Introduction to Social Computing

Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

Social Networking

Billionaires’ Shuffle

2007

Facebook in 2004.02

2008

at **23** and $**1.5** billion later...

2008

Alexa as of Nov. 2008	USA	CHINA	Global
1	Google	Baidu	Yahoo
2	Yahoo	**QQ**	Google
3	**Myspace**	Sina	**YouTube**
4	**YouTube**	Google.cn	Windows Live
5	**Facebook**	Taobao	**Facebook**
6	Windows Live	163	MSN
7	MSN	Yahoo	**Myspace**
8	**Wikipedia**	Google	**Wikipedia**
9	EBay	Sohu	**Blogger**
10	AOL	Youku	Yahoo.jp

Outline

- Introduction to Social Computing
- Query Suggestion
- Collaborative Filtering
- Human Computation
- Privacy and Trust in Social Network
- Social Computing in Education

Social Computing Road Map

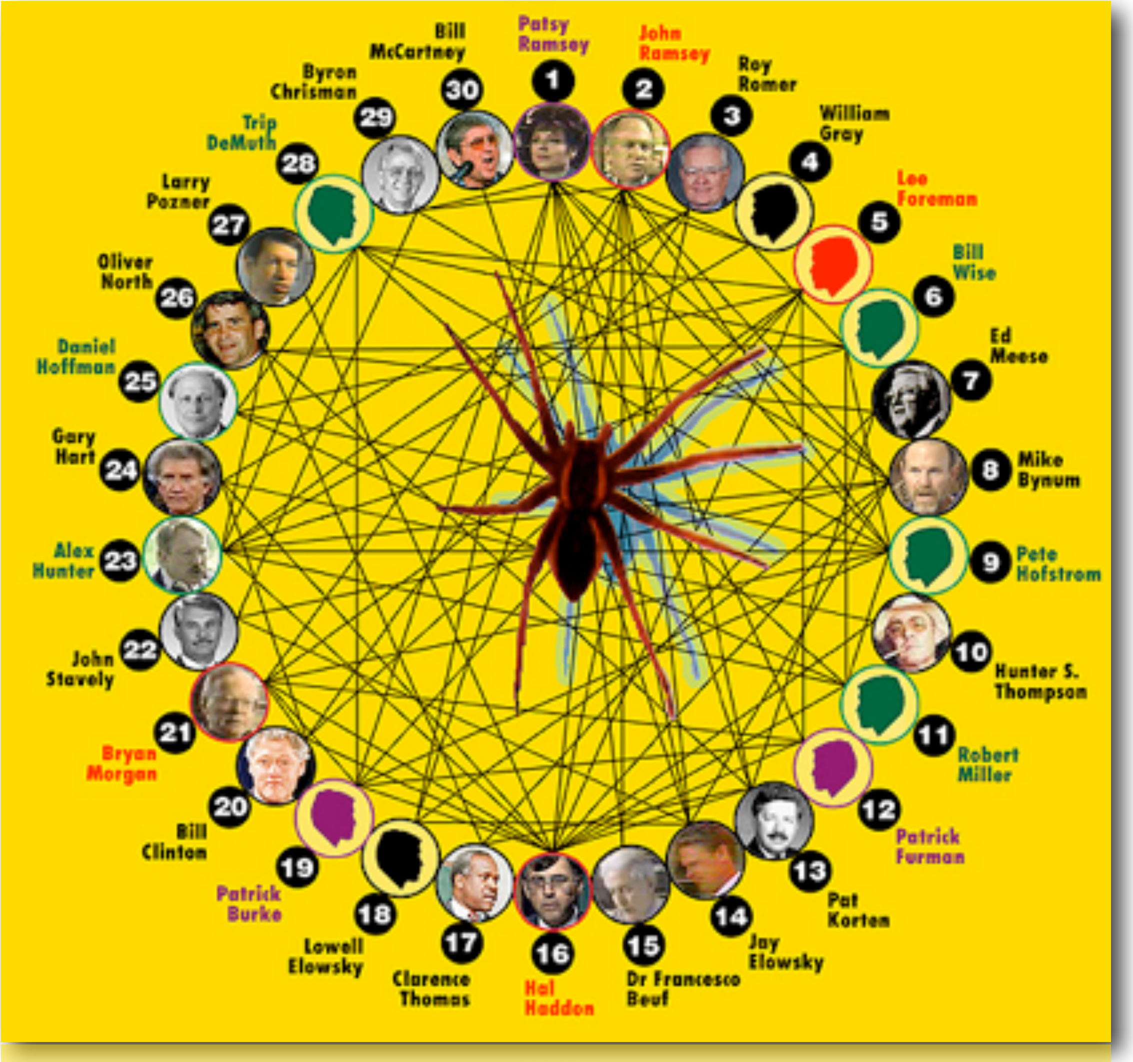
- Social Platforms
    - Social Network
    - Social Media
    - Social games
    - Social bookmarking
    - Social News and Social Knowledge Sharing
- Social Computing Introduction
- Techniques in Social Computing
- Summary

Web 2.0

- Web as a medium vs. **Web as a platform**
- Read-Only Web vs. **Read-and-Write Web**
- Static vs. **Dynamic**
- Restrictive vs. **Freedom & Empowerment**
- Technology-centric vs. **User-centric**
- Limited vs. **Rich User Experience**
- Individualistic vs. **Group/Collective Behavior**
- Consumer vs. **Producer**
- Transactional vs. **Relational**
- Top-down vs. **Bottom-up**
- People-to-Machine vs. **People-to-People**
- Search & browse vs. **Publish & Subscribe**
- Closed application vs. **Service-oriented Services**
- Functionality vs. **Utility**
- Data vs. **Value**

Social Networks

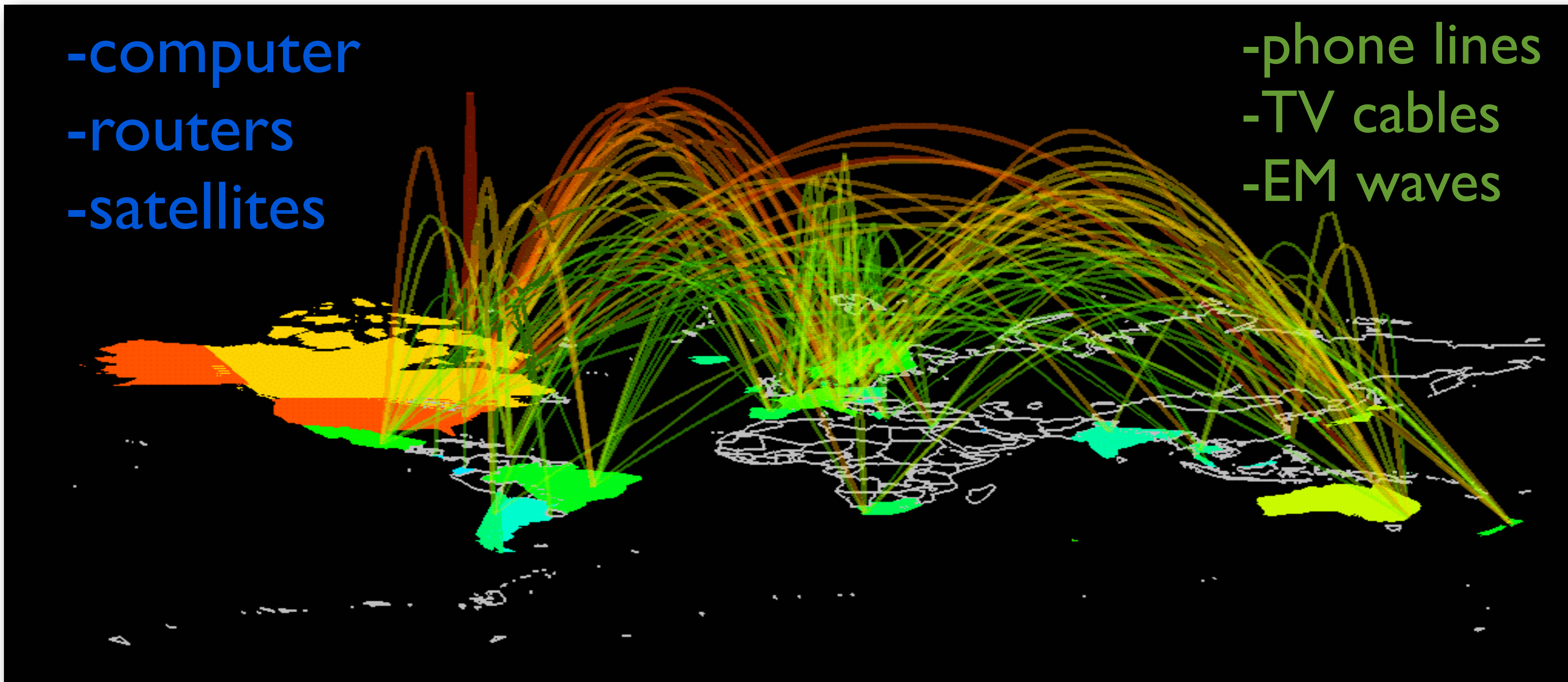
Society:
Nodes: individuals
Links: social relationship (family/work/friendship/etc.)



S. Milgram and John Guare: Six Degree of Separation.
Social networks: Many individuals with diverse social interactions between them.

Social Networks

- The Earth is developing an electronic nervous system, a network with diverse nodes and links.



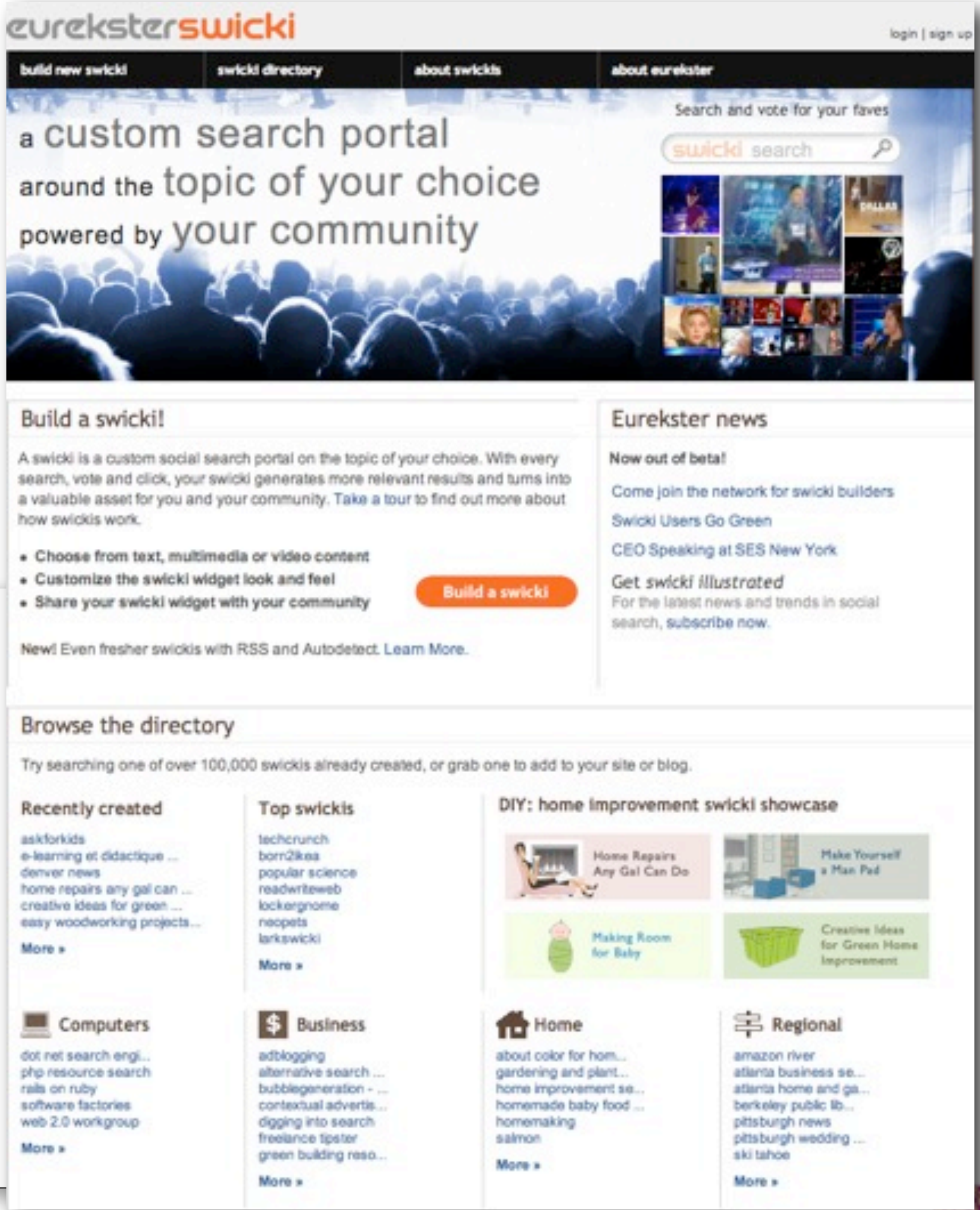
Communication networks: many non-identical components with diverse connections between them.

Social Networking Sites

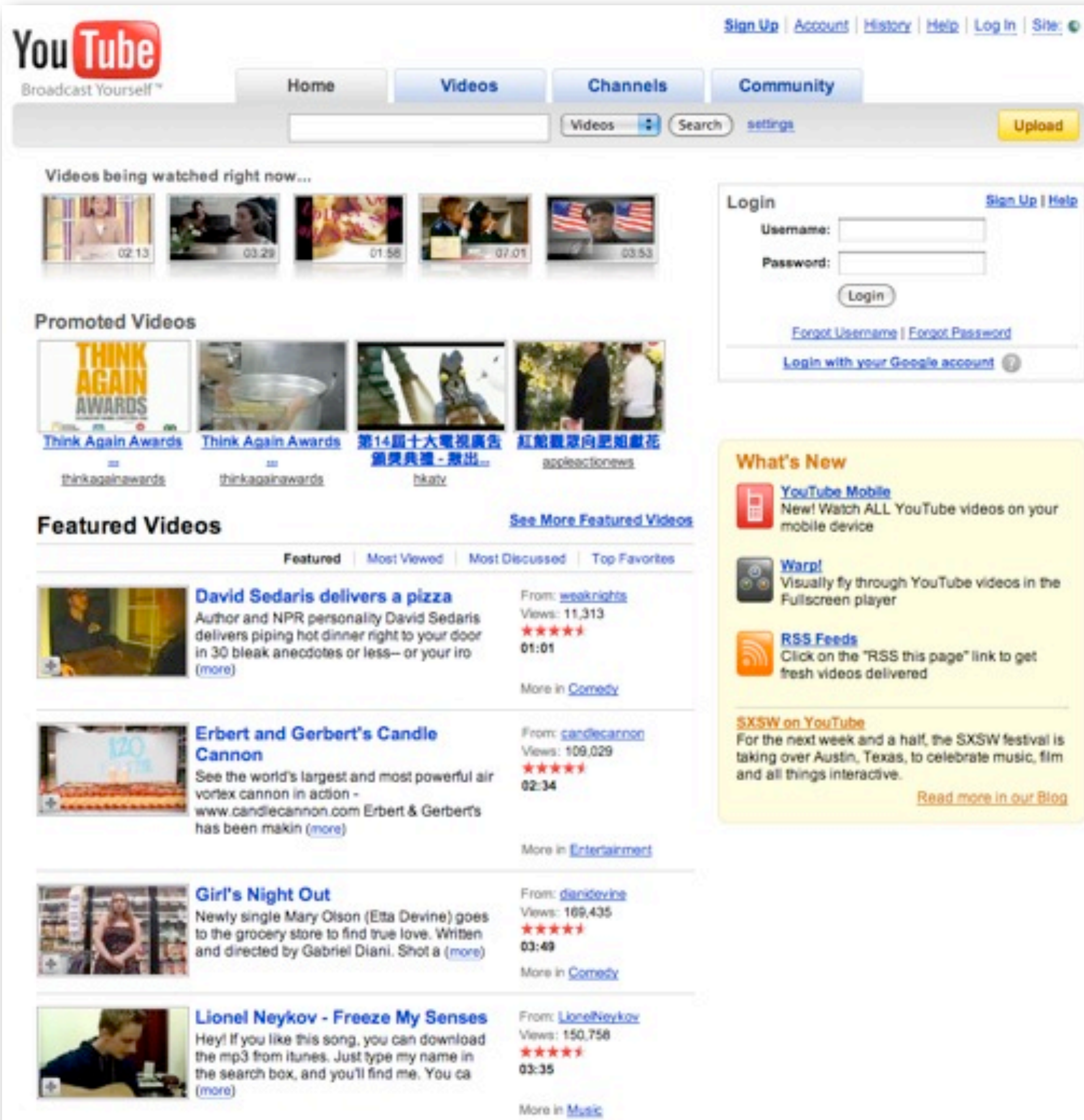
- Example of Social Networking Sites: FaceBook, MySpace, Blogger, QQ, etc.

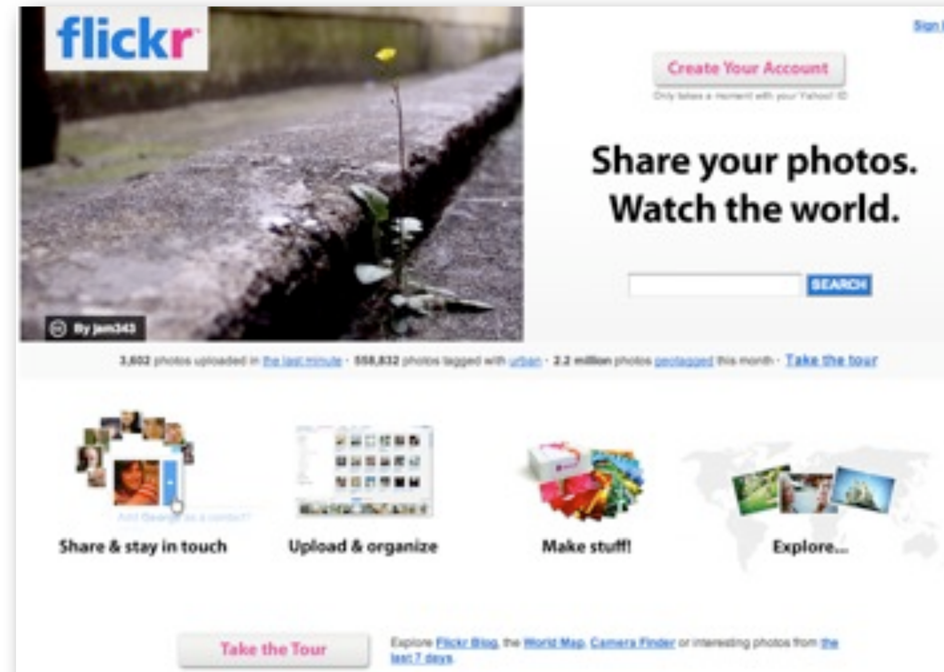
Social Search

- Social Search Engine
- Leveraging your social networks for searching

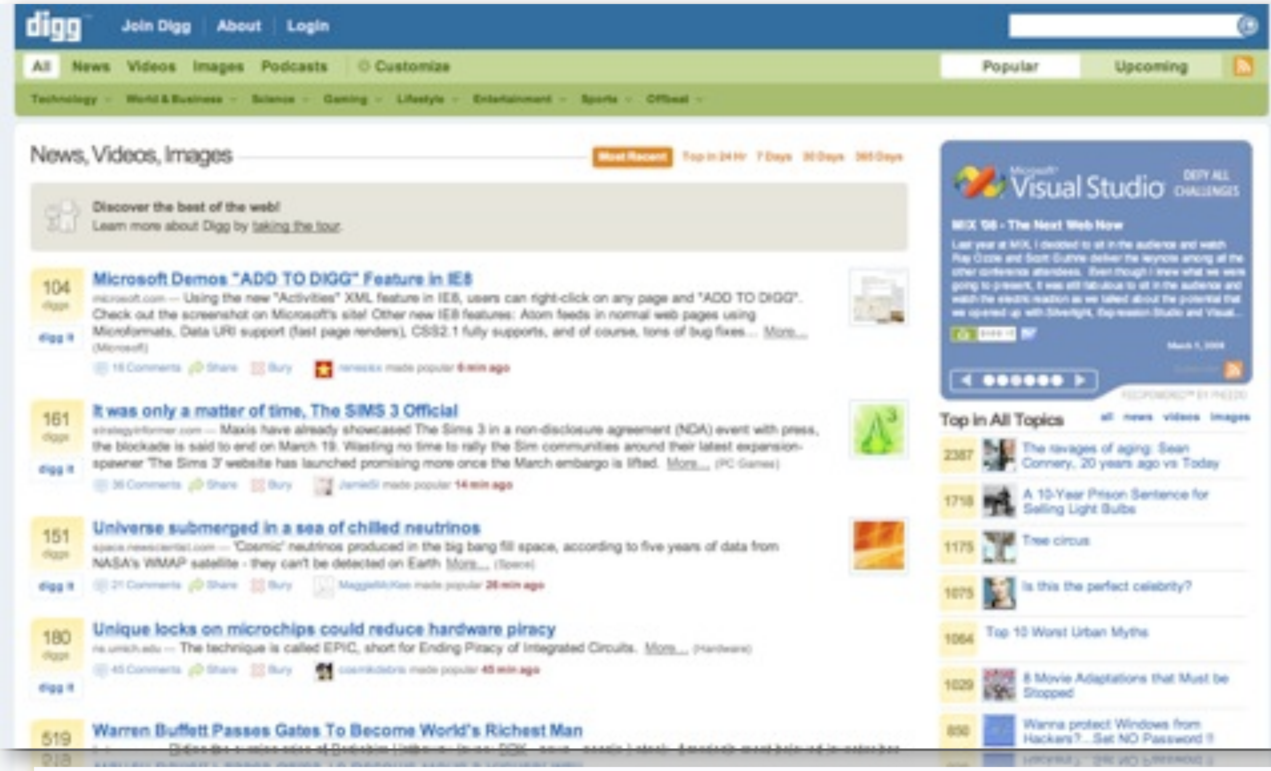


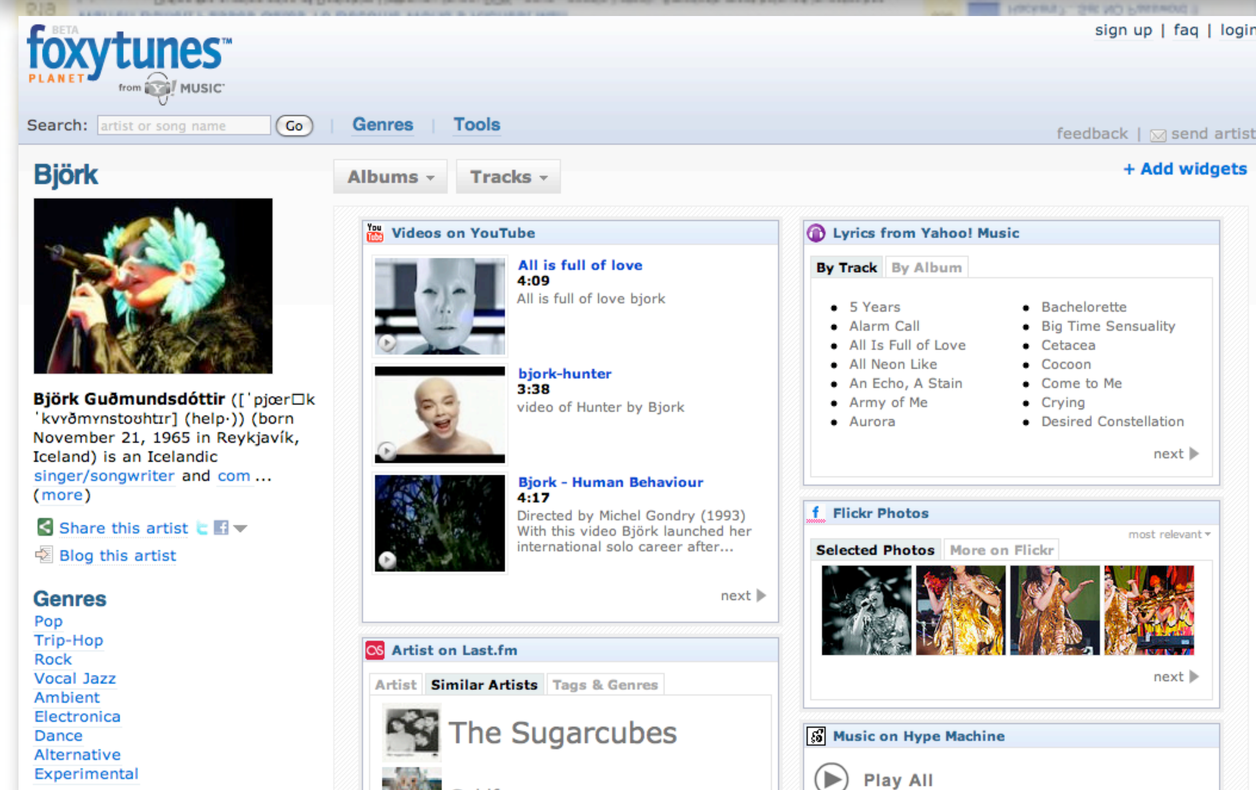
Social Media

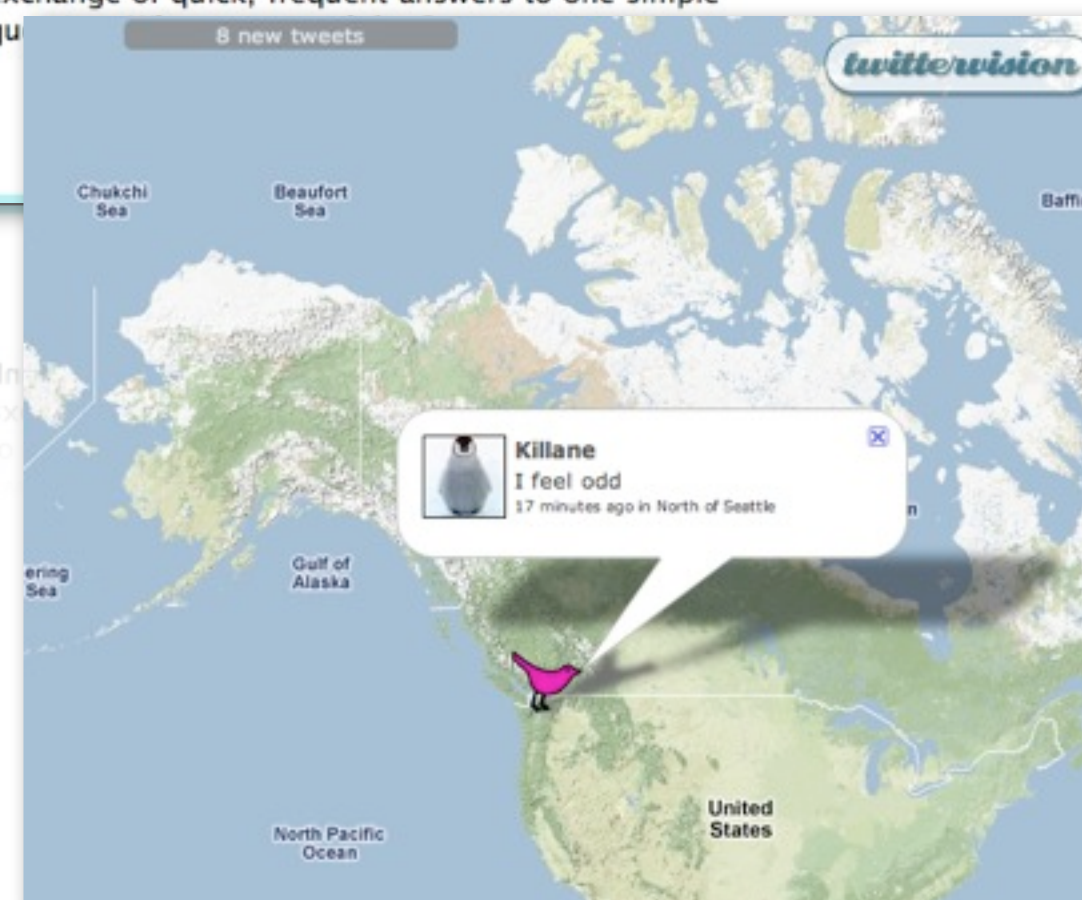




Social News/Mash Up

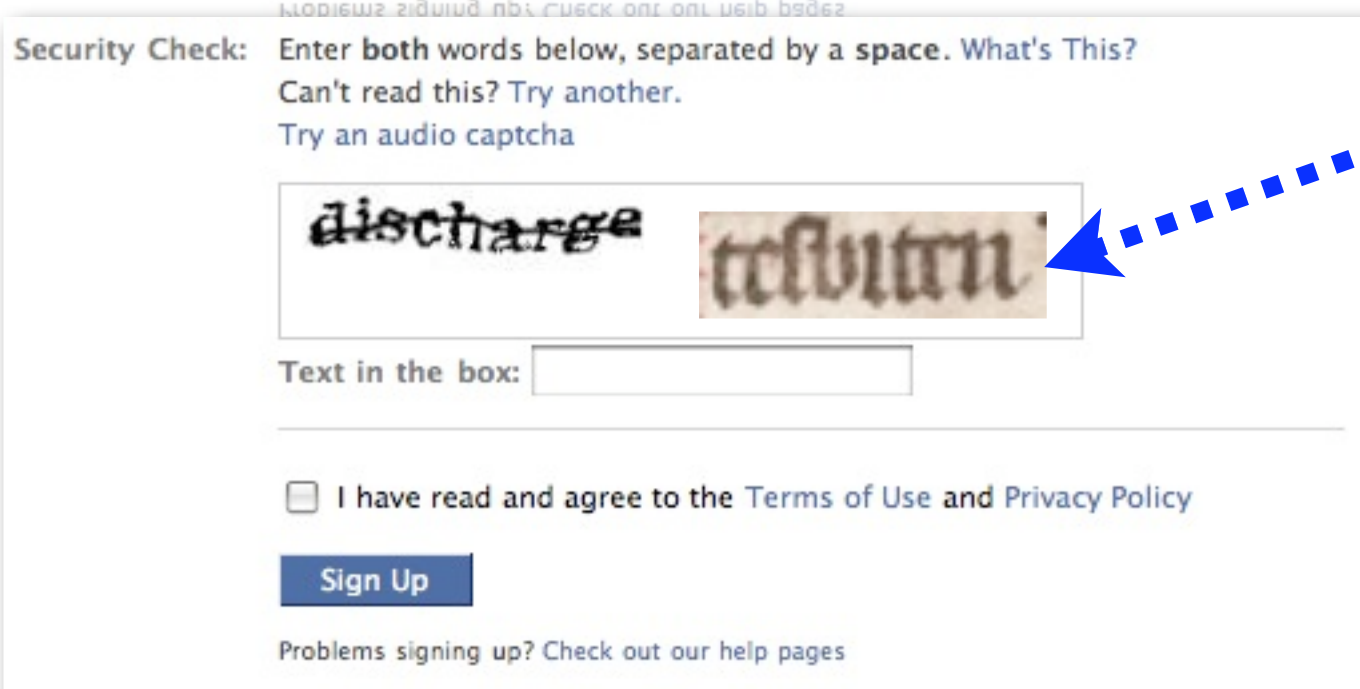


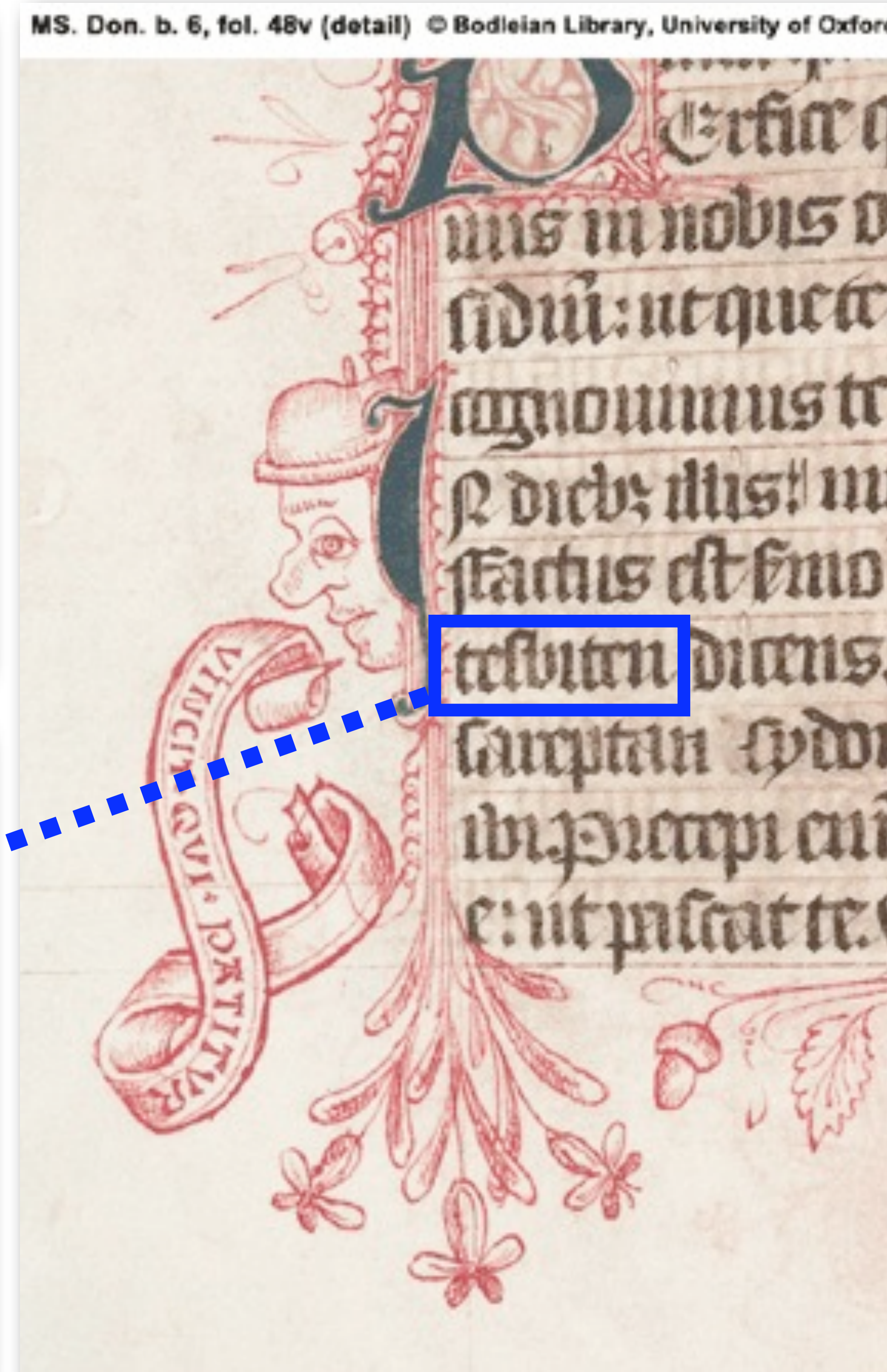




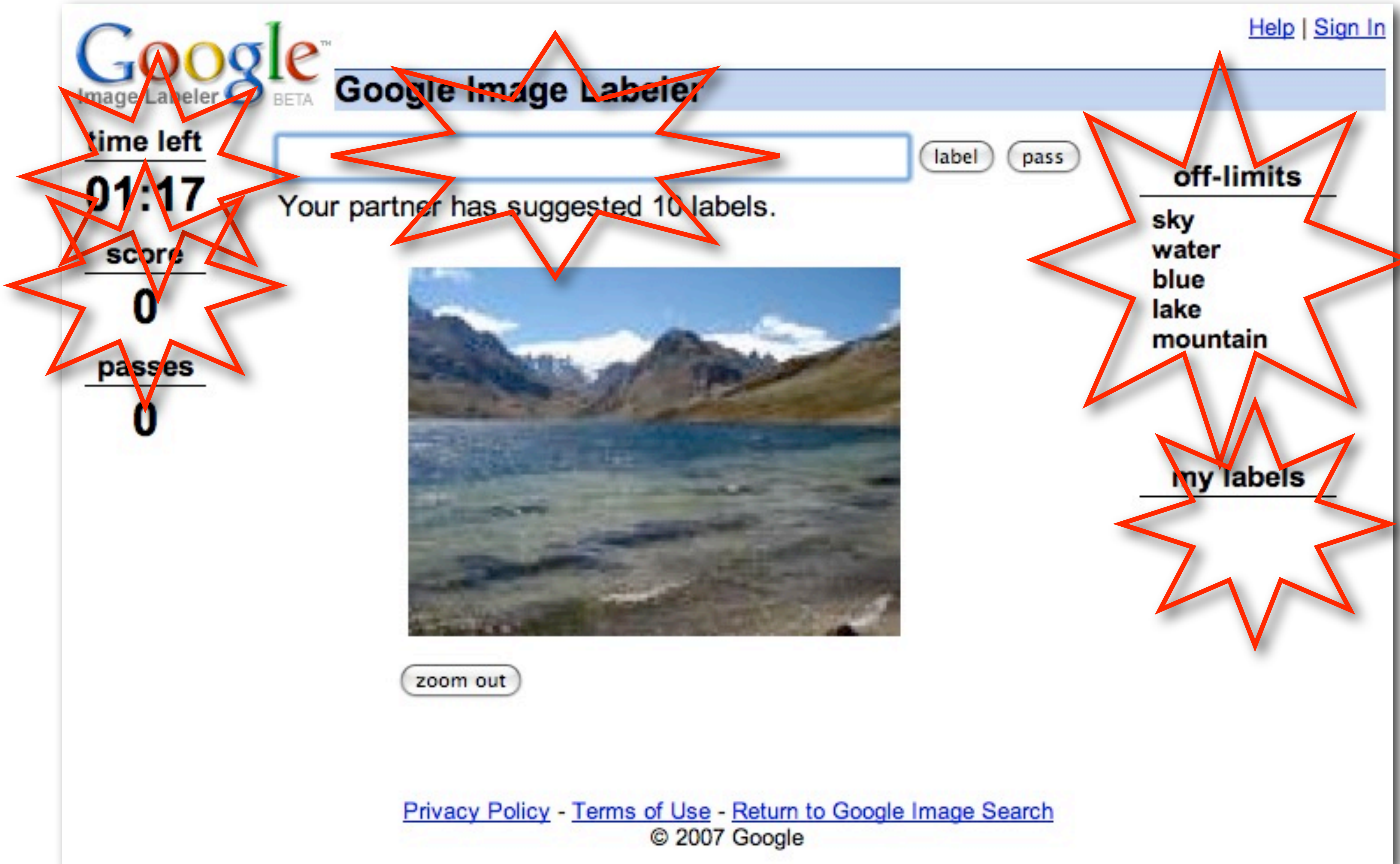
Social Knowledge Sharing

Social/Human Computation





Human Computation



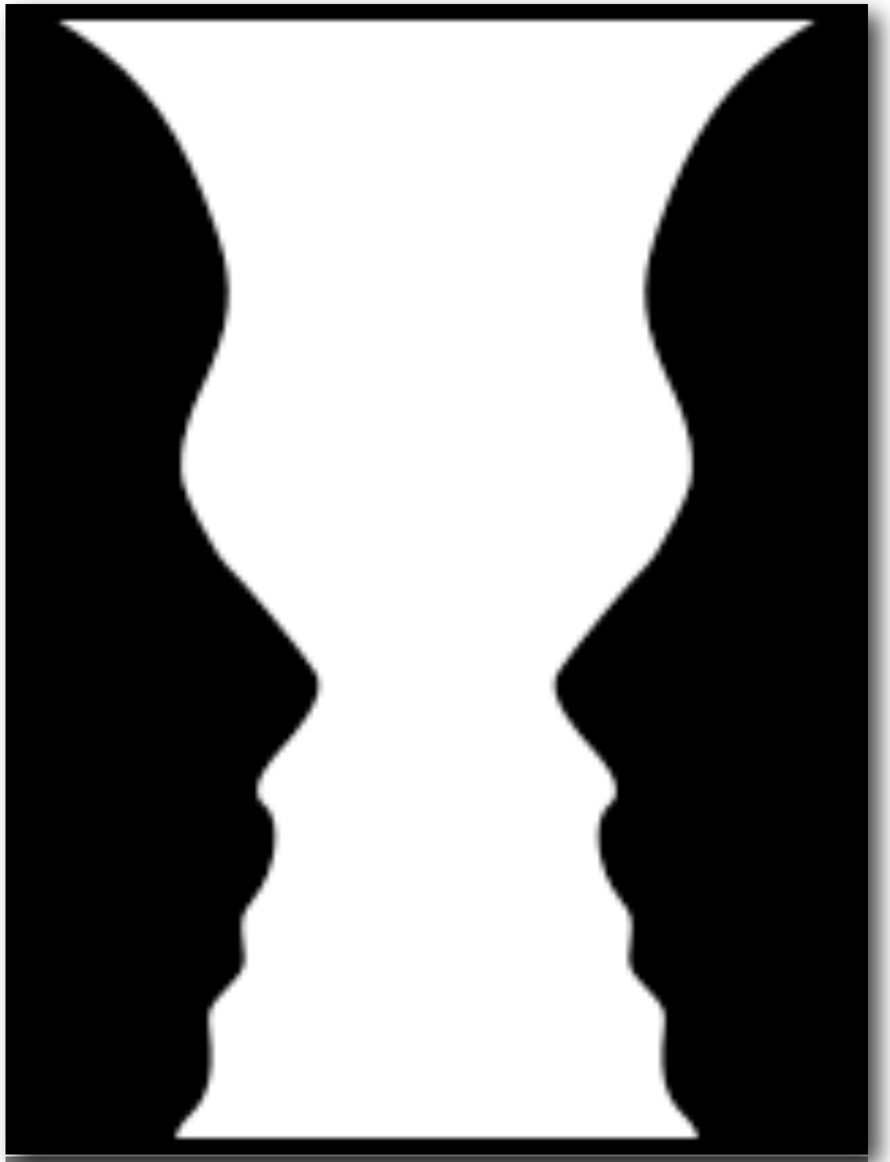
Web 2.0 Revolution

- Glocalization-think globally and act locally!
- Weblication-Web is the application!
- Three C's

