Google News Personalization: Scalable Online Collaborative Filtering

Abhinandan Das, Mayur Datar, Ashutosh Garg

WWW 2007, May 8-12, 2007

Presented by: Jerry Fu

4/24/2008

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

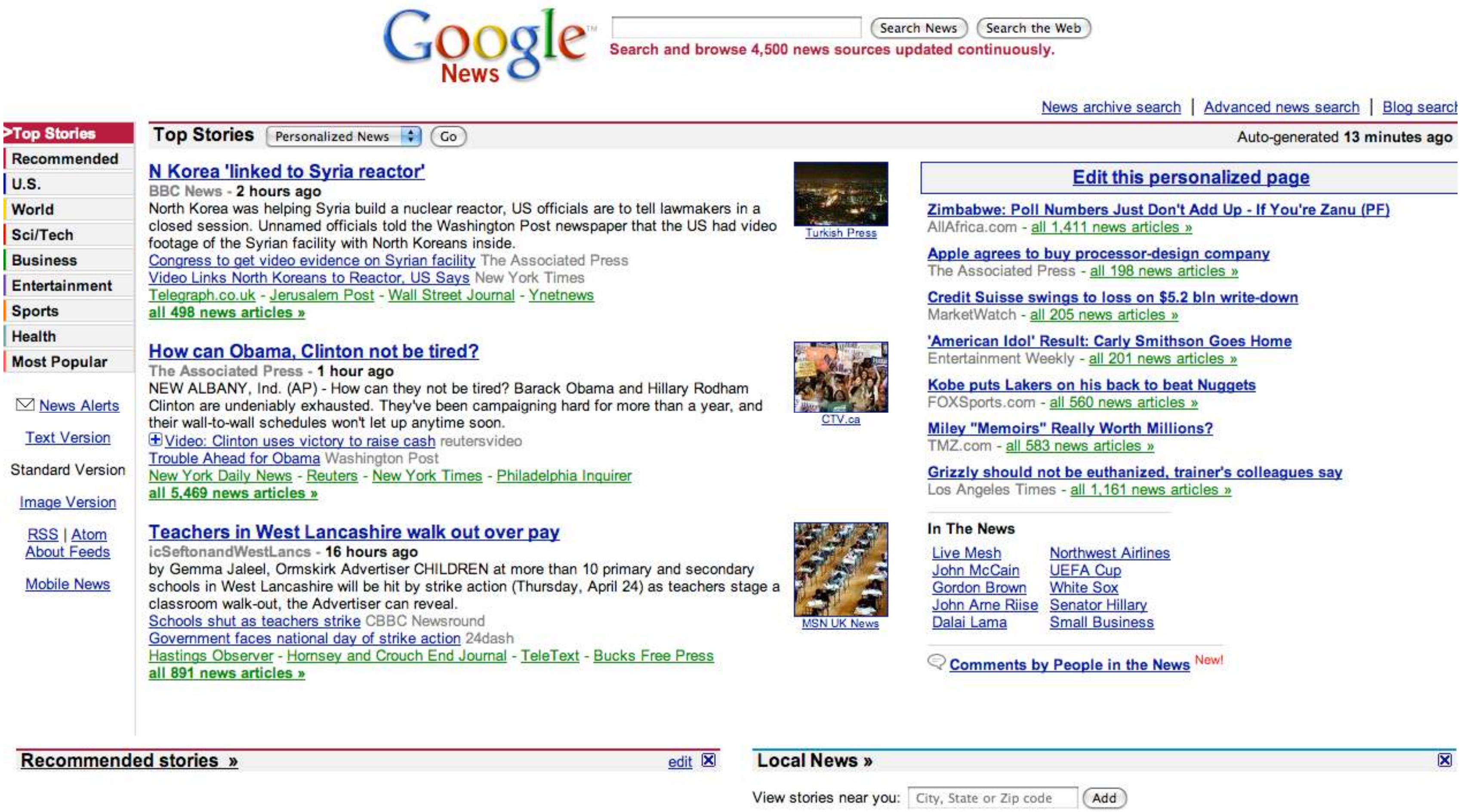
Problem Setting

- Google news aggregrates articles from several thousand news sources daily
- Users do not know what they want, but want to see something “interesting”
- Present several articles that are recommended specifically for user based on:
  - User click history
  - Community click history

