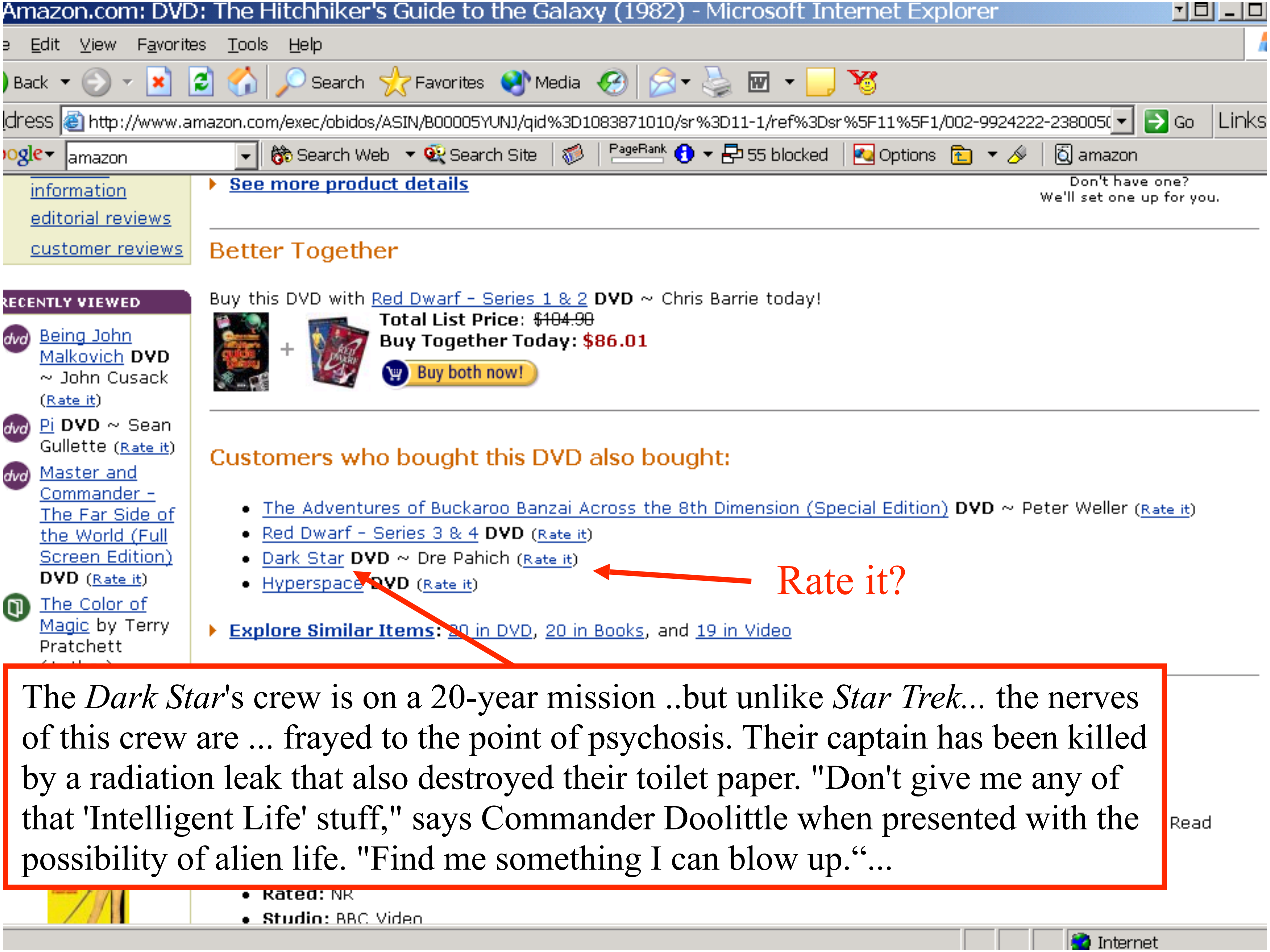
Collaborative Filtering

William W. Cohen
Center for Automated Learning and Discovery
Carnegie Mellon University

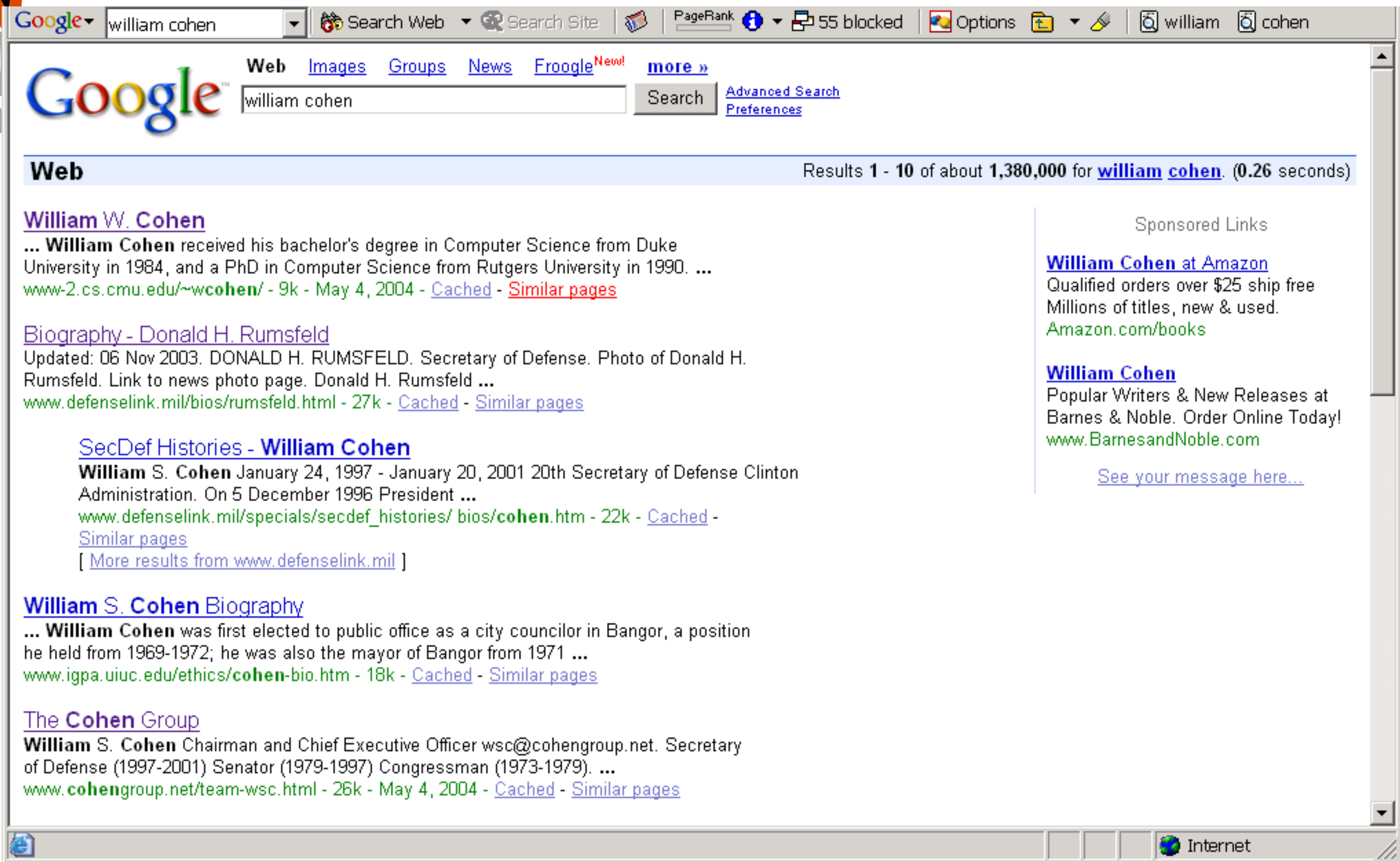
Everyday Examples of Collaborative Filtering...



Rate it?

The *Dark Star*'s crew is on a 20-year mission ..but unlike *Star Trek*... the nerves of this crew are ... frayed to the point of psychosis. Their captain has been killed by a radiation leak that also destroyed their toilet paper. "Don't give me any of that 'Intelligent Life' stuff," says Commander Doolittle when presented with the possibility of alien life. "Find me something I can blow up.“...

Everyday Examples of Collaborative Filtering...



Biography - Donald H. Rumsfeld - Microsoft Internet Explorer

Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.defenselink.mil/bios/rumsfeld.html Go Links

Google william cohen Search Web Search Site PageRank 55 blocked Options william cohen

PageRank is Google's measure of the importance of this page (6/10)

UNITED STATES DEPARTMENT OF DEFENSE

Search GO

Updated: 06 Nov 2003

DONALD H. RUMSFELD

Secretary of Defense

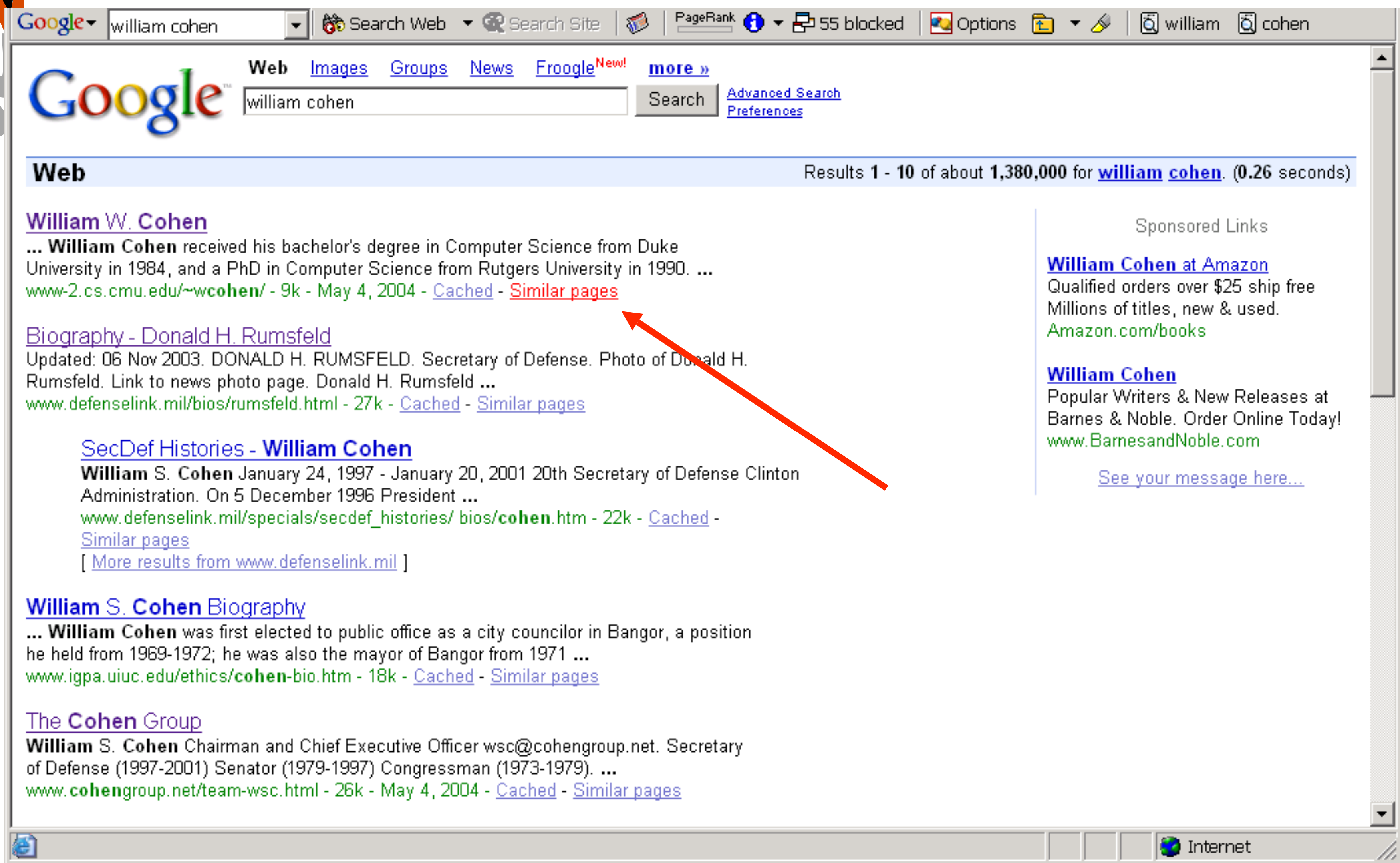
Donald H. Rumsfeld was sworn in as the 21st Secretary of Defense on January 20, 2001. Before assuming his present post, the former Navy pilot had also served as the 13th Secretary of Defense, White House Chief of Staff, U.S. Ambassador to NATO, U.S. Congressman and chief executive officer of two Fortune 500 companies.