Connectivity

Collaboration

Communities



Social Relations

presence

crew

identity

binary

teams

populations

social role

squad

cardinal

reputation

organizations

expertise

integer

trust

cohorts

markets

ownership

real

communities

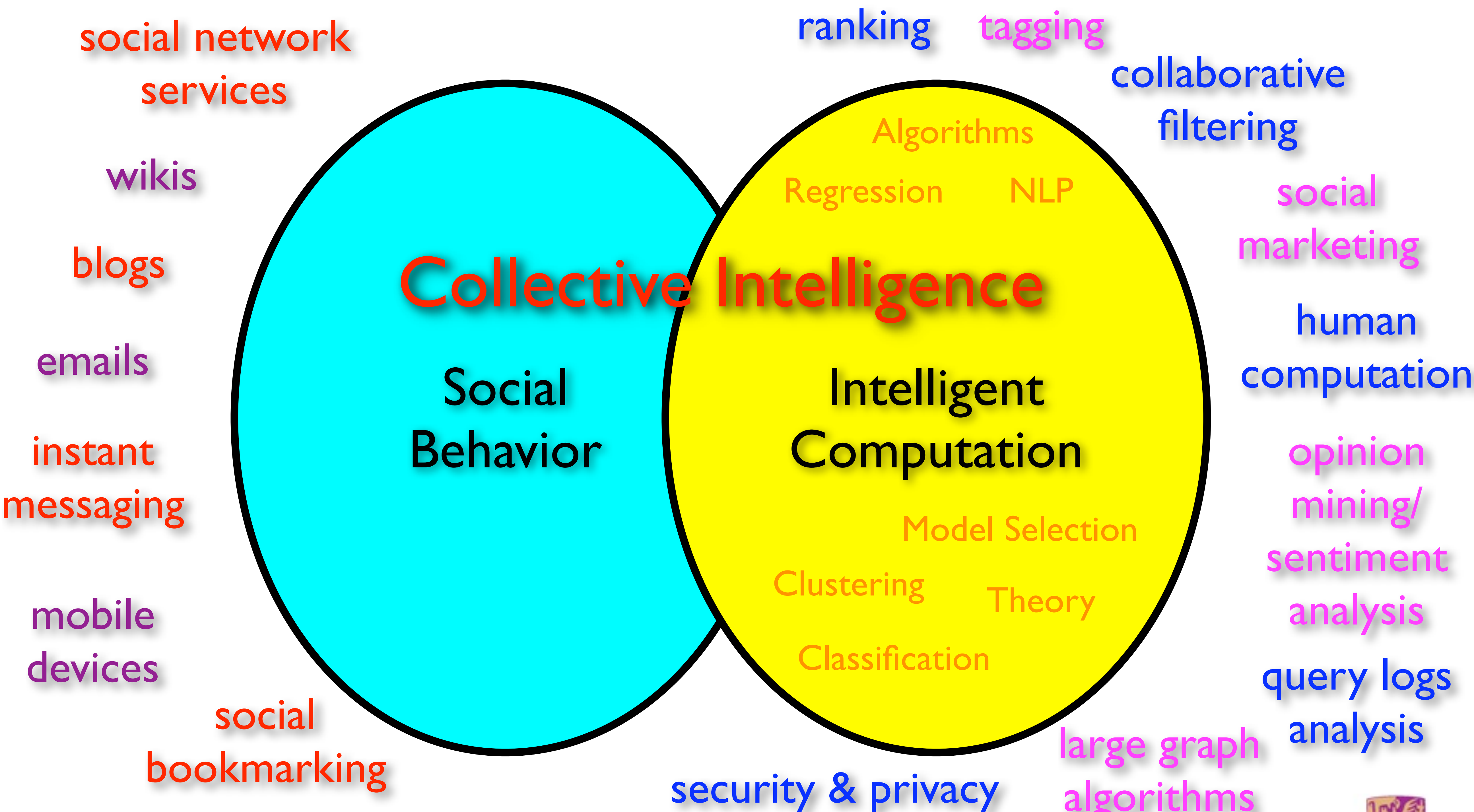
accountability

partners

knowledge

groups

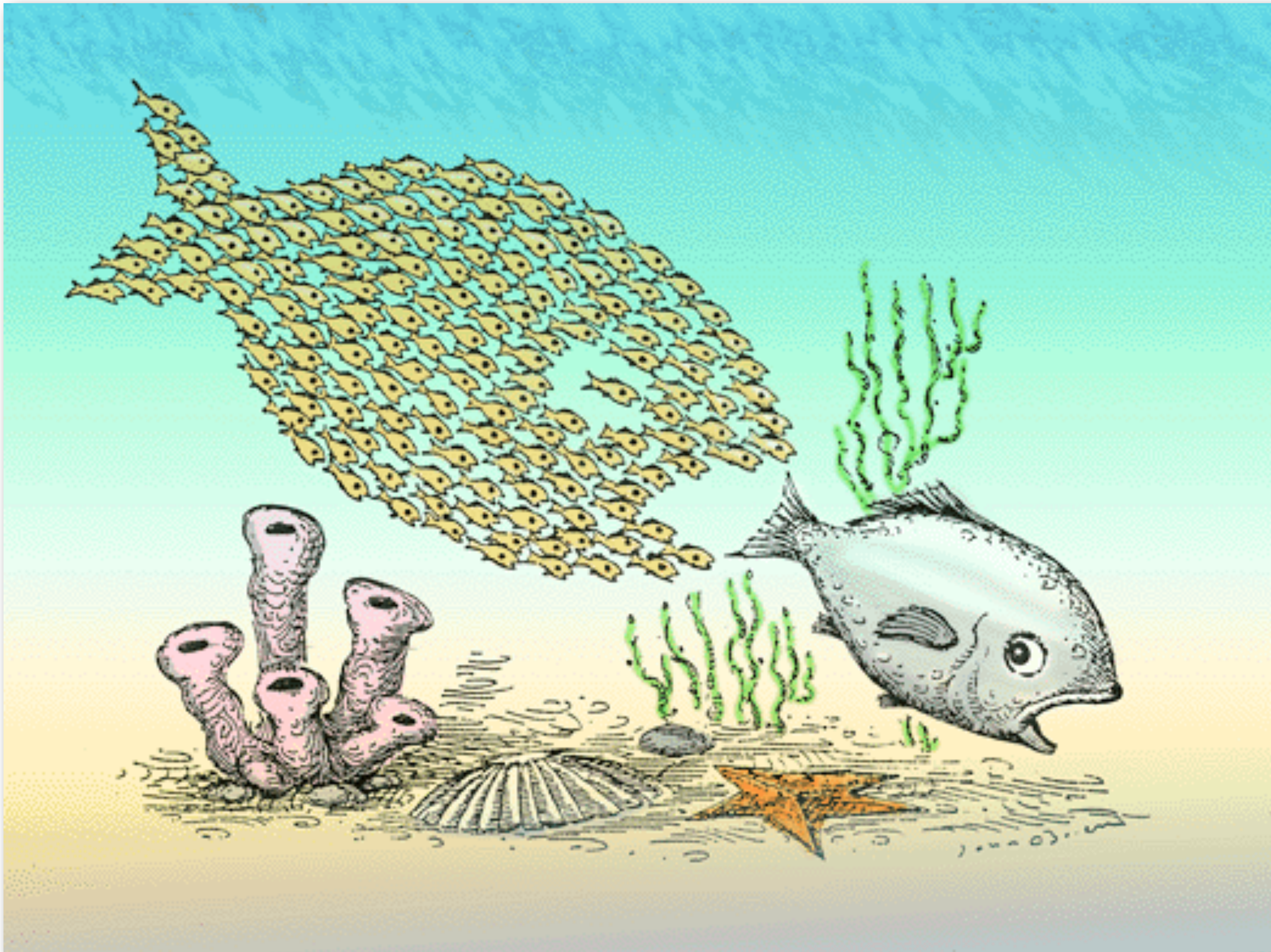
Social Computing



Definition of Social Computing

[wiki]

- Any Computer-mediated communication and interaction
- In the weaker sense: supporting any sort of social behavior
  - blogs, email, instant messaging, wiki, social network services, social bookmarking
- In the stronger sense: supporting "computations" that are carried out by a group of people
  - collaborative filtering, online auctions, prediction markets, reputation systems, tagging, verification games



Emerging Issues

- Theory and models
- Seach, mining, and ranking of existing information, e.g., spatial (relations) and temporal (time) domains
    - Dealing with partial and incomplete information, e.g., collaborative filtering, ranking, tagging, etc.
- Scalability and algorithmic issues
- Security and privacy issues
- Monetization of social interactions

Computational Perspective

- Classification, clustering, regression, etc.
- New insights on the data
  - Social relations are often hidden (latent)
  - Change data from (x, y) to $(x, c_1(x), c_2(x), \cdots, y)$
- $c(x)$ = context in *tags*, *relations*, *ratings*, etc.
- data type = *binary*, *integer*, *real*, *cardinal*, etc.

Social Computing Road Map

- Social Computing Introduction
- Social Platforms
- Techniques in Social Computing
  - Social Network Theory, Modeling and Analysis
  - Graph/Link Mining
  - Learning to Rank
  - Query Log Processing
  - Web Spam Detection
  - Collaborative Filtering
  - Opinion Mining
  - Privacy and Trust
- Summary

Social Network Theory

- Consider many kinds of networks:
  - social, technological, business, economic, content, ...
- These networks tend to share certain informal properties:
  - large scale; continual growth
  - distributed, organic growth: vertices "decide" who to link to
  - interaction restricted to links
  - mixture of local and long-distance connections
  - abstract notions of distance: geographical, content, social,…

Social Network Theory

- Do these networks share more quantitative universals?
- What would these "universals" be?
- How can we make them precise and measure them?
- How can we explain their universality?
- This is the domain of social network theory

Some Interesting Quantities

- Connected components
  - how many, and how large?
- Network diameter
  - maximum (worst-case) or average?
  - exclude infinite distances? (disconnected components)
  - the small-world phenomenon
- Clustering
  - to what extent that links tend to cluster "locally"?
  - what is the balance between local and long-distance connections?
  - what roles do the two types of links play?
- Degree distribution
  - what is the typical degree in the network?
  - what is the overall distribution?

Graph/Link Mining

- Heterogeneous, multi-relational data represented as a graph or network
  - Nodes are objects
    - Objects have attributes
    - Objects may have labels or classes
  - Edges are links
    - Links may have attributes
    - Links may be directed
- Links represent relationships and interactions between objects -- rich content for mining

What Is New For Mining

- Traditional machine learning and data mining approaches assume:
  - A random sample of homogeneous objects from single relation
- Real world data sets:
  - Multi-relational, heterogeneous and semi-structured
- Link Mining
  - Newly emerging research area at the intersection of research in social network and link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming

What is a Link in Link Mining

- Link: relationship among data
- Homogeneous networks
  - Single object type and single link type
  - Single model social networks (e.g., friends)
  - WWW: a collection of linked Web pages
- Heterogeneous networks
  - Multiple object and link types
  - Medical network: patients, doctors, disease, contacts, treatments
  - Bibliographic network: publications, authors, venues

Learning to Rank

- Booming Search Industry

Learning to Rank

- Given query q and set of docs $d_1, \ldots d_n$
  - Find documents relevant to q
  - Typically expressed as a ranking on $d_1, \ldots d_n$
  - Are social signals important?

Widely-used Judgement

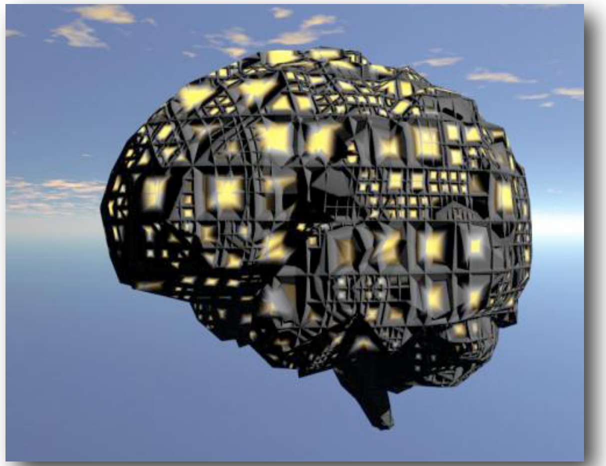
- Pointwise
  - Binary judgment (Relevant vs. Irrelevant)
  - Multi-valued discrete (Perfect > Excellent > Good > Fair > Bad)
- Pairwise
  - Pairwise preference
    - Document A is more relevant than document B w.r.t. query q
- Listwise
  - Partial or total orders
  - Could be mined from click-through logs

Conventional Ranking Models

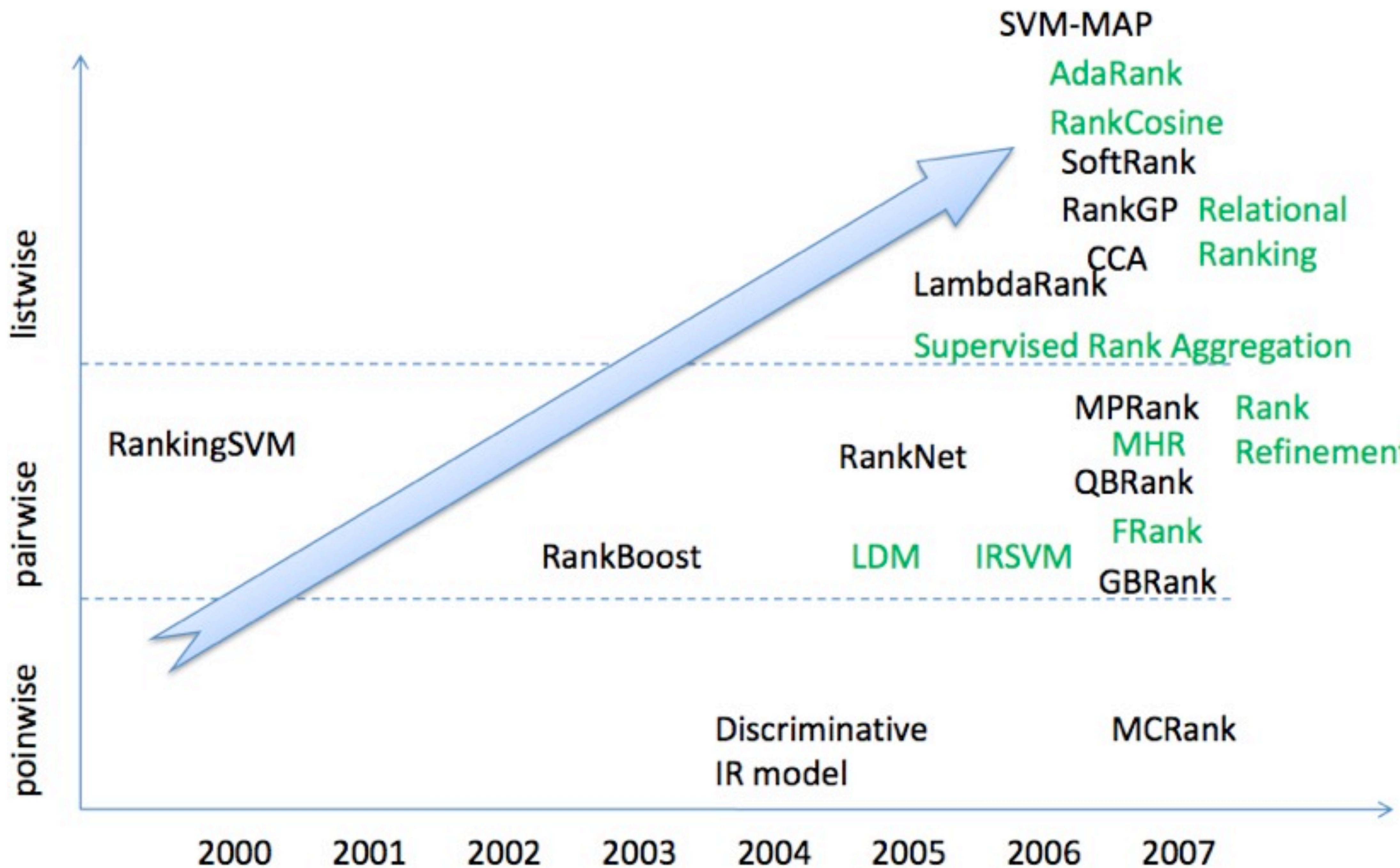
- Content relevance
  - Boolean model, extended Boolean model, etc.
  - Vector space model, latent semantic indexing (LSI), etc.
  - BM25 model, statistical language model, etc.
  - Span based model, distance aggregation model, etc.
- Page Quality
  - Link analysis: HITS, PageRank, TrustRank, etc.
  - Log mining: DirectHITS, BrowseRank, etc

Machine Learning Can Help

- Machine learning is an effective tool
  - To automatically tune parameters
  - To combine multiple evidences
  - To avoid over-fitting (by means of regularization, etc.)
- Learning to Rank
  - Use machine learning technologies to train the ranking model
  - A hot research topic these years



Learning To Rank Techniques



http://research.microsoft.com/en-us/people/tyliu/default.aspx

Resources

- LETOR benchmark: a package of benchmark data sets for learning to rank, released by Microsoft Research Asia.
- Current LETOR baselines
    - Ranking SVM
    - RankBoost
    - AdaRank
    - Multiple hyperline ranker
    - FRank
    - ListNet

Real life Example for Collaborative Filtering

- User's perspective
  - Lots of online products, books, movies, etc
  - Reduce my choices

- Manager's perspective

"if I have 3 million customers on the web, I should have 3 million stores on the web."

CEO of Amazon.com

More Examples

- Movielens: movies
- Moviecritic: movies again
- My launch: music
- Gustos starrater: web pages
- Jester: Jokes
- TV Recommender: TV shows
- Suggest 1.0: different products
- And much more…

How it Works?

- Each user has a profile
- Users rate items
    - Explicitly: score from 1..5
    - Implicitly: web usage mining
        - Time spent in viewing the item
        - Navigation path, etc…
- System does the rest, How?
    - Look at users collective behavior
    - Look at the active user history

Techniques

- User-User Methods
  - Identify like-minded users
  - Memory-based: K-NN
  - Model-based: Clustering
- Item-Item Method
  - Identify buying patterns
  - Correlation Analysis
  - Linear Regression
  - Association Rule Mining
  - Belief Network

Query Log Processing

- Search engines and social network sites collect a voluminous amount of query log or click-through data from their users

- These logs can be used to improve retrieval results

What is Clickthrough Data

- Query logs recorded by search engines

$$\langle u, q, l, r, t \rangle$$

Table 1: Samples of search engine clickthrough data

ID	Query	URL	Rank	Time
358	facebook	http://www.facebook.com	1	2008-01-01 07:17:12
358	facebook	http://en.wikipedia.org/wiki/Facebook	3	2008-01-01 07:19:18
3968	apple iphone	http://www.apple.com/iphone/	1	2008-01-01 07:20:36
...	...	...	...	...

  - Server logs
  - Error logs
  - Cookie logs
  - Query data
  - Web meta data

- User's relevance feedback to indicate desired/preferred/target results

Techniques in Query Log Processing

- Association Rules
  - a priori, is-a hierarchical, …
- Discovery of sequential patterns
  - modified a priori -- order is important
- Classification and clustering
  - k-means, birch, …
  - SVM

Opinion Mining

- Two main types of textual information on the Web
  - Facts vs. Opinions
  - Current search engines search for facts (assume they are true)
    - Facts can be expressed with topic keywords.
  - Search engines do not search for opinions
    - Opinions are hard to express with a few keywords
      - E.g. *How do people think of Motorola Cell phones?*
    - Current search ranking strategy is not appropriate for opinion retrieval/search.

Opinion Mining

- What is an opinion?
  - A person's ideas and thoughts towards something.
  - It is an assessment, judgment or evaluation of something.
  - An opinion is not a fact, because opinion has not been proven or verified.
  - All information on the web is better described as opinion rather than fact.
- Opinion Mining
  - It is text mining and computational linguistics which tries to detect the opinions expressed in the nature language texts.
    - Opinion Extraction: specified method of information extraction, delivering inputs for opinion mining
    - Sentiment analysis and sentiment classification

Applications

- Businesses and organizations: product, service benchmarking, market intelligence
  - Business spends a huge amount of money to find consumer sentiments and opinions
- Individuals: interested in other's opinions when
  - Purchasing a product or using a service
  - Finding opinions on political topics
- Ads placements: Placing ads in the user-generated content
  - Place an ad when one praises a product
  - Place an ad from a competitor if one criticizes a product
- Opinion retrieval/search: providing general search for opinions

Research Topics

- Development of linguistic resources for opinion mining
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document focuses on a single object
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
- At the feature level
  - Identify and extract commented features
- Comparative opinion mining
  - Identify comparative sentences

Summary

- Social Platforms
  - Social Network
  - Social Media
  - Social games
  - Social bookmarking
  - Social News and Social Knowledge Sharing
- Techniques in Social Computing
  - Social Network Theory, Modeling and Analysis
  - Graph/Link Mining
  - Learning to Rank
  - Query Log Processing
  - Web Spam Detection
  - Collaborative Filtering
  - Opinion Mining

A brief overview of the emerging field of social computing.

References

- https://agora.cs.illinois.edu/display/cs512/home.
- J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In ICML, 2004.
- J. S. Breese, D. Heckerman, and C. M. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In UAI, pages 43–52, 1998.
- M. Deshpande and G. Karypis. Item-based top- recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143–177, 2004.
- J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In SIGIR, pages 230–237. ACM, 1999.
- J. L. Herlocker, J. A. Konstan, and J. Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Inf. Retr., 5(4):287–310, 2002.
- R. Jaschke, M. Grahl, A. Hotho, B. Krause, C. Schmitz, and G. Stumme. Organizing publications and bookmarks in bibsonomy. In CKC, 2007.
- L. von Ahn. Games with a purpose. IEEE Computer, 39(6):92–94, 2006.

References

- C. S. Andreas Hotho, Robert Jaschke and G. Stumme1. Bibsonomy: A social bookmark and publication sharing system. In CS-TIW'06. Aalborg University Press, 2006.

- G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. A. Golder, K. Fox, M. Davis, C. Marlow, and M. Naaman. Why do tagging systems work? In CHI Extended Abstracts, pages 36–39, 2006.

- P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In WSDM, pages 195–206, 2008.

- R. Jaschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In A. Hinneburg, editor, LWA, pages13–20. 2007.

- B. Krause, A. Hotho, and G. Stumme. A comparison of social bookmarking with traditional search. In ECIR, pages 101–113, 2008.

- L. Specia and E. Motta. Integrating folksonomies with the semantic web. In ESWC, pages 624–639, 2007.

References

- H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. IEEE Trans. Knowl. Data Eng., 15(4):829–839, 2003.
- W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon. Cross-lingual query suggestion using query logs of different languages. In SIGIR, pages 463–470, 2007.
- R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors, WWW, pages 387–396. ACM, 2006.
- H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In CIKM, pages 709–718, 2008.
- Q. Mei, D. Zhou, and K. W. Church. Query suggestion using hitting time. In CIKM, pages 469–478, 2008.
- J. Xu and W. B. Croft. Query expansion using local and global document analysis. In SIGIR, pages 4–11, 1996.

References

- S. Cucerzan and R. W. White. Query suggestion based on user landing pages. In SIGIR, pages 875–876, 2007.
- R. B. D. M. C. Edith L. M. Law, Luis von Ahn. Tagatune: A game for music and sound annotation. ISMIP, 2007.
- L. von Ahn and L. Dabbish. Labeling images with a computer game. In CHI, pages 319–326, 2004.
- L. von Ahn and L. Dabbish. Designing games with a purpose. Commun. ACM, 51(8): 58–67, 2008.
- L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In CHI, pages 79–82, 2006.
- L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In CHI, pages 75–78, 2006.
- L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In CHI '06, pages 55–64, New York, NY, USA, 2006. ACM.

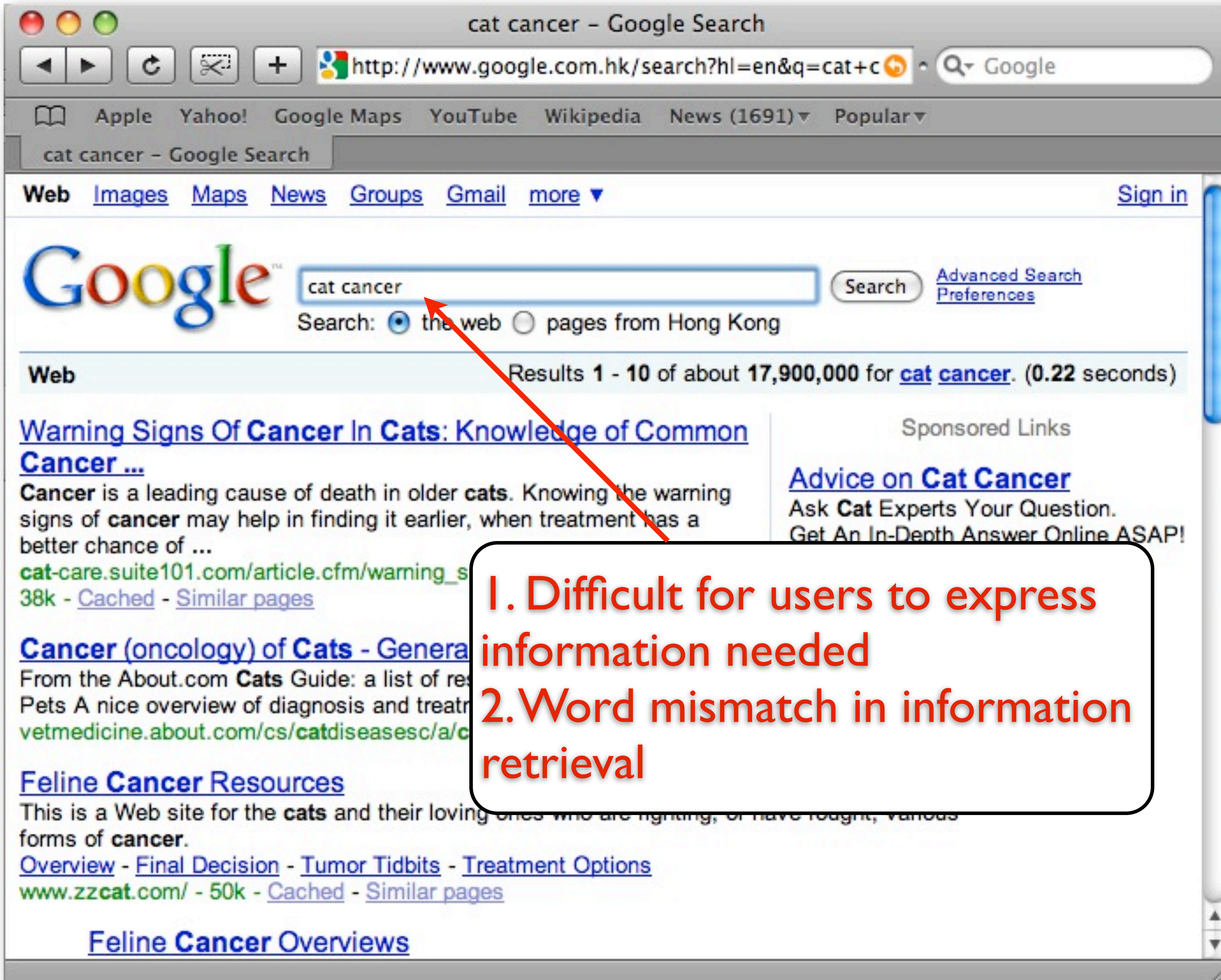
References

- Ashwin Machanavajjhala , Daniel Kifer , Johannes Gehrke , Muthuramakrishnan Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, TKDD, 2007
- Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, ICDE, 2007.
- Xiao, X., Tao, Y, Dynamic Anonymization: Accurate Statistical Analysis with Privacy Preservation, SIGMOD, 2008.
- Michael Hay, Gerome Miklau, David Jensen, Don Towsley and Philipp Weis, Resisting Structural Re-identification in Anonymized Social Networks, PVLDB, 2008
- Lars Backstrom, Cynthia Dwork and Jon Kleinberg, Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, WWW, 2007
- Kun liu and Evimaria Terzi, Towards Identity Anonymization on Graphs. SIGMOD, 2008
- Bin Zhou and Jian Pei, Preserving Privacy in Social Networks Against Neighborhood Attacks, ICDE, 2008

Query Suggestion

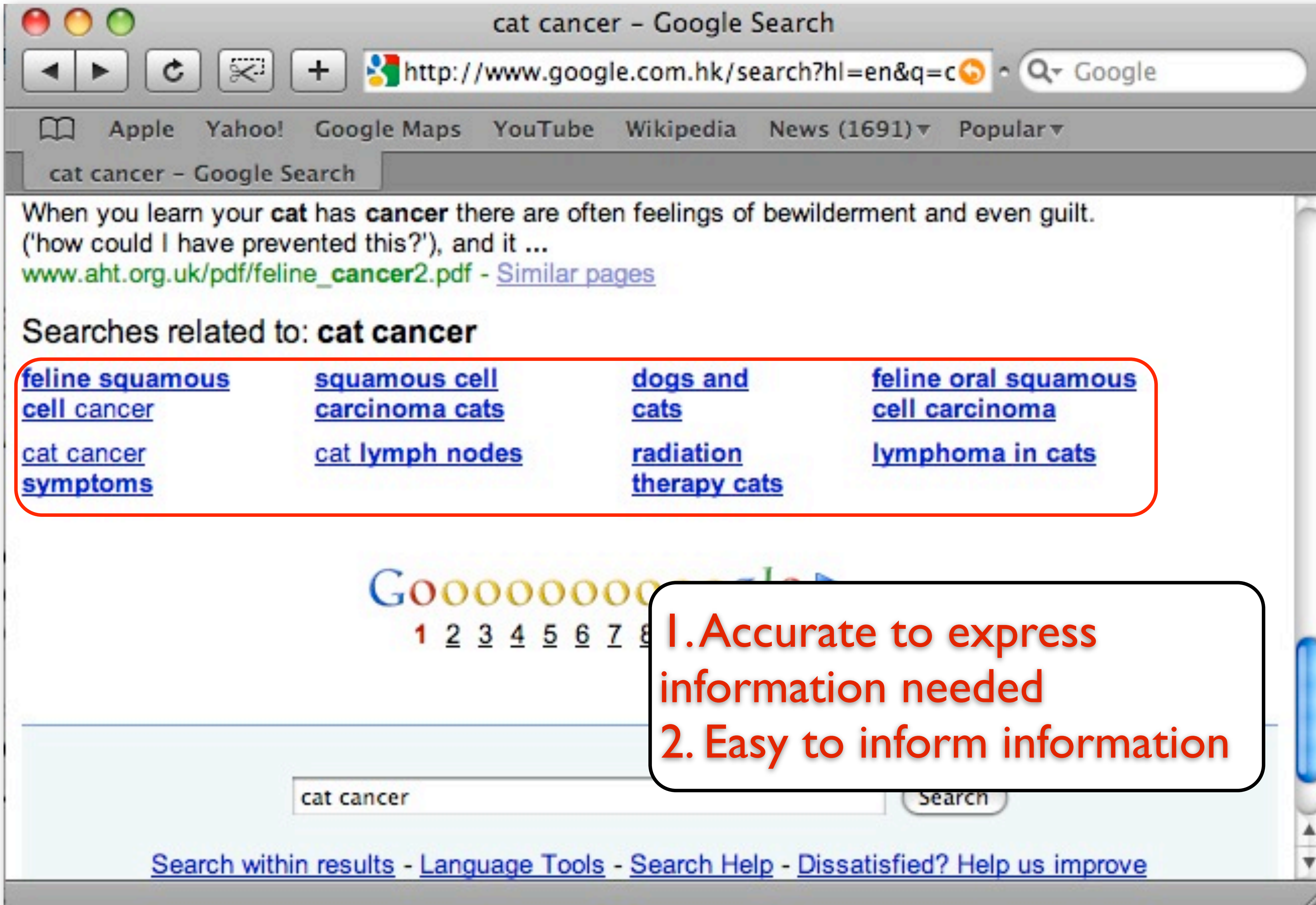
Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