Problem Statement

- Given:
  - N users $U = u_1, u_2, ..., u_N$
  - M news articles $S = s_1, s_2, ..., s_M$
  - For each user u, click history $C_u = h_1, h_2, ..., h_{|C_u|}$, where $h_i \in S$
- Recommend K stories to user u, within a few hundred milliseconds
- Approach: collaborative filtering
- Treat user clicks as noisy positive votes

4


Google News
Search News
Search the Web
Search and browse 4,500 news sources updated continuously.
News archive search | Advanced news search | Blog searc
Top Stories
Recommended
U.S.
World
Sci/Tech
Business
Entertainment
Sports
Health
Most Popular
News Alerts
Text Version
Standard Version
Image Version
RSS | Atom
About Feeds
Mobile News
Top Stories
Personalized News
Go
Auto-generated 13 minutes ago
N Korea 'linked to Syria reactor'
BBC News - 2 hours ago
North Korea was helping Syria build a nuclear reactor, US officials are to tell lawmakers in a closed session. Unnamed officials told the Washington Post newspaper that the US had video footage of the Syrian facility with North Koreans inside.
Congress to get video evidence on Syrian facility The Associated Press
Video Links North Koreans to Reactor, US Says New York Times
Telegraph.co.uk - Jerusalem Post - Wall Street Journal - Ynetnews
all 498 news articles »
Turkish Press
How can Obama, Clinton not be tired?
The Associated Press - 1 hour ago
NEW ALBANY, Ind. (AP) - How can they not be tired? Barack Obama and Hillary Rodham Clinton are undeniably exhausted. They've been campaigning hard for more than a year, and their wall-to-wall schedules won't let up anytime soon.
Video: Clinton uses victory to raise cash reutersvideo
Trouble Ahead for Obama Washington Post
New York Daily News - Reuters - New York Times - Philadelphia Inquirer
all 5,469 news articles »
CTV.ca
Teachers in West Lancashire walk out over pay
icSeftonandWestLancs - 16 hours ago
by Gemma Jaleel, Ormskirk Advertiser CHILDREN at more than 10 primary and secondary schools in West Lancashire will be hit by strike action (Thursday, April 24) as teachers stage a classroom walk-out, the Advertiser can reveal.
Schools shut as teachers strike CBBC Newsround
Government faces national day of strike action 24dash
Hastings Observer - Hornsey and Crouch End Journal - TeleText - Bucks Free Press
all 891 news articles »
MSN UK News
Edit this personalized page
Zimbabwe: Poll Numbers Just Don't Add Up - If You're Zanu (PF)
AllAfrica.com - all 1,411 news articles »
Apple agrees to buy processor-design company
The Associated Press - all 198 news articles »
Credit Suisse swings to loss on $5.2 bln write-down
MarketWatch - all 205 news articles »
'American Idol' Result: Carly Smithson Goes Home
Entertainment Weekly - all 201 news articles »
Kobe puts Lakers on his back to beat Nuggets
FOXSports.com - all 560 news articles »
Miley "Memoirs" Really Worth Millions?
TMZ.com - all 583 news articles »
Grizzly should not be euthanized, trainer's colleagues say
Los Angeles Times - all 1,161 news articles »
In The News
Live Mesh
John McCain
Gordon Brown
John Arne Riise
Dalai Lama
Northwest Airlines
UEFA Cup
White Sox
Senator Hillary
Small Business
Comments by People in the News New!
Recommended stories »
edit
Local News »
View stories near you: City, State or Zip code
Add

5

A tough problem indeed

Recommended stories » edit

Please try again shortly.

We apologize that this section is currently unavailable.