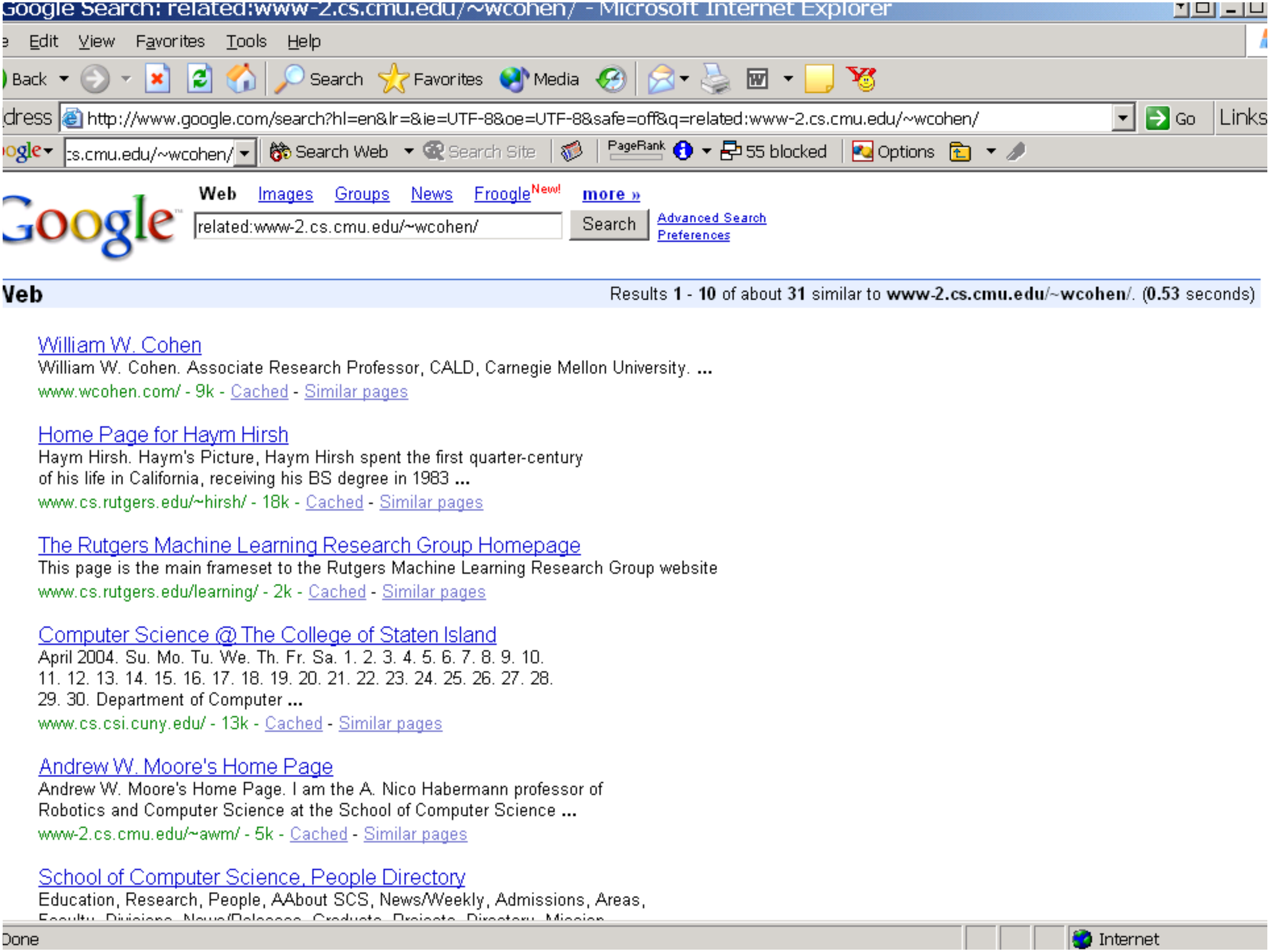
Secretary Rumsfeld is responsible for directing the actions of the Defense Department in response to the terrorist attacks on September 11, 2001. The war is being waged against a backdrop of major change within the Department of Defense. The department has developed a new defense strategy and replaced the old model for sizing forces with a newer approach more relevant to the 21st century. Secretary Rumsfeld proposed and the President approved a significant reorganization of the worldwide command structure, known as the Unified Command Plan, that resulted in the establishment of the U.S. Northern Command and the U.S. Strategic Command, the latter charged with the responsibilities formerly held by the Strategic and Space Commands which were disestablished.

The Department also has refocused its space capabilities and fashioned a new concept of strategic deterrence that

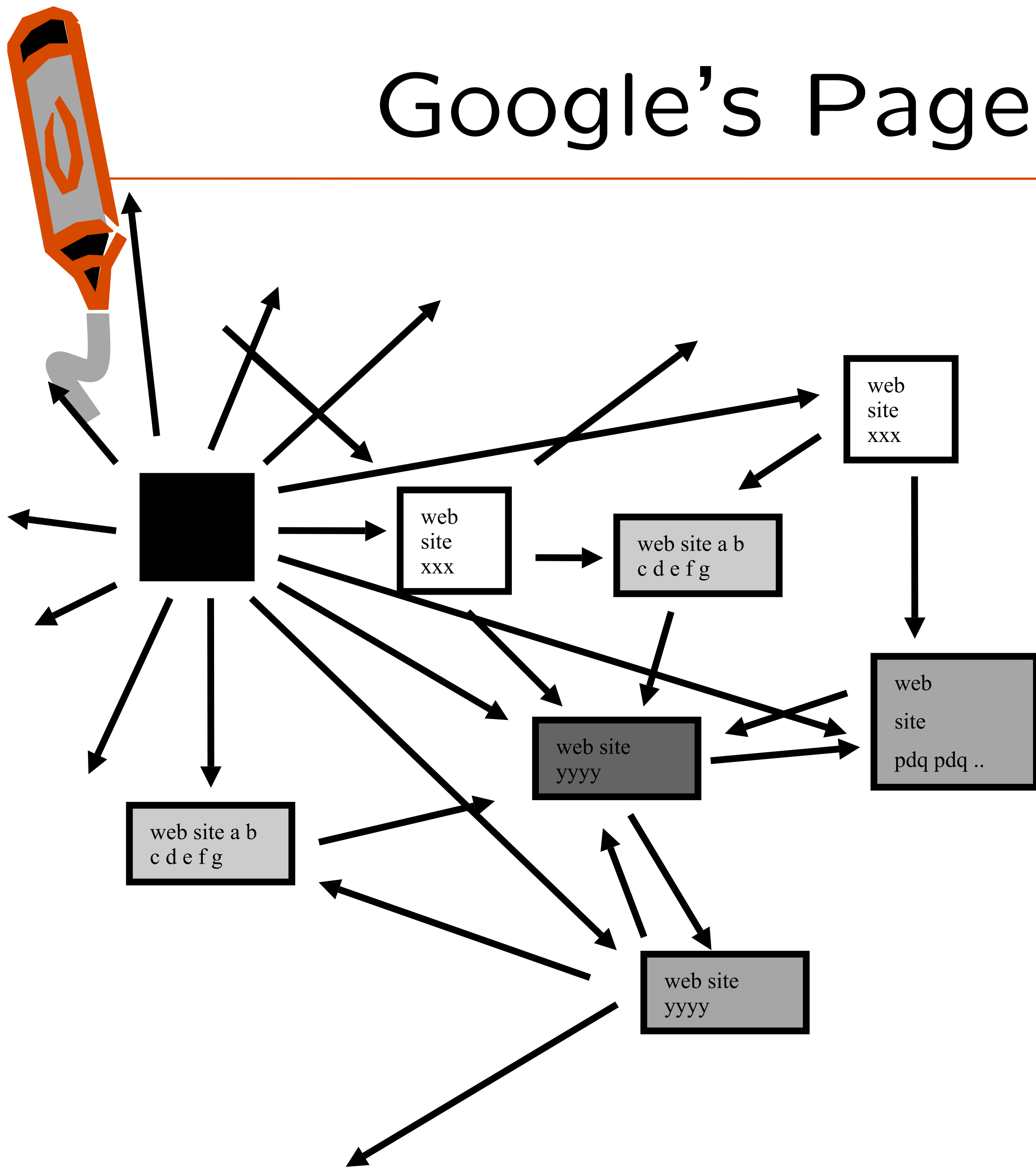
Internet

Everyday Examples of Collaborative Filtering...




Google Search: related:www-2.cs.cmu.edu/~wcohen/ - Microsoft Internet Explorer
Edit View Favorites Tools Help
Back Search Favorites Media
http://www.google.com/search?hl=en&lr=&ie=UTF-8&oe=UTF-8&safe=off&q=related:www-2.cs.cmu.edu/~wcohen/
Go Links
Search Web Search Site PageRank 55 blocked Options
Web Images Groups News Froogle New! more »
related:www-2.cs.cmu.edu/~wcohen/
Search
Advanced Search
Preferences
Results 1 - 10 of about 31 similar to www-2.cs.cmu.edu/~wcohen/. (0.53 seconds)
William W. Cohen
William W. Cohen. Associate Research Professor, CALD, Carnegie Mellon University. ...
www.wcohen.com/ - 9k - Cached - Similar pages
Home Page for Haym Hirsh
Haym Hirsh. Haym's Picture, Haym Hirsh spent the first quarter-century
of his life in California, receiving his BS degree in 1983 ...
www.cs.rutgers.edu/~hirsh/ - 18k - Cached - Similar pages
The Rutgers Machine Learning Research Group Homepage
This page is the main frameset to the Rutgers Machine Learning Research Group website
www.cs.rutgers.edu/learning/ - 2k - Cached - Similar pages
Computer Science @ The College of Staten Island
April 2004. Su. Mo. Tu. We. Th. Fr. Sa. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10.
11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28.
29. 30. Department of Computer ...
www.cs.csi.cuny.edu/ - 13k - Cached - Similar pages
Andrew W. Moore's Home Page
Andrew W. Moore's Home Page. I am the A. Nico Habermann professor of
Robotics and Computer Science at the School of Computer Science ...
www-2.cs.cmu.edu/~awm/ - 5k - Cached - Similar pages
School of Computer Science, People Directory
Education, Research, People, AAbout SCS, News/Weekly, Admissions, Areas,
Done
Internet