Motivation



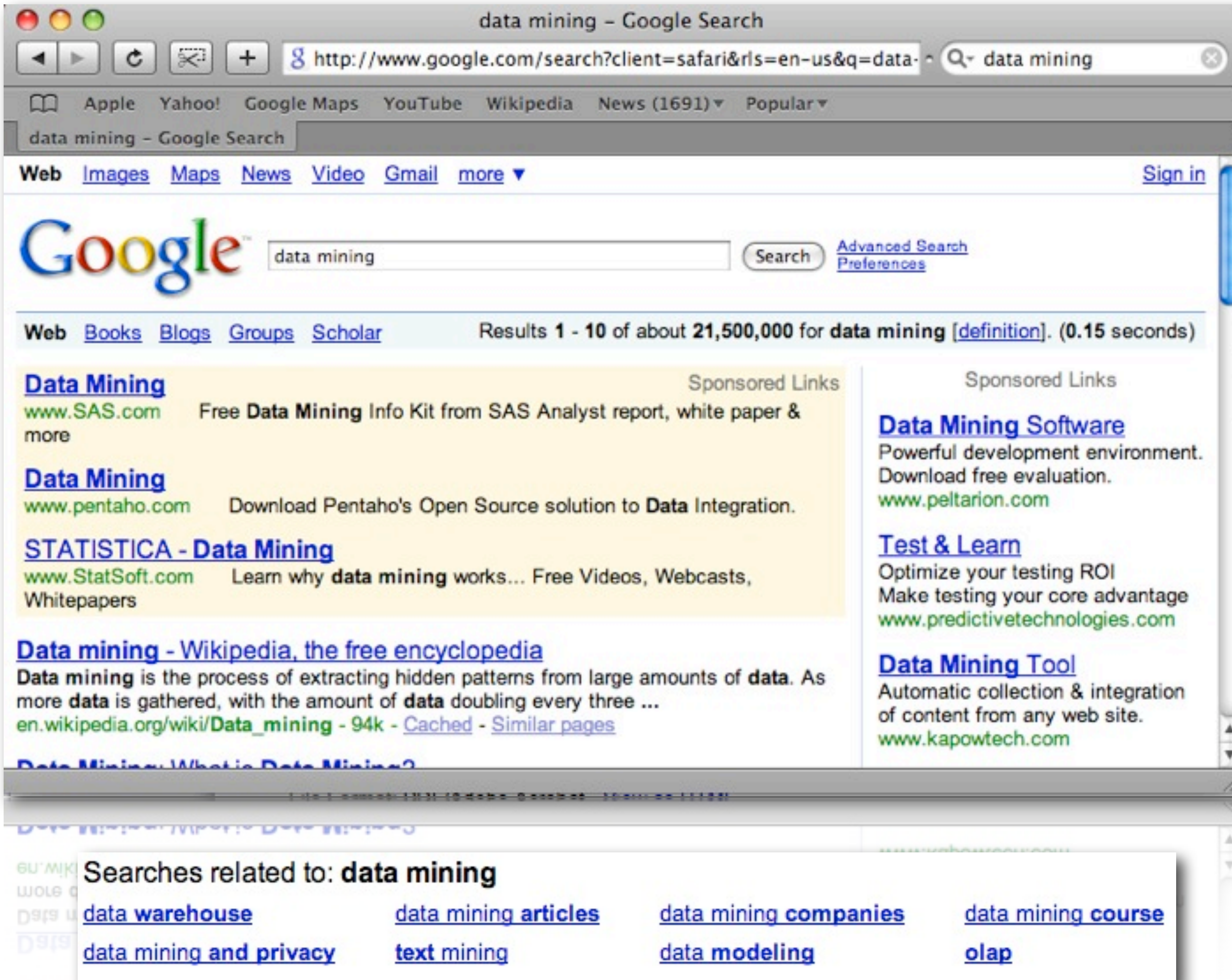
1. Difficult for users to express information needed
2. Word mismatch in information retrieval

Motivation



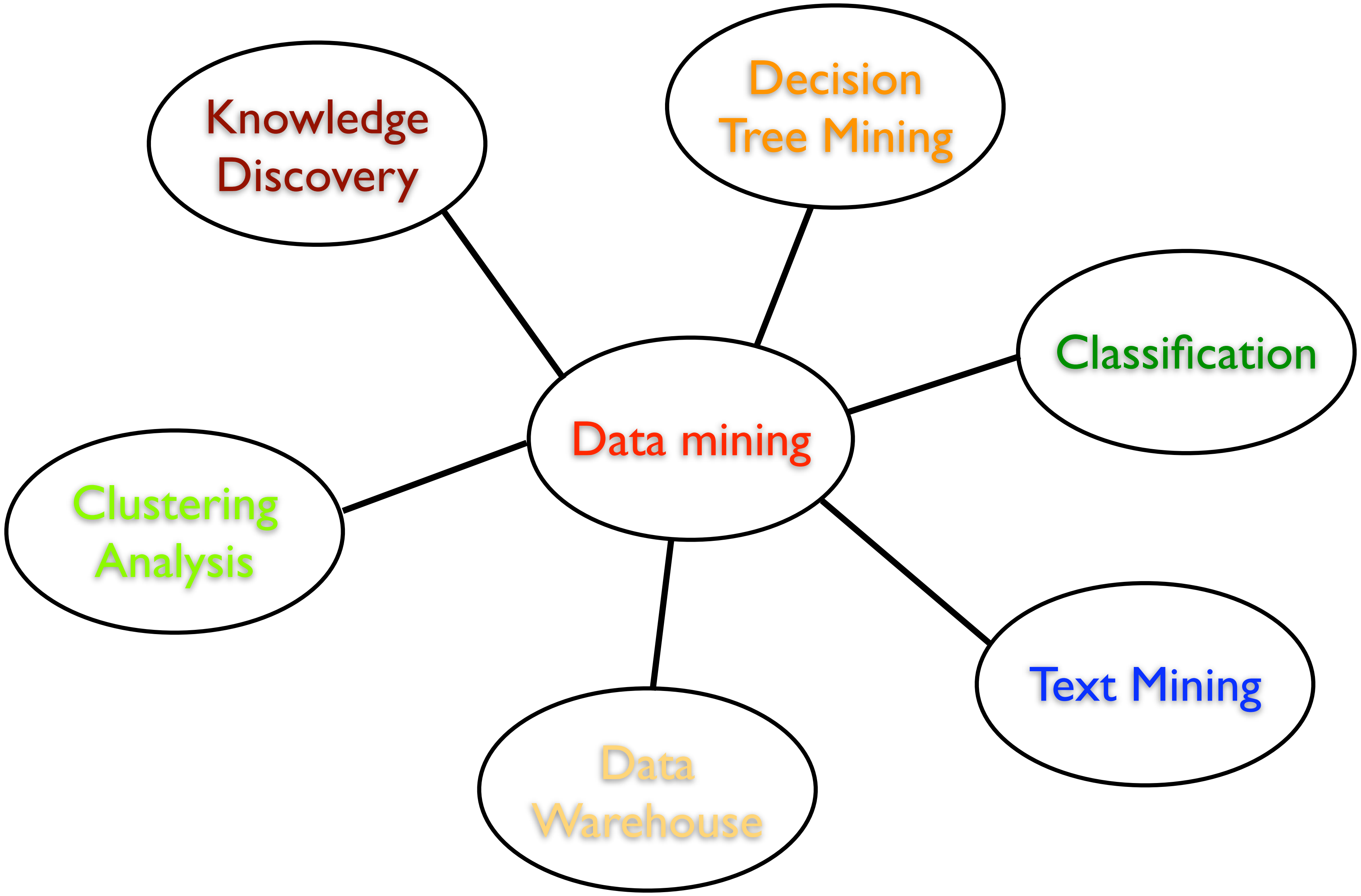
1. Accurate to express information needed
2. Easy to inform information

Motivation


data mining – Google Search
http://www.google.com/search?client=safari&rls=en-us&q=data-
data mining
Apple Yahoo! Google Maps YouTube Wikipedia News (1691) Popular
data mining – Google Search
Web Images Maps News Video Gmail more
Sign in
Google
data mining
Search
Advanced Search
Preferences
Web Books Blogs Groups Scholar
Results 1 - 10 of about 21,500,000 for data mining [definition]. (0.15 seconds)
Sponsored Links
Data Mining
www.SAS.com Free Data Mining Info Kit from SAS Analyst report, white paper & more
Data Mining
www.pentaho.com Download Pentaho's Open Source solution to Data Integration.
STATISTICA - Data Mining
www.StatSoft.com Learn why data mining works... Free Videos, Webcasts, Whitepapers
Data mining - Wikipedia, the free encyclopedia
Data mining is the process of extracting hidden patterns from large amounts of data. As more data is gathered, with the amount of data doubling every three ...
en.wikipedia.org/wiki/Data_mining - 94k - Cached - Similar pages
Sponsored Links
Data Mining Software
Powerful development environment. Download free evaluation.
www.peltarion.com
Test & Learn
Optimize your testing ROI
Make testing your core advantage
www.predictivetechnologies.com
Data Mining Tool
Automatic collection & integration of content from any web site.
www.kapowtech.com
Searches related to: data mining
data warehouse
data mining articles
data mining companies
data mining course
data mining and privacy
text mining
data modeling
olap

Challenges

- Word mismatch: people often use different words to describe concepts in their queries than authors use to describe the same concepts in their documents.



Challenges

- Queries contain ambiguous and new terms
  - apple: "apple computer" or "apple pie"?
  - NDCG:?
- Users tend to submit short queries consisting of only one or two words
  - almost 20% one-word queries
  - almost 30% two-word queries

- Users may have little or even no knowledge about the topic they are searching for!

Classes of Suggestion Relevance

[Jones, 2006]

- Precise rewriting
  - The rewritten form of query matches user's intent
- Approximate rewriting
  - The rewritten form has a direct close relationship to the topic described by the initial query
- Possible rewriting
  - The rewritten form either has some categorical relationship to the initial query or describes a complementary product
- Clear mismatch
  - The rewritten form has no clear relationship to user's intent

Example Queries and Query-suggestion

Class	Score	Examples		
Precise rewriting	1	automotive insurance	↦	automobile insurance
		corvette car	↦	chevrolet corvette
		apple music player	↦	apple ipod
		apple music player	↦	ipod
		cat cancer	↦	feline cancer
		help with math homework	↦	math homework help
Approximate rewriting	2	apple music player	↦	ipod shuffle
		personal computer	↦	compaq computer
		hybrid car	↦	toyota prius
		aeron chair	↦	office furniture
Possible rewriting	3	onkyo speaker system	↦	yamaha speaker system
		eye-glasses	↦	contact lenses
		orlando bloom	↦	johnny depp
		cow	↦	pig
		ibm thinkpad	↦	laptop bag
Clear mismatch	4	jaguar xj6	↦	os x jaguar
		time magazine	↦	time and date magazine

Typical Query Suggestion

[Jinxi Xu, 1996]

- Global analysis
  - Selects expansion terms on the basis of the information on the whole document set
  - Relatively robust
  - Expensive in terms of disk space and computer time
- Local analysis
  - Formulate expansion terms based on top-ranked results
  - Relatively efficient
  - Perform badly for queries with few relevant documents

Query Expansion by Mining Query Log

[Hang Cui, 2003]

- TF-iDF
  - Each document is represented as a document vector $\{W_1^{(d)}, W_2^{(d)}, \ldots W_N^{(d)}\}$, where $W_i^{(d)}$ is the weight of the *i*th item in a document, defined as

$$W_i^{(d)} = \frac{\ln(1 + tf_i^{(d)}) \times idf_i^{(d)}}{\sqrt{\sum \ln^2(1 + tf_i^{(d)}) \times \sum (idf_i^{(d)})^2}},$$

$$idf_i^{(d)} = \ln \frac{N}{n_i},$$

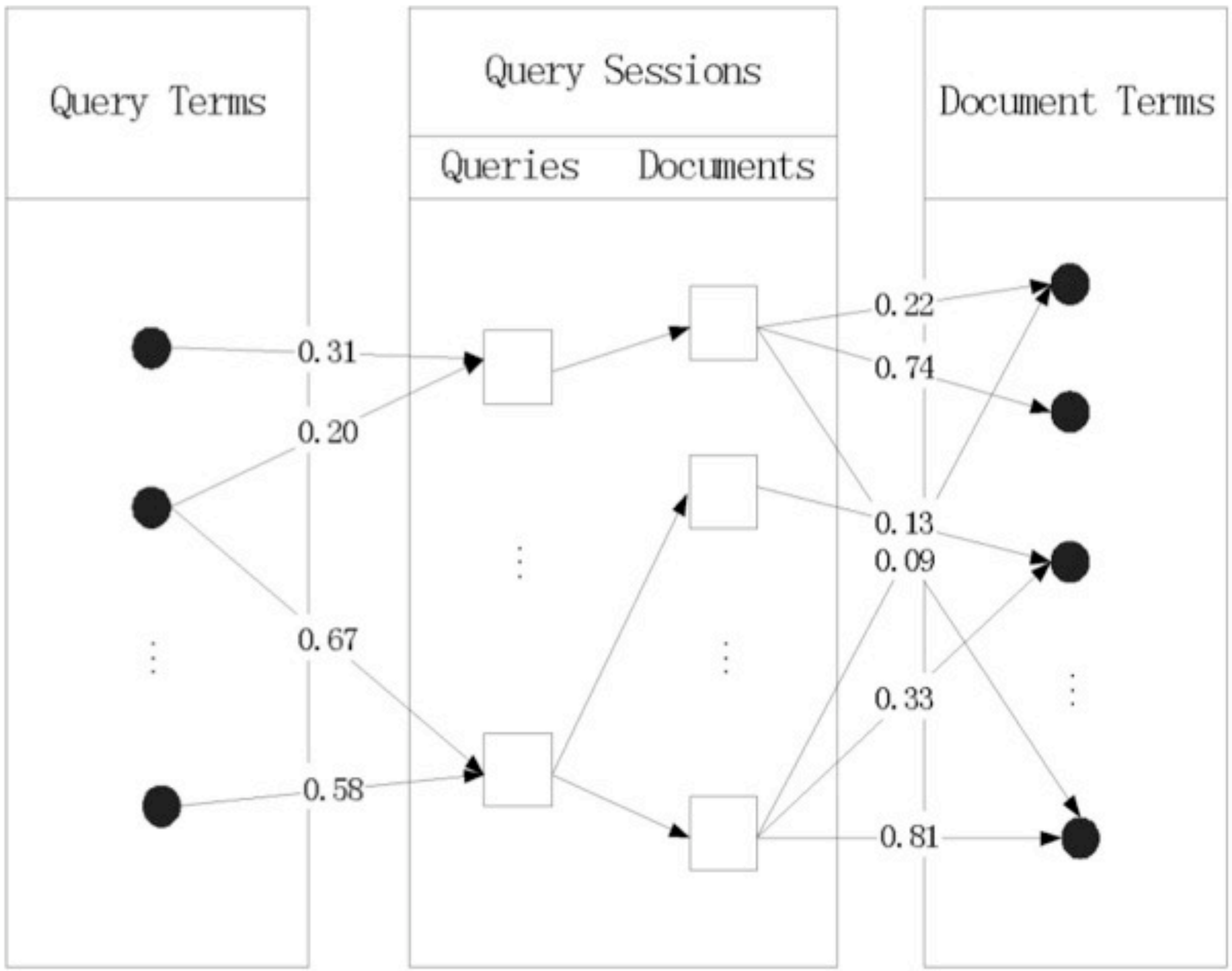
  - Similarity between query terms and document terms

$$Similarity = \frac{\sum_{i=1}^{N} W_i^{(q)} W_i^{(d)}}{\sqrt{\sum_{i=1}^{N} (W_i^{(q)})^2} \sqrt{\sum_{i=1}^{N} (W_i^{(d)})^2}}.$$

Query Expansion by Mining Query Log

[Hang Cui, 2003]

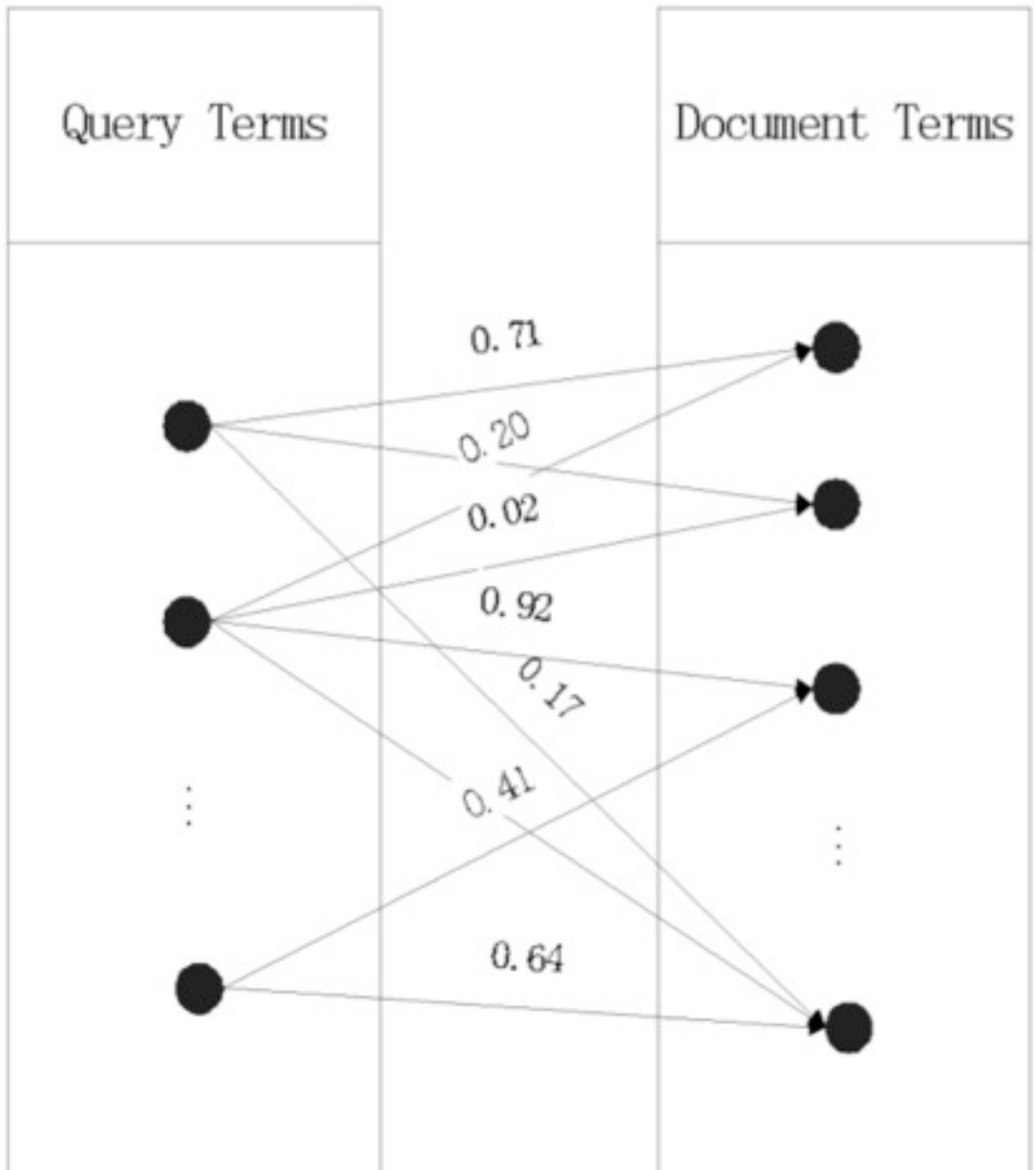
- The general idea



Query Expansion by Mining Query Log

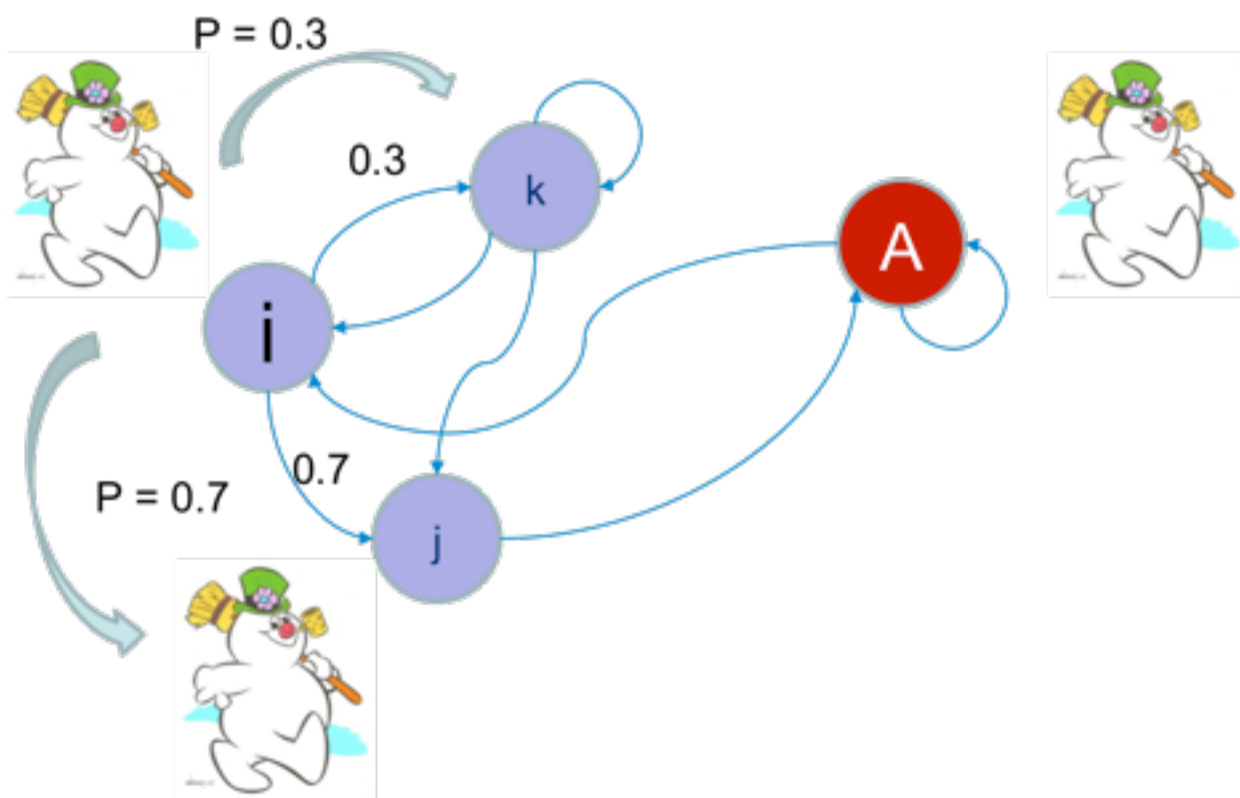
[Hang Cui, 2003]

- Establishing correlation between query terms and document terms via query sessions



Query Suggestion Using Hitting Time

[Qiaozhu Mei, 2008]

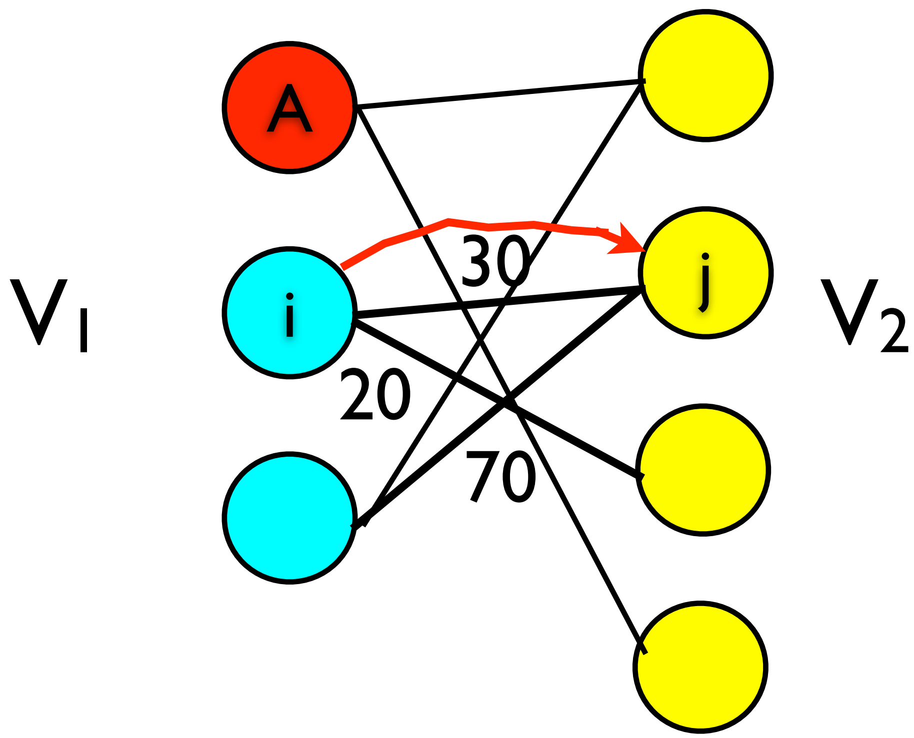


- Hitting time T^A: the first time that the random walk is at vertex A
- Mean hitting time h_i^A: expectation of T^A given that the walk starts from i

Query Suggestion Using Hitting Time

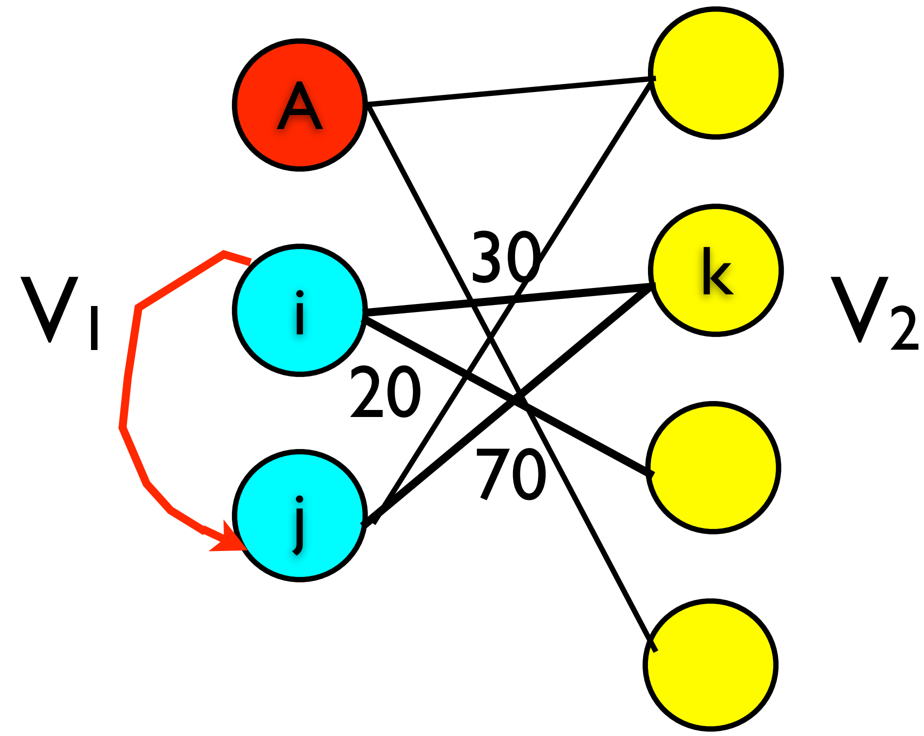
[Qiaozhu Mei, 2008]

- Calculating the transition probability



i, j at different side

$$p_{ij} = \frac{w(i,j)}{d_i}$$



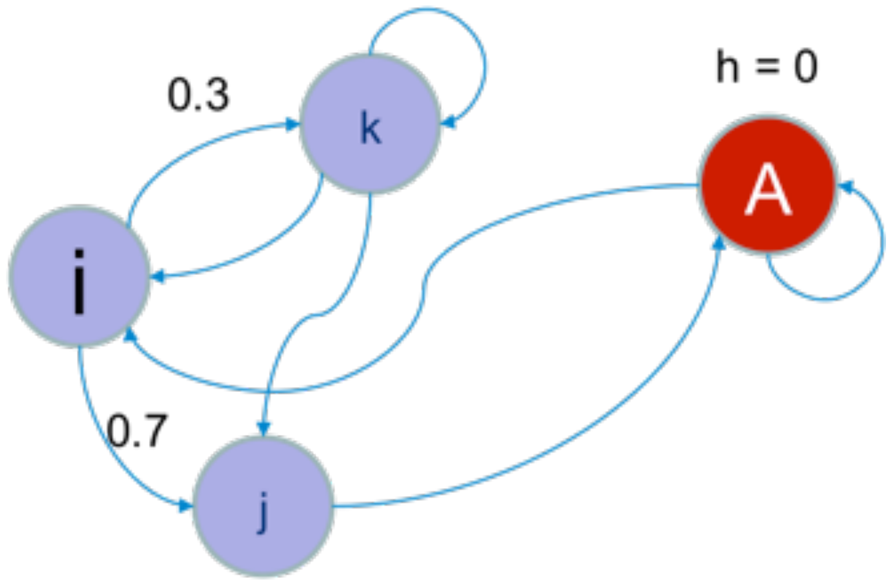
i, j at the same side

$$p_{ij} = \sum_{k \in V_2} \frac{w(i,k)}{d_i} \frac{w(k,j)}{d_k}$$

Query Suggestion Using Hitting Time

[Qiaozhu Mei, 2008]

- Computing hitting time $$h_i^A = \sum_{j \in V} p_{ij} h_j^A + 1$$



$h_i^A = 0.7\, h_j^A + 0.3\, h_k^A + 1$

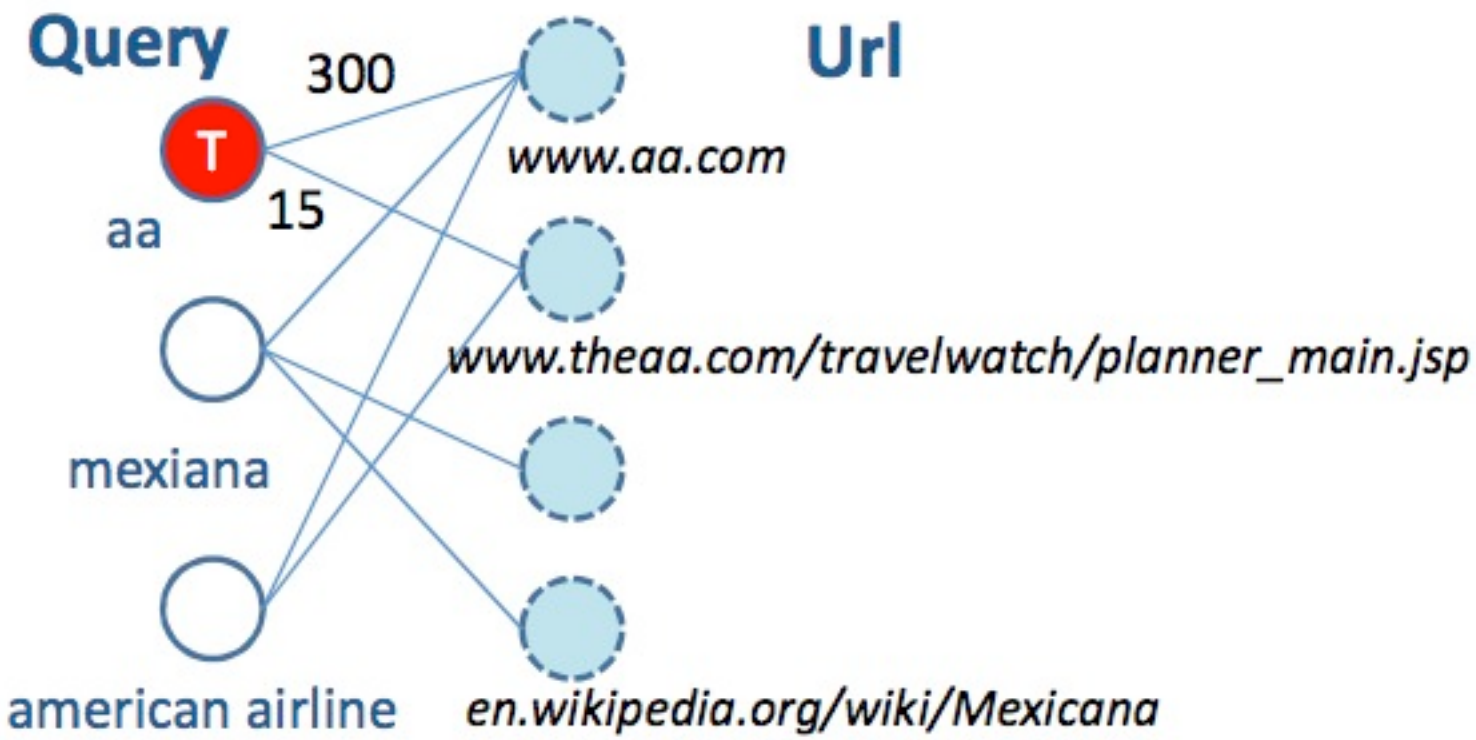
$$\begin{cases} h_i^A = 0 & \text{for } i \in A \\ h_i^A = \sum_{j \notin A} p_{ij} h_j^A + 1 & \text{for } i \notin A \end{cases}$$

Iteratively!

Query Suggestion Using Hitting Time

[Qiaozhu Mei, 2008]

- Generate query suggestion



Query-URL bipartite graph

Step 1: construct a (kNN) subgraph from query log data

Step 2: compute transition probabilities p_{ij}

Step 3: compute hitting time h_i^A

Step 4: rank candidate queries using h_i^A

Query Suggestion Using Hitting Time

[Qiaozhu Mei, 2008]

Query = msg			
HititngTime	Google	Yahoo	Live
msg facts food msg poisoning of america msg in fast food ... msg network madison square garden	msg chinese food msg health other names for msg msg duty msg symptoms marine security guard michael schenker	msg error msg network msg seating chart valentine msg foods with msg yahoo msg verizon text msg	Madison Square Garden Msg Allergy MSN Msg Food Monosodium Glutamate Ticketmaster Msg Tickets
Query = friends			
HittingTime	Google	Yahoo	Live
wikipedia friends friends tv show wikipedia friends home page friends warner bros the friends series friends official site friends(1994)	friendship friends poem friendster friends episode guide friends scripts how to make friends true friends	secret friends friends reunited hide friends hi 5 friends find friends poems for friends friends quotes	Find Friend Friendship Friends TV Show Best Friends Secret Friends Jennifer Aniston Friendster
Query = aa			
HittingTime	Google	Yahoo	Live
alcoholics anonymous automobile association theaa american airlines american air american airline- ticket reservations	N/A	aa route planner aa route finder aa airlines aa meetings aa autoroute aa road map aa 12 shotgun	AA Route Finder AA Route Planner AA Airlines American Airlines American Airlines AA Meetings

Query suggestions generated using hitting time on Query-URL graph

Query Suggestion Using Clickthrough Data

- Query logs recorded by search engines

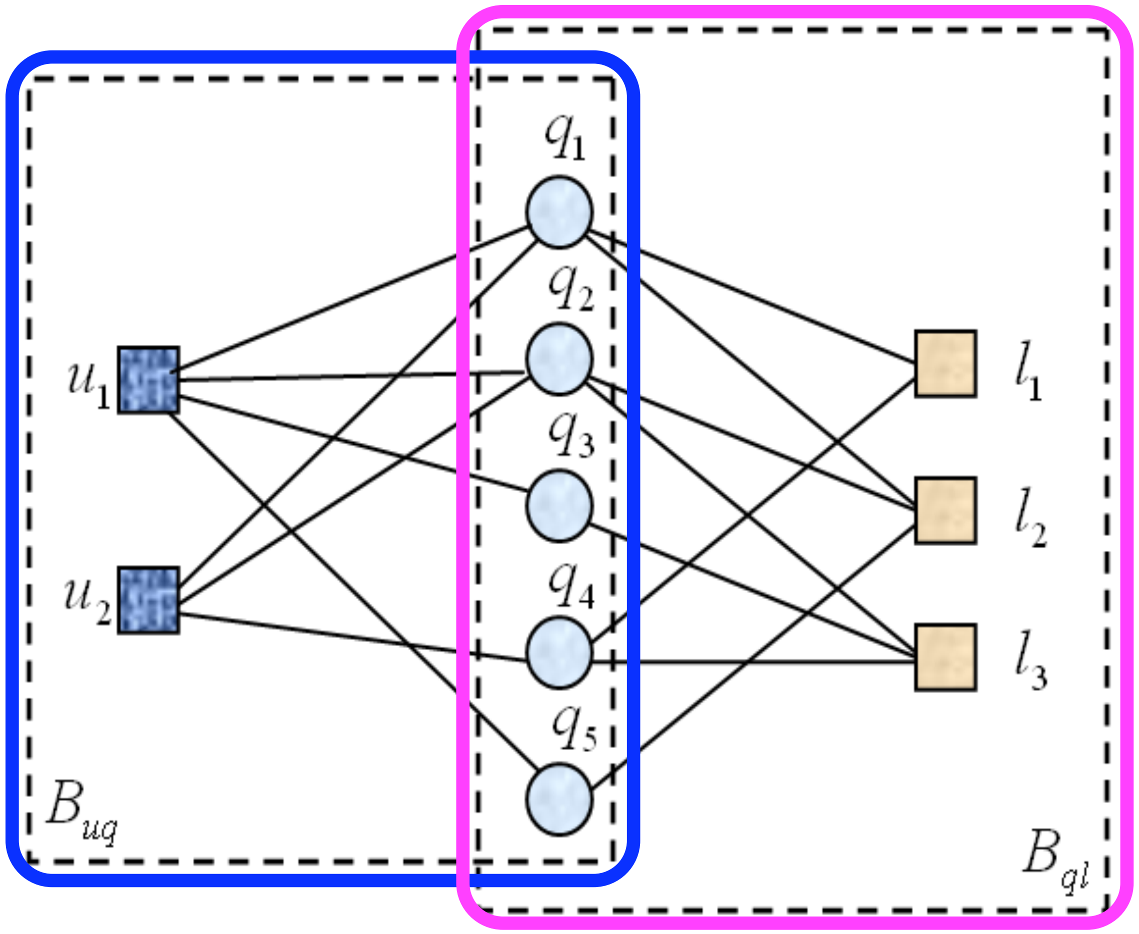
$$\langle u, q, l, r, t \rangle$$

Table 1: Samples of search engine clickthrough data

ID	Query	URL	Rank	Time
358	facebook	http://www.facebook.com	1	2008-01-01 07:17:12
358	facebook	http://en.wikipedia.org/wiki/Facebook	3	2008-01-01 07:19:18
3968	apple iphone	http://www.apple.com/iphone/	1	2008-01-01 07:20:36
...	...	...	...	...