World » edit

Zimbabwe: Poll Numbers Just Don't Add Up - If You're Zanu (PF)

CBC.ca

AllAfrica.com - **56 minutes ago**
THE pure statistics of Zimbabwe's contested election demonstrate clearly that the present recount of ballots from 23 constituencies is a sham designed to improve the performance of Robert Mugabe and his Zanu (PF) party, a senior opposition MP has said.
Brown and Zuma call for Zimbabwe election results Reuters UK
Arm Zimbabwe's Opposition Wall Street Journal
Aljazeera.net - CNN International - Washington Post - BBC News
all 1,411 news articles »

Local News »

View stories near you: City, State or Zip code Add

U.S. » edit

Petraeus named to Central Command; will face Afghanistan, Iran threats

Malaysia Star

New York Daily News - **1 hour ago**
BY RICHARD SISK WASHINGTON - President Bush's favorite general has been handed the daunting task of winning the wars in Afghanistan and Iraq as well as confronting the threat from Iran, the Pentagon announced Wednesday.
Video: US Iraq Commander gets promotion RussiaToday
Petraeus' promotion tied to future war policy Houston Chronicle
USA Today - International Herald Tribune - Washington Post - Washington Times
all 962 news articles »

Israelis Claim Secret Agreement With US

CCTV

Washington Post - **3 hours ago**
By Glenn Kessler A letter that President Bush personally delivered to then-Israeli Prime Minister Ariel Sharon four years ago has emerged as a significant obstacle to the president's efforts to forge a peace deal between the Israelis and Palestinians ...
Abbas asks White House for help in Mideast peace talks Boston Globe

6

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

Memory-based algorithms

- Maintain similarity between users (common measures include Pearson correlation coefficient and cosine similarity)
- For a story *s,* calculate recommendation by weighing other user ratings with similarity
- “Ratings” in this case are binary (click or not clicked)

Model-based algorithms

- Create model for each user based on past ratings
- Use model to predict ratings on new items
- Recent work captures multiple interests of users
- Approaches: Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI), Markov Decision Process, Latent Dirichlet Allocation

9

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

Combined Algorithm for Google News

- Use combined memory-based and model-based algorithms
- Here, model-based approaches are
  - MinHash
  - Probabilistic latent semantic indexing (PLSI)
- Memory-based approach is item covisitation

MinHash Algorithm

- Clustering method that assigns users to clusters based on their overlapping set of clicked articles
- Uses Jaccard coefficient, with every user represented by click history

$$\mathbf{S(u,v)} = \frac{|\mathbf{C_u} \cup \mathbf{C_v}|}{|\mathbf{C_u} \cap \mathbf{C_v}|}$$

- Recommend stores clicked on by user *v* to user *u* with weight *S(u,v)*

Probabilistic latent semantic indexing (PLSI)

- Users ($u \in U$) and news stories ($s \in S$) are random variables
- Z is a hidden variable models the relationship between U and S as follows

$$\textbf{Model: } \mathbf{p(s|u;\theta) = \sum_{z=1}^{L} p(z|u)p(s|z)}$$

- Z represents user and item communities
- Generative model of stories s for user u

Recommendations based on covisitation

- Covisitation is defined as two stories clicked by the same user within a given time interval
- Store as a graph with nodes at stories, edges as age discounted covisitation counts
- Update graph (using user history) whenever we receive a click

14

Combined Algorithm for Google News

- Combined memory-based and model-based algorithms
- Here, model-based approaches are
  - MinHash
  - Probabilistic latent semantic indexing (PLSI)
- Memory-based approach is item covisitation

15

Algorithm scores

For clustering (model) algorithms:
Score of story s for user u

$$\mathbf{r}_{u,s} \propto \sum_{c:u:\in c} \underbrace{w(u,c)}_{\text{fractional membership in cluster}} \sum_{v:v\in c} I(v,s)$$

For covisitation (memory) algorithm:

$$\mathbf{r_{u,s} \propto \sum_{t \in C_u} I(s,t)}$$

I(s,t) indicates whether stories s and t were covisited

16

Combined Scores

Scores for stories combined by:

$\sum_a w_a r_{s,a}$

$w_a =$ **weight for algorithm** a
$r_{s,a} =$ **score for** s **from algorithm** a

Appropriate weights are learned experimentally.