Google's PageRank



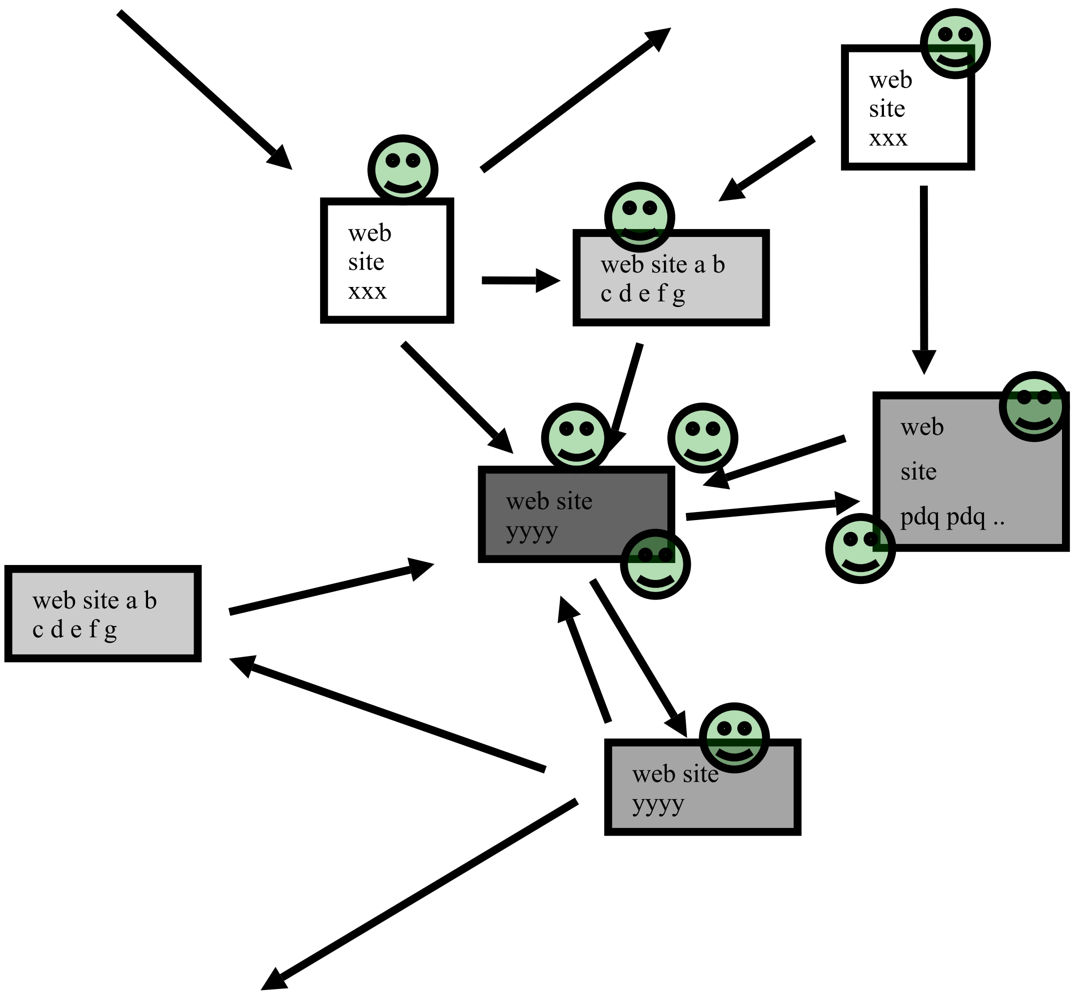
Inlinks are "good" (recommendations)

Inlinks from a "good" site are better than inlinks from a "bad" site

but inlinks from sites with many outlinks are not as "good"...

"Good" and "bad" are relative.

Google's PageRank

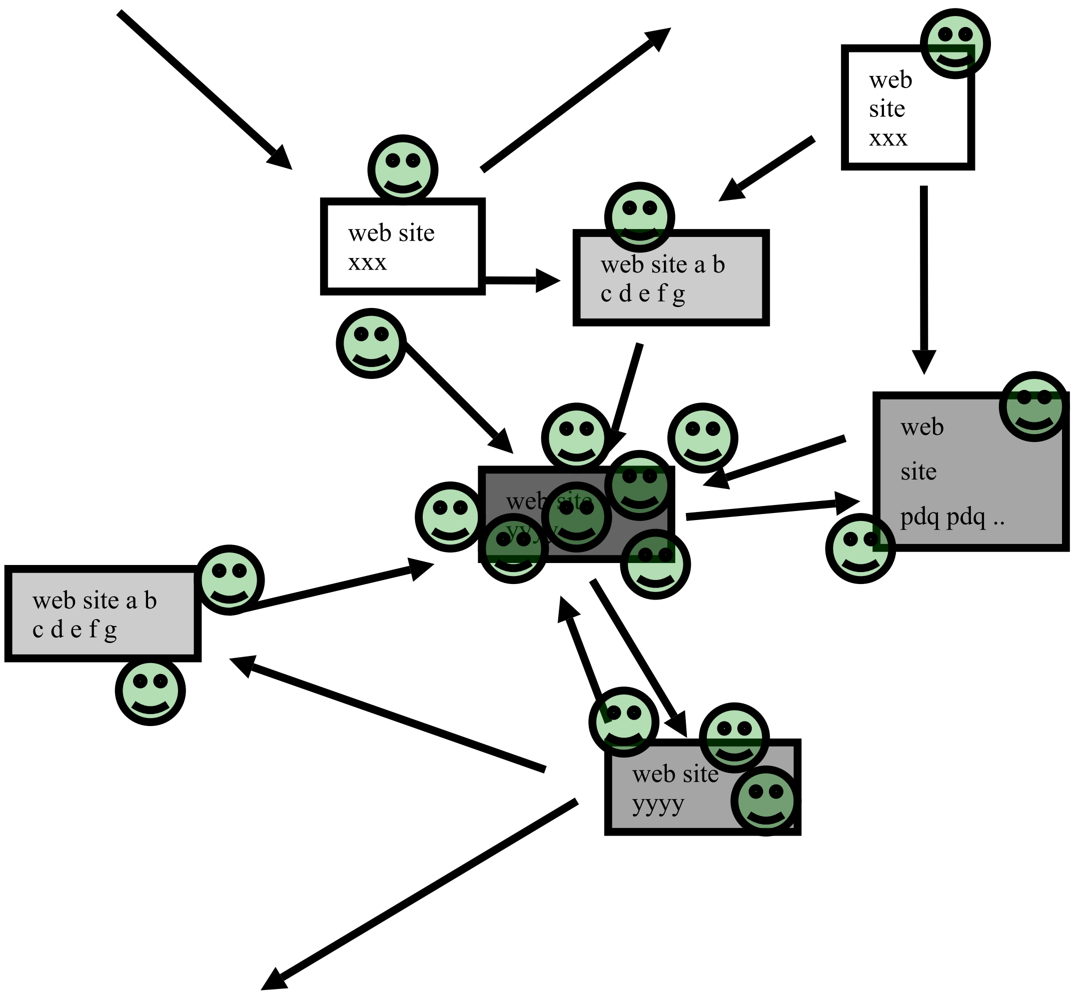


Imagine a "pagehopper" that always either

- follows a random link, or
- jumps to random page

Google's PageRank

(Brin & Page, http://www-db.stanford.edu/~backrub/google.html)



Imagine a "pagehopper" that always either

- follows a random link, or
- jumps to random page

PageRank ranks pages by the amount of time the pagehopper spends on a page:

- or, if there were many pagehoppers, PageRank is the expected "crowd size"

Everyday Examples of Collaborative Filtering...

- Bestseller lists
- Top 40 music lists
- The “recent returns” shelf at the library
- Unmarked but well-used paths thru the woods
- The printer room at work
- Many weblogs
- “Read any good books lately?”
- ....
- Common insight: personal tastes are correlated:
  - If Alice and Bob both like X and Alice likes Y then Bob is more likely to like Y
  - especially (perhaps) if Bob knows Alice

Outline

- Non-systematic survey of some CF systems
  - CF as basis for a virtual community
  - memory-based recommendation algorithms
  - visualizing user-user via item distances
  - CF versus content filtering
- Algorithms for CF
- CF with different inputs
  - true ratings
  - assumed/implicit ratings
- Conclusions/Summary

BellCore's MovieRecommender

- Recommending And Evaluating Choices In A Virtual Community Of Use. Will Hill, Larry Stead, Mark Rosenstein and George Furnas, Bellcore; CHI 1995

By **virtual community** we mean "a group of people who share characteristics and interact in essence or effect only". In other words, people in a Virtual Community influence each other *as though* they interacted but they *do not interact*. Thus we ask: "Is it possible to arrange for people to share some of the personalized informational benefits of community involvement without the associated communications costs?"

MovieRecommender Goals

Recommendations should:

- simultaneously ease and encourage rather than replace social processes....should make it easy to participate while leaving in hooks for people to pursue more personal relationships if they wish.
- be for sets of people not just individuals...multi-person recommending is often important, for example, when two or more people want to choose a video to watch together.
- be from people not a black box machine or so-called ”agent”.
- tell how much confidence to place in them, in other words they should include indications of how accurate they are.

BellCore's MovieRecommender

- Participants sent email to videos@bellcore.com
- System replied with a list of 500 movies to rate on a 1-10 scale (250 random, 250 popular)
  - Only subset need to be rated
- New participant P sends in rated movies via email
- System compares ratings for P to ratings of (a random sample of) previous users
- Most similar users are used to predict scores for unrated movies (more later)
- System returns recommendations in an email message.

Suggested Videos for: John A. Jamus.

Your must-see list with predicted ratings:

- 7.0 "Alien (1979)"
- 6.5 "Blade Runner"
- 6.2 "Close Encounters Of The Third Kind (1977)"

Your video categories with average ratings:

- 6.7 "Action/Adventure"
- 6.5 "Science Fiction/Fantasy"
- 6.3 "Children/Family"
- 6.0 "Mystery/Suspense"
- 5.9 "Comedy"
- 5.8 "Drama"

The viewing patterns of 243 viewers were consulted. Patterns of 7 viewers were found to be most similar. Correlation with target viewer:

- 0.59 viewer-130 (unlisted@merl.com)
- 0.55 bullert,jane r (bullert@cc.bellcore.com)
- 0.51 jan_arst (jan_arst@khdld.decnet.philips.nl)
- 0.46 Ken Cross (moose@denali.EE.CORNELL.EDU)
- 0.42 rskt (rskt@cc.bellcore.com)
- 0.41 kkgg (kkgg@Athena.MIT.EDU)
- 0.41 bnn (bnn@cc.bellcore.com)

By category, their joint ratings recommend:

- Action/Adventure:
  - "Excalibur" 8.0, 4 viewers
  - "Apocalypse Now" 7.2, 4 viewers
  - "Platoon" 8.3, 3 viewers
- Science Fiction/Fantasy:
  - "Total Recall" 7.2, 5 viewers
- Children/Family:
  - "Wizard Of Oz, The" 8.5, 4 viewers
  - "Mary Poppins" 7.7, 3 viewers

Mystery/Suspense:
- "Silence Of The Lambs, The" 9.3, 3 viewers

Comedy:
- "National Lampoon's Animal House" 7.5, 4 viewers
- "Driving Miss Daisy" 7.5, 4 viewers
- "Hannah and Her Sisters" 8.0, 3 viewers

Drama:
- "It's A Wonderful Life" 8.0, 5 viewers
- "Dead Poets Society" 7.0, 5 viewers
- "Rain Man" 7.5, 4 viewers

Correlation of predicted ratings with your actual ratings is: 0.64 This number measures ability to evaluate movies accurately for you. 0.15 means low ability. 0.85 means very good ability. 0.50 means fair ability.