- Users' relevance feedback to indicate desired/preferred/target results

Joint Bipartite Graph



$B_{uq} = (V_{uq}, E_{uq})$
$V_{uq} = U \cup Q$
$U = \{u_1, u_2, ..., u_m\}$
$Q = \{q_1, q_2, ..., q_n\}$
$E_{uq} = \{(u_i, q_j)|$ there is an edge from u_i to $q_j\}$ is the set of all edges.
The edge (u_i, q_j) exists in this bipartite graph if and only if a user u_i issued a query q_j.

$B_{ql} = (V_{ql}, E_{ql})$
$V_{ql} = Q \cup L$
$Q = \{q_1, q_2, ..., q_n\}$
$L = \{l_1, l_2, ..., l_p\}$
$E_{ql} = \{(q_i, l_j)|$ there is an edge from q_i to $l_j\}$ is the set of all edges.
The edge (q_j, l_k) exists if and only if a user u_i clicked a URL l_k after issuing an query q_j.

Key Points

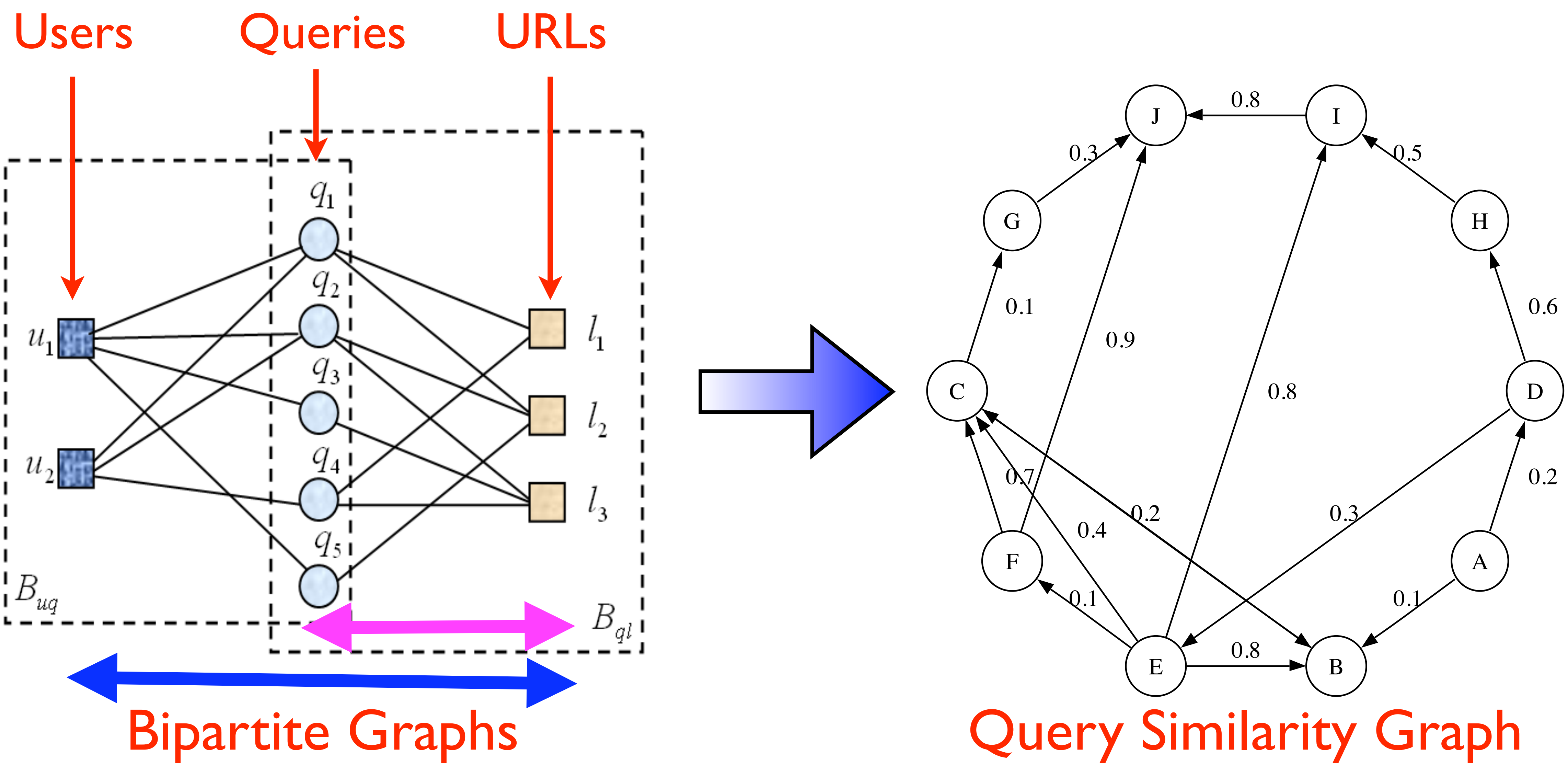
- Two-level latent semantic analysis

Level 1

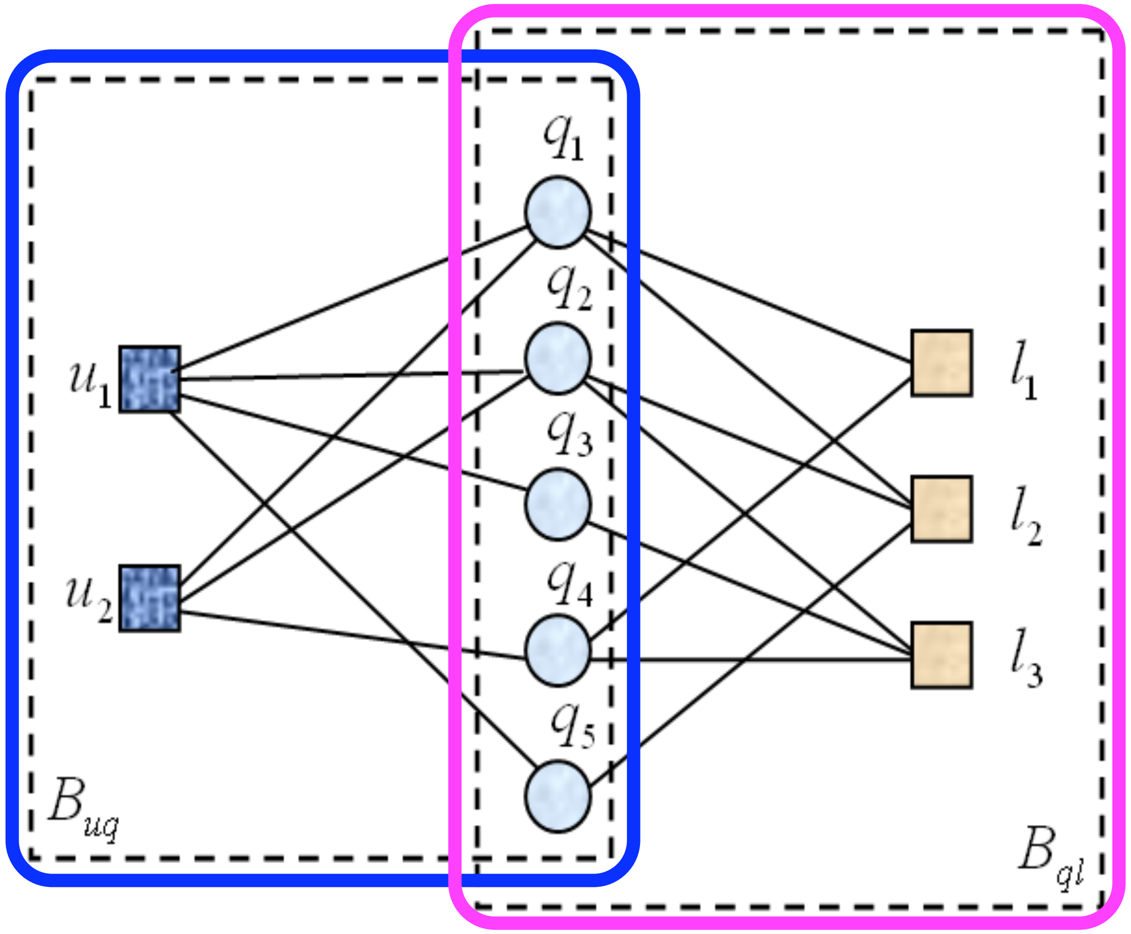
- Consider the use of a joint user-query and query-URL bipartite graphs for query suggestion
- Use matrix factorization for learning query features in constructing the Query Similarity Graph

Level 2

- Use heat diffusion for similarity propagation for query suggestions



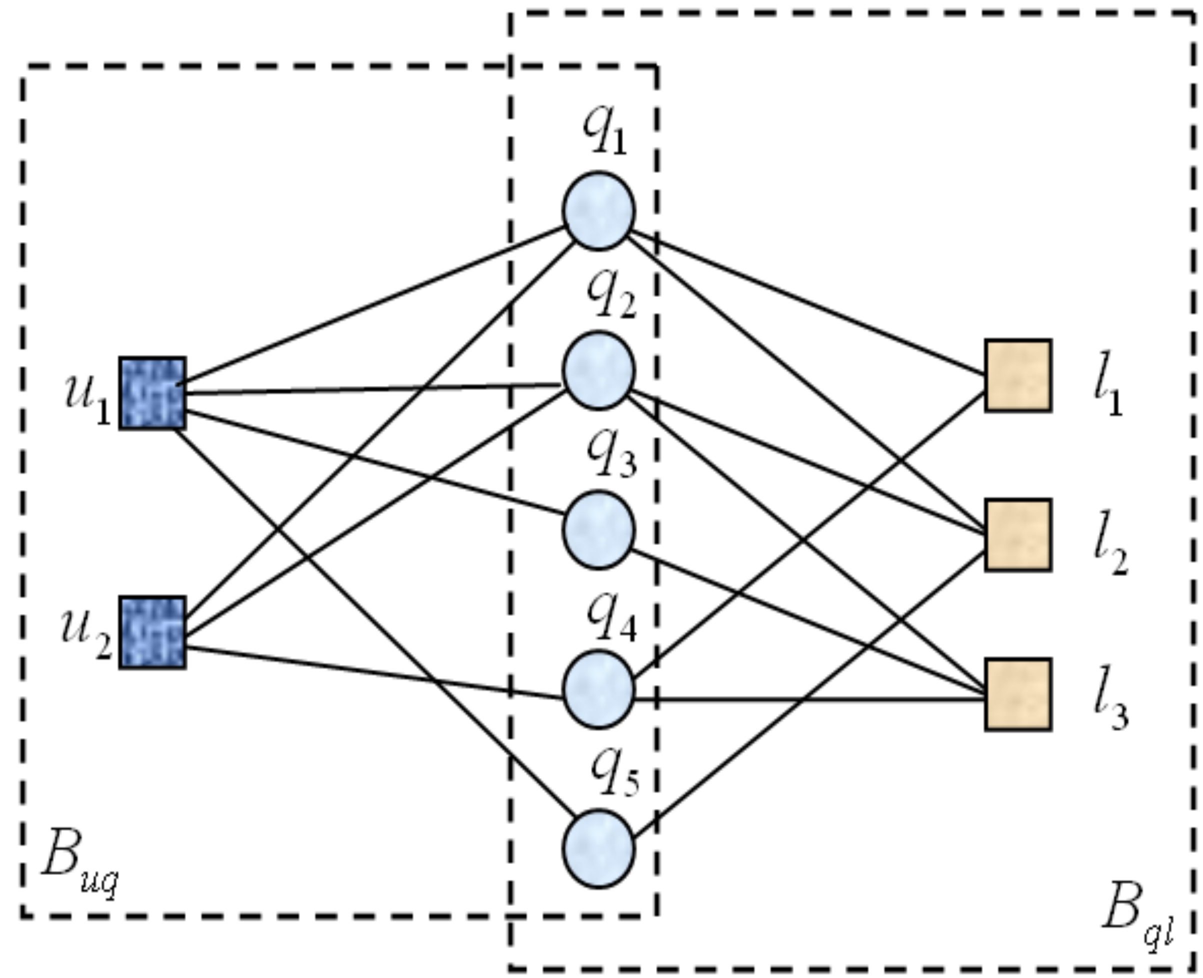
- Queries are issued by the users, and which URLs to click are also decided by the users
- Two distinct users are similar if they issued similar queries
- Two queries are similar if they are issued by similar users



r^*_{ij}	Normalized weight, how many times u_i issued q_j
s^*_{jk}	Normalized weight, how many times q_j is linked to l_k
U_i	L-dimensional vector of user u_i
Q_j	L-dimensional vector of query q_j
L_k	L-dimensional vector of URL l_k

$$\begin{aligned}\mathcal{H}(R,U,Q) &= \min_{U,Q}\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I^R_{ij}(r^*_{ij} - g(U_i^T Q_j))^2 \\ &+ \frac{\alpha_u}{2}\|U\|_F^2 + \frac{\alpha_q}{2}\|Q\|_F^2\end{aligned}$$

$$\begin{aligned}\mathcal{H}(S,Q,L) &= \min_{Q,L}\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{p} I^S_{jk}(s^*_{jk} - g(Q_j^T L_k))^2 \\ &+ \frac{\alpha_q}{2}\|Q\|_F^2 + \frac{\alpha_l}{2}\|L\|_F^2\end{aligned}$$



$$\mathcal{H}(S,R,U,Q,L) =$$
$$\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{p} I_{jk}^{S}(s_{jk}^{*} - g(Q_j^T L_k))^2 + \frac{\alpha_r}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}(r_{ij}^{*} - g(U_i^T Q_j))^2$$
$$+\frac{\alpha_u}{2}\|U\|_F^2 + \frac{\alpha_q}{2}\|Q\|_F^2 + \frac{\alpha_l}{2}\|L\|_F^2,$$

- A local minimum can be found by performing gradient descent in U_i, Q_j and L_k

Gradient Descent Equations

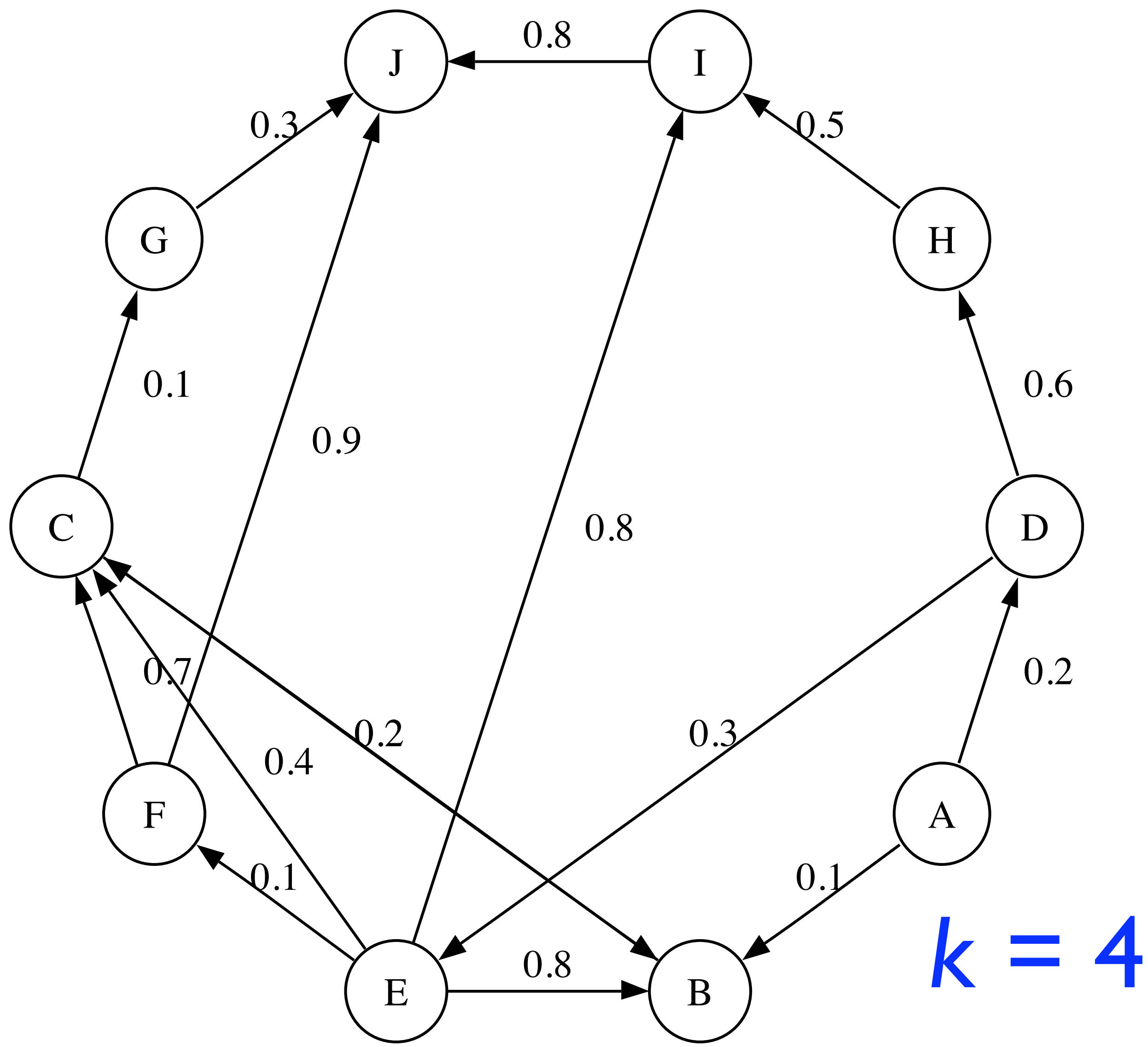
$$\frac{\partial \mathcal{H}}{\partial U_i} = \alpha_r \sum_{j=1}^{n} I_{ij}^{R} g'(U_i^T Q_j)(g(U_i^T Q_j) - r_{ij}^{*}) Q_j + \alpha_u U_i,$$

$$\boxed{\begin{aligned}\frac{\partial \mathcal{H}}{\partial Q_j} &= \sum_{k=1}^{p} I_{jk}^{S} g'(Q_j^T L_k)(g(Q_j^T L_k) - s_{jk}^{*}) L_k \\ &+ \alpha_r \sum_{i=1}^{m} I_{ij}^{R} g'(U_i^T Q_j)(g(U_i^T Q_j) - r_{ij}^{*}) U_i + \alpha_q Q_j,\end{aligned}}$$

$$\frac{\partial \mathcal{H}}{\partial L_k} = \sum_{j=1}^{n} I_{jk}^{S} g'(Q_j^T L_k)(g(Q_j^T L_k) - s_{jk}^{*}) Q_j + \alpha_l L_k,$$

Only the Q matrix, the queries' latent features,
is being used to generate the query similarity graph!

Query Similarity Graph



- Similarities are calculated using queries' latent features
- Only the top-*k* similar neighbors (terms) are kept

Similarity Propagation

- Based on the Heat Diffusion Model
- In the query graph, given the heat sources and the initial heat values, start the heat diffusion process and perform *P* steps
- Return the Top-*N* queries in terms of highest heat values for query suggestions

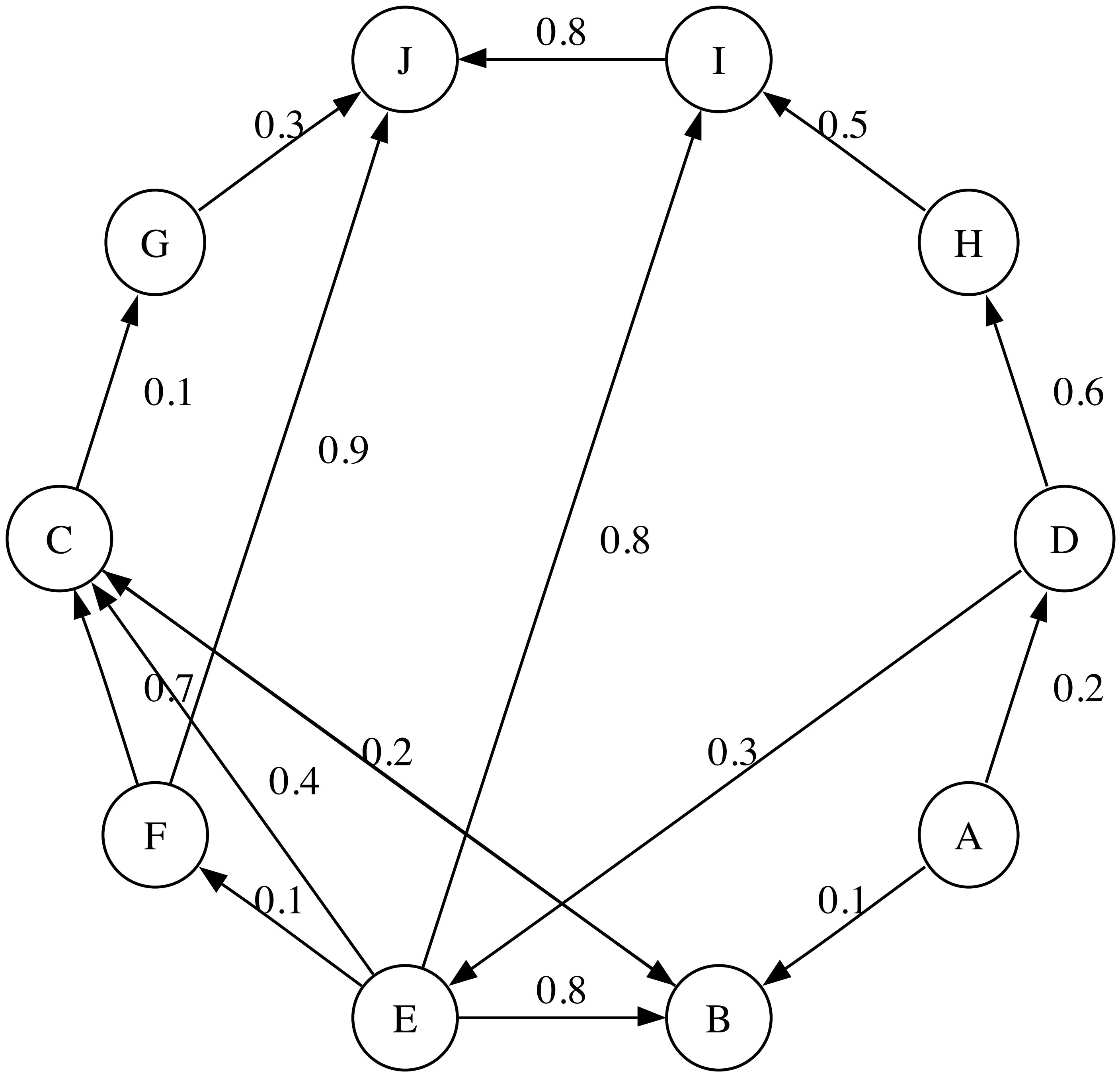
Heat Diffusion Model

- Heat diffusion is a physical phenomena
- Heat flows from high temperature to low temperature in a medium
- Heat kernel is used to describe the amount of heat that one point receives from another point
- The way that heat diffuse varies when the underlying geometry varies

$$\boxed{\rho C_P \frac{\partial T}{\partial t} = Q + \nabla \cdot (k \nabla T)}$$

ρ	Density
C_P	Heat capacity and constant pressure
$\frac{\partial T}{\partial t}$	Change in temperature over time
Q	Heat added
k	Thermal conductivity
∇T	Temperature gradient
$\nabla \cdot \mathbf{v}$	Divergence

Heat Diffusion Process



Similarity Propagation Model

$$\frac{f_i(t+\Delta t) - f_i(t)}{\Delta t} = \alpha\left(-\frac{\tau_i}{d_i} f_i(t) \sum_{k:(q_i,q_k)\in E} w_{ik} + \sum_{j:(q_j,q_i)\in E} \frac{w_{ji}}{d_j} f_j(t)\right) \quad (1)$$

$$\mathbf{f}(1) = e^{\alpha \mathbf{H}} \mathbf{f}(0) \quad (2)$$

$$H_{ij} = \begin{cases} w_{ji}/d_j, & (q_j, q_i) \in E, \\ -(\tau_i/d_i) \sum_{k:(i,k)\in E} w_{ik}, & i = j, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

$$\mathbf{f}(1) = e^{\alpha \mathbf{R}} \mathbf{f}(0), \quad \boxed{\mathbf{R} = \gamma \mathbf{H} + (1-\gamma)\mathbf{g}\mathbf{1}^T} \quad (4)$$

α	Thermal conductivity
d_i	Heat value of node i at time t
$f_i(t)$	Heat value of node i at time t
w_{ik}	Weight between node i and node k
$\mathbf{f}(0)$	Vector of the initial heat distribution
$\mathbf{f}(1)$	Vector of the heat distribution at time 1
τ_i	Equal to 1 if node i has outlinks, else equal to 0
γ	Random jump parameter, and set to 0.85
$\mathbf{g}$	Uniform stochastic distribution vector

Discrete Approximation

- Compute $e^{\alpha \mathbf{R}}$ is time consuming
- We use the discrete approximation to substitute

$$\mathbf{f}(1) = \left(\mathbf{I} + \frac{\alpha}{P}\mathbf{R}\right)^P \mathbf{f}(0)$$

- For every heat source, only diffuse heat to its neighbors within *P* steps
- In our experiments, *P* = 3 already generates fairly good results

Query Suggestion Procedure

- For a given query *q*

1. Select a set of *n* queries, each of which contains at least one word in common with *q*, as heat sources
2. Calculate the initial heat values by

$$f_{\hat{q}_i}(0) = \frac{|\mathcal{W}(q) \cap \mathcal{W}(\hat{q}_i)|}{|\mathcal{W}(q) \cup \mathcal{W}(\hat{q}_i)|}$$

q = "Sony"
"Sony" = 1
"Sony Electronics" = 1/2
"Sony Vaio Laptop" = 1/3

3. Use $\mathbf{f}(1) = e^{\alpha \mathbf{R}} \mathbf{f}(0)$ to diffuse the heat in graph
4. Obtain the Top-*N* queries from $\mathbf{f}(1)$

Physical Meaning of α

- If set α to a large value
  - The results depend more on the query graph, and more semantically related to original queries, e.g., travel => lowest air fare
- If set α to a small value
  - The results depend more on the initial heat distributions, and more literally similar to original queries, e.g., travel => travel insurance

Query Suggestions

Table 2: Examples of LSQS Query Suggestion Results ($k = 50$)

Testing Queries	Suggestions				
	$\alpha = 10$			$\alpha = 1000$	
	Top 1	Top 2	Top 3	Top 4	Top 5
michael jordan	michael jordan shoes	michael jordan bio	pictures of michael jordan	nba playoff	nba standings
travel	travel insurance	abc travel	travel companions	hotel tickets	lowest air fare
java	sun java	java script	java search	sun microsystems inc	virtual machine
global services	ibm global services	global technical services	staffing services	temporary agency	manpower professional
walt disney land	world of disney	disney world orlando	disney world theme park	disneyland grand hotel	disneyland in california
intel	intel vs amd	amd vs intel	pentium d	pentium	centrino
job hunt	jobs in maryland	monster job	jobs in mississippi	work from home online	monster board
photography	photography classes	portrait photography	wedding photography	adobe elements	canon lens
internet explorer	ms internet explorer	internet explorer repair	internet explorer upgrade	microsoft com	security update
fitness	fitness magazine	lifestyles family fitness	fitness connection	womens health magazine	family fitness
m schumacher	schumacher	red bull racing	formula one racing	ferrari cars	formula one
solar system	solar system project	solar system facts	solar system planets	planet jupiter	mars facts
sunglasses	replica sunglasses	cheap sunglasses	discount sunglasses	safilo	marhon
search engine	audio search engine	best search engine	search engine optimization	song lyrics search	search by google
disease	grovers disease	liver disease	morgellons disease	colic in babies	oklahoma vital records
pizzahut	pizza hut menu	pizza coupons	pizza hut coupons	papa johns pizza coupon	papa johns
health care	health care proxy	universal health care	free health care	great west healthcare	uhc
flower delivery	global flower delivery	online florist	flowers online	send flowers	virtual flower
wedding	wedding guide	wedding reception ideas	wedding decoration	unity candle	centerpiece ideas
astronomy	astronomy magazine	astronomy pic of the day	star charts	space pictures	comet

References

- S. Cucerzan and R. W. White. Query suggestion based on user landing pages. In SIGIR, pages 875–876, 2007.
- H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. IEEE Trans. Knowl. Data Eng., 15(4):829–839, 2003.
- W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon. Cross-lingual query suggestion using query logs of different languages. In SIGIR, pages 463–470, 2007.
- R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors, WWW, pages 387–396. ACM, 2006.
- H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In CIKM, pages 709–718, 2008.
- Q. Mei, D. Zhou, and K. W. Church. Query suggestion using hitting time. In CIKM, pages 469–478, 2008.
- J. Xu and W. B. Croft. Query expansion using local and global document analysis. In SIGIR, pages 4–11, 1996.

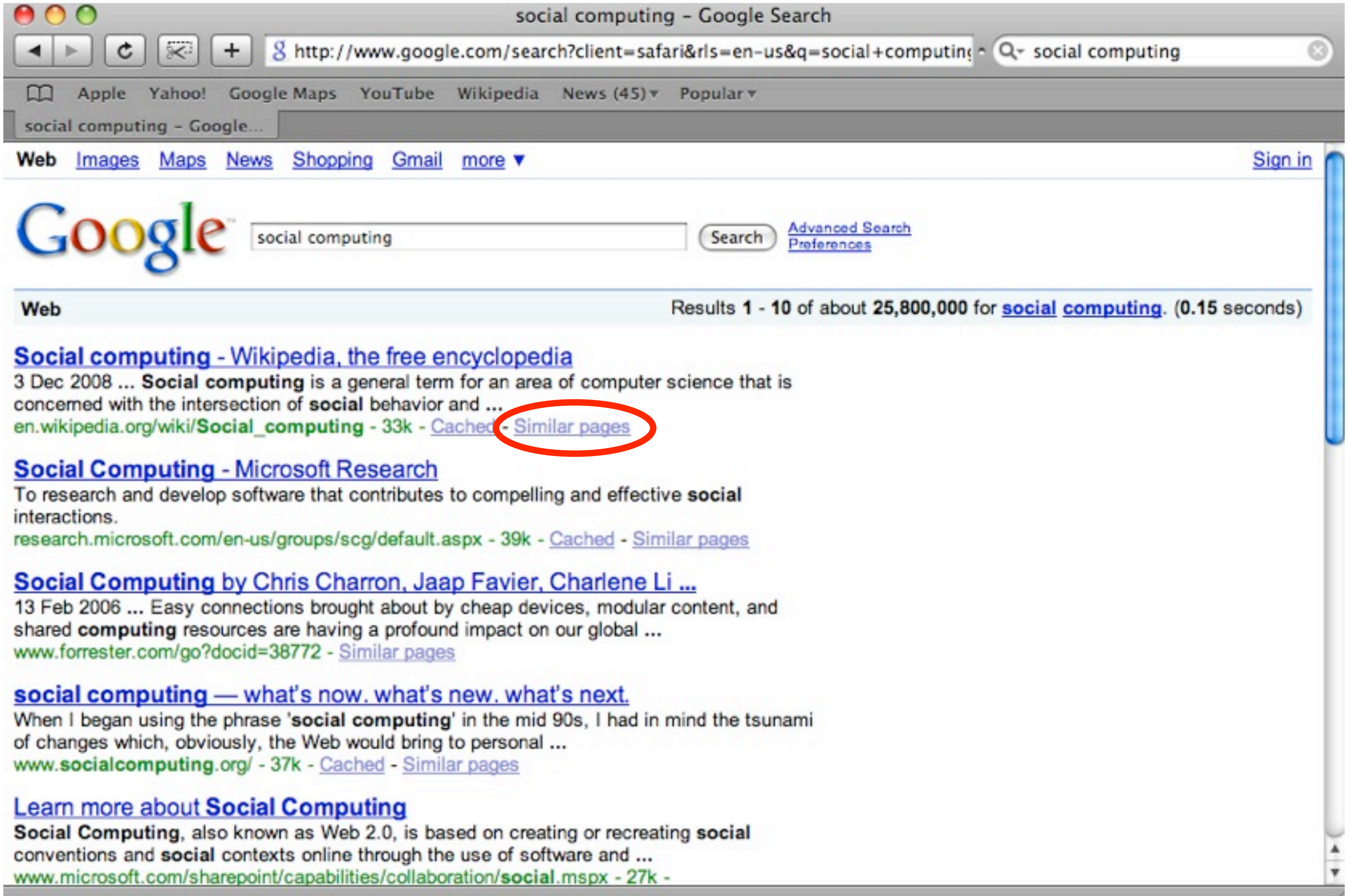
Collaborative Filtering

Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

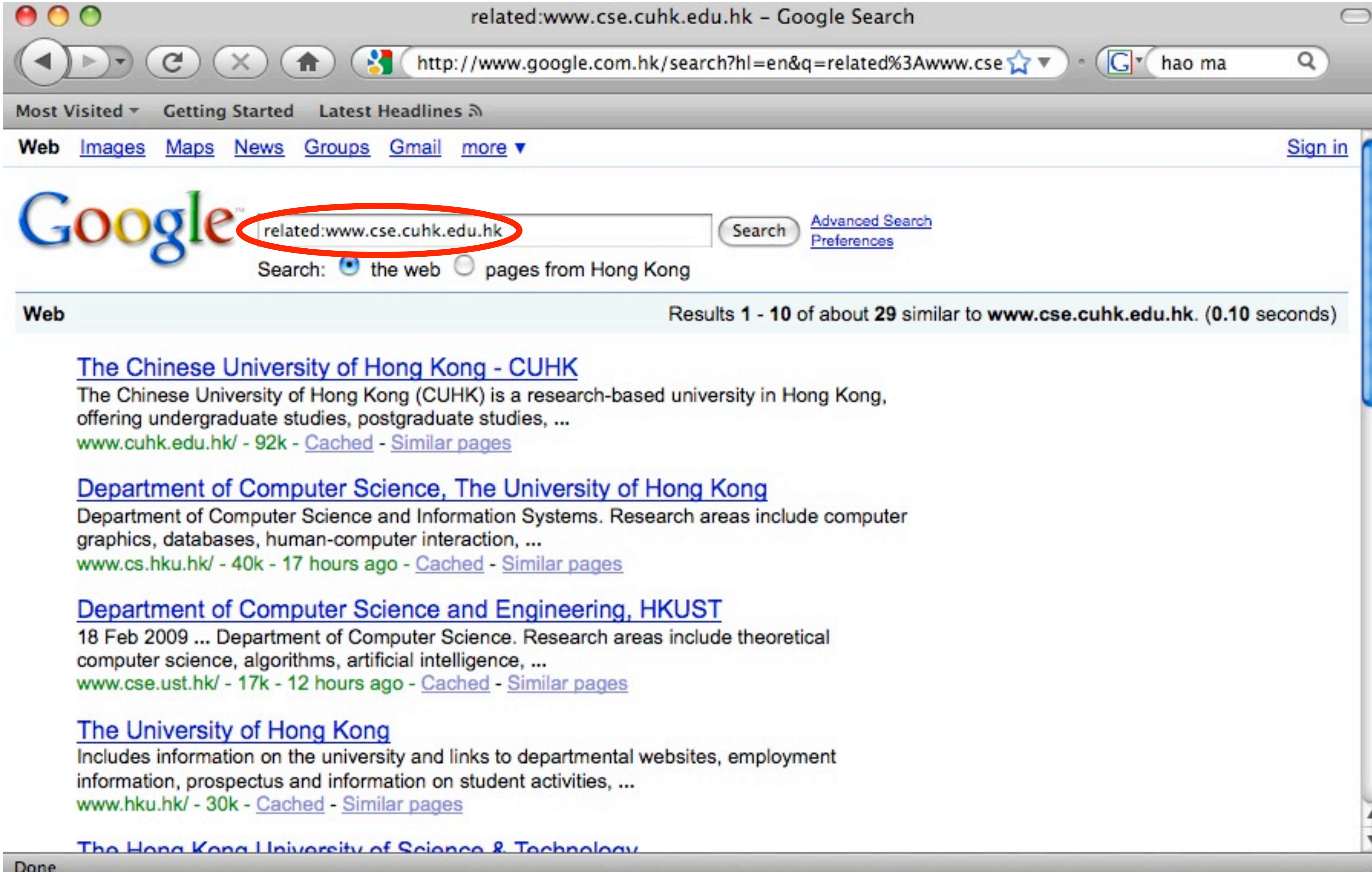
Outline

- Introduction
- The Framework
- User-User Method
  - Memory-based
  - Model-based
- Item-Item Method
  - Correlation Analysis
  - Association Rule Mining