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

MapReduce Overview

- MapReduce is a method to process large amounts of data in a cluster
- Inspired by *Map* and *Reduce* in Lisp
- Data set split across machines (shards)
- *Map* produces key/value pairs
- Key space partitioned into regions (hashed)
- *Reduce* merges values for key

MapReduce Overview

- MapReduce is a method to process large amounts of data in a cluster
- Inspired by *Map* and *Reduce* in Lisp
- Data set split across machines (shards)
- *Map* produces key/value pairs
  - Ex. Counting web page acceses
  - *Emit(URL, “1”)*

MapReduce Overview (cont.)

- Key space partitioned into regions, or shards, so that *Reduce* can be performed across many machines
- *Reduce* merges the values that share same key
  - Combines the data derived in Map in an appropriate manner
  - Ex. for web page accesses, sum all values for a given URL

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

MinHash implementation

- As presented before, Jaccard similarity is infeasbile to implement in this setting
- Apply Locality Sensitive Hashing (LSH), or MinHashing
- Create random permutation *P of S* (set of news articles)
- Calculate user hash value as index of first item in user’s click history
- Users *u, v* in same cluster with probability equal to their similarity, $S(u, v)$

MinHash Impl (cont.)

- To further refine clusters, concatenate *p* hash keys for each user. *u,v* in same cluster with probability $S(u, v)^p$
- High precision, low recall
- Can improve recall by hashing user to *q* clusters
- Typical values: *p* ranges from 2 to 4, *q* ranges from 10-20
- Instead of permuting *S*, generate random seed value for each of the *p* X *q* hash functions

MinHash and MapReduce

- Iterate over user click history, and calculate $p \times q$ MinHash values
- Group calculated values into q groups of p hashes
- Concatenate p MinHash values to get cluster-id
- cluster-id = key, user-id = value

MinHash and MapReduce

- Split key-value pairs into shards by hashing keys
- Sort shard by key (cluster-id), so all users mapped into same cluster appear together
- In Reduce phase, obtain cluster membership list, and inverse list (user membership in clusters)
- Prune away low membership clusters
- Store user history and cluster-id's together

26

PLSI Model

Model: $p(s|u;\theta) = \sum_{z=1}^{L} p(z|u)p(s|z)$

- *Z* represents user communities and like-minded users
- Generative model of stories from users with conditional probability distributions (CPDs) *p (z* | *u) and p (s* | *z)*
- Learn CPDs using Expectation Maximization (EM)

PLSI EM Algorithm

- Estimate CPDs
- Minimize $L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log(p(s_t|u_t;\theta))$
- Calculate distribution of hidden variable *Z*

$$\textbf{E-step:}\ \ q^*(z;u,s;\hat{\theta}) = p(z|u,s;\hat{\theta}) = \frac{\hat{p}(s|z)\hat{p}(z|u)}{\sum_{z\in Z}\hat{p}(s|z)\hat{p}(z|u)}$$

- Use distribution as "weights" for calculating CPDs

$$\textbf{M-step:}\ \ p(s|z) = \frac{\sum_u q^*(z;u,s;\hat{\theta})}{\sum_s\sum_u q^*(z;u,s;\hat{\theta})}$$

$$p(z|u) = \frac{\sum_s q^*(z;u,s;\hat{\theta})}{\sum_z\sum_s q^*(z;u,s;\hat{\theta})}$$