BellCore's MovieRecommender

- Evaluation:
  - Withhold 10% of the ratings of each user to use as a test set
  - Measure correlation between predicted ratings and actual ratings for test-set movie/user pairs

Figure 3 *Two Scatterplots of Actual Ratings by Predicted Ratings. Plot on left shows movie critics as predictor (r=0.22). Plot on right shows virtual community as predictor (r=0.62) (all values are jittered for the purpose of visual presentation, 3269 predictions each for 291 users)*

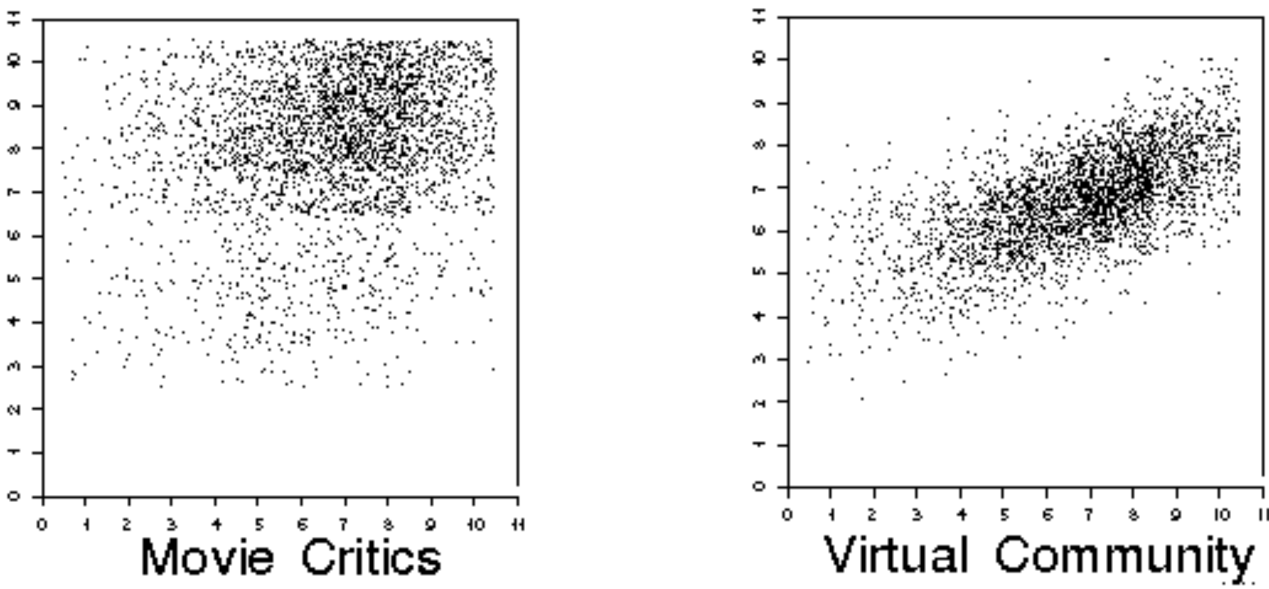
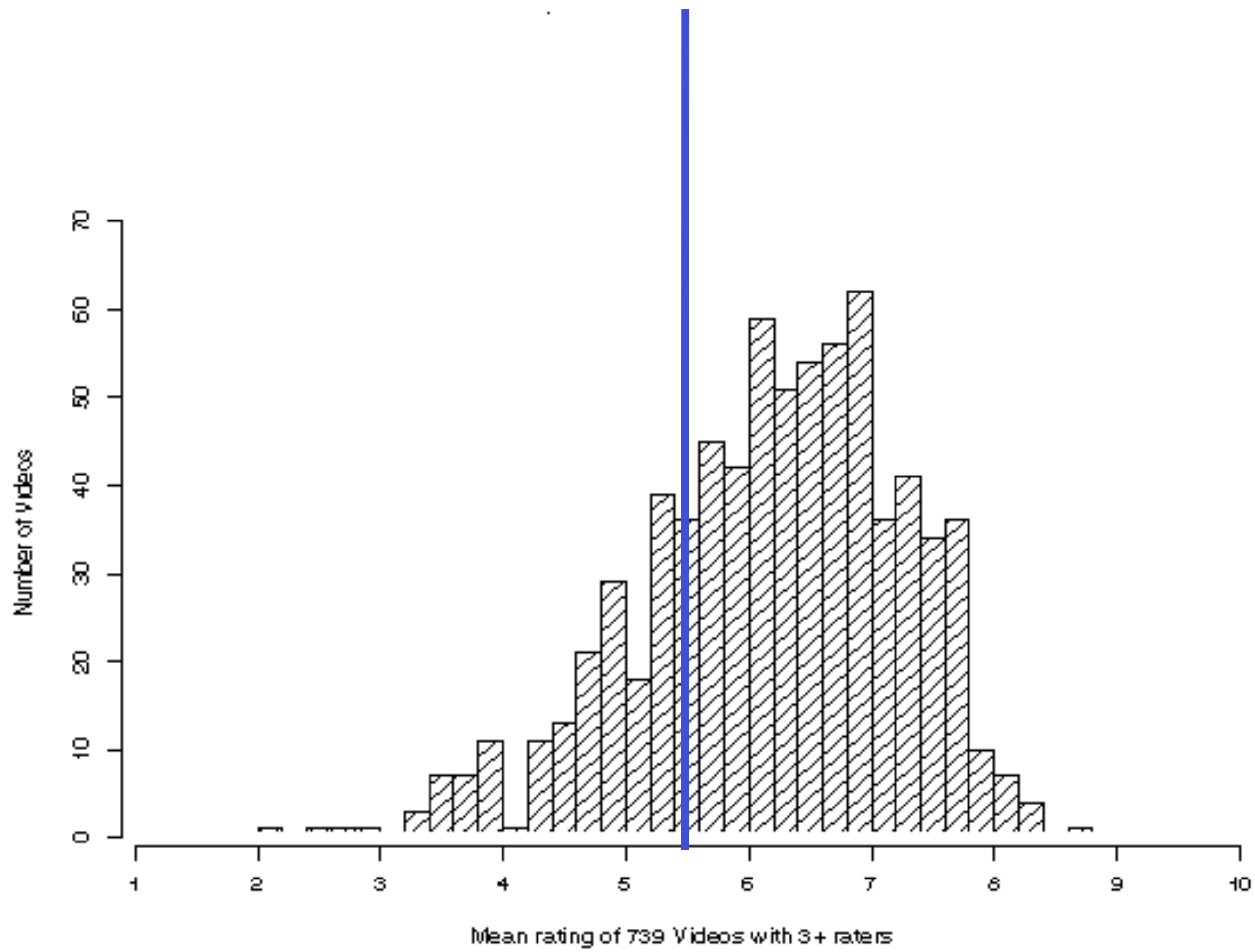


Figure 2 *Distribution of Video Mean Ratings*



Another key observation: *rated movies* tend to have *positive* ratings:

i.e., people rate what they watch, and watch what they like

Question: Can observation replace explicit rating?

BellCore's MovieRecommender

- Participants sent email to videos@bellcore.com
- System replied with a list of 500 movies to rate
- New participant P sends in rated movies via email
- System compares ratings for P to ratings of (a random sample of) previous users
- Most similar users are used to predict scores for unrated movies
  - Empirical Analysis of Predictive Algorithms for Collaborative Filtering Breese, Heckerman, Kadie, UAI98
- System returns recommendations in an email message.

Algorithms for Collaborative Filtering 1: Memory-Based Algorithms (Breese et al, UAI98)

- $v_{i,j}$ = vote of user i on item j
- I_i = items for which user i has voted
- Mean vote for i is

$$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

- Predicted vote for "active user" a is weighted sum

$$p_{a,j} = \overline{v}_a + \kappa \sum_{i=1}^{n} \underbrace{w(a,i)}_{\text{weights of } n \text{ similar users}} (v_{i,j} - \overline{v}_i)$$

normalizer

Algorithms for Collaborative Filtering 1: Memory-Based Algorithms (Breese et al, UAI98)

- K-nearest neighbor
- Pearson correlation coefficient (Resnick '94, Grouplens):

$$w(a,i) = \begin{cases} 1 & \text{if } i \in \text{neighbors}(a) \\ 0 & \text{else} \end{cases}$$

- Cosine distance (from IR)

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}}$$

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

Algorithms for Collaborative Filtering 1: Memory-Based Algorithms (Breese et al, UAI98)

- Cosine with "inverse user frequency" $f_i = \log(n/n_j)$, where n is number of users, n_j is number of users voting for item j

$$w(a,i) = \frac{\sum_j f_j \sum_j f_j v_{a,j} v_{i,j} - (\sum_j f_j v_{a,j})(\sum_j f_j v_{i,j}))}{\sqrt{UV}}$$

where

$$U = \sum_j f_j (\sum_j f_j v_{a,j}^2 - (\sum_j f_j v_{a,j})^2)$$

$$V = \sum_j f_j (\sum_j f_j v_{i,j}^2 - (\sum_j f_j v_{i,j})^2)$$

Algorithms for Collaborative Filtering 1: Memory-Based Algorithms (Breese et al, UAI98)

- Evaluation:
  - split users into train/test sets
  - for each user a in the test set:
    - split a's votes into observed (I) and to-predict (P)
    - measure average absolute deviation between predicted and actual votes in P
    - predict votes in P, and form a ranked list
    - assume (a) utility of k-th item in list is max($v_{a,j}$-d,0), where d is a "default vote" (b) probability of reaching rank k drops exponentially in k. Score a list by its expected utility R_a
  - average R_a over all test users

Algorithms for Collaborative Filtering 1: Memory-Based Algorithms (Breese et al, UAI98)

soccer score

	EachMovie, Rank Scoring			
Algorithm	Given2	Given5	Given10	AllBut1
CR+	**41.60**	42.33	**41.46**	**23.16**
VSIM	**42.45**	**42.12**	**40.15**	22.07
BC	38.06	36.68	34.98	21.38
BN	28.64	30.50	33.16	23.49
POP	30.80	28.90	28.01	13.94
RD	*0.75*	*0.75*	*0.78*	*0.78*

Why are these numbers **worse**?

golf score

	EachMovie, Absolute Deviation			
Algorithm	Given2	Given5	Given10	AllBut1
CR	**1.257**	1.139	**1.069**	**0.994**
BC	1.127	1.144	1.138	1.103
BN	1.143	1.154	1.139	1.066
VSIM	**2.113**	**2.177**	**2.235**	**2.136**
RD	*0.022*	*0.023*	*0.025*	*0.043*