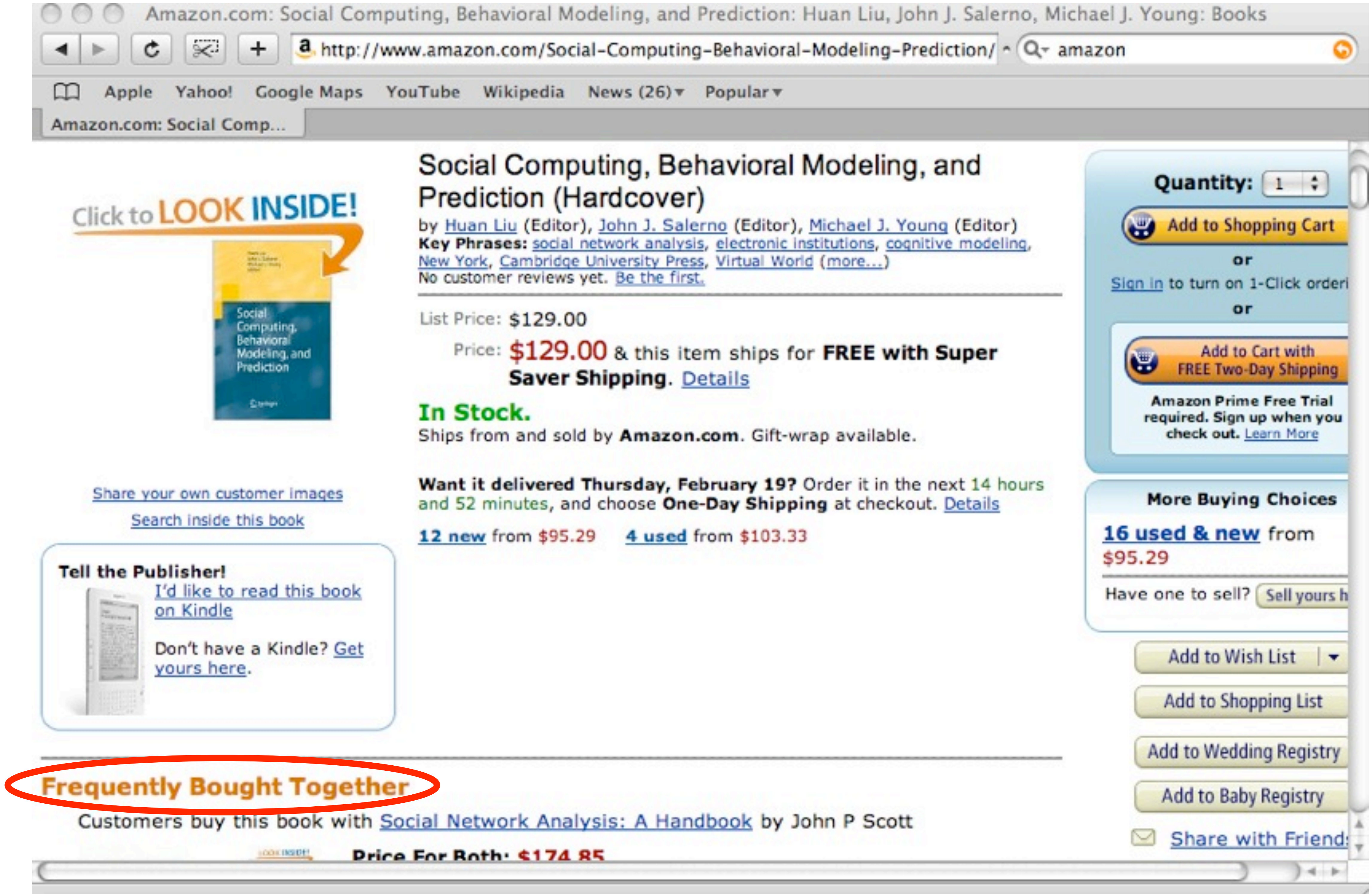
Real Life Examples



Real Life Examples



Real Life Examples



Real Life Examples

Five scales rating

★ I hate it

★★ I don't like it

★★★ It's ok

★★★★ I like it

★★★★★ I love it

Motivation

- User Perspective
  - Lots of online products, books, movies, news, web pages, etc.
  - Reduce my choices…please…

- Manager Perspective

" *if I have 3 million customers on the web, I should have 3 million stores on the web."*

CEO of Amazon.com [SCH01]

Basic Approaches

- Content-based Filtering
  - Recommend items based on key-words
  - More appropriate for information retrieval
- Collaborative Filtering (CF)
  - Look at users with similar rating styles
  - Look at similar items for each item

Underling assumption: personal tastes are correlated--
Active user will prefer those items which the similar users prefer.

Framework

Items

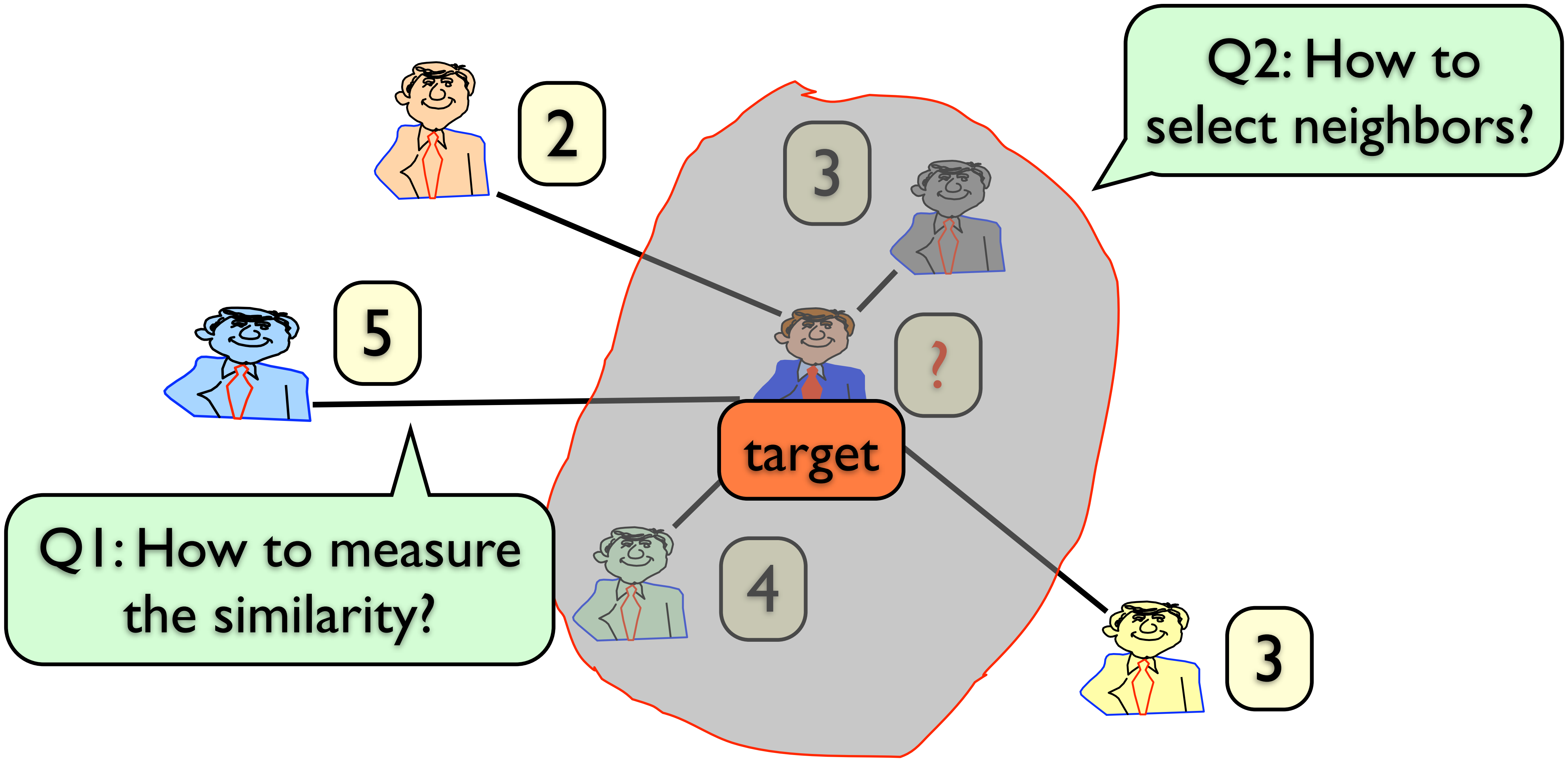
Users	i_1	i_2				i_j							i_m
u_1													
u_2	1	3		4		2		5			3	4	
u_i		3		4		r_{ij}	3	4		3	4		4
u_n	1			3	5	2		4	1			3	

- The tasks
  - Find the unknown rating?
  - Which item should be recommended?

Collaborative Filtering

- User-User Methods
  - Memory-based
  - Model-based
- Item-Item Method
  - Identify buying patterns
  - Correlation Analysis
  - Linear Regression
  - Belief Network
  - Association Rule Mining

User-User Similarity



User-based Collaborative Filtering

Items

Users

u_1													
u_2	1	3		4		2		5			3	4	
u_3													
u_4		3		4		$\boxed{}$	3	4		3	4		4
u_5													
u_6	1			3	5	2		4	1			3	

User-based Collaborative Filtering

Items

Users													
u_1													
u_2	1	3		4		2		5			3	4	
u_3													
u_4		3		4			3	4		3	4		4
u_5													
u_6	1			3	5	2		4	1			3	

User-based Collaborative Filtering

Items

Users

u_1													
u_2	1	3		4		2		5			3	4	
u_3													
u_4		3		4			3	4		3	4		4
u_5													
u_6	1			3	5	2		4	1			3	

User-based Collaborative Filtering

Items

Users

u_1													
u_2	1	3		4		2		5			3	4	
u_3													
u_4		3		4			3	4		3	4		4
u_5													
u_6	1			3	5	2		4	1			3	

User-based Collaborative Filtering

Items

Users

u_1													
u_2	1	3		4		2		5			3	4	
u_3													
u_4		3		4			3	4		3	4		4
u_5													
u_6	1			3	5	2		4	1			3	

User-based Collaborative Filtering

- Predict the ratings of active users based on the ratings of similar users found in the user-item matrix
  - Pearson correlation coefficient

$$w(a,i) = \frac{\sum_j (r_{aj} - \bar{r}_a)(r_{ij} - \bar{r}_i)}{\sqrt{\sum_j (r_{aj} - \bar{r}_a)^2 \sum_j (r_{ij} - \bar{r}_i)^2}} \quad j \in I(a) \cap I(i)$$

  - Cosine measure

$$c(a,i) = \frac{r_a \cdot r_i}{||r_a||_2 * ||r_i||_2}$$

u_i	1	3		4		2		5			3	4	
u_a		3		4			3	4		3	4		4
	1			3	5	2		4	1			3	

Nearest Neighbor Approaches

[Sarwar, 00a]

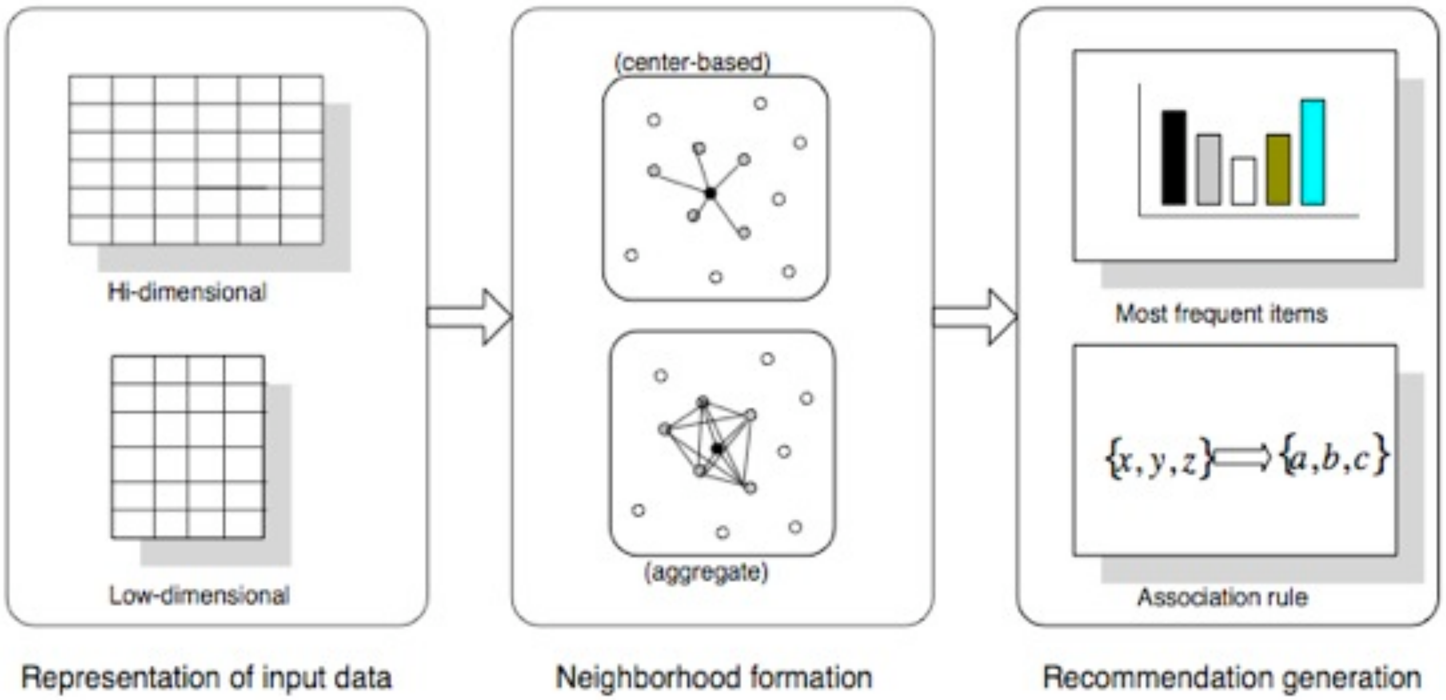


Figure 1: Three main parts of a Recommender System.

- Identify highly similar users to the active one
  - All with a measure greater than a threshold
  - Best K ones
- Prediction $$r_{aj} = \bar{r}_a + \frac{\sum_i w(a,i)(r_{ij} - \bar{r}_i)}{\sum_i w(a,i)}$$

Clustering

[Breese, 98]

- Build clusters: k-mean, k-medoid, etc. (offline)
- Identify the nearest cluster to the active user
- Prediction:
  - Use the center of the cluster

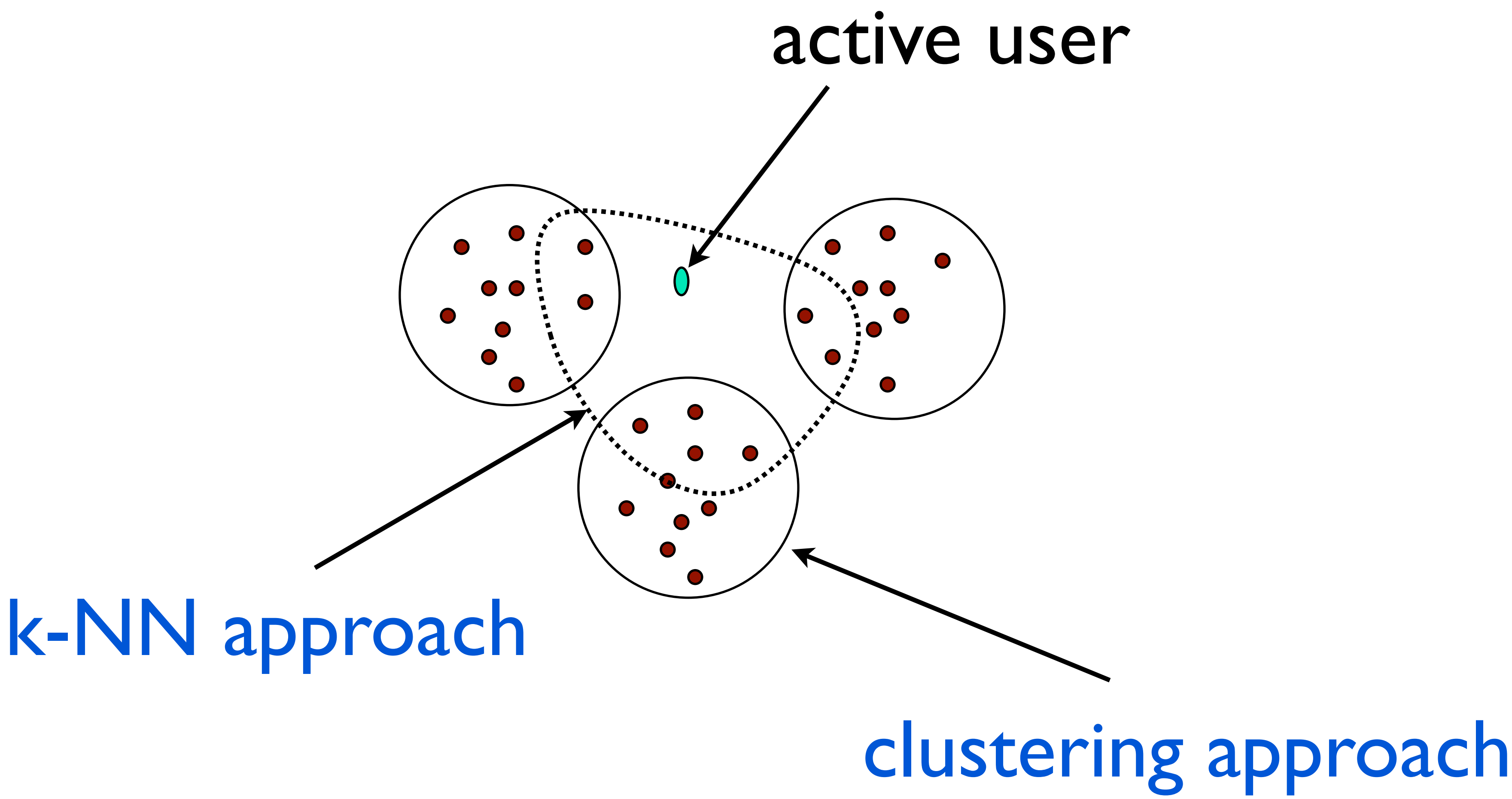
    faster

  - Weighted average between cluster members
    - Weights depend on the active user

    slower but more accurate

Clustering vs. k-NN Approaches

- K-NN using Pearson measure is slower but more accurate
- Clustering is more scalable



Data Sparsity

- Similarity: $w(a,i) = \frac{\sum_j (r_{aj} - \bar{r}_a)(r_{ij} - \bar{r}_i)}{\sqrt{\sum_j (r_{aj} - \bar{r}_a)^2 \sum_j (r_{ij} - \bar{r}_i)^2}}$

In the Amazon.com:
• 2 million books
• Active users may have rated < 1% of the product (a large set of 20,000 books)
• Pearson nearest neighbor algorithm may unable to make any product recommendation for a particular user

example from [Sarwar, 00a]

- Suffer from sparsity
  - Not enough common items
  - Implies spurious neighbors and hence bad recommendations

Selecting Relevant Instances

[Kai Yu, 2001]

	Superman	Titanic	Dance with Wolves	Batman
Jason	5			5
Karen			3	4
Fred	2	5		2
Tom	4	3	4	?

Predict this

- Superman and Batman are correlated
- Titanic and Batman are negatively correlated
- "Dances with Wolves" has nothing to do with Batman's rating
- Karen is not a good instance to consider

- Formalize: $MI(X;Y) = H(X) - H(X|Y)$

Selecting Relevant Instances

[Kai Yu, 2001]

- Offline phase:
  - Estimate mutual information between items
  - For each item:
    - Find users who rated it
    - Compute their strength of description (how many relevant items they also rated)
    - Retain subset of them (10% works fine)
- Online phase:
  - To predict the target item's rating, run k-NN on its reduced instance space

Better results with less data… quality not quantity is what matter

Collaborative Filtering

- User-User Methods
  - Memory-based
  - Model-based
- Item-Item Method
  - Correlation Analysis
  - Linear Regression
  - Belief Network
  - Association Rule Mining

Item-Item Similarity

- Search for similarities among items
- Item-Item similarity is more stable that user-user similarity
- First Order Models
  - Correlation Analysis
  - Linear Regression
- Higher Order Models
  - Belief Network
  - Association Rule Mining

Correlation-based Methods

[Sarwar, 2001]

- Same as in user-user similarity but on item vectors
- Pearson correlation coefficient
  - Look for users who rated both items

$$s_{ij} = \frac{\sum_u (r_{uj} - \bar{r}_j)(r_{ui} - \bar{r}_i)}{\sqrt{\sum_u (r_{uj} - \bar{r}_j)^2 \sum_u (r_{ui} - \bar{r}_i)^2}}$$

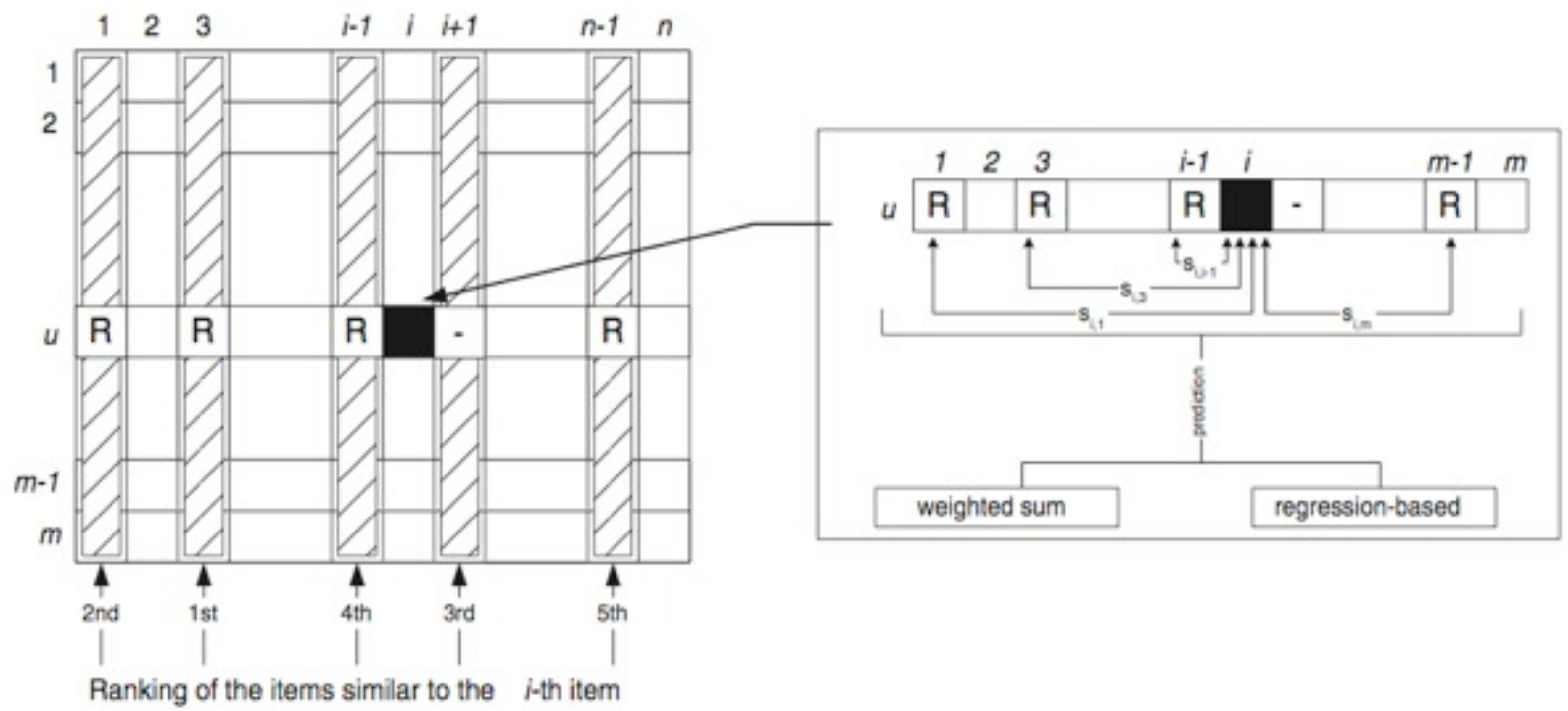
  - u: users rated both items

	i_1	i_2				i_i		i_j					i_m
u_1													
u_2	1	3		4		2		5			3	4	
u_i		3		4			3	4		3	4		4
u_n	1			3	5	2		4	1			3	

Correlation-based Method

[Sarwar, 2001]

- Calculate item similarity, then determine its k-most similar items



- Predict rating for a given user-item pair as a weighted sum over similar items that he rated

$$r_{ai} = \frac{\sum_j s_{ij} r_{aj}}{\sum_j s_{ij}}$$

					3		
u_a	2		3		?		4
					i		

Association Rule Mining

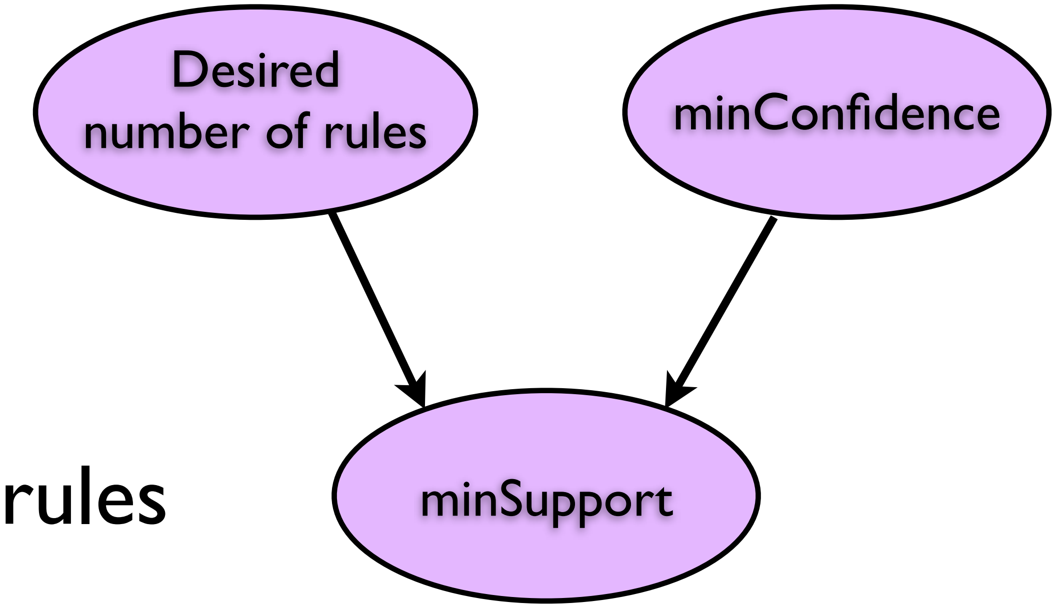
- Offline processing
  - Work on the binary level (like, dislike)
  - View user as market basket containing items liked by user
  - Discover association rules between items
- Online processing:
  - Match items that the active user like with rules left hand side
  - Recommend rules' consequent based on support and confidence

Association Rule Mining : Problems

- High support threshold leads to low coverage and may eliminate important, but infrequent items from consideration
- Low support thresholds result in very large model sizes, computationally expensive offline pattern discovery phase and slower online matching phase
- Solution:
    - Adaptive Association Rule Mining

Adaptive Association Rule Mining

[Lin, 2001]



- Given:
  - transaction dataset
  - target item
  - desired range for number of rules
  - specified minimum confidence
- Find: set S of association rules for target item such that
  - number of rules in S is in given range
  - rules in S satisfy minimum confidence constraint
  - rules in S have higher support than rules not in S that satisfy above constraints

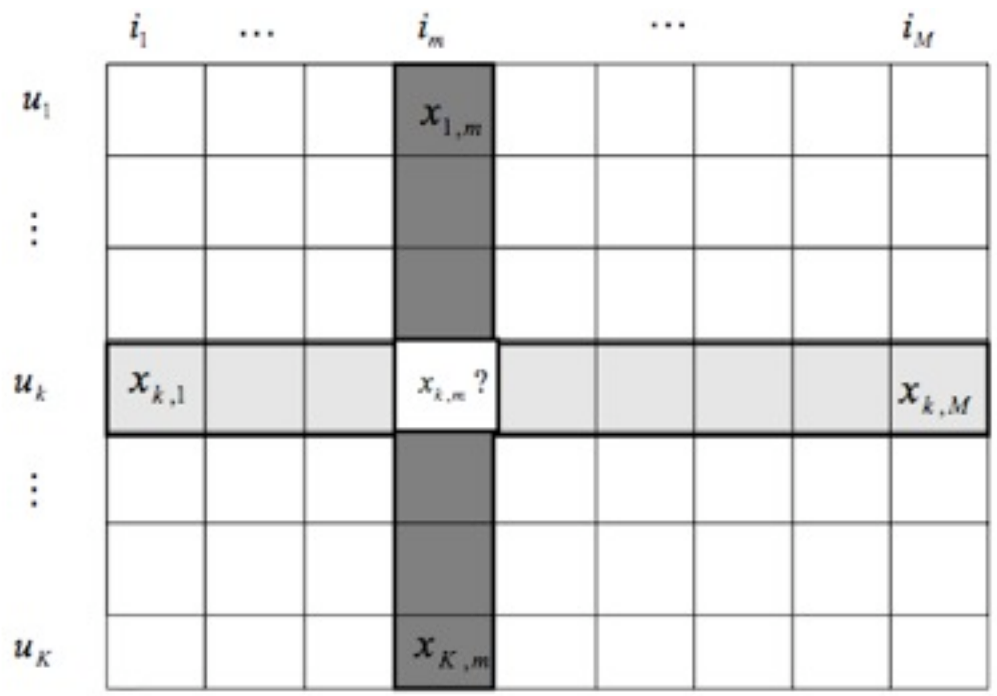
Adaptive Association Rule Mining

[Lin, 2000]

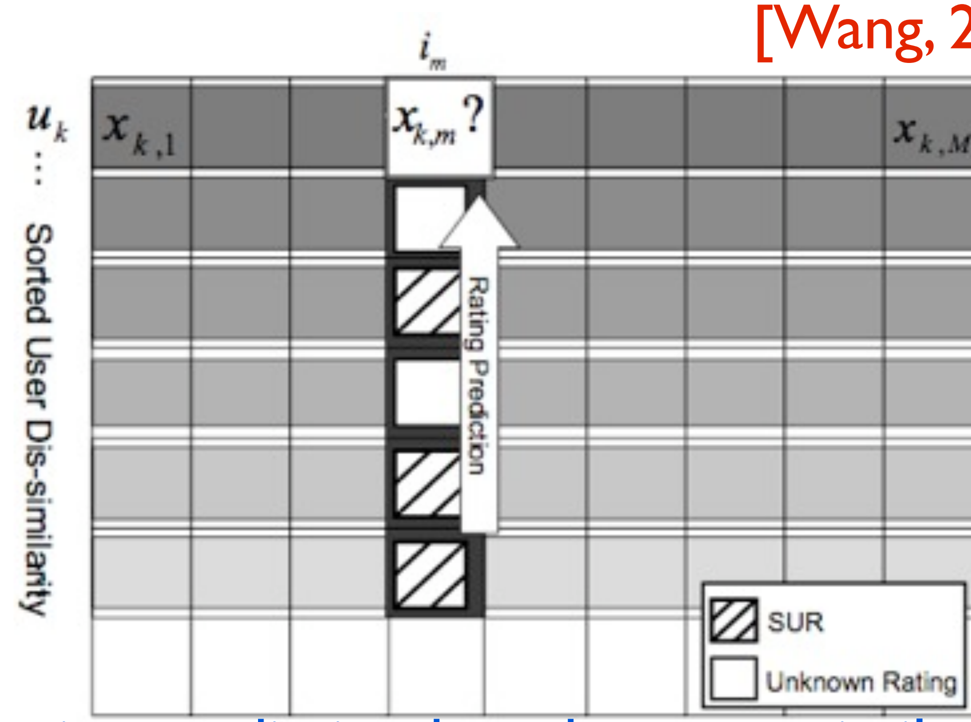
- Discover rules with one item on the head
  - Like (x, item1) ^ Like (x, item2)=>Like(x, target)
- The miner discovers association rules iteratively (for each target item) until the desired number of rules are extracted
- Support is adjusted per-item

Unifying User-based and Item-based CF

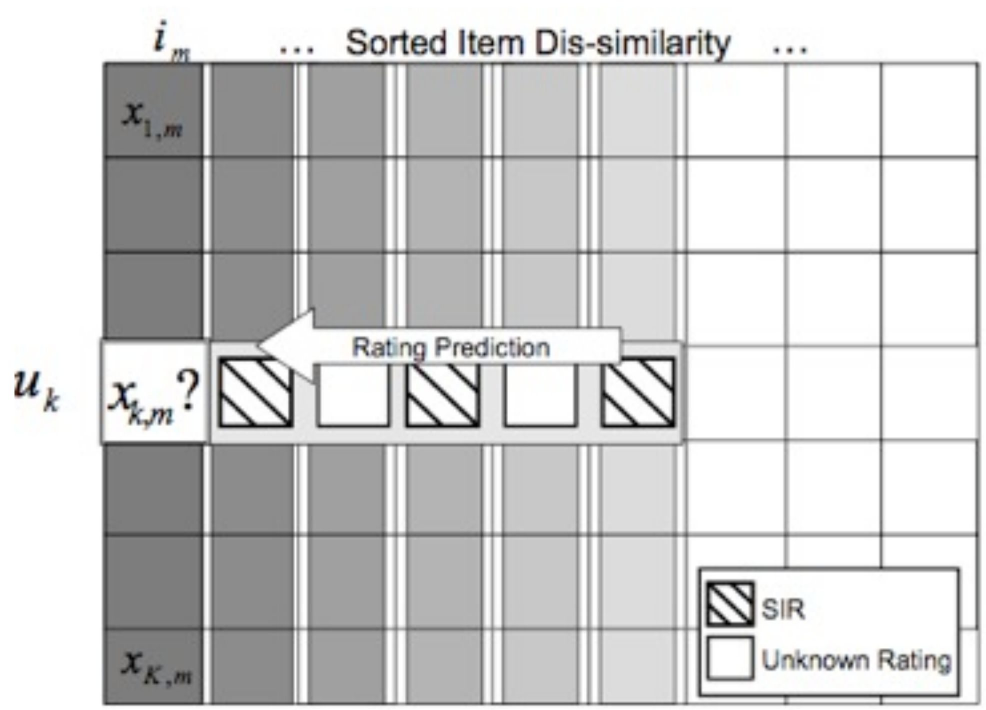
[Wang, 2006]



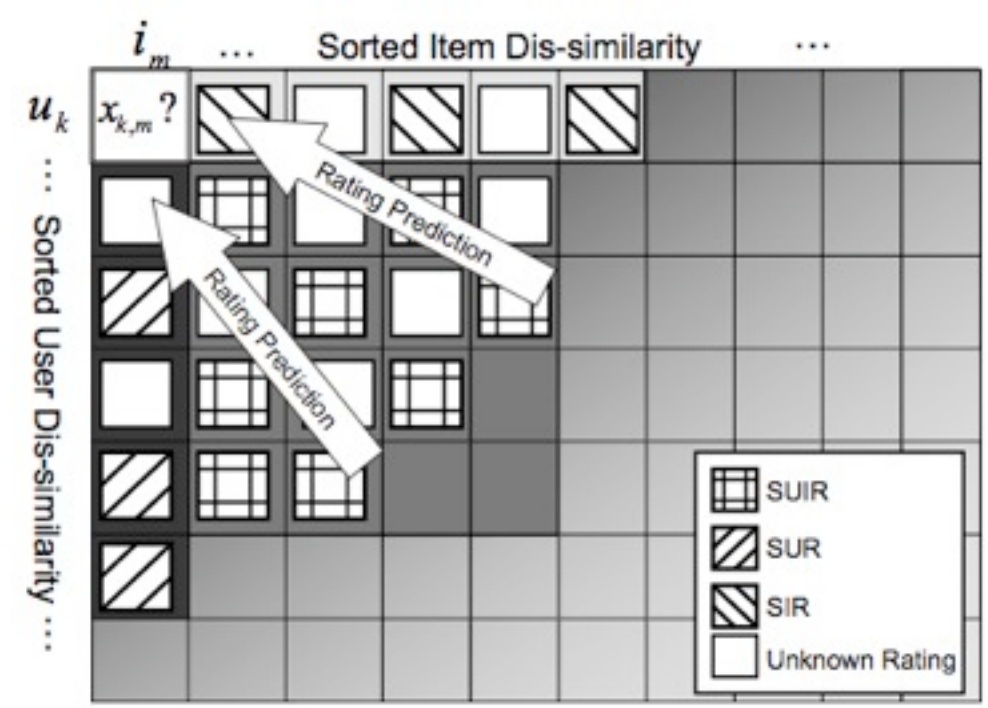
User-item matrix



Rating prediction based on user similarity



Rating prediction based on item similarity



Rating prediction based on rating similarity

Unifying User-based and Item-based CF

[Wang, 2006]

- The final rating is estimated by fusing predictions from three sources:
  - Similar user ratings: $\mathrm{SUR}_{k,m} = \{x_{a,m} | u_a \in S_u(u_k)\}$
    - Predictions based on ratings of the same item by other users
  - Similar item ratings: $\mathrm{SIR}_{k,m} = \{x_{k,b} | i_b \in S_i(i_m)\}$
    - Predictions based on different item ratings made by the same user
  - Similar user item ratings:

    $$\mathrm{SUIR}_{k,m} = \{x_{a,b} | u_a \in S_u(u_k), i_b \in S_i(i_m), a \neq k, b \neq m\}$$

    - Predictions based on similar item ratings made by similar users

Unifying User-based and Item-based CF

[Wang, 2006]

- Unify weight matrix to combine the predictors from three different sources

$$W_{k,m}^{a,b} = \begin{cases} \dfrac{s_{\mathbf{u}}(\mathbf{u}_k, \mathbf{u}_a)}{\sum\limits_{x_{a,b} \in SUR} s_{\mathbf{u}}(\mathbf{u}_k, \mathbf{u}_a)} \lambda (1-\delta) & x_{a,b} \in SUR \\ \dfrac{s_{\mathbf{i}}(\mathbf{i}_m, \mathbf{i}_b)}{\sum\limits_{x_{a,b} \in SIR} s_{\mathbf{i}}(\mathbf{i}_m, \mathbf{i}_b)} (1-\lambda)(1-\delta) & x_{a,b} \in SIR \\ \dfrac{s_{\mathbf{ui}}(x_{k,m}, x_{a,b})}{\sum\limits_{x_{a,b} \in SUIR} s_{\mathbf{ui}}(x_{k,m}, x_{a,b})} \delta & x_{a,b} \in SUIR \\ 0 & otherwise \end{cases}$$

$$\widehat{x}_{k,m} = \sum_{x_{a,b}} p_{k,m}(x_{a,b}) W_{k,m}^{a,b} \qquad \sum_{x_{a,b}} W_{k,m}^{a,b} = 1.$$

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

Items

Users													
u_1													
u_2	1	3	2	1	2	3							
u_3													
u_4						3	1	4	1	2	3	1	2
u_5													
u_6													

Does these two users really have the same taste?