MapReduce for EM

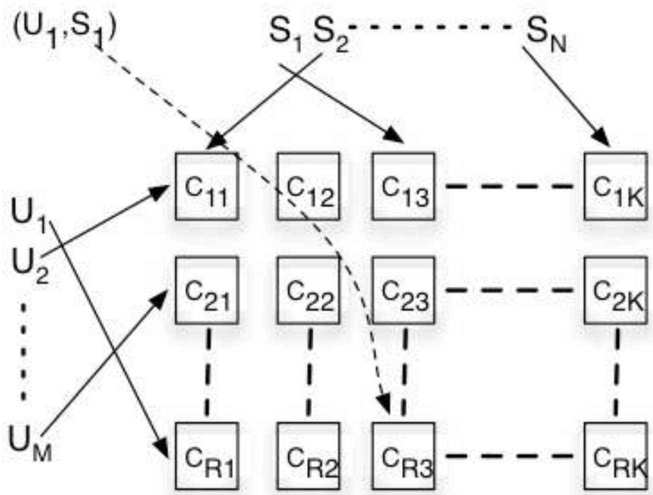
- Rewrite EM equations - replace *p (s / z)*
- E-step: $\mathbf{q^*(z; u, s; \hat{\theta}) = p(z|u, s; \hat{\theta}) = \frac{\frac{N(z,s)}{N(z)}\hat{p}(z|u)}{\sum_{z \in Z} \frac{N(z,s)}{N(z)}\hat{p}(z|u)}}$

$$\mathbf{N(z, s) = \textstyle\sum_u q^*(z; u, s; \hat{\theta})}$$

$$\mathbf{N(z) = \textstyle\sum_s \sum_u q^*(z; u, s; \hat{\theta})}$$

- Calculating *q** can be performed in independently for every *(u,s)* pair in click logs
- *Map* loads CPDs from a single user shard and a single item shard - key

Sharding for EM



- Users and items hashed into *R* and *K* groups
- *Map* loads needed CPDs, calculates q*
- key-value: (*u,q*), (s,q*), (z,q*)*

- Depending on key-value pair received, reduce calculates
  - *N(z,s)* if it receives (s,q*)
  - *p(z | u)* if it receives (u, q*), or *N(z)* for z
  - *N(z)* if it receives (z, q*)

30

PLSI on a dynamic dataset

- Model needs to be retrained whenever there are new users/items
- Approximate model by using learned values of $P(z \mid u)$
- $P(s \mid z)$ can be updated in real time by updating user clusters on a click
- New users get recommendations from covisitation algorithm

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

Making recommendations by algorithm

- Refined clusters from MinHash, weighted clusters from PLSI
- For each story in cluster, calculate score by counting clicks discounted by age
- For covisitation, recommend article *s* by for user *u* adding covisitation entry for each item in C_u and normalizing