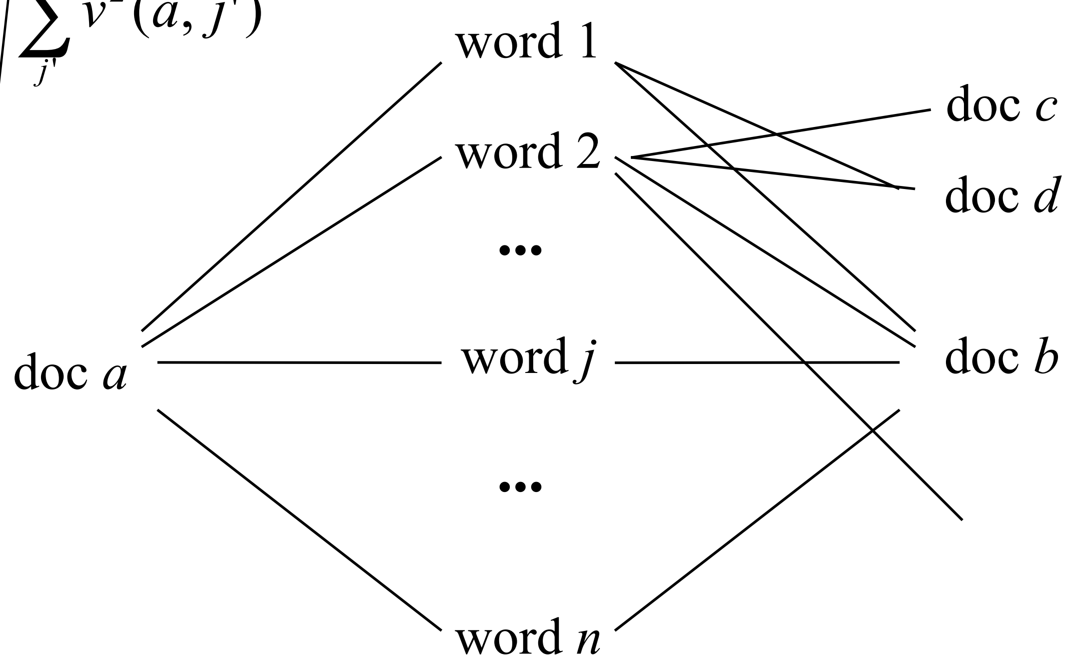
Visualizing Cosine Distance

similarity of doc a to doc b = $sim(a,b) = \sum_{\text{word } i} \frac{v(a,j)}{\sqrt{\sum_{j'} v^2(a,j')}} \times \frac{v(b,j)}{\sqrt{\sum_{j'} v^2(b,j')}}$

$$= A' \times B'$$

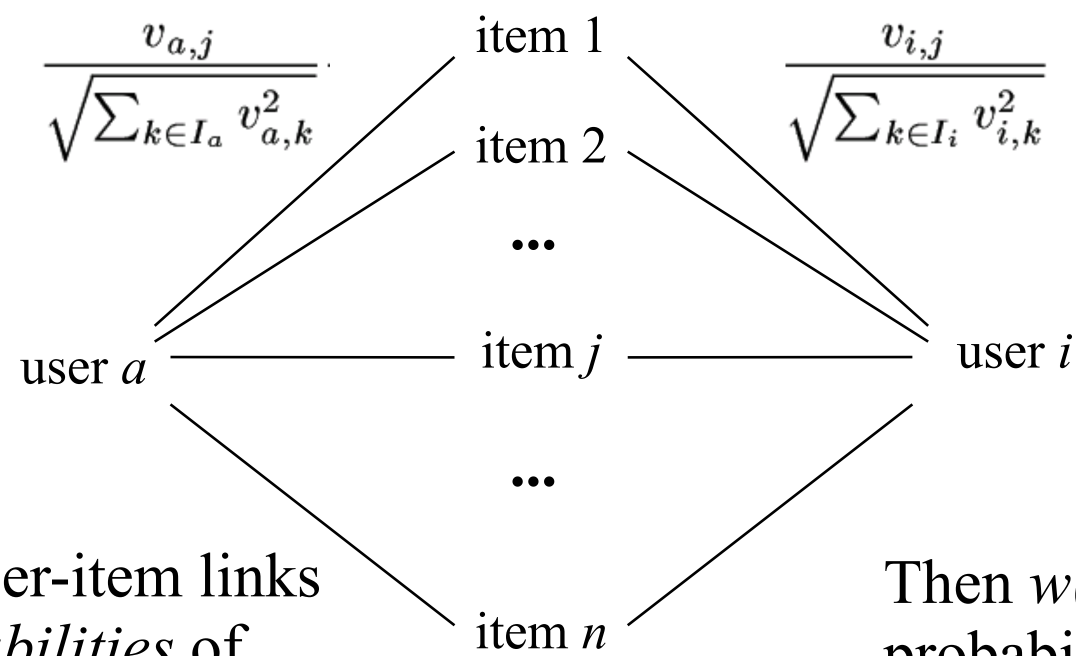
Let $\vec{A} = < \dots, v(a,j), \dots >$

Let $\vec{A}' = \frac{\vec{A}}{\| \vec{A} \|} = \frac{\vec{A}}{\sqrt{\sum_{j'} v^2(a,j')}}$



Visualizing Cosine Distance

distance from user a to user i = $$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}}$$

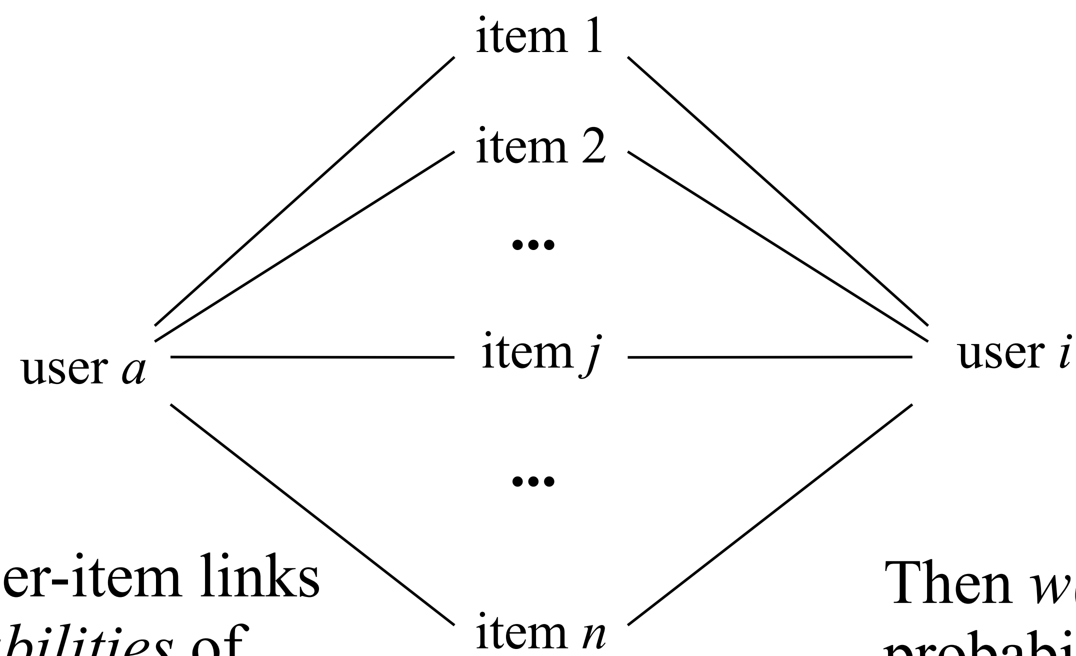


Suppose user-item links were *probabilities* of following a link

Then *w(a,i)* is probability of a and i "meeting"

Visualizing Cosine Distance

Approximating Matrix Multiplication for Pattern Recognition Tasks, Cohen & Lewis, SODA 97—explores connection between cosine distance/inner product and random walks



Suppose user-item links were *probabilities* of following a link

Then *w(a,i)* is probability of *a* and *i* "meeting"

Outline

- Non-systematic survey of some CF systems
  - CF as basis for a virtual community
  - memory-based recommendation algorithms
  - visualizing user-user via item distances
  - CF versus content filtering
- Algorithms for CF
- CF with different inputs
  - true ratings
  - assumed/implicit ratings

LIBRA Book Recommender

Content-Based Book Recommending Using Learning for Text Categorization. Raymond J. Mooney, Loriene Roy, Univ Texas/Austin; DL-2000

[CF] assumes that a given user's tastes are generally the same as another user ... Items that have not been rated by a sufficient number of users cannot be effectively recommended. Unfortunately, statistics on library use indicate that most books are utilized by very few patrons. ... [CF] approaches ... recommend popular titles, perpetuating homogeneity.... this approach raises concerns about privacy and access to proprietary customer data.

LIBRA Book Recommender

- Database of textual descriptions + meta-information about books (from Amazon.com's website)
  - title, authors, synopses, published reviews, customer comments, related authors, related titles, and subject terms.
- Users provides 1-10 rating for training books
- System learns a model of the user
  - Naive Bayes classifier predicts Prob(user rating>5|book)
- System explains ratings in terms of "informative features" and explains features in terms of examples

LIBRA Book Recommender

The Fabric of Reality:
The Science of Parallel Universes- And Its Implications
by David Deutsch recommended because:

Slot	Word	Strength
DESCRIPTION	MULTIVERSE	75.12
DESCRIPTION	UNIVERSES	25.08
DESCRIPTION	REALITY	22.96
DESCRIPTION	UNIVERSE	15.55
DESCRIPTION	QUANTUM	14.54
DESCRIPTION	INTELLECT	13.86
DESCRIPTION	OKAY	13.75
DESCRIPTION	RESERVATIONS	11.56

The word UNIVERSES is positive due to your ratings:

Title	Rating	Count
The Life of the Cosmos	10	15
Before the Beginning : Our Universe and Others	8	7
Unveiling the Edge of Time	10	3
Black Holes : A Traveler's Guide	9	3
The Inflationary Universe	9	2

LIBRA Book Recommender

Key differences from MovieRecommender:

• *vs* collaborative filtering, recommendation is based on properties of the *item being recommended,* not tastes of other users

• *vs* memory-based techniques, **LIBRA** builds an *explicit model* of the user's tastes (expressed as weights for different words)

The Fabric of Reality:
The Science of Parallel Universes- And Its Implications
by David Deutsch recommended because:

Slot	Word	Strength
DESCRIPTION	MULTIVERSE	75.12
DESCRIPTION	UNIVERSES	25.08
DESCRIPTION	REALITY	22.96
DESCRIPTION	UNIVERSE	15.55
DESCRIPTION	QUANTUM	14.54
DESCRIPTION	INTELLECT	13.86
DESCRIPTION	OKAY	13.75
DESCRIPTION	RESERVATIONS	11.56

….