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- We use the following equation to solve this problem:

$$Sim'(a,u) = \frac{Min(|I_a \cap I_u|, \gamma)}{\gamma} \cdot Sim(a,u)$$

  - $|I_a \cap I_u|$ is the number of items which user a and user u rated in common

- Then the similarity between items could be defined as:

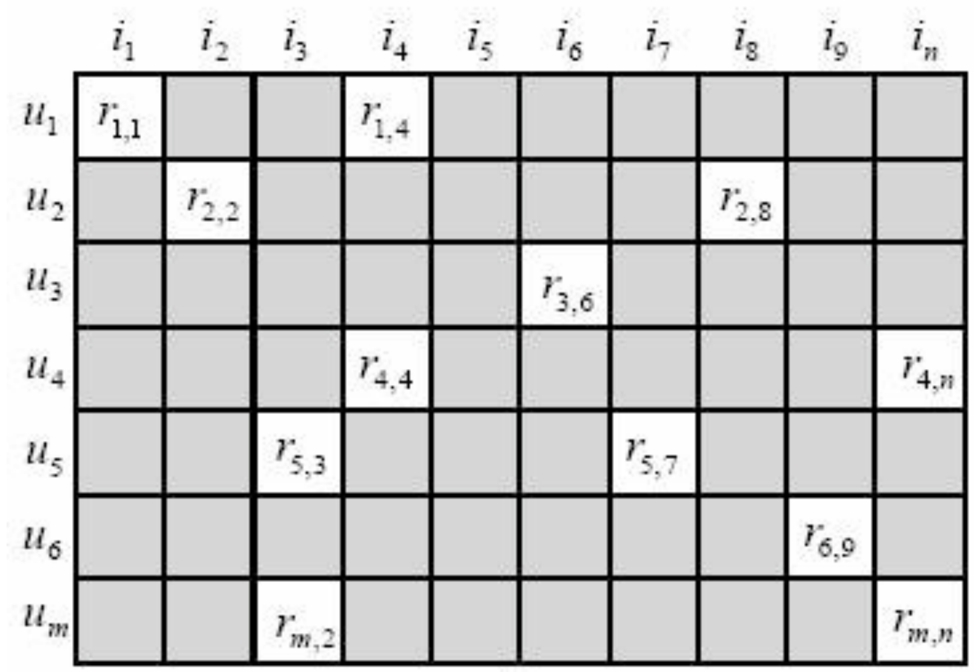
$$Sim'(a,u) = \frac{Min(|I_a \cap I_u|, \gamma)}{\gamma} \cdot Sim(a,u)$$

  - $|U_i \cap U_j|$ is the number of users who rated both item i and item j

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

User-Item Matrix



(a)

- Challenges of collaborative Filtering
  - Data sparsity
  - Prediction accuracy
  - Scalability

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- Data sparsity
    - Propose an algorithm to increase the density of User-Item Matrix
    - Only predict some of the missing data
- Prediction accuracy
    - Adopt significance weighting
    - Linearly combine user information with item information
    - Predict the missing data with high confidence

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_n
u_1	$r_{1,1}$			$r_{1,4}$						
u_2		$r_{2,2}$						$r_{2,8}$		
u_3						$r_{3,6}$				
u_4				$r_{4,4}$						$r_{4,n}$
u_5			$r_{5,3}$				$r_{5,7}$			
u_6									$r_{6,9}$	
u_m			$r_{m,2}$							$r_{m,n}$

(a)

User-Item Matrix

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_n
u_1	$r_{1,1}$	0	$\hat{r}_{1,3}$	$r_{1,4}$	0	$\hat{r}_{1,6}$	0	$\hat{r}_{1,8}$	$\hat{r}_{1,9}$	0
u_2	0	$r_{2,2}$	0	$\hat{r}_{2,4}$	$\hat{r}_{2,5}$	0	$\hat{r}_{2,7}$	$r_{2,8}$	0	$\hat{r}_{2,n}$
u_3	$\hat{r}_{3,1}$	0	$\hat{r}_{3,3}$	$\hat{r}_{3,4}$	$\hat{r}_{3,5}$	$r_{3,6}$	0	$\hat{r}_{3,8}$	$\hat{r}_{3,9}$	0
u_4	$\hat{r}_{4,1}$	$\hat{r}_{4,2}$	0	$r_{4,4}$	$\hat{r}_{4,5}$	$\hat{r}_{4,6}$	$\hat{r}_{4,7}$	0	$\hat{r}_{4,9}$	$r_{4,n}$
u_5	$\hat{r}_{5,1}$	$\hat{r}_{5,2}$	$r_{5,3}$	0	$\hat{r}_{5,5}$	0	$r_{5,7}$	$\hat{r}_{5,8}$	$\hat{r}_{5,9}$	$\hat{r}_{5,n}$
u_6	$\hat{r}_{6,1}$	$\hat{r}_{6,2}$	0	$\hat{r}_{6,4}$	$\hat{r}_{6,5}$	$\hat{r}_{6,6}$	$\hat{r}_{6,7}$	0	$r_{6,9}$	$\hat{r}_{6,n}$
u_m	$\hat{r}_{m,1}$	0	$r_{m,2}$	$\hat{r}_{m,4}$	0	$\hat{r}_{m,6}$	0	$\hat{r}_{m,8}$	$\hat{r}_{m,9}$	$r_{m,n}$

(b)

Predicted User-Item Matrix

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- For every missing data $r_{u,i}$, a set of similar users $S(u)$ towards user u can be generated according to:

$$S(u) = \{u_a | Sim'(u_a, u) > \eta, u_a \neq u\}$$

  - $\mathrm{Sim'}(u_a, u)$ is computed using Significance Weighting

- At the same time, for every missing data $r_{u,i}$, a set of similar items i can be generated according to:

$$S(i) = \{i_k | Sim'(i_k, i) > \theta, i_k \neq i\}$$

  - θ is the item similarity threshold

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- Given the missing data $\mathrm{r}_{u,i}$, if $\mathrm{S(u)} \neq \emptyset \wedge S(i) \neq \emptyset$ the prediction of missing data $\mathrm{P}(\mathrm{r}_{u,i})$ is define as:
- $$P(r_{u,i}) = \lambda \times (\overline{u} + \frac{\sum_{u_a \in S(u)} Sim'(u_a, u) \cdot (r_{u_a,i} - \overline{u}_a)}{\sum_{u_a \in S(u)} Sim'(u_a, u)}) + (1-\lambda) \times (\overline{i} + \frac{\sum_{i_k \in S(i)} Sim'(i_k, i) \cdot (r_{u,i_k} - \overline{i}_k)}{\sum_{i_k \in S(i)} Sim'(i_k, i)})$$

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- If $\mathrm{S(u)} \neq \emptyset \wedge S(i) = \emptyset$, prediction of missing data $\mathrm{P(r}_{u,i})$ is defined as:

$$P(r_{u,i}) = \overline{u} + \frac{\sum_{u_a \in S(u)} Sim'(u_a, u) \cdot (r_{u_a,i} - \overline{u}_a)}{\sum_{u_a \in S(u)} Sim'(u_a, u)}$$

- If $\mathrm{S(u)} = \emptyset \wedge S(i) \neq \emptyset$, prediction of missing data $\mathrm{P(r}_{u,i})$ is define as:

$$P(r_{u,i}) = \overline{i} + \frac{\sum_{i_k \in S(i)} Sim'(i_k, i) \cdot (r_{u,i_k} - \overline{i}_k)}{\sum_{i_k \in S(i)} Sim'(i_k, i)}$$

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- If $\mathrm{S(u)} \neq \emptyset \wedge S(i) = \emptyset$, predication of missing data $\mathrm{P(r}_{u,i})$ is defined as:

$$P(r_{u,i}) = 0$$

- This consideration is different from all other existing predication or smoothing methods -- they always try to predict all the missing data in the user-item matrix, which will predict some missing data with bad quality

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

- Table: Mean Absolute Error (MAE) comparison with other approaches (A smaller MAE value means a better performance)

Training Users	Methods	Given5	Given10	Given20
MovieLens 300	EMDP	**0.784**	**0.765**	**0.755**
	UPCC	0.838	0.814	0.802
	IPCC	0.870	0.838	0.813
MovieLens 200	EMDP	**0.796**	**0.770**	**0.761**
	UPCC	0.843	0.822	0.807
	IPCC	0.855	0.834	0.812
MovieLens 100	EMDP	**0.811**	**0.778**	**0.769**
	UPCC	0.876	0.847	0.811
	IPCC	0.890	0.850	0.824

Dataset: 100,000 ratings (1-5 scales) rated by 943 users on 1,682 movies

Effective Missing Data Prediction for CF

[Hao Ma, 2007]

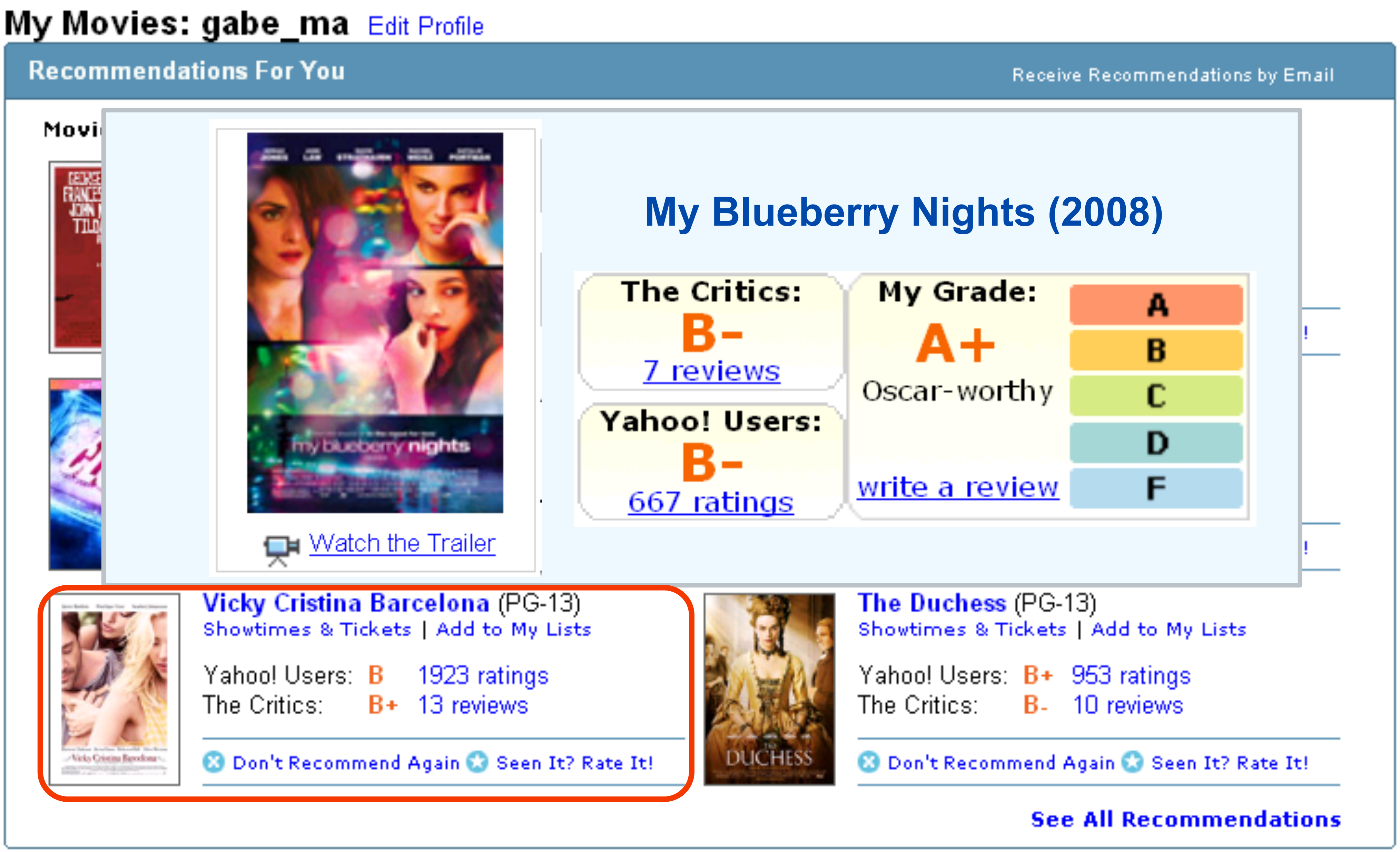
- Table: MAE comparison with state-of-the arts algorithms (A smaller MAE value means a better performance)

Num. of Training Users	100			200			300		
Ratings Given	5	10	20	5	10	20	5	10	20
EMDP	**0.807**	**0.769**	**0.765**	**0.793**	**0.760**	**0.751**	**0.788**	**0.754**	**0.746**
SF	0.847	0.774	0.792	0.827	0.773	0.783	0.804	0.761	0.769
SCBPCC	0.848	0.819	0.789	0.831	0.813	0.784	0.822	0.810	0.778
AM	0.963	0.922	0.887	0.849	0.837	0.815	0.820	0.822	0.796
PD	0.849	0.817	0.808	0.836	0.815	0.792	0.827	0.815	0.789
PCC	0.874	0.836	0.818	0.859	0.829	0.813	0.849	0.841	0.820

SoRec: Social Recommendation

[Hao Ma, 2008]

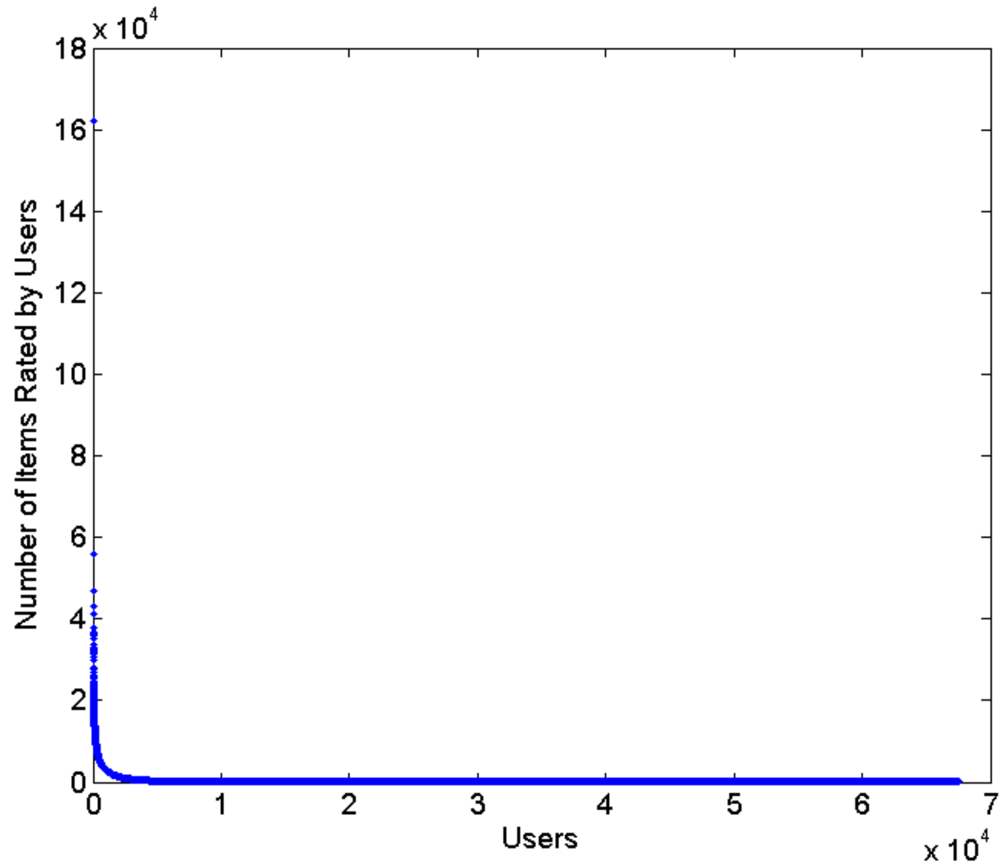
- Challenges: Data Sparsity problem

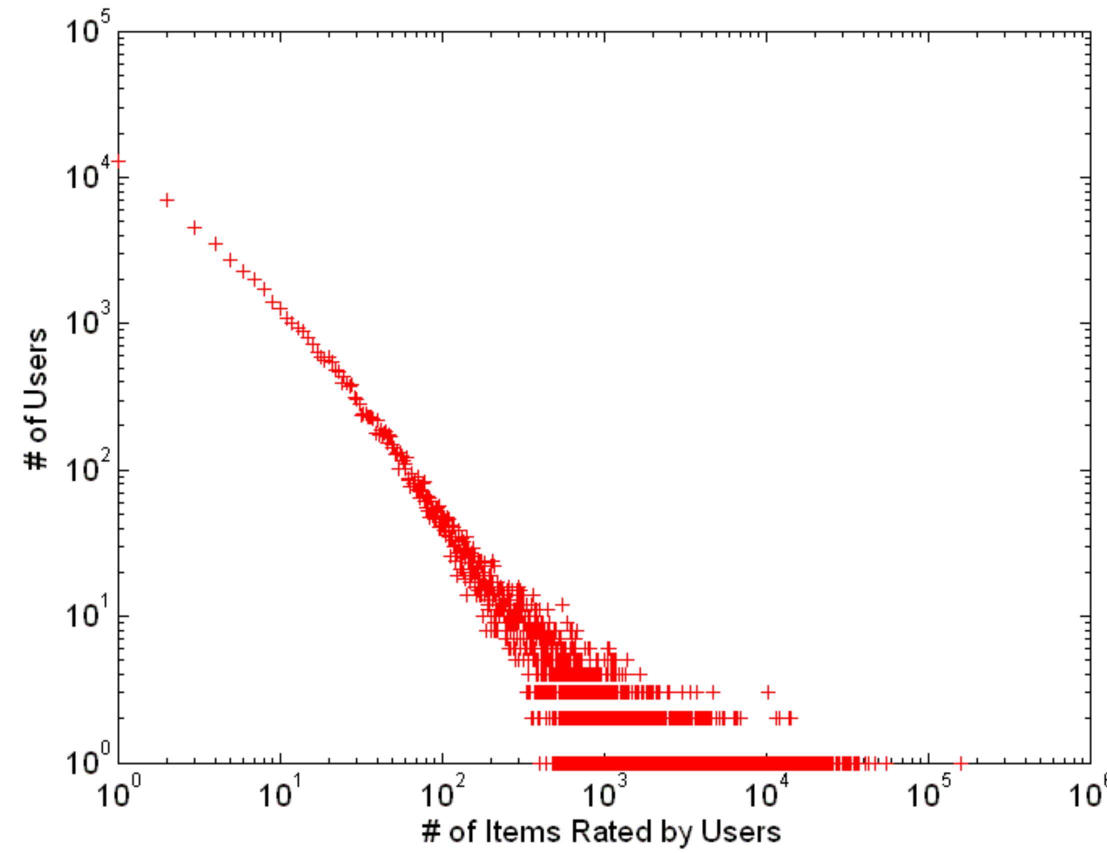


SoRec: Social Recommendation

[Hao Ma, 2008]

- Challenge: Number of rating per user





Extracted From Epinions.com

114,222 users, 754,987 items and 13,385,713 ratings

SoRec: Social Recommendation

[Hao Ma, 2008]

- Challenges: Traditional recommender systems ignore the social connections between users

Recommendations from friends

SoRec: Social Recommendation

[Hao Ma, 2008]

- "Yes, there is a correlation - from social networks to personal behavior on the web"

  Parag Singla and Matthew Richardson (WWW'08)

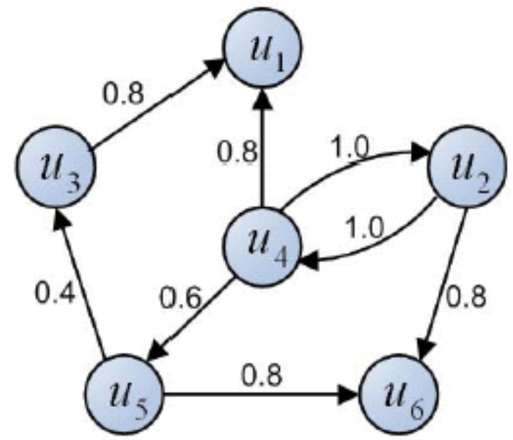
  - Analyze the who talks to whom social network over 10 million people with their related search results
  - People who chat with each other are more likely to share the same or similar interests

- To improve the recommendation accuracy and solve the data sparsity problem, users' social network should be taken into consideration

SoRec: Social Recommendation

[Hao Ma, 2008]

- Problem definition



(a) Social Network Graph

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	5	2		3		4		
u_2	4	3			5			
u_3	4		2				2	4
u_4								
u_5	5	1	2		4	3		
u_6	4	3		2	4		3	5

(b) User-Item Matrix

$$U = \begin{bmatrix} 1.55 & 1.22 & 0.37 & 0.81 & 0.62 & -0.01 \\ 0.36 & 0.91 & 1.21 & 0.39 & 1.10 & 0.25 \\ 0.59 & 0.20 & 0.14 & 0.83 & 0.27 & 1.51 \\ 0.39 & 1.33 & -0.43 & 0.70 & -0.90 & 0.68 \\ 1.05 & 0.11 & 0.17 & 1.18 & 1.81 & 0.40 \end{bmatrix},$$

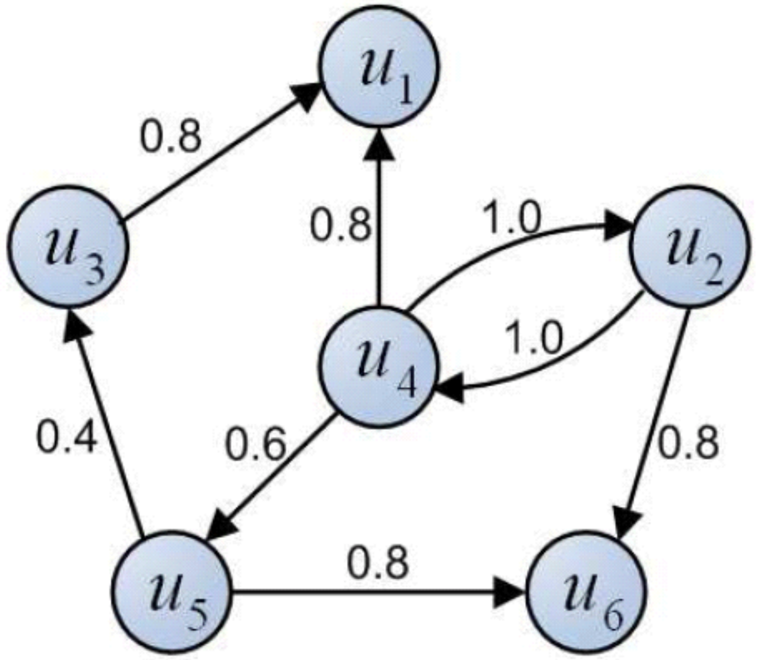
$$V = \begin{bmatrix} 1.00 & -0.05 & -0.24 & 0.26 & 1.28 & 0.54 & -0.31 & 0.52 \\ 0.19 & -0.86 & -0.72 & 0.05 & 0.68 & 0.02 & -0.61 & 0.70 \\ 0.49 & 0.09 & -0.05 & -0.62 & 0.12 & 0.08 & 0.02 & 1.60 \\ -0.40 & 0.70 & 0.27 & -0.27 & 0.99 & 0.44 & 0.39 & 0.74 \\ 1.49 & -1.00 & 0.06 & 0.05 & 0.23 & 0.01 & -0.36 & 0.80 \end{bmatrix},$$

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	5	2	**2.5**	3	**4.8**	4	**2.2**	**4.8**
u_2	4	3	**2.4**	**2.9**	5	**4.1**	**2.6**	**4.7**
u_3	4	**1.7**	2	**3.2**	**3.9**	**3.0**	2	4
u_4	**4.8**	**2.1**	**2.7**	**2.6**	**4.7**	**3.8**	**2.4**	**4.9**
u_5	5	1	2	**3.4**	4	3	**1.5**	**4.6**
u_6	4	3	**2.9**	2	4	**3.4**	3	5

SoRec: Social Recommendation

[Hao Ma, 2008]

- Social network graph matrix factorization



(a) Social Network Graph

$$p(C|U, Z, \sigma_C^2) = \prod_{i=1}^{m} \prod_{k=1}^{m} \mathcal{N}\left[\left(c_{ik}|g(U_i^T Z_k), \sigma_C^2\right)\right]^{I_{ik}^C}$$

SoRec: Social Recommendation

[Hao Ma, 2008]

- User-item rating matrix factorization

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	5	2		3		4		
u_2	4	3			5			
u_3	4		2				2	4
u_4								
u_5	5	1	2		4	3		
u_6	4	3		2	4		3	5

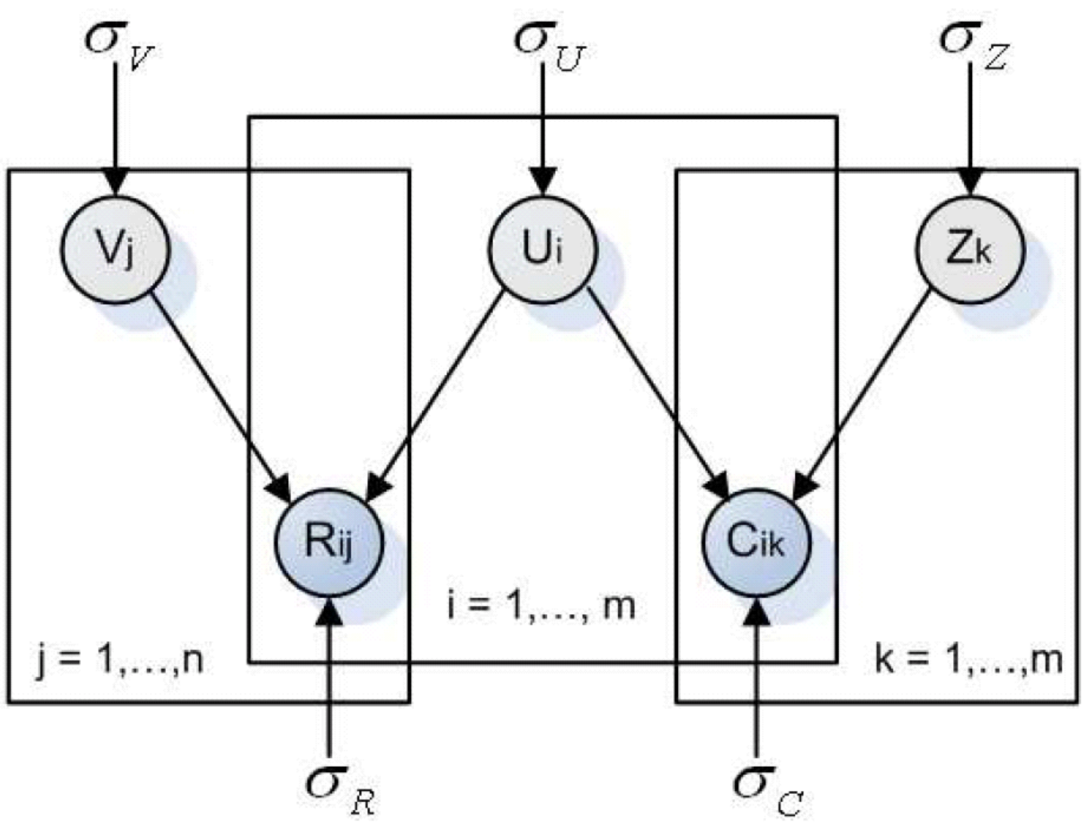
(b) User-Item Matrix

$$p(C|U,V,\sigma_R^2) = \prod_{i=1}^{m}\prod_{j=1}^{n} \mathcal{N}\left[\left(r_{ij}|g(U_i^T V_j),\sigma_R^2\right)\right]^{I_{ij}^R}$$

SoRec: Social Recommendation

[Hao Ma, 2008]

- Social recommendation



$$\mathcal{L}(R, C, U, V, Z) =$$
$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}(r_{ij} - g(U_i^T V_j))^2 + \frac{\lambda_C}{2}\sum_{i=1}^{m}\sum_{k=1}^{m} I_{ik}^{C}(c_{ik}^{*} - g(U_i^T Z_k))^2$$
$$+ \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2 + \frac{\lambda_Z}{2}\|Z\|_F^2,$$

SoRec: Social Recommendation

[Hao Ma, 2008]

- Gradient descent

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial U_i} &= \sum_{j=1}^{n} I_{ij}^{R} g'(U_i^T V_j)(g(U_i^T V_j) - r_{ij})V_j \\
&+ \lambda_C \sum_{j=1}^{m} I_{ik}^{C} g'(U_i^T Z_k)(g(U_i^T Z_k) - c_{ik}^{*})Z_k + \lambda_U U_i, \\
\frac{\partial \mathcal{L}}{\partial V_j} &= \sum_{i=1}^{m} I_{ij}^{R} g'(U_i^T V_j)(g(U_i^T V_j) - r_{ij})U_i + \lambda_V V_j, \\
\frac{\partial \mathcal{L}}{\partial Z_k} &= \lambda_C \sum_{i=1}^{m} I_{ik}^{C} g'(U_i^T Z_k)(g(U_i^T Z_k) - c_{ik}^{*})U_i + \lambda_Z Z_k,
\end{aligned}
$$

SoRec: Social Recommendation

[Hao Ma, 2008]

- Table: MAE comparison with other approaches (A smaller MAE value means a better performance)

Training Data	Dimensionality = 5				Dimensionality = 10			
	MMMF	PMF	CPMF	SoRec	MMMF	PMF	CPMF	SoRec
99%	1.0008	0.9971	0.9842	**0.9018**	0.9916	0.9885	0.9746	**0.8932**
80%	1.0371	1.0277	0.9998	**0.9321**	1.0275	1.0182	0.9923	**0.9240**
50%	1.1147	1.0972	1.0747	**0.9838**	1.1012	1.0857	1.0632	**0.9751**
20%	1.2532	1.2397	1.1981	**1.1069**	1.2413	1.2276	1.1864	**1.0944**

MMMF:
J.D.M Rennie and N. Srebro (ICML'05)

PMF & CPMF:
R. Salakhutdinov and A. Mnih (NIPS'08)

Epinions: 40,163 users who rated 139,529 items with totally 664,824 ratings

References

- https://agora.cs.illinois.edu/display/cs512/home.
- J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In ICML, 2004.
- J. S. Breese, D. Heckerman, and C. M. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In UAI, pages 43–52, 1998.
- M. Deshpande and G. Karypis. Item-based top- recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143–177, 2004.
- J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In SIGIR, pages 230–237. ACM, 1999.
- J. L. Herlocker, J. A. Konstan, and J. Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Inf. Retr., 5(4):287–310, 2002.

References

- J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. Commun. ACM, 40(3):77–87, 1997.
- G. Linden, B. Smith, and J. York. Industry report: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Distributed Systems Online, 4(1), 2003.
- H. Ma, I. King, and M. R. Lyu. Effective missing data prediction for collaborative filtering. In SIGIR, pages 39–46, 2007.
- H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In CIKM, pages 931–940, 2008.
- B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In WWW, pages 285–295, 2001.
- J. Wang, A. P. de Vries, and M. J. T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In SIGIR, 2006.

Human Computation

Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

Motivation

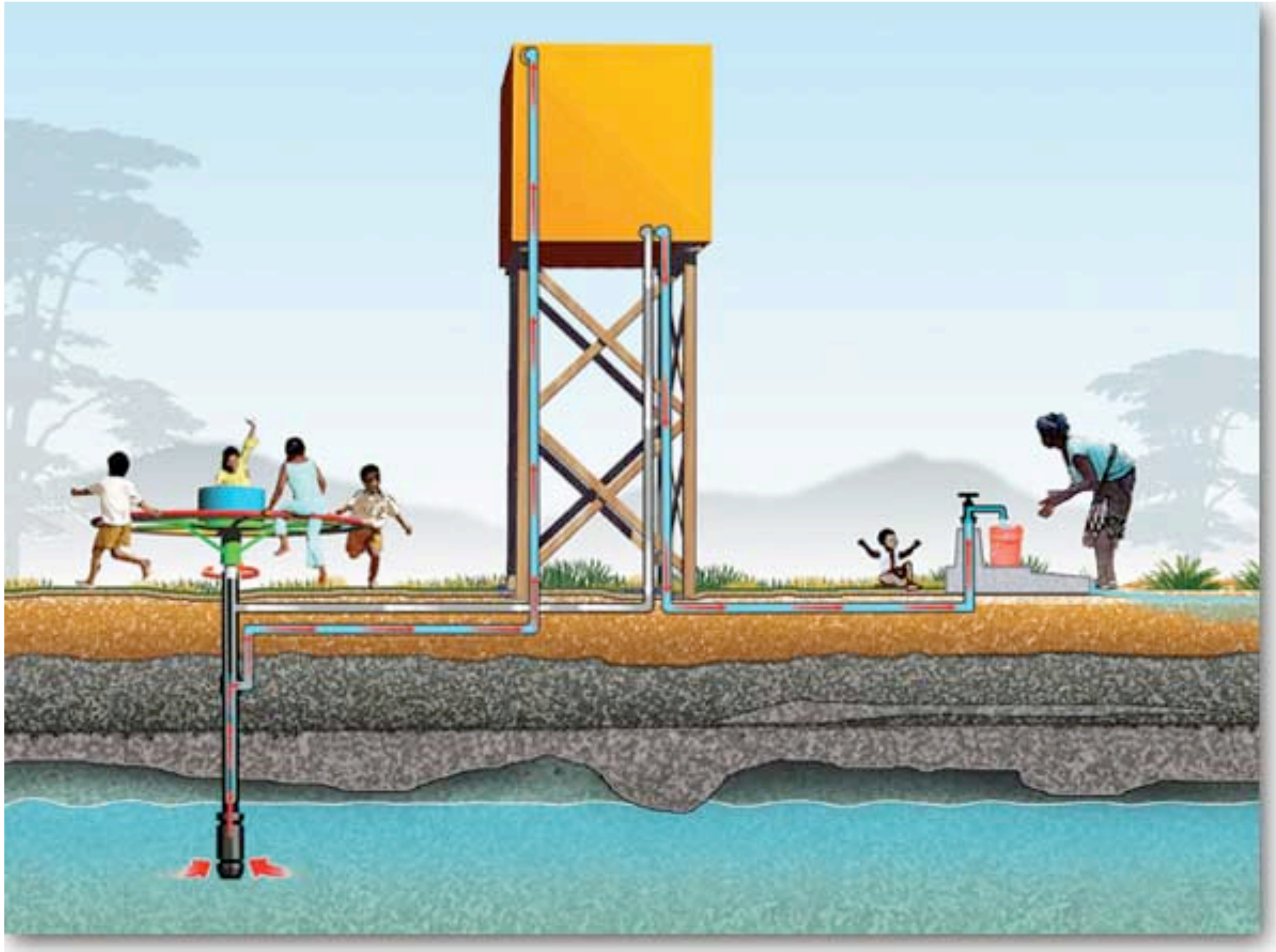
- Many tasks are trivial for human, but continue to challenge even the most sophisticated computer

What's this? Apple is a kind of ??? sad? happy?