Generating candidates for recommendation

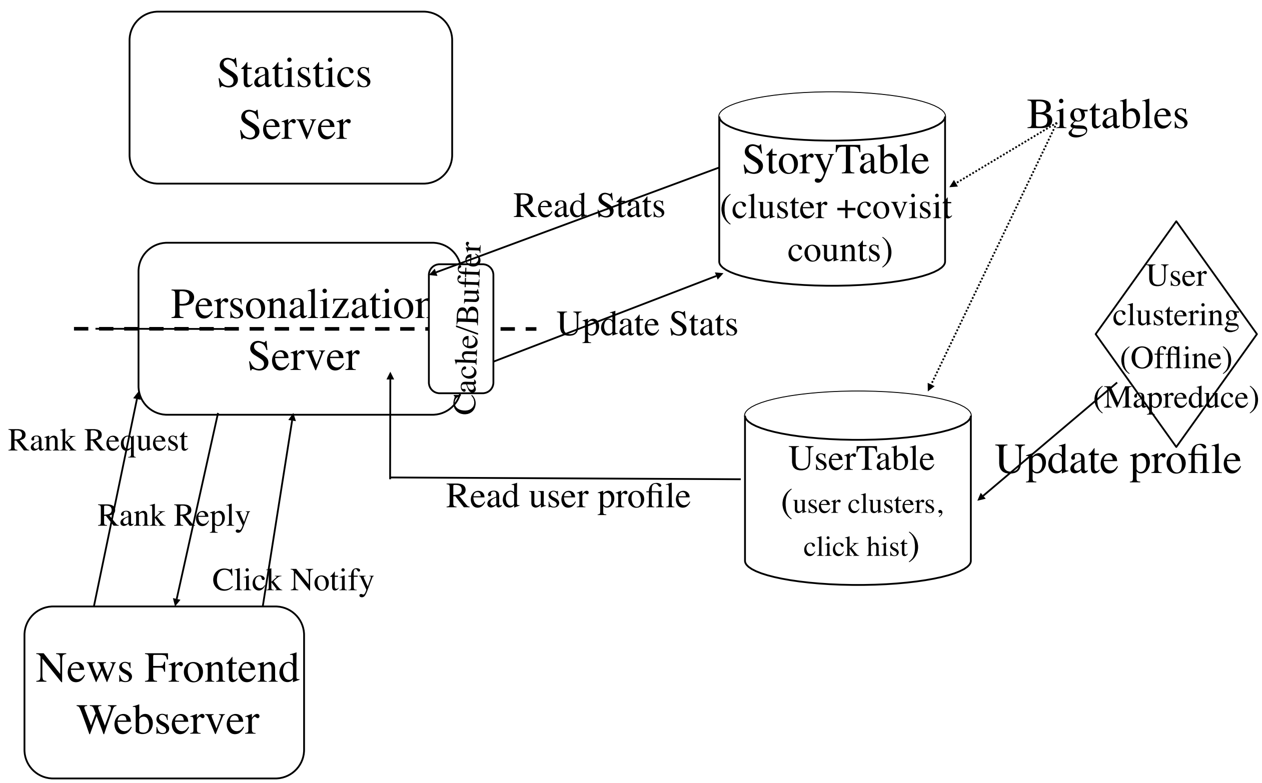
- Use stories from news frontend, based on story freshness, news sections, language, etc.
- Alternatively, use all stories from relevant clusters and covisitation
- Benefits of each set

34

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

System Architecture



*Taken from http://www.sfbayacm.org/events/slides/2007-10-10-google.ppt

System Workflow

- On recommend request - FrontEnd contacts Personalization Server
  - Fetch user clusters and click history from UT
  - Fetch cluster click counts from ST
  - Calculate score for each candidate story *s*
- On story click - FrontEnd contacts Statistics Server
  - Update click histories in UT for every user cluster
  - Update covisitation counts for recent click history

37

Outline

- Introduction and problem
- Related work on recommendation algorithms
- Overview of combined recommendation algorithm
- Overview of MapReduce
- Algorithm implementation details
- Generation of recommendations
- System architecture
- Evaluation of system

Summary of Algorithms

- MinHash
  - Each user clustered into 100 clusters
  - Calculate user u's score for an item s using:

    $\sum_{v \neq u} w(u, v) I_{v,s}$
    where v = all users except for u,
    $w(u, v)$ = similarity between u and v based on cluster membership
    I = indicator of whether v clicked on s