LIBRA Book Recommender

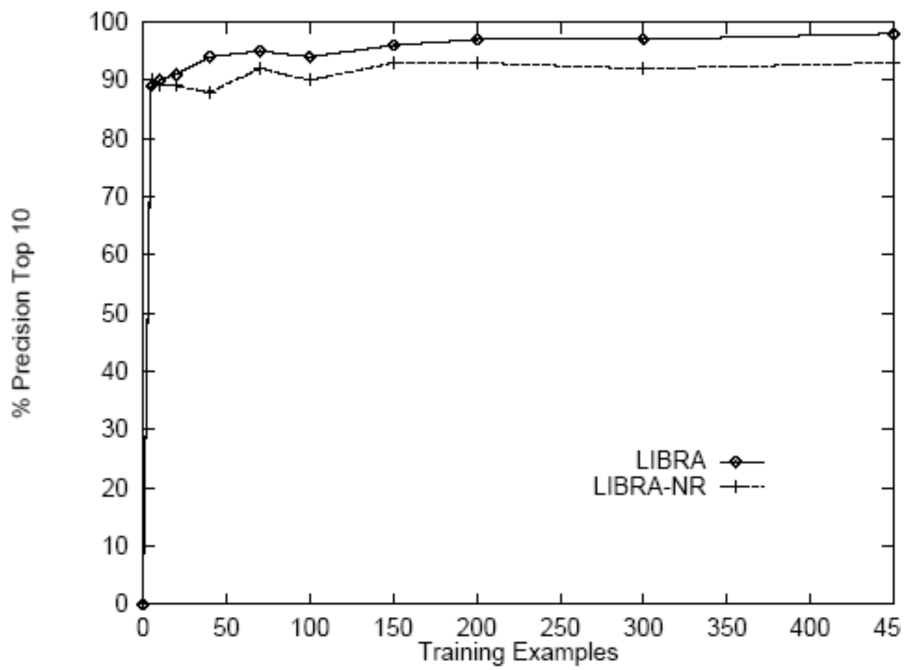


Figure 2: MYST **Precision at Top 10**

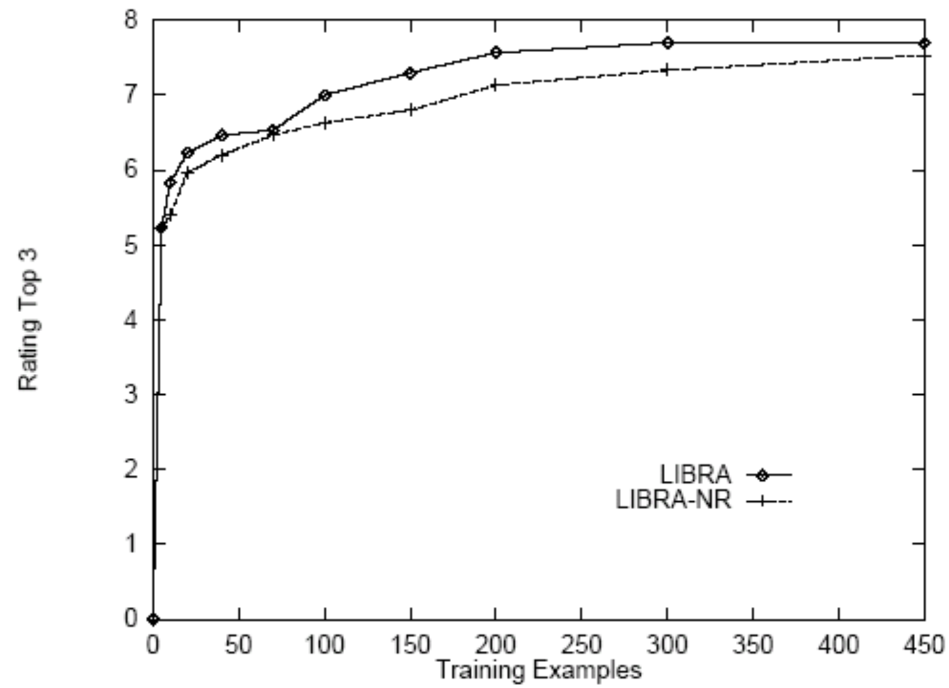
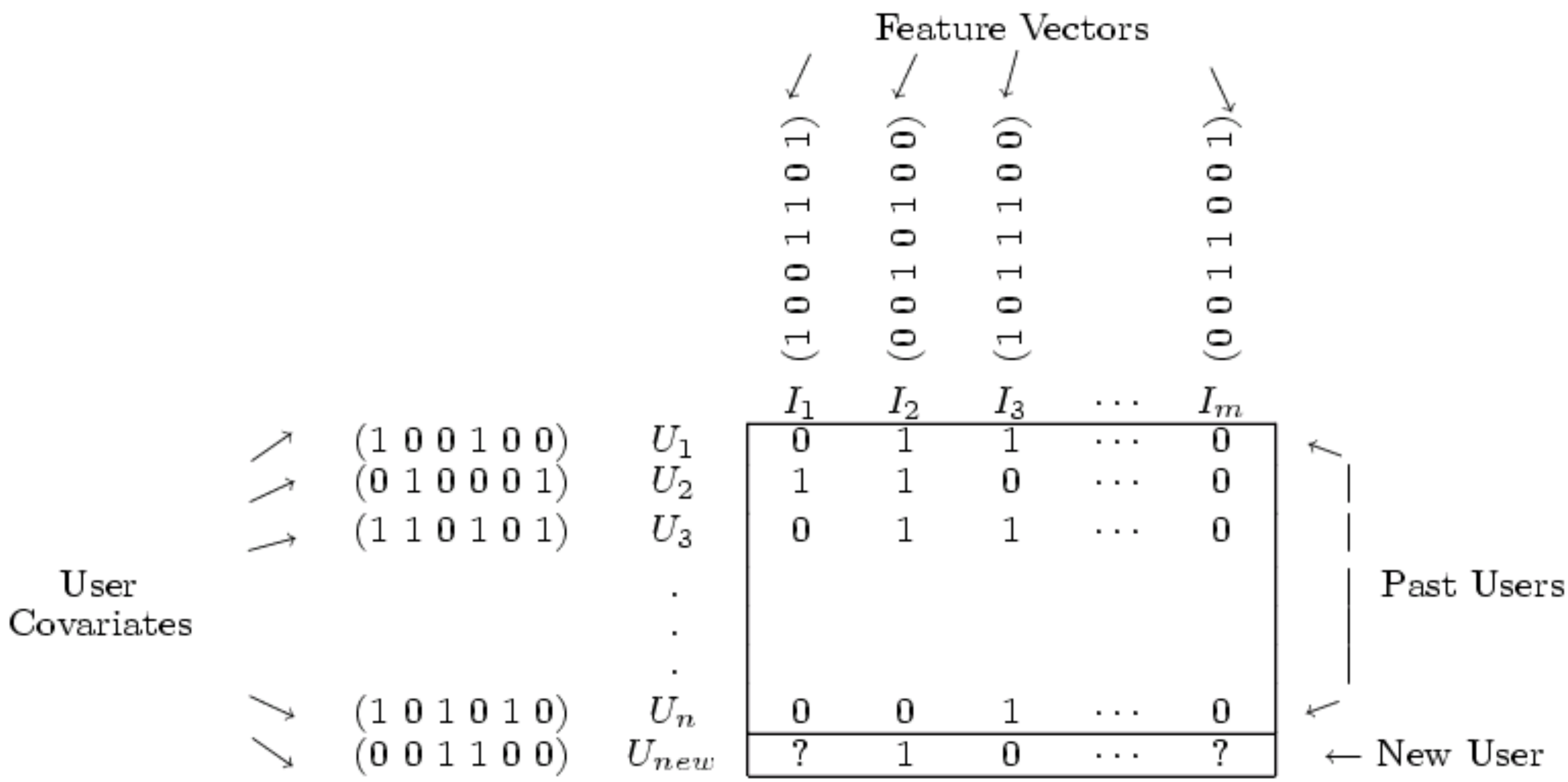


Figure 3: SF **Average Rating of Top 3**

LIBRA-NR = no related author/title features

Collaborative + Content Filtering

(Basu et al, AAAI98; Condliff et al, AI-STATS99)



Collaborative + Content Filtering

(Basu et al, AAAI98; Condliff et al, AI-STATS99)

		Airplane	Matrix	Room with a View	...	Hidalgo
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	9	7	2		7
Carol	*53,F,20k*	8		9		
...						
Kumar	*25,M,22k*	9	3			6
U_a	*48,M,81k*	4	7	?	?	?

Collaborative + Content Filtering As Classification (Basu, Hirsh, Cohen, AAAI98)

Classification task: map **(user,movie)** pair into **{likes,dislikes}**

Training data: known likes/dislikes

Test data: active users

Features: **any** properties of user/movie pair

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
...						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	0	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

Examples: *genre(U,M), age(U,M), income(U,M),...*

- *genre(Carol,Matrix) = action*
- *income(Kumar,Hidalgo) = 22k/year*

Features: **any** properties of user/movie pair *(U,M)*

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
...						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	0	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

Examples: *usersWhoLikedMovie(U,M):*

- *usersWhoLikedMovie(Carol,Hidalgo) = {Joe,…,Kumar}*
- *usersWhoLikedMovie(U_a, Matrix) = {Joe,…}*

Features: **any** properties of user/movie pair *(U,M)*

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
…						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	0	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

Examples: *moviesLikedByUser(M,U):*

- *moviesLikedByUser(*,Joe) = {Airplane,Matrix,...,Hidalgo}*
- *actionMoviesLikedByUser(*,Joe)={Matrix,Hidalgo}*

Features: **any** properties of user/movie pair *(U,M)*

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
…						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	0	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

genre={romance}, age=48, sex=male, income=81k,
usersWhoLikedMovie={Carol}, moviesLikedByUser={Matrix,Airplane}, ...

Features: **any** properties of user/movie pair *(U,M)*

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
...						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	1	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

genre={romance}, age=48, sex=male, income=81k, usersWhoLikedMovie={Carol}, moviesLikedByUser={Matrix,Airplane}, ...

genre={action}, age=48, sex=male, income=81k, usersWhoLikedMovie = {Joe,Kumar}, moviesLikedByUser={Matrix,Airplane},...

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	1	1	0		1
Carol	*53,F,20k*	1		1		0
...						
Kumar	*25,M,22k*	1	0	0		1
U_a	*48,M,81k*	1	1	?	?	?

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

genre={romance}, age=48, sex=male, income=81k, usersWhoLikedMovie={Carol}, moviesLikedByUser={Matrix,Airplane}, ...

genre={action}, age=48, sex=male, income=81k, usersWhoLikedMovie = {Joe,Kumar}, moviesLikedByUser={Matrix,Airplane},...

- Classification learning algorithm: rule learning (RIPPER)
  - If *NakedGun33/13* ∈*moviesLikedByUser* and *Joe*∈ *usersWhoLikedMovie* and *genre=comedy* then predict *likes(U,M)*
  - If *age>12* and *age<17* and *HolyGrail* ∈ *moviesLikedByUser* ***and*** *director=MelBrooks* then predict *likes(U,M)*
  - If *Ishtar* ∈ *moviesLikedByUser* then predict *likes(U,M)*

Collaborative + Content Filtering As Classification (Basu et al, AAAI98)

- Classification learning algorithm: rule learning (RIPPER)

  - If *NakedGun33/13* $\in$ *moviesLikedByUser* and *Joe* $\in$ *usersWhoLikedMovie* and *genre=comedy* then predict *likes(U,M)*

  - If *age>12* and *age<17* and *HolyGrail* $\in$ *moviesLikedByUser* ***and*** *director=MelBrooks* then predict *likes(U,M)*

  - If *Ishtar* $\in$ *moviesLikedByUser* then predict *likes(U,M)*

- Important difference from memory-based approaches:

  - again, Ripper builds an explicit model—of how user's tastes relate items, and to the tastes of other users

Basu et al 98 - results

- Evaluation:
  - Predict liked(U,M)="M in top quartile of U's ranking" from features, evaluate recall and precision
  - Features:
    - Collaborative: UsersWhoLikedMovie, UsersWhoDislikedMovie, MoviesLikedByUser
    - Content: Actors, Directors, Genre, MPAA rating, ...
    - Hybrid: ComediesLikedByUser, DramasLikedByUser, UsersWhoLikedFewDramas, ...
- Results: at same level of recall (about 33%)
  - Ripper with collaborative features only is worse than the original MovieRecommender (by about 5 pts precision – 73 vs 78)
  - Ripper with hybrid features is better than MovieRecommender (by about 5 pts precision)

Technical Paper Recommendation

(Basu, Hirsh, Cohen, Neville-Manning, JAIR 2001)

A **special case** of CF is when items and users can both be represented over the **same** feature set (e.g., with text)

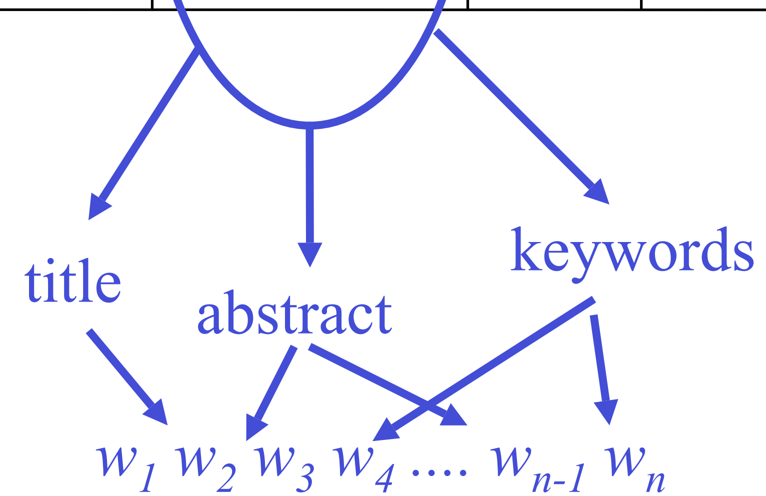
		Shallow parsing with conditional random fields.Sha and Pereira, …	Hidden Markov Support Vector Machines, Altun et al, …	…	Large Margin Classification Using the Perceptron Algorithm, Freund and Schapire
Haym	*cs.rutgers.edu/ ~hirsh*				
William	*cs.cmu.edu/ ~wcohen*				
…					
Soumen	*cs.ucb.edu/ ~soumen*				

How similar are these two documents?