- People spend a lot of time playing games

Having Fun = Work

Idea of Human Computation



- Take advantage of people's desire to be entertained and perform useful tasks as a side effect

Image Labeling

- Determine the contents of images by providing meaningful labels for them

The ESP Game

[Luis von Ahn, 2004]

Use computational power of humans to label images.

The ESP Game

[Luis von Ahn, 2004]

Player 1

Guessing: Car
Guessing: Hat
Guessing: Kid

Success!
You agree on car.

Player 2

Guessing: Boy
Guessing: Car

Success!
You agree on car.

The ESP Game

[Luis von Ahn, 2004]

The ESP Game

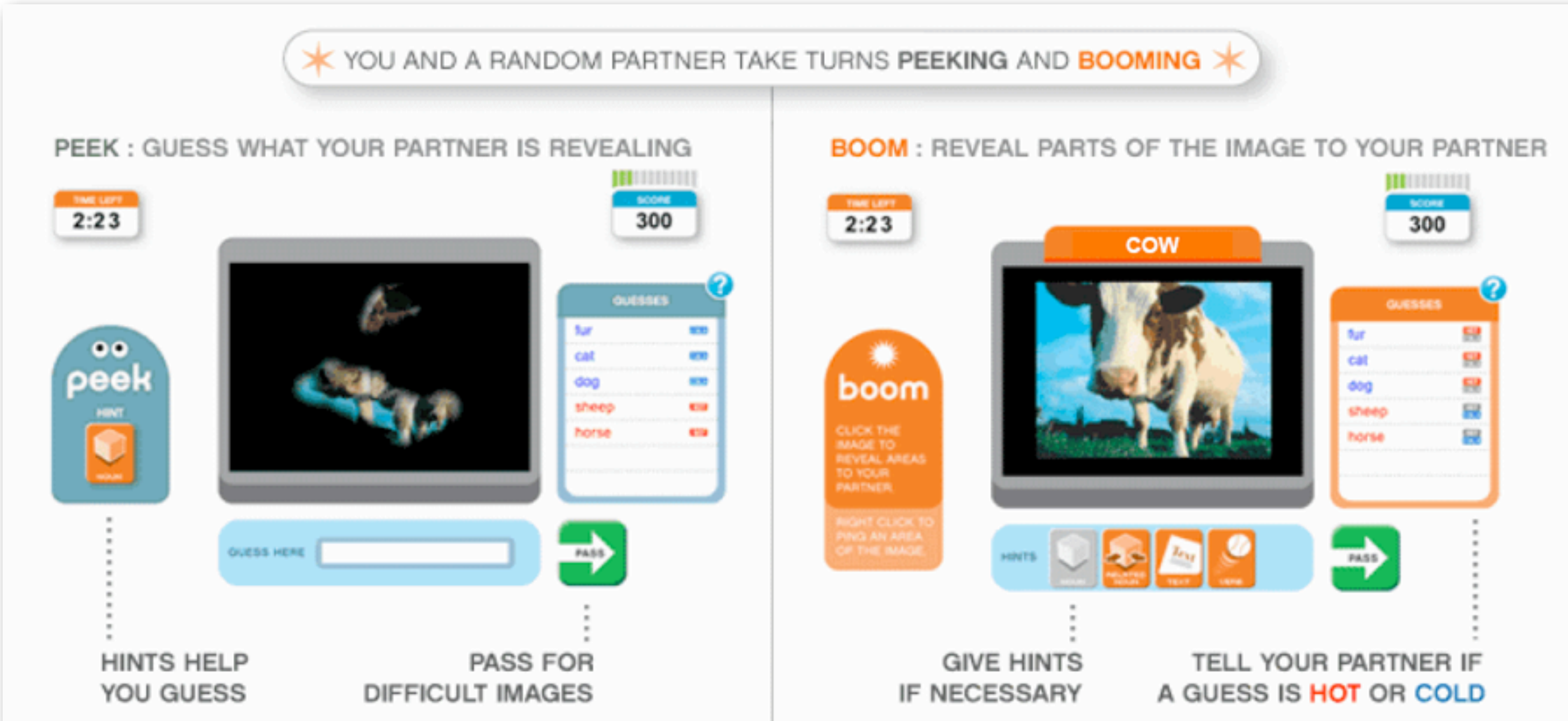
[Luis von Ahn, 2004]

- 5,000 people continuously playing the game could assign a label to all image indexed by Google in 31 days

- The game was posted on website on August 9, 2003. For the first 4 months, a total of 13,630 people played the game, generating 1,271,451 labels for 293,760 different images

The Peekaboom Game

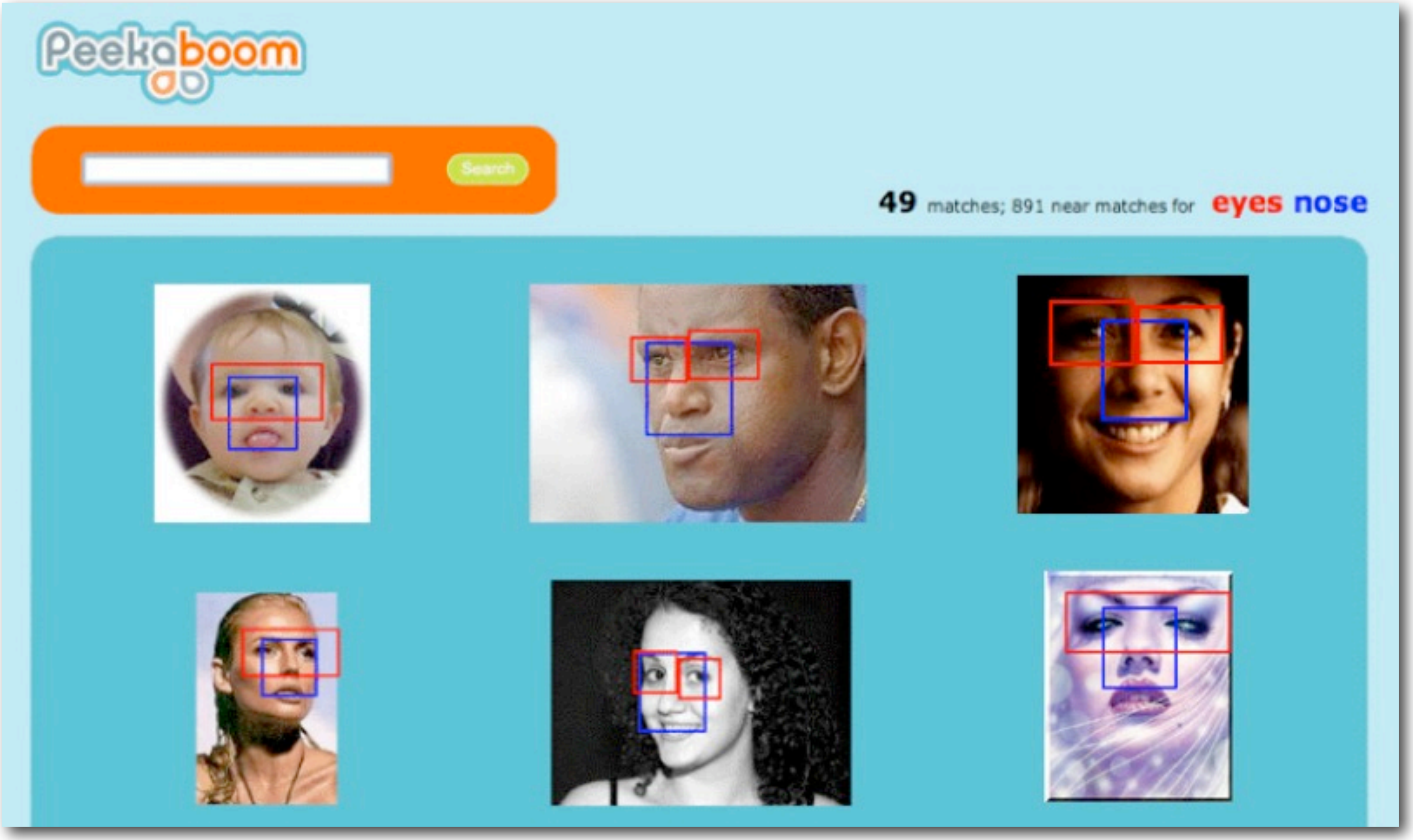
[Luis von Ahn, 2006]



Locating objects in the image

The Peekaboom Game

[Luis von Ahn, 2006]



Object bounding-boxes obtained from Peekaboom data

The Verbosity Game

[Luis von Ahn, 2006]

Collecting common-sense facts

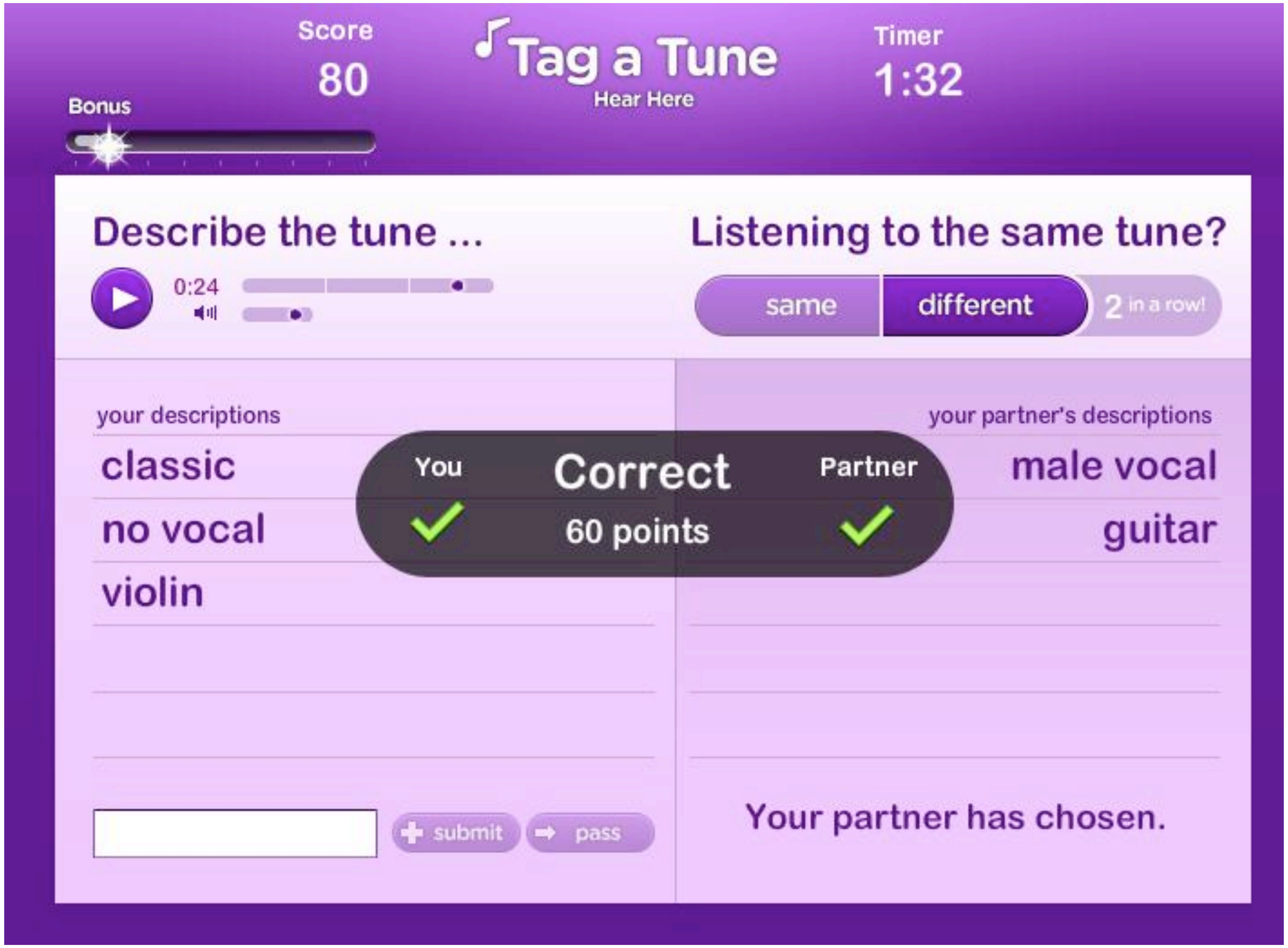
The Verbosity Game

[Luis von Ahn, 2006]

Part of the narrator's screen

The TagATune Game

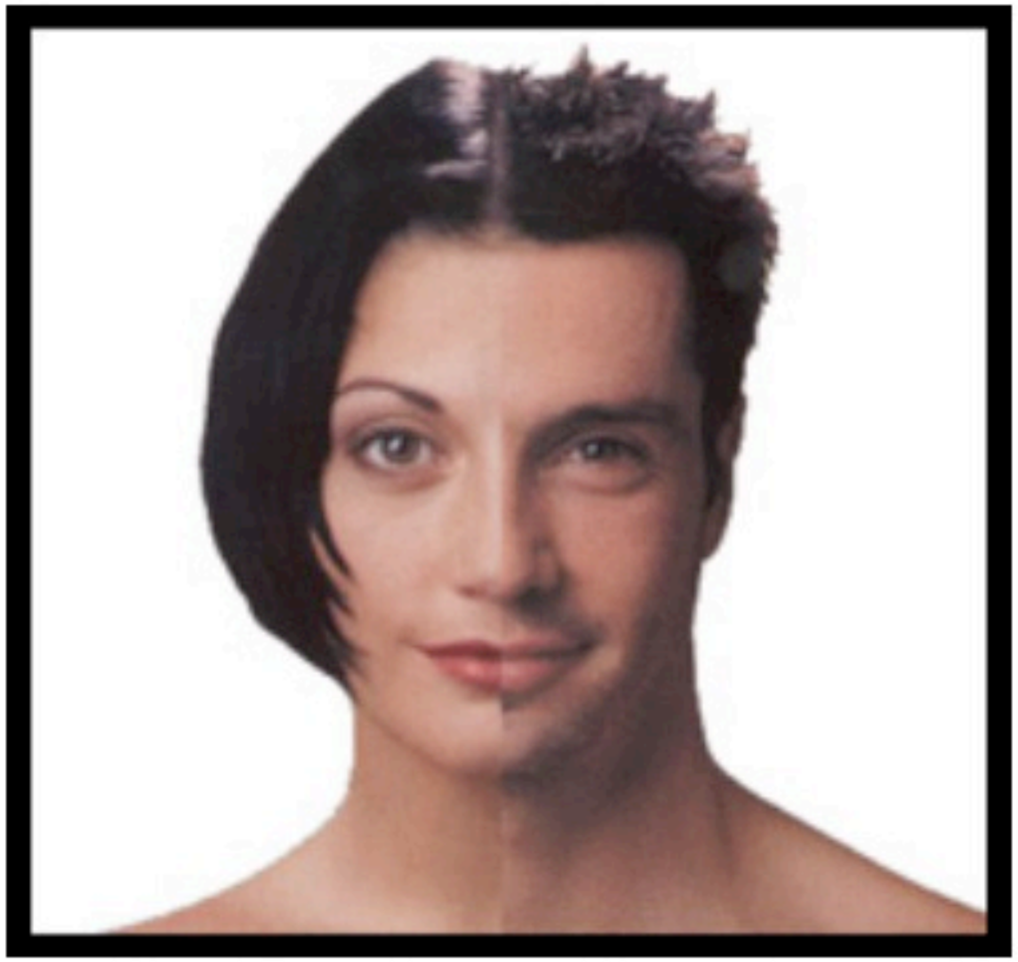
[Edith L.M. Law, 2006]



Annotations of audio files

The Phetch Game

[Luis von Ahn, 2006]



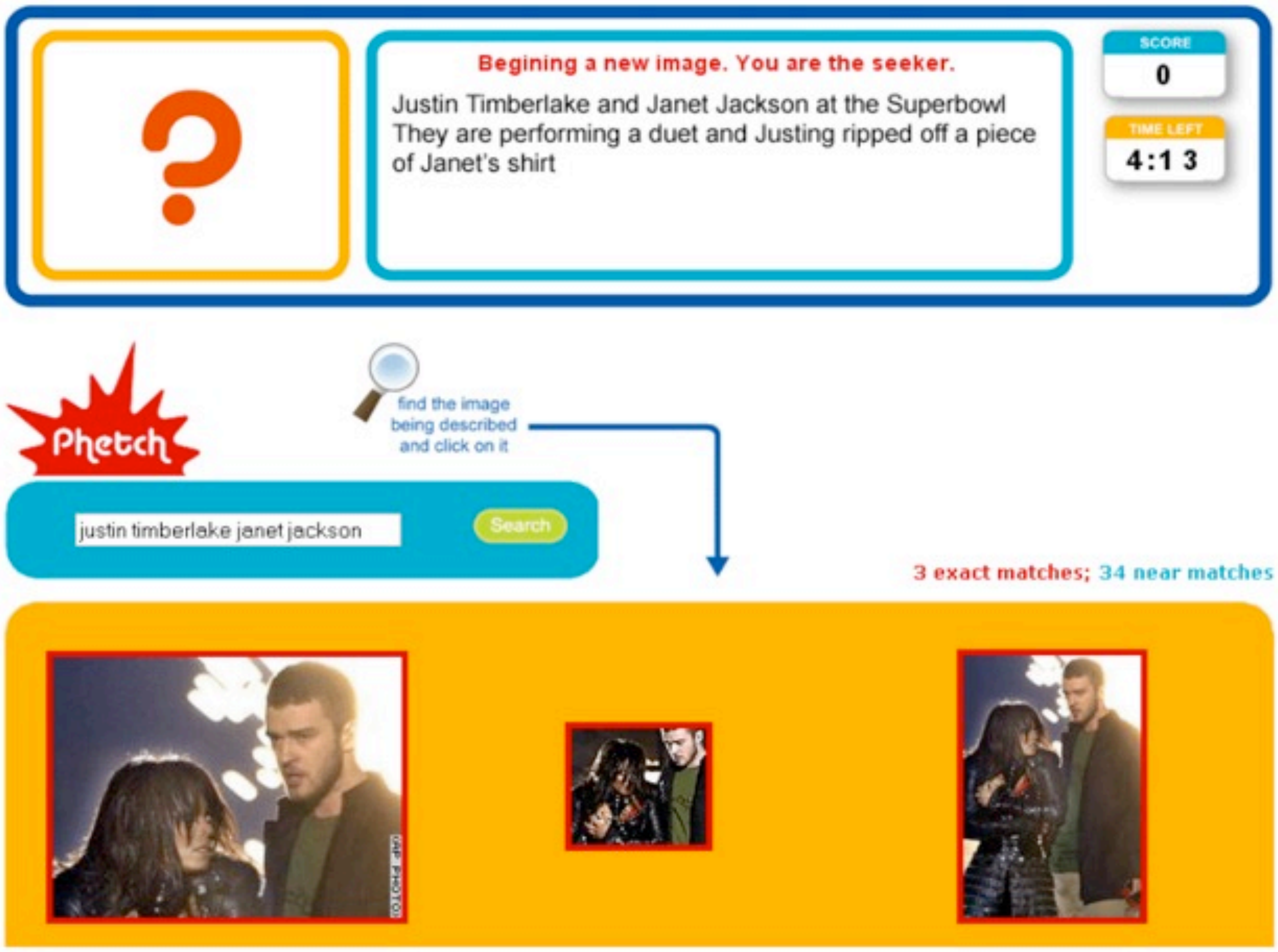
Two inherently different images share the same ESP labels: "man" and "woman"

The Phetch Game

[Luis von Ahn, 2006]

The Phetch Game

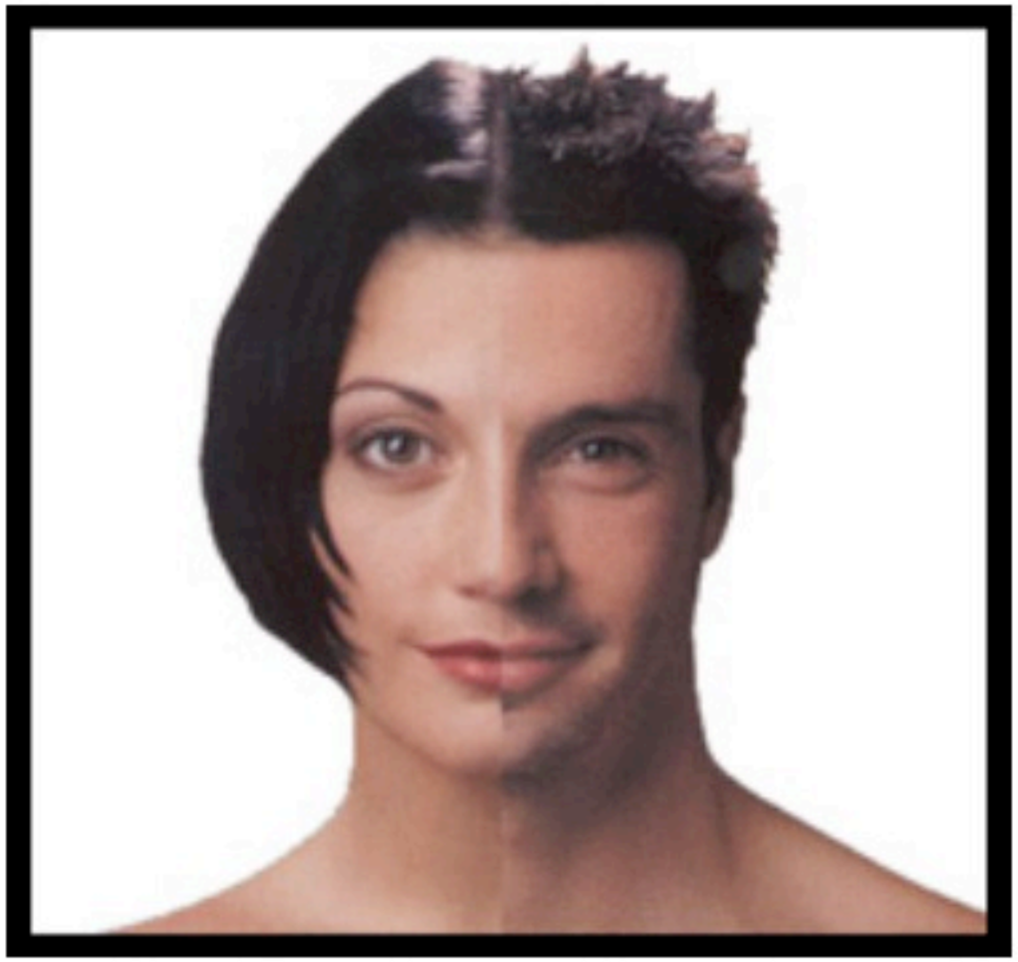
[Luis von Ahn, 2006]



Screen of the seeker's interface

The Phetch Game

[Luis von Ahn, 2006]



The Phetech description are different: "half-man half-woman with black hair" and "an abstract line drawing of a man with a violin and a woman with a flute"

Designing Games with a Purpose

[Luis von Ahn, 2008]

- Social Games or Game with A purpose is an innovative idea that make use of human brain power to solve difficult problems
  - Output-agreement games
  - Inversion-problem games
  - Input-agreement games

Output-Agreement Games

[Luis von Ahn, 2008]

Figure 1: In this output-agreement game, players are given the same input and must agree on an appropriate output.

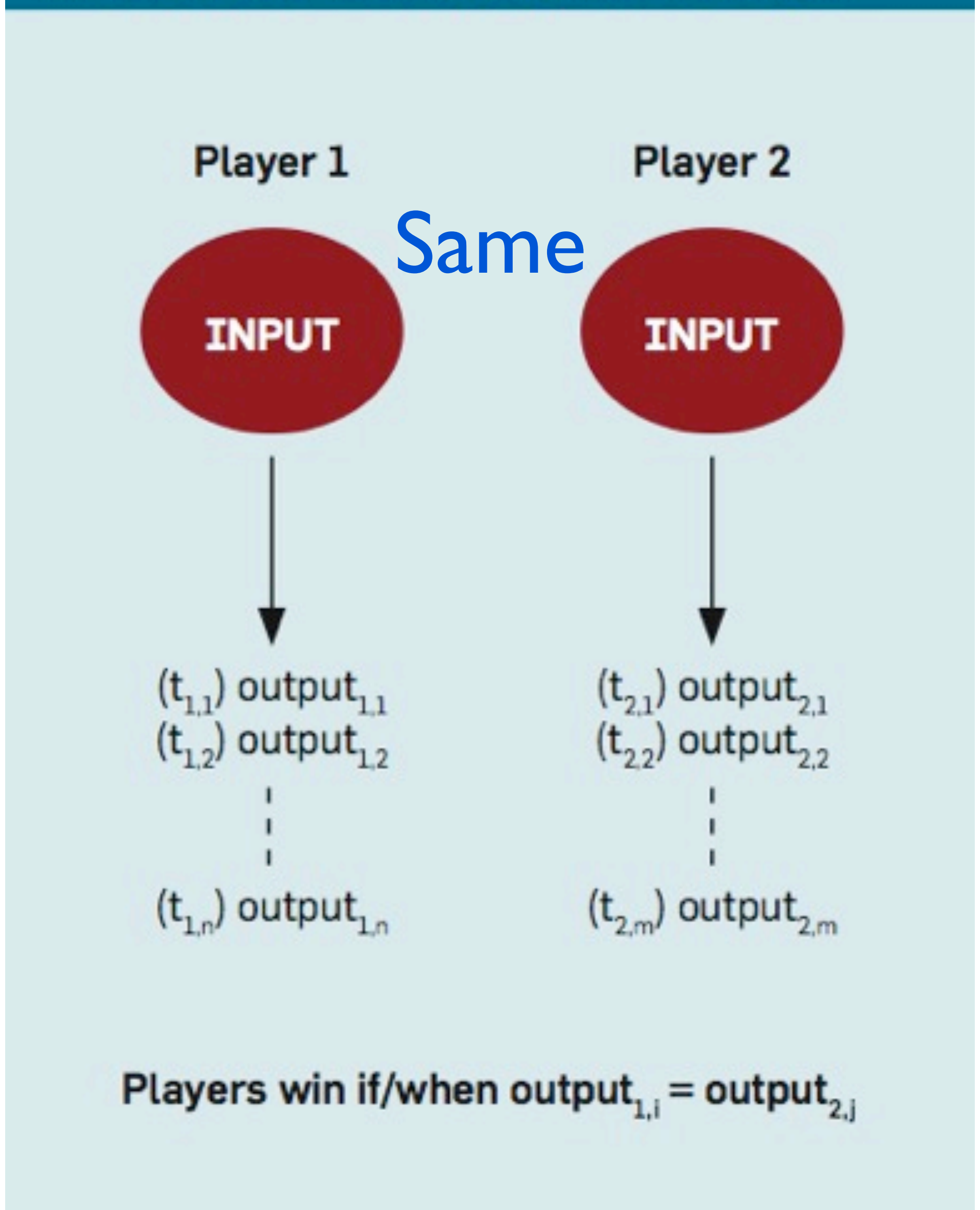
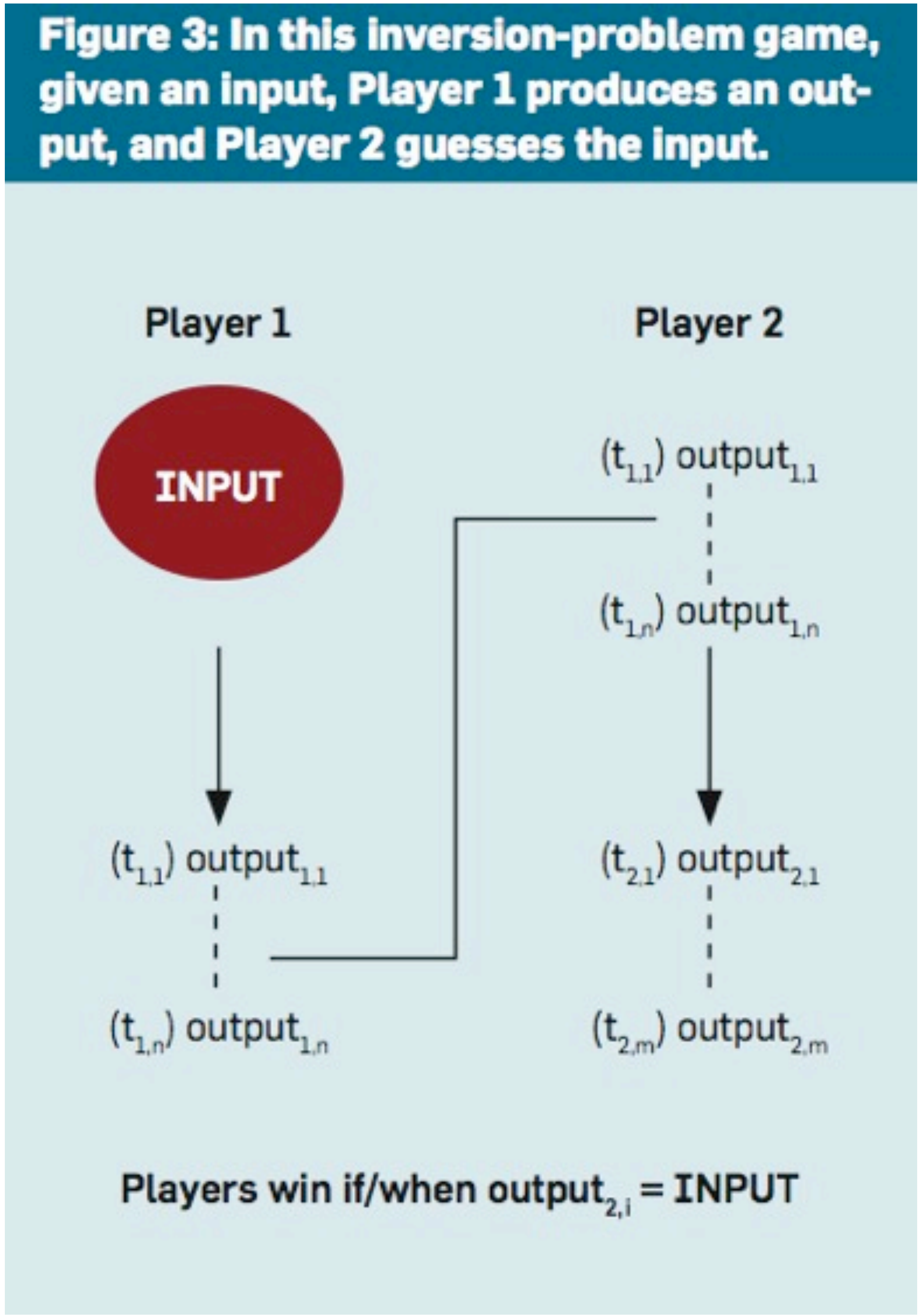


Figure 2: In this output-agreement game, the partners are agreeing on a label.

Inversion-Problem Games

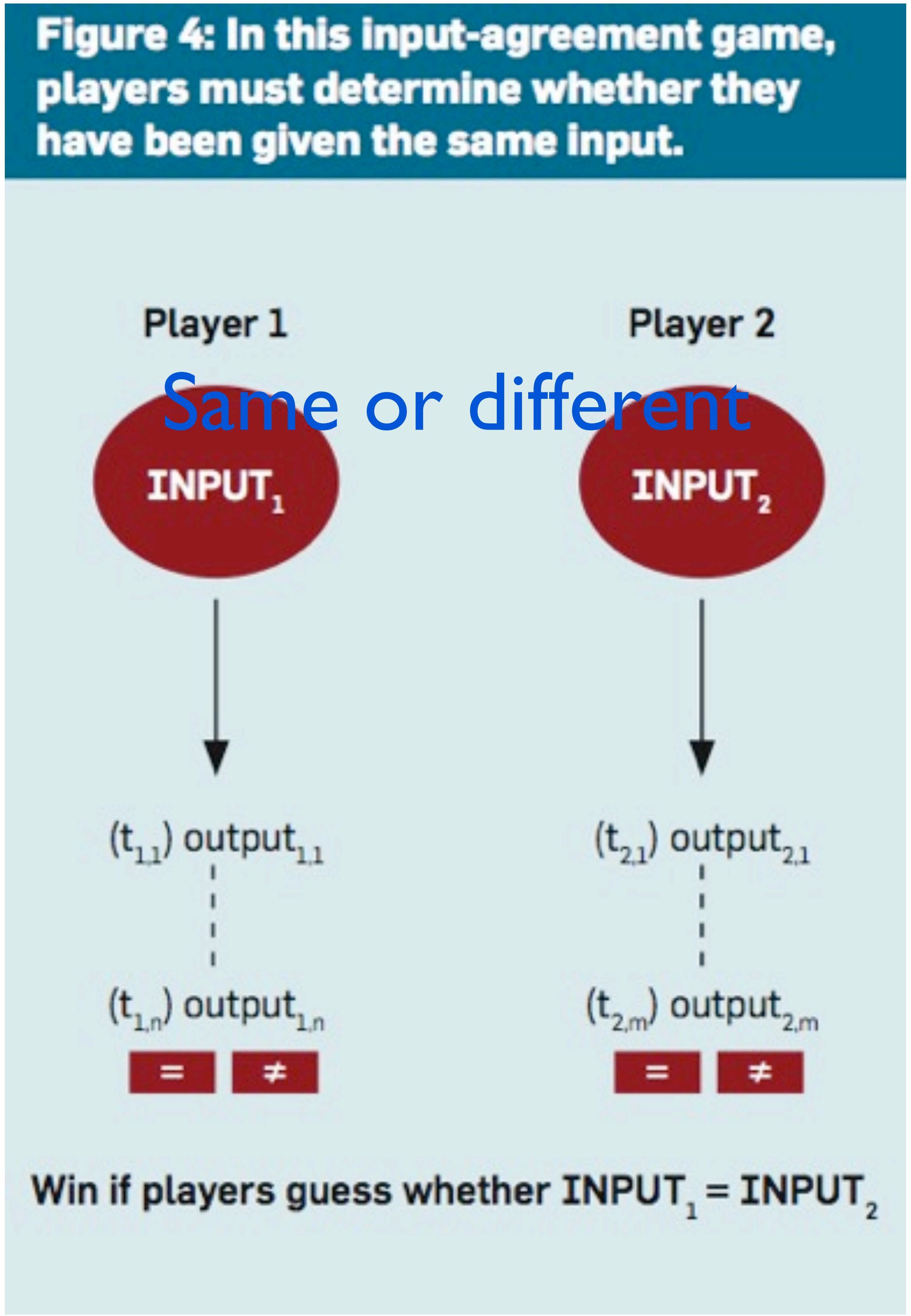
[Luis von Ahn, 2008]



- One "describers"
- Others "guesser"
- The output given by describers should help the guesser produce the original input

Input-Agreement Games

[Luis von Ahn, 2008]



- Both players produce output describing their input, to help their partners to determine whether their inputs are the same or different

Designing Games with a Purpose

[Luis von Ahn, 2008]

	Input-agreement	Inversion-problem	Output-agreement
Initial setup	Two random strangers	Two (or more) random strangers	Two random strangers
Rules	Same input; Both produce output	Describer sees the input and produces output; Guesser(s) searches for input	Same or different input; Both produce output and guess whether input are the same
Winning-condition	Same output	Guesser produce the same input	Both correctly determine

Make Games More Entertaining

[Luis von Ahn, 2008]

- Introduce challenge
  - Timed response, score keeping, player skill level, high score lists, and randomness
- Introduce competition
- Introduce variation
- Introduce communication

Ensue Output Accuracy

[Luis von Ahn, 2008]

- Random matching
  - Cannot collaborate to cheat
- Player testing
  - Quality of intelligent
- Repetition
  - Probabilistic correct
- Taboo output
  - Eliminate obvious answers, increase diversity

Why is it important?