- Correlation
  - Calculate score using same equation as MinHash

39

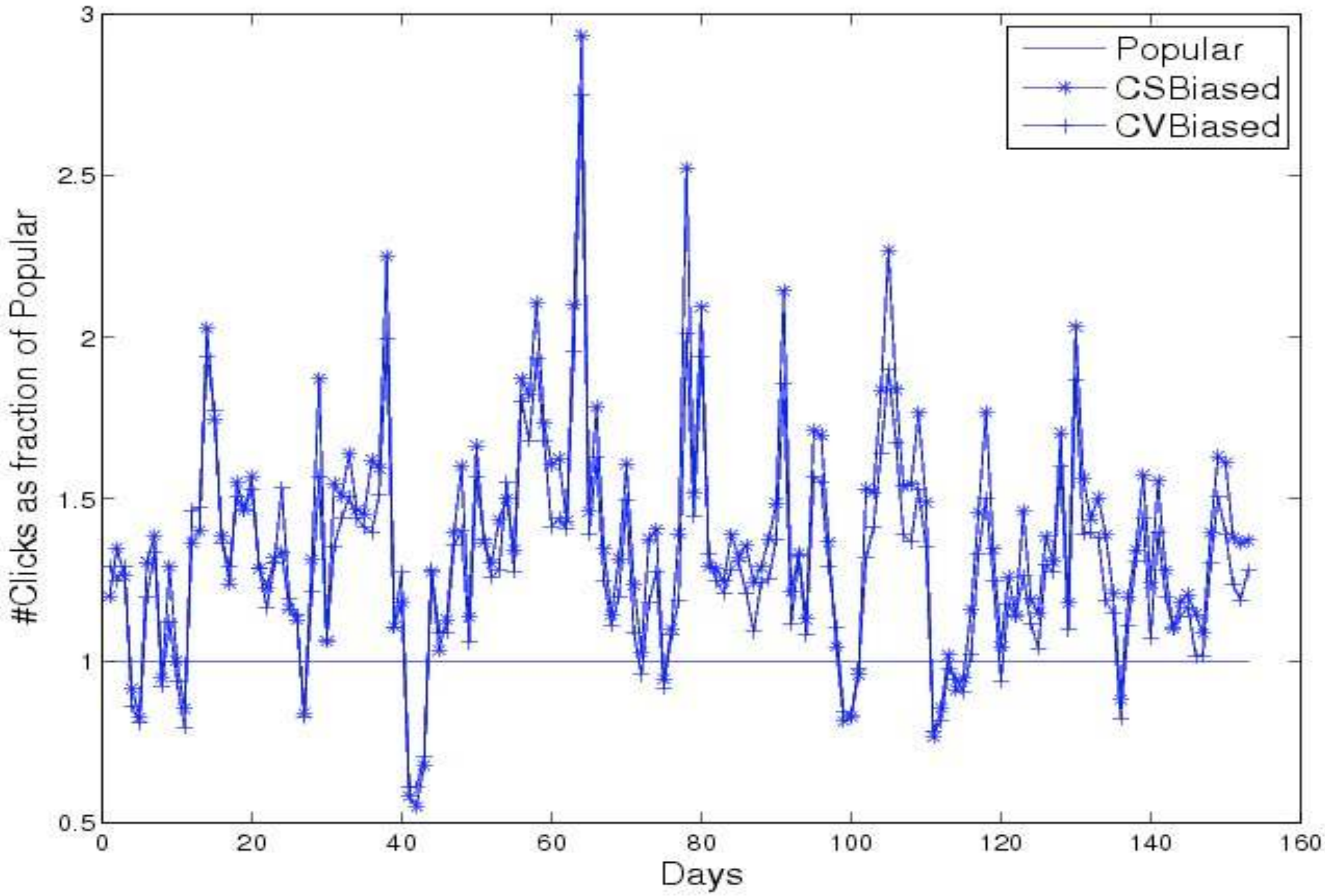
Summary of Algorithms (cont.)

- PLSI
  - Rating is conditional likelihood calculated from $p(s|u) = \sum_z p(z|u)p(s|z)$
  - $p(z|u)$ and $p(s|z)$ estimated using EM
- Rating always falls between 0 and 1, binarized using a threshold