Technical Paper Recommendation
(Basu et al, JAIR 2001)

A **special case** of CF is when items and users can both be represented over the **same** feature set (e.g., with text)

		Shallow parsing with conditional random fields.Sha and Pereira, …	Hidden Markov Support Vector Machines, Altun et al, **…**	**…**	Large Margin Classification Using the Perceptron Algorithm, Freund and Schapire
Haym	*cs.rutgers.edu/ ~hirsh*				
William	*cs.cmu.edu/ ~wcohen*				
…					
Soumen	*cs.ucb.edu/ ~soumen*				

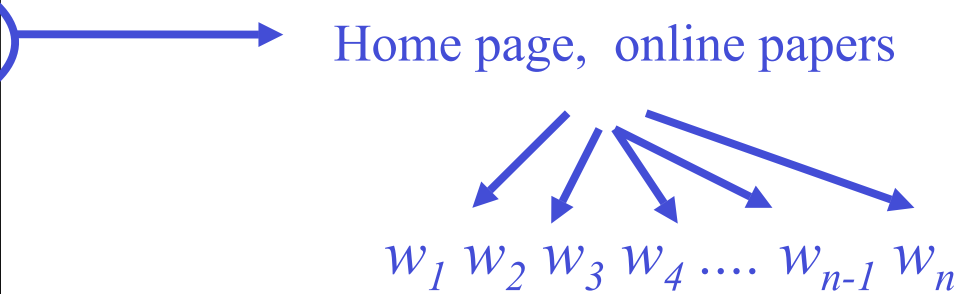


Technical Paper Recommendation

(Basu et al, JAIR 2001)

A **special case** of CF is when items and users can both be represented over the **same** feature set (e.g., with text)

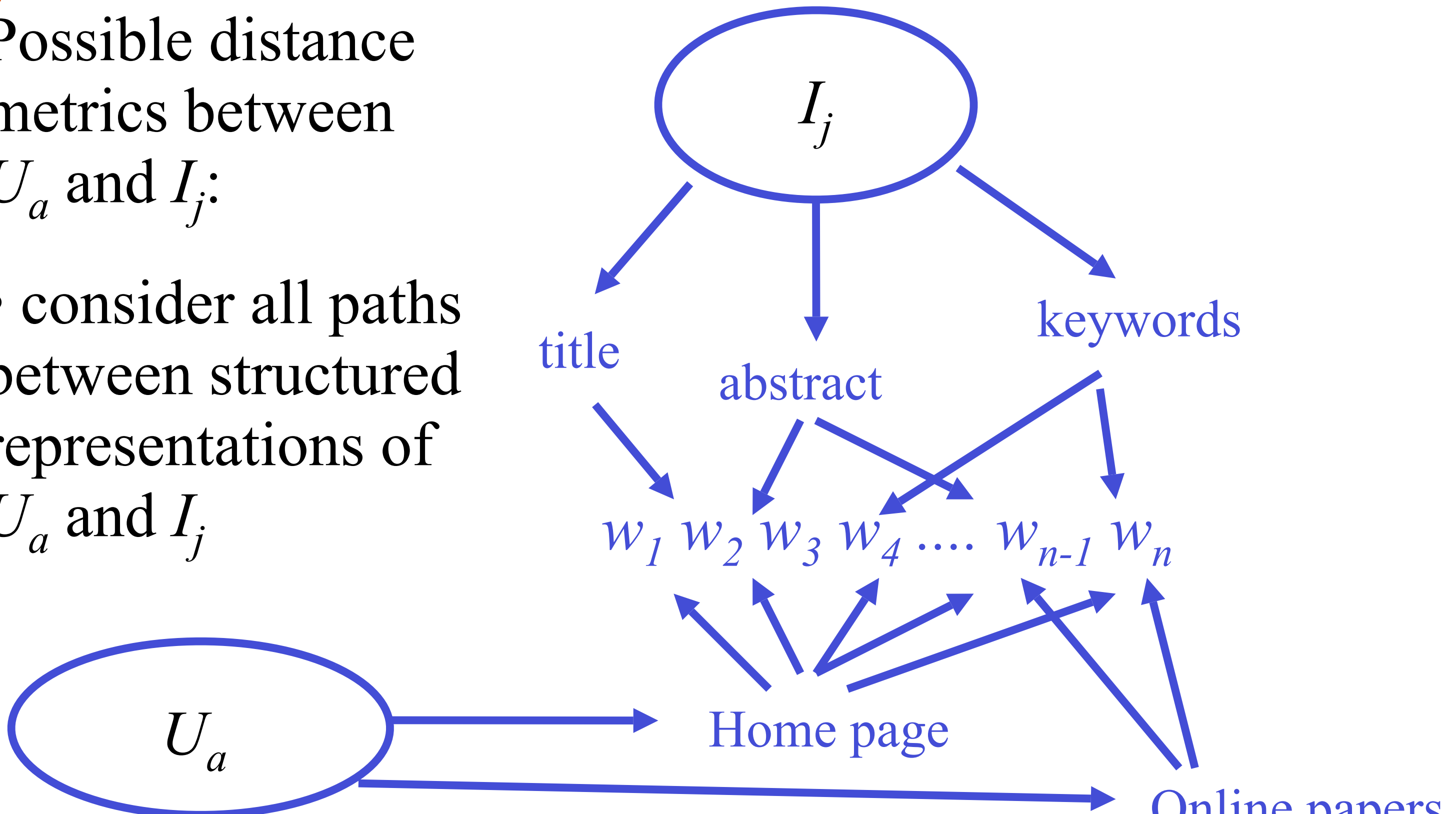
		Shallow parsing with conditional random fields.Sha and Pereira, …	Hidden Markov Support Vector Machines, Altun et al, **…**	**…**	Large Margin Classification Using the Perceptron Algorithm, Freund and Schapire
Haym	*cs.rutgers.edu/~hirsh*				
William	*cs.cmu.edu/~wcohen*				
…					
Soumen	*cs.ucb.edu/~soumen*				



Technical Paper Recommendation
(Basu et al, JAIR 2001)

Possible distance metrics between U_a and I_j:

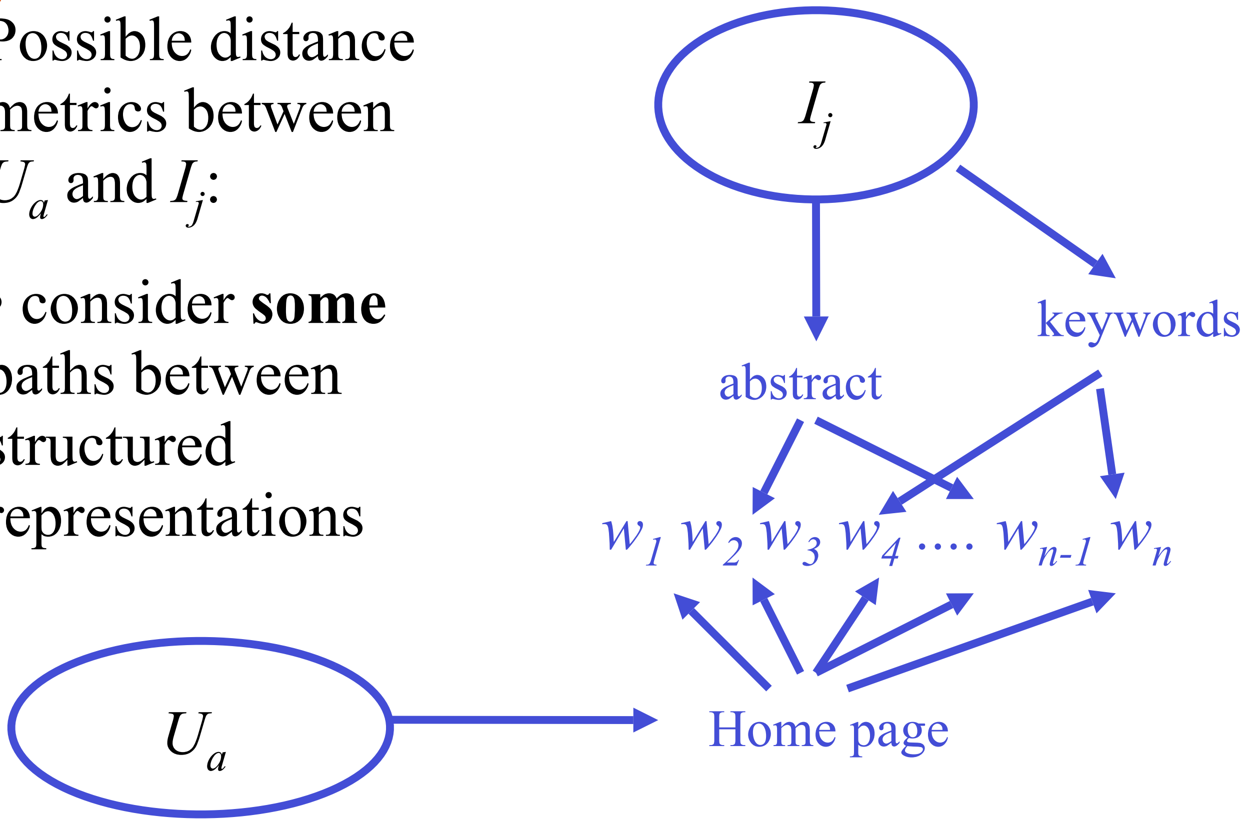
• consider all paths between structured representations of U_a and I_j



Technical Paper Recommendation
(Basu et al, JAIR 2001)

Possible distance metrics between U_a and I_j:

