino Man Ghostbusters Major League **Man hattan** There Yet? Gia One Hour G Cheaper by Flubber Evolution Twister Photo A Fish White Chicks Hardball This is the Dozen The Office **Rush Hour** Called Wanda Spinal Tap **Rush Hour Beverly Hills Bad Boys Uncle Buck** Pearl Harbor Save the Special 2 Spellb A Cinderella Cop II 11 Message in Last Dance Story Adventures in Titanic Ghost Gone in a Bottle Peter Pan Crocodile Dundee Babysitting K-9 Sister Act A Knight's In Good South Park: 60 Seconds **Bringing Down** Tale Primary Colors Season 1 Company **Bad Boys** the House Lethal Weapon Cocktail Hard to Lethal Weapon Volcano **Red Dragon** Kill Con Air

ANATOMY OF THE LONG TAIL

Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.

6.100



THE NEW GROWTH MARKET: **OBSCURE PRODUCTS YOU CAN'T GET ANYWHERE BUT ONLINE**



Sources: Erik Brynjolfason and Jeffrey Hu, MIT, and Michael Smith, Cemegie Mellon; Barnes & Noble; Netflix; RealNetworks

Effect of Inventory Costs

Physical retailers

Profit threshold for physical stores (like Tower Records)

Hybrid retailers

Titles

Profit threshold for stores with no retail overhead (like Amazon.com)

Pure digital retailers

Profit threshold for stores with no physical goods (like Rhapsody)

Spam Filtering

• Adversarial IR

- Targeting, probing, spam traps, adaptation cycle

- Compression-based techniques
- Blacklists and whitelists
 - Members-only mailing lists, zombies
- Identity authentication
 - Sender ID, DKIM, key management