- Some statistics for ESP game (July 2008)
  - 200,000+ players have contributed 50+ million labels
  - Each player plays for a total of 91 minutes
  - The throughput is about 233 labels/player/hour (i.e., one label every 15 seconds)
- Idea behind
  - Solve some problems which are difficult to be solved by computers
  - Take advantage of people's desire to be entertained
  - Produce useful metadata as a by-product

Modeling Human Computation

- Three challenging issues to consider
  - Game integrity issues
    - How do we make the game do what we want to do?
  - Quality assurance issues
    - How do we know the results are correct and useful?
  - Game design issues
    - How do we make the system interesting to play?

Summary

- Human computation opens a new frontier!
- Further exploration of human cognitive abilities
- Theoretical modeling and analysis of social gaming
- Software platforms to support quick prototyping

References

- R. B. D. M. C. Edith L. M. Law, Luis von Ahn. Tagatune: A game for music and sound annotation. ISMIP, 2007.
- L. von Ahn. Games with a purpose. IEEE Computer, 39(6):92–94, 2006.
- L. von Ahn and L. Dabbish. Labeling images with a computer game. In CHI, pages 319–326, 2004.
- L. von Ahn and L. Dabbish. Designing games with a purpose. Commun. ACM, 51(8):58–67, 2008.
- L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In CHI, pages 79–82, 2006.
- L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In CHI, pages 75–78, 2006.
- L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In CHI ’06, pages 55–64, New York, NY, USA, 2006. ACM.

Privacy and Trust in Social Network

Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

Outline

- What is privacy and trust?
- Privacy in social network
  - Basic privacy requirement
  - Privacy in graph
- Trust in social network
- Reference

What is Privacy

- Privacy is the ability of an individual or group to seclude themselves or information about themselves and thereby reveal themselves selectively.
  - Different privacy boundaries and content
  - Voluntarily sacrificed
  - Uniquely identifiable data relating to a person or persons

What is Trust

- Trust is a relationship of reliance
  - Not related to good character or morals
  - Trust does not need to include an action that you and the other party are mutually engaged in
  - Trust is a prediction of reliance on an action
  - Conditional

Privacy and Trust Tradeoff

- Privacy
- Need legal rights
- Reveal more data to trustworthy people

- Trust
- Provide access rights
- Gain trust through open sensitive data

Outline

- What is privacy and trust?
- Privacy in social network
  - Basic privacy requirement
  - Privacy in graph
- Trust in social network
- Reference

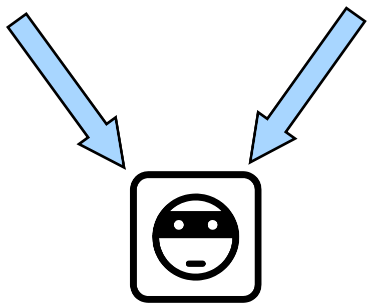
Motivation

Published table

Age	Zip.	Salary
17	12k	1000
19	13k	1010
20	14k	1020
24	16k	50000
29	21k	16000
34	24k	24000
39	36k	33000
45	39k	31000

Voter registration list

Name	Age	Zip.
Andy	17	12k
Bill	19	13k
Ken	20	14k
Jane	23	15k
Nash	24	16k
Joe	29	21k
Sam	34	24k
Linda	39	36k
Mary	45	39k



An adversary

Fact: 87% of Americans can be uniquely identified by {Zipcode, gender, date-of-birth}.

k-anonymity

[Sweeney, 2001]

	Age	Zip.	Salary
Andy	17	12k	1000
	19	13k	1010
	20	14k	1020
	24	16k	50000
	29	21k	16000
	34	24k	24000
	39	36k	33000
	45	39k	31000

(a) The microdata

Group ID	Age	Zip.	Salary
1	[17,24]	[12k,16k]	1000
1	[17,24]	[12k,16k]	1010
1	[17,24]	[12k,16k]	1020
1	[17,24]	[12k,16k]	50000
2	[29,34]	[21k,24k]	16000
2	[29,34]	[21k,24k]	24000
3	[39,45]	[36k,39k]	33000
3	[39,45]	[36k,39k]	31000

(b) Generalization

A group

Not sure about the salary of Andy now!

- *k*-anonymity
  - Divide tuples into groups
  - Each group has at least *k* tuples

Problem with *k*-anonymity

[Machanavajjhala, 2001]

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

A 4-anonymous table

What about we know a person's Zip Code = 13053 and Age = 31?

In this case, we can conclude his/her disease is Cancer.

l-diversity

[Machanavajjhala, 2001]

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

A 3-diverse table

- *l*-diversity
  - Divide tuples into groups
  - Each group has at least *l* different sensitive values

(*k*, e)-anonymity

[Zhang, 2007]

	ID	Quasi-identifiers			Sensitive
tuple ID	name	age	zipcode	gender	salary
1	Alex	35	27101	M	$54,000
2	Bob	38	27120	M	$55,000
3	Carl	40	27130	M	$56,000
4	Debra	41	27229	F	$65,000
5	Elain	43	27269	F	$75,000
6	Frank	47	27243	M	$70,000
7	Gary	52	27656	M	$80,000
8	Helen	53	27686	F	$75,000
9	Jason	58	27635	M	$85,000

Microdata

		Quasi-identifiers			Sensitive
group ID	tuple ID	age	zipcode	gender	salary
1	1	[31-40]	271*	*	$56,000
1	2	[31-40]	271*	*	$54,000
1	3	[31-40]	271*	*	$55,000
2	4	[41-50]	272*	*	$65,000
2	5	[41-50]	272*	*	$75,000
2	6	[41-50]	272*	*	$70,000
3	7	[51-60]	276*	*	$80,000
3	8	[51-60]	276*	*	$75,000
3	9	[51-60]	276*	*	$85,000

A 3-diverse table

Though the salary in group 1 is different, we are sure that Alex's salary is around 55,000

- (*k*, e)-anonymity
  - Each group has at least *k* tuples
  - Difference between the maximum and minimum values must be at least e

Outline

- What is privacy and trust?
- Privacy in social network
  - Basic privacy requirement
  - Privacy in graph
- Trust in social network
- Reference

Possible Attacks on Anonymized Graphs

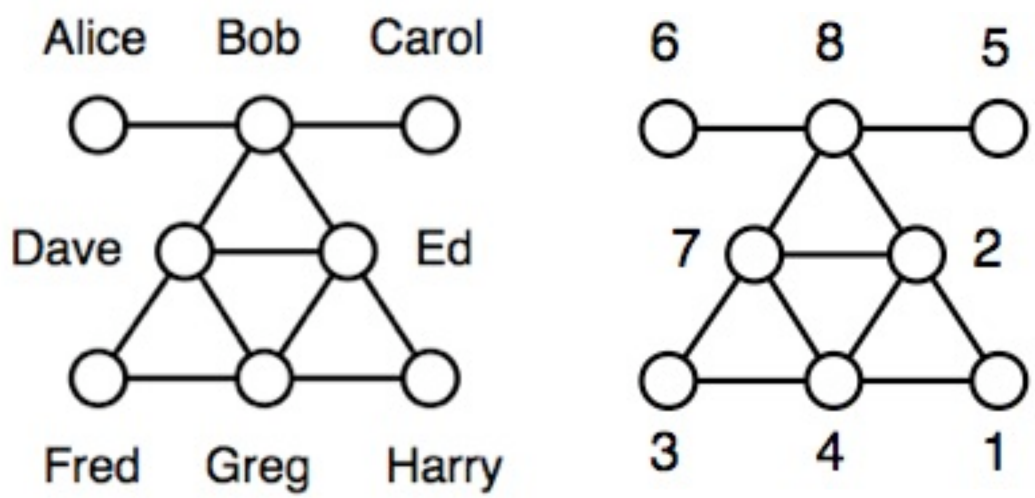
- Attack method [Michael Hay, 2008]
  - Identify by neighborhood information
  - It includes
    - Vertex Refinement Queries
    - Sub-graph Queries
    - Hub Fingerprint Queries

Possible Attacks on Anonymized Graphs

- Attack types [Lars Backstrom, 2008]
  - Active Attacks
    - Create a small number of new user accounts linking with other users before the anonymized graph is generated
  - Passive Attacks
    - Identify themselves in the published graph
  - Semi-passive Attacks
    - Create necessary link with other users

Vertex Refinement Queries

[Michael Hay, 2008]



Alice	6
Bob	8
Carol	5
Dave	7
Ed	2
Fred	3
Greg	4
Harry	1

(a) graph

Node ID	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{H}_2$
Alice	ϵ	1	$\{4\}$
Bob	ϵ	4	$\{1, 1, 4, 4\}$
Carol	ϵ	1	$\{4\}$
Dave	ϵ	4	$\{2, 4, 4, 4\}$
Ed	ϵ	4	$\{2, 4, 4, 4\}$
Fred	ϵ	2	$\{4, 4\}$
Greg	ϵ	4	$\{2, 2, 4, 4\}$
Harry	ϵ	2	$\{4, 4\}$

(b) vertex refinements

Equivalence Relation	Equivalence Classes
$\equiv_{\mathcal{H}_0}$	$\{A, B, C, D, E, F, G, H\}$
$\equiv_{\mathcal{H}_1}$	$\{A, C\}$ $\{B, D, E, G\}$ $\{F, H\}$
$\equiv_{\mathcal{H}_2}$	$\{A, C\}\{B\}\{D, E\}\{G\}\{F, H\}$
$\equiv_{A}$	$\{A, C\}\{B\}\{D, E\}\{G\}\{F, H\}$

(c) equivalence classes

H^*'s computation is linear in the number of edges in the graph!

Summary

- Data privacy and security is a real and serious issue
- *k*-Anonymity and *l*-Diversity could help but may not be watertight
- Anonymizing graphs through graph generalization, node partitioning, and graph summarization

References

- L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002
- Ashwin Machanavajjhala , Daniel Kifer , Johannes Gehrke , Muthuramakrishnan Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, TKDD, 2007
- Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, ICDE, 2007.
- Xiao, X., Tao, Y, Dynamic Anonymization: Accurate Statistical Analysis with Privacy Preservation, SIGMOD, 2008.
- Michael Hay, Gerome Miklau, David Jensen, Don Towsley and Philipp Weis, Resisting Structural Re-identification in Anonymized Social Networks, PVLDB, 2008
- Lars Backstrom, Cynthia Dwork and Jon Kleinberg, Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, WWW, 2007
- Kun liu and Evimaria Terzi, Towards Identity Anonymization on Graphs. SIGMOD, 2008
- Bin Zhou and Jian Pei, Preserving Privacy in Social Networks Against Neighborhood Attacks, ICDE, 2008

Social Computing in Education

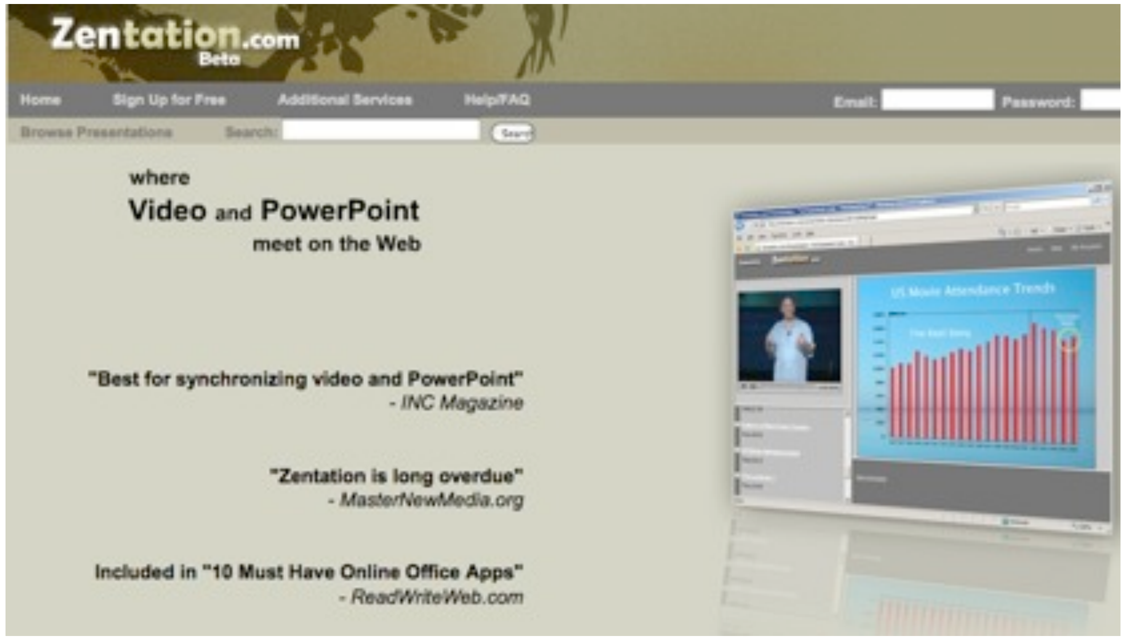
Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://wiki.cse.cuhk.edu.hk/irwin.king/home

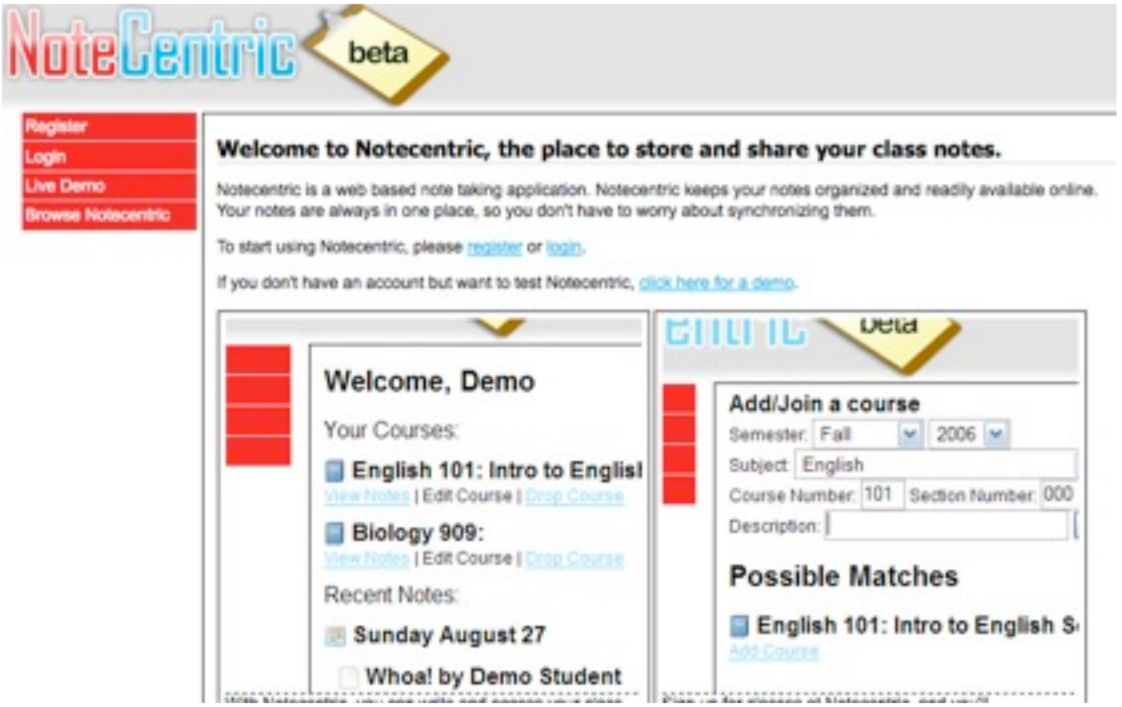
Categories of Educational Activities

- Media sharing
- Media manipulation
- Conversational arenas
- Online games and virtual worlds
- Social networking
- Blogging
- Social bookmarking
- Recommender systems
- Collaborative editing
- Wikis
- Syndication

Media Sharing

General	Educational
Uploading and downloading media files for audience or exchange	Sites have emerged that welcome creative digital material organized by educators





Zentation: Share video and powerpoint

NoteCentric: Share university class notes

Media Manipulation

General	Educational
Use web-accessible tools to design and edit digital media files	Provide graphical representations education materials

Thumbstacks: Allow presentations to be built and played online

Googlelittrips: Link literature to places or maps

Conversational Arenas

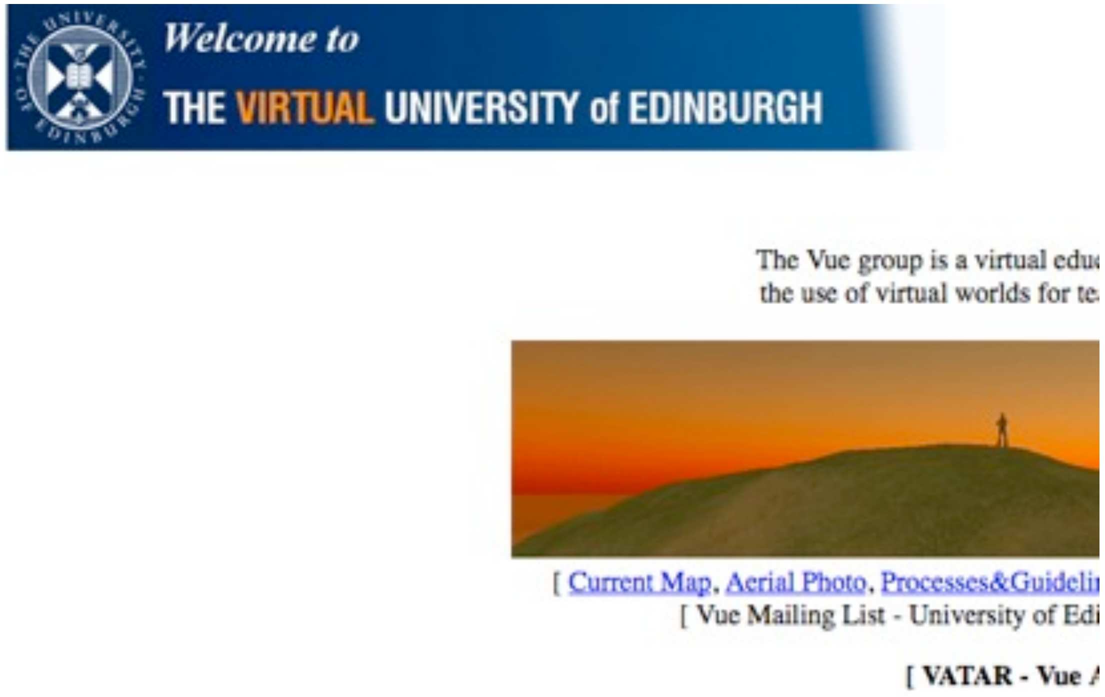
General	Educational
One-to-one or one-to-many conversations between internet users	Support educational conversations by a variety of tools

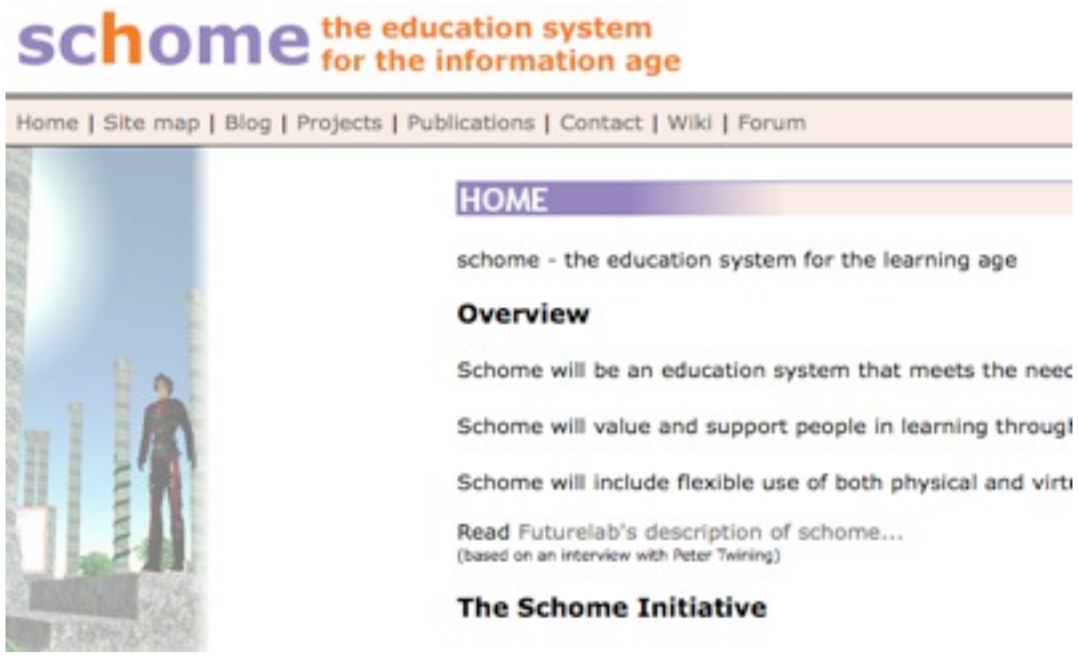
Think: Teachers and students create learning projects, participate in a website competition...

Chatmaker: Users can create chat rooms for personal websites, blogs, newsgroups...

Online Games and Virtual Worlds

General	Educational
Rule-governed games or themed environments that invite live interaction with other users	Develop multi-player online games for educational purpose



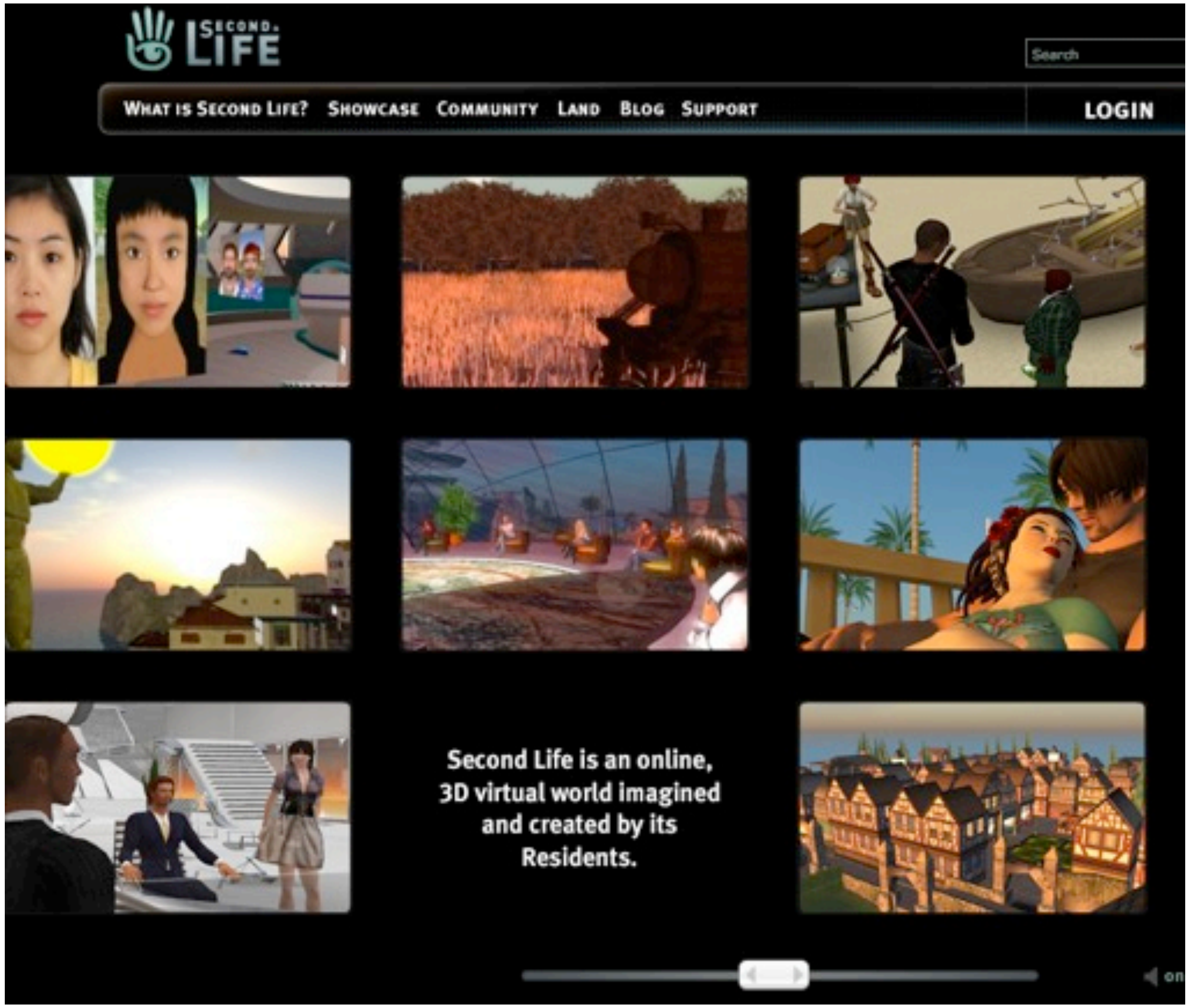


Vue: Provide a virtual educational and research institute

Schome: An education system to support people in learning throughout their lives

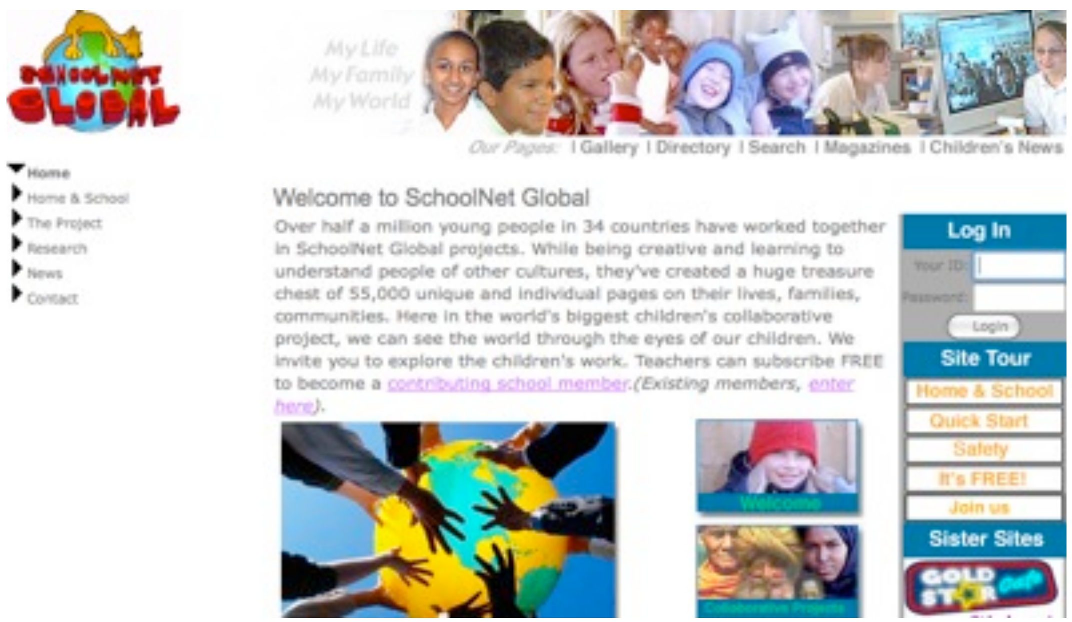
Online Games: Second Life

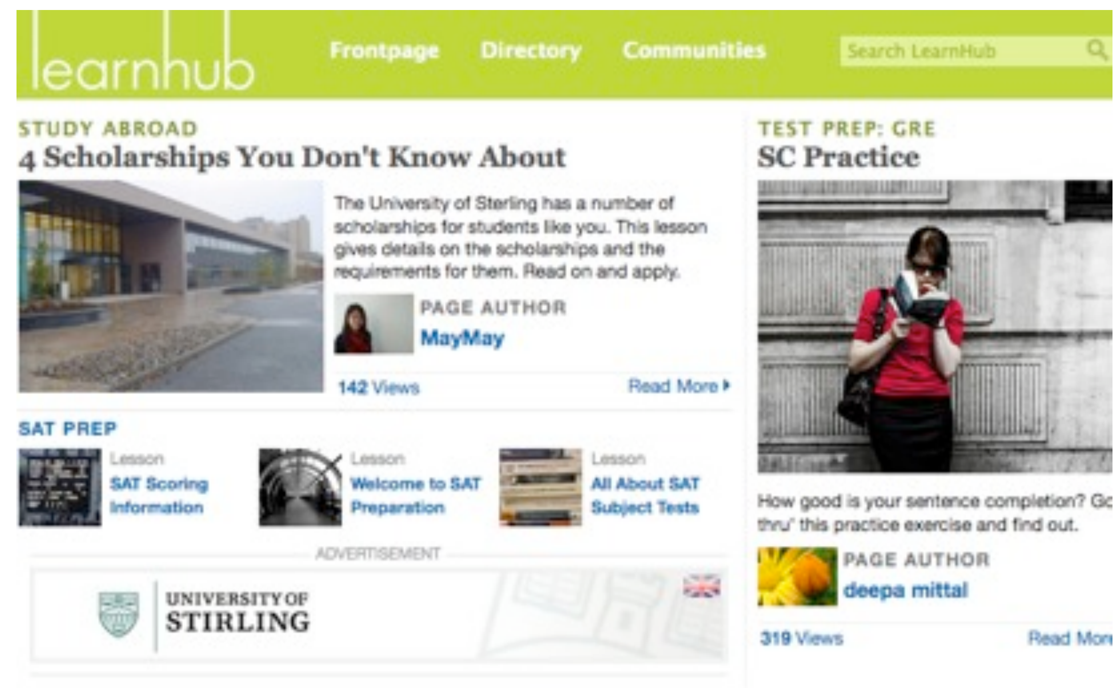
- Second Life: The Second Life Grid platform provides a powerful platform for interactive experiences
- Use it for classes, research, learning and projects
- University have set up virtual campuses where students can meet, attend classes, and create content together



Social Networking

General	Educational
Websites that structure social interaction between members who form subgroups of 'friends'	Typically include education-oriented friendship groups



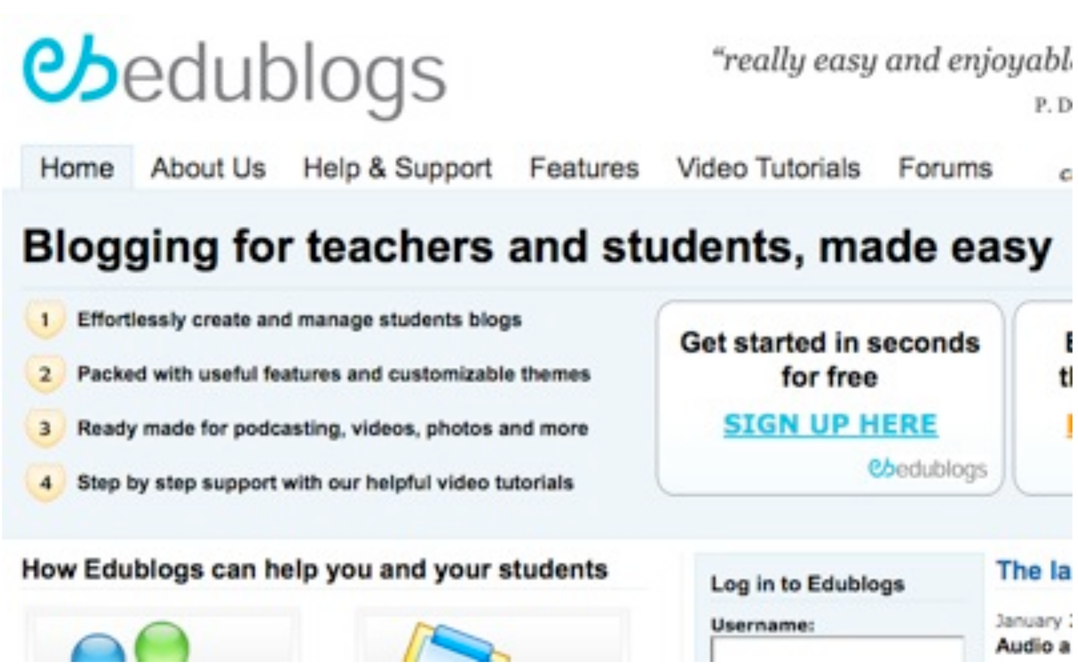


Schoolnetglobal: Provides a child-oriented design and security service for cross-site collaboration

Learnhub: Teachers can create learning communities.

Blogging

General	Educational
An on-line journal or diary in which a user can post text and digital material while others can view and comment	Blog sites exist especially for students and teachers



Edublogs: Blogging for teachers and students

Nature: Encourages scientific authors to blog around their findings

Wikis

General	Educational
Web-based services allow users unrestricted access to create, edit and link pages	Sites that allow students and teachers to establish their own wiki with an educational slant

Pbwiki: students and teacher can create their own wiki

Wikiversity: devoted to learning resources, learning projects, and research for use in all levels, types, and styles of education

Social Bookmarking

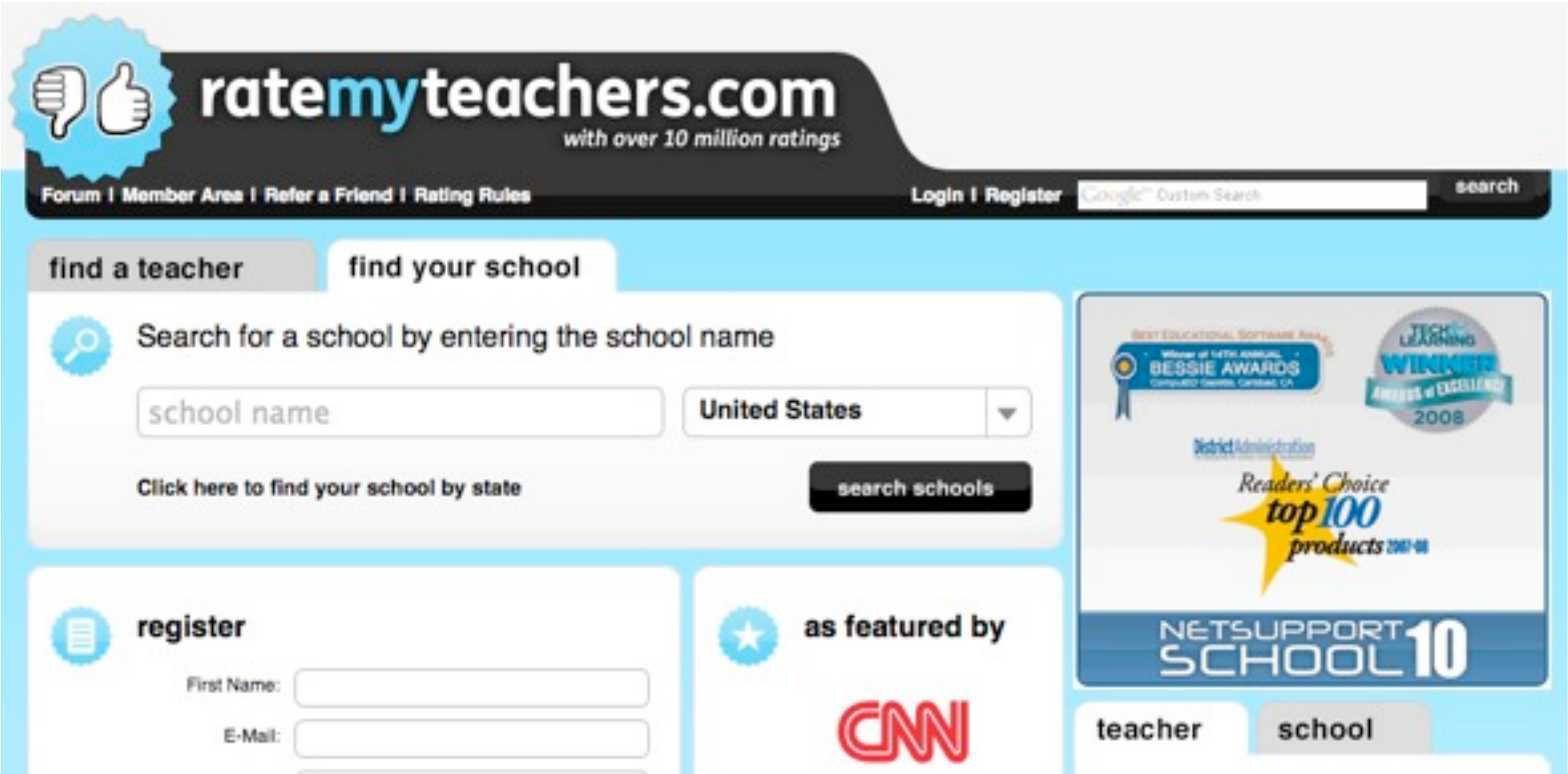
General	Educational
Allow users to submit their bookmarked web pages to a central site where they can be tagged and found by others	Bookmarks sharing systems designed for research and education users

BibSonomy: A system for sharing bookmarks and list of literature

Citeulike: A website for the collecting and sharing research publications

Recommender Systems