Evaluation on Live Traffic

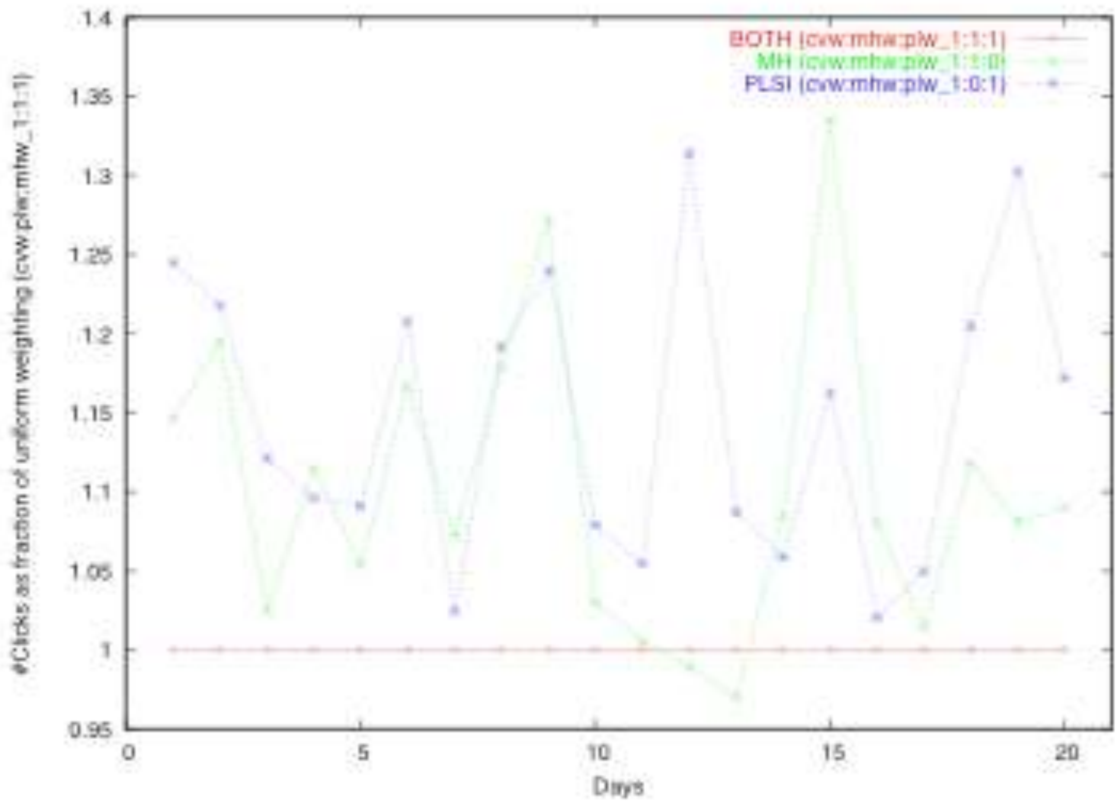
- Compare three algorithms
  - Covisitation - CVBiased
  - Combined PLSI/MinHash - CSBiased
  - Popular
- To test on live traffic
  - Generate recommendation list from each algorithm.
  - Create combined interleaved list alternating the order of the algorithms
  - Count clicks on each algorithms recommendations

Model-based algorithms win



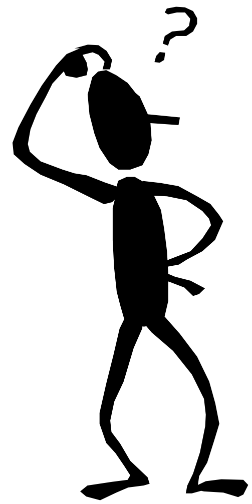
*Taken from http://www.sfbayacm.org/events/slides/2007-10-10-google.ppt

Comparison of models



43

Questions?



44

Equations

E-step: $\mathbf{q^*(z; u, s; \hat{\theta}) = p(z|u, s; \hat{\theta}) = \frac{\frac{N(z,s)}{N(z)}\hat{p}(z|u)}{\sum_{z \in Z} \frac{N(z,s)}{N(z)}\hat{p}(z|u)}}$

$\mathbf{N(z, s) = \sum_u q^*(z; u, s; \hat{\theta})}$

$\mathbf{N(z) = \sum_s \sum_u q^*(z; u, s; \hat{\theta})}$

$\mathbf{p(z|u) = \frac{\sum_s q^*(z;u,s;\hat{\theta})}{\sum_z \sum_s q^*(z;u,s;\hat{\theta})}}$

$$r_{u_a,s_k} = \sum_{i \neq a} I_{u_i,s_k} w(u_a, u_i)$$

w similarity measure, such as Pearson correlation coefficient or cosine similarity

I_{u_i,s_k} indicates whether user i clicked on story k

45