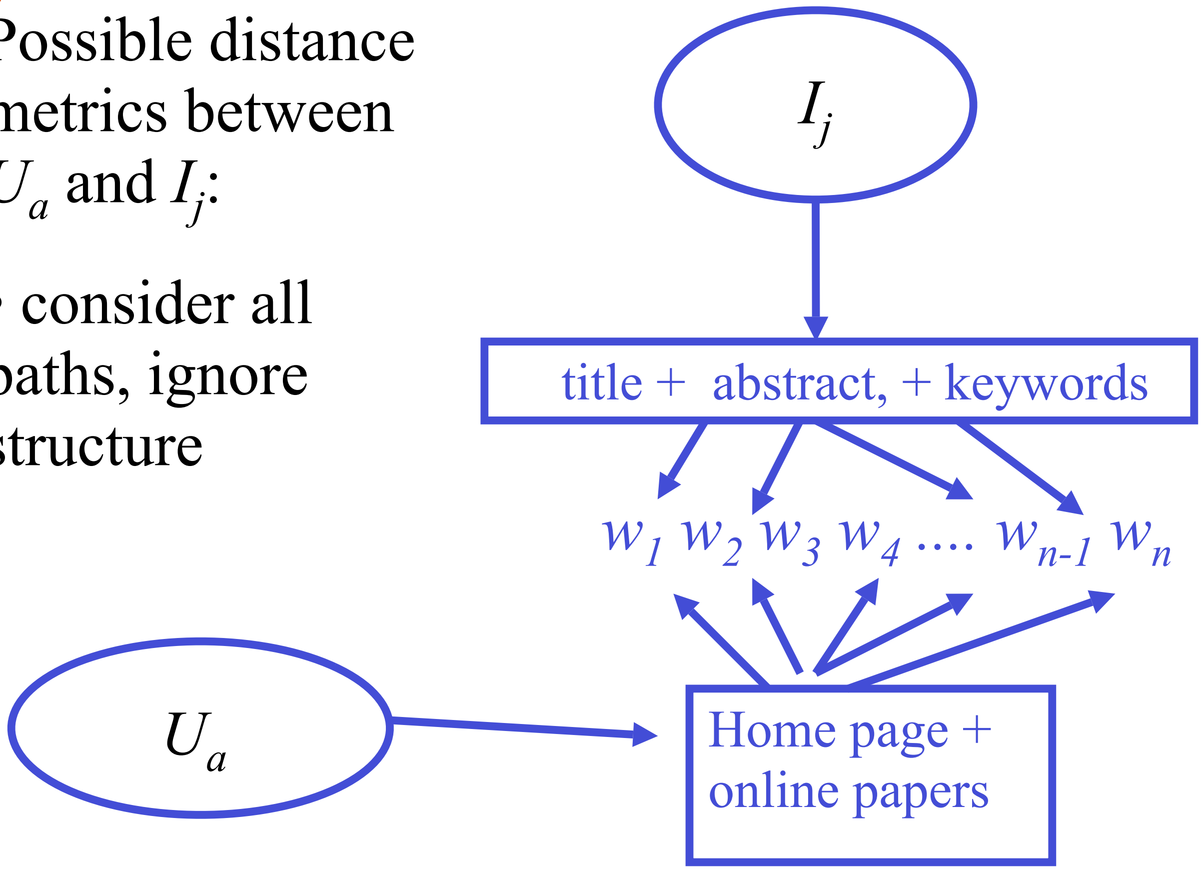
- consider **some** paths between structured representations



Technical Paper Recommendation
(Basu et al, JAIR 2001)

Possible distance metrics between U_a and I_j:

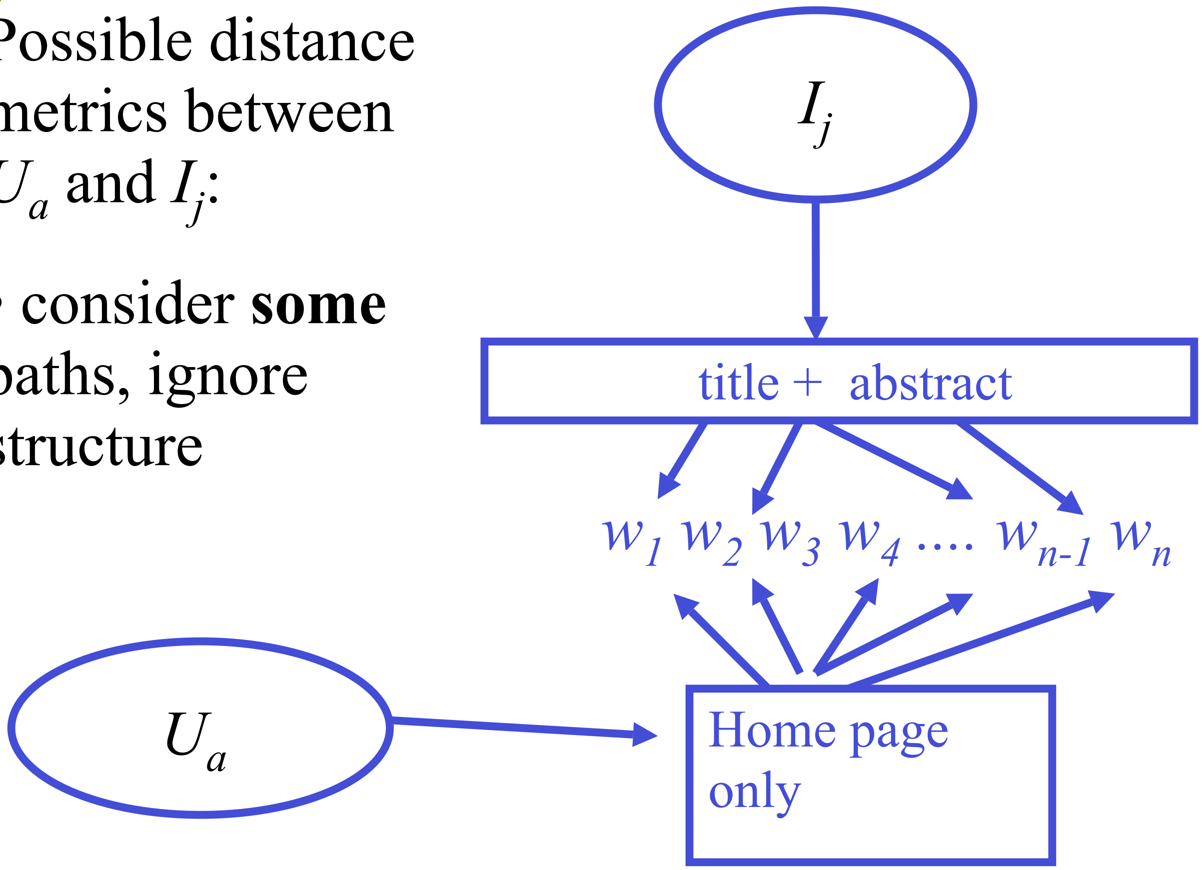
- consider all paths, ignore structure



Technical Paper Recommendation
(Basu et al, JAIR 2001)

Possible distance metrics between U_a and I_j:

• consider **some** paths, ignore structure



Technical Paper Recommendation
(Basu et al, JAIR 2001)

- Use WHIRL (Datalog + built-in cosine distances) to formulate structure similarity queries
  - Product of TFIDF-weighted cosine distances over each part of structure
- Evaluation
  - Try and predict stated reviewer preferences in AAAI self-selection process
    - Noisy, since not all reviewers examine all papers
  - Measure precision in top 10, and top 30

Technical Paper Recommendation

2001)

Methods(s)	*Top 10*	*Top 30*
kNN	0.294	0.154
$ExtendedDirectBayes$	0.300	0.129

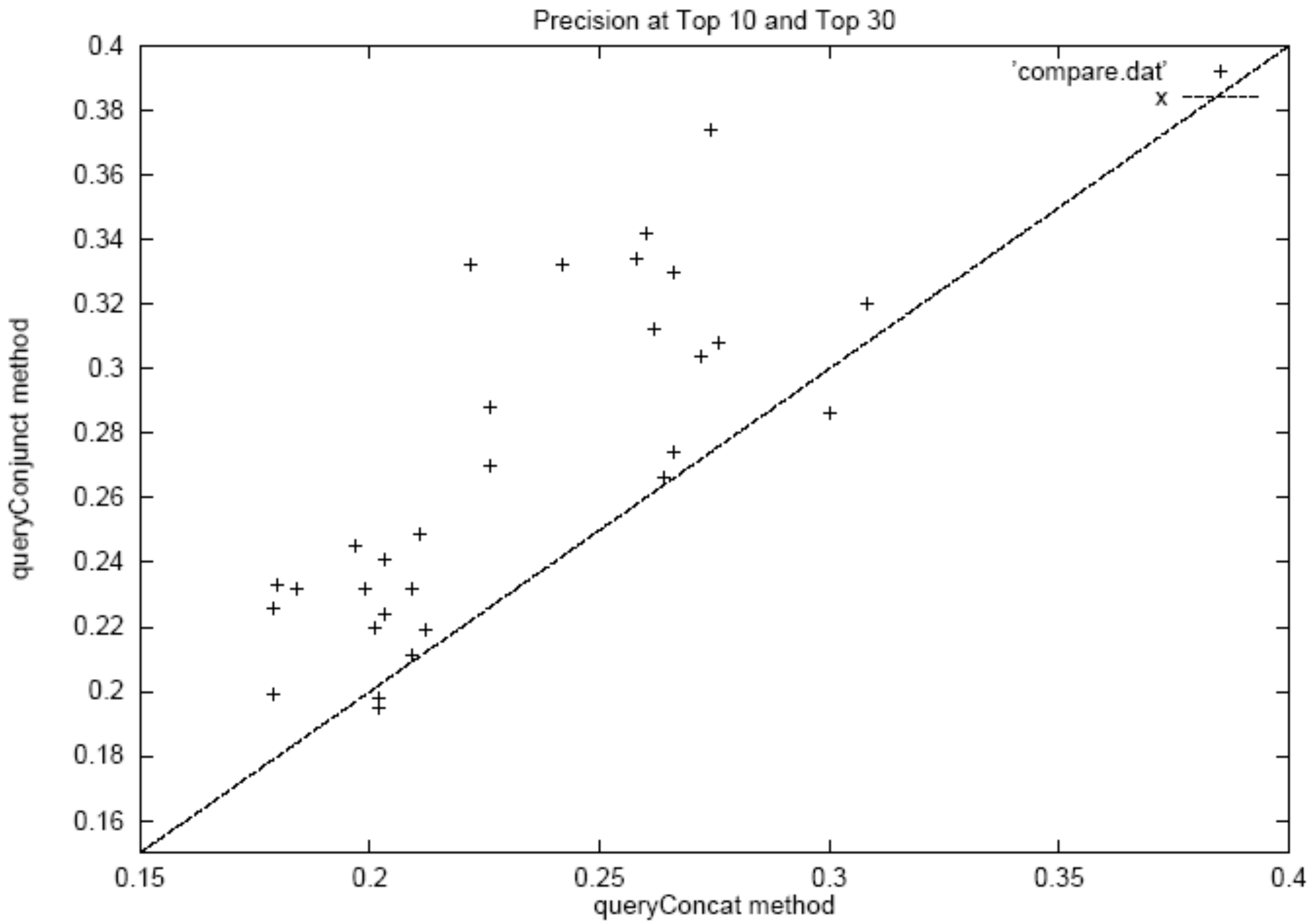
Source(s)	A	K	T	AK	AT	KT	AKT
p(Top10)	0.248	0.260	0.234	0.266	0.274	0.308	0.330
h(Top10)	0.210	0.284	0.232	0.288	0.270	0.320	0.332
ph(Top10)	0.334	0.304	0.332	0.312	0.342	0.286	0.374
p(Top30)	0.194	0.201	0.177	0.198	0.195	0.220	0.232
h(Top30)	0.169	0.217	0.183	0.226	0.199	0.232	0.232
ph(Top30)	0.245	0.219	0.233	0.224	0.241	0.211	0.249

p=papers, h=homePage

A=abstract, K=keywords, T=title

structured similarity queries with WHIRL

Technical Paper Recommendation
(Basu et al, JAIR 2001)



Structure *vs* no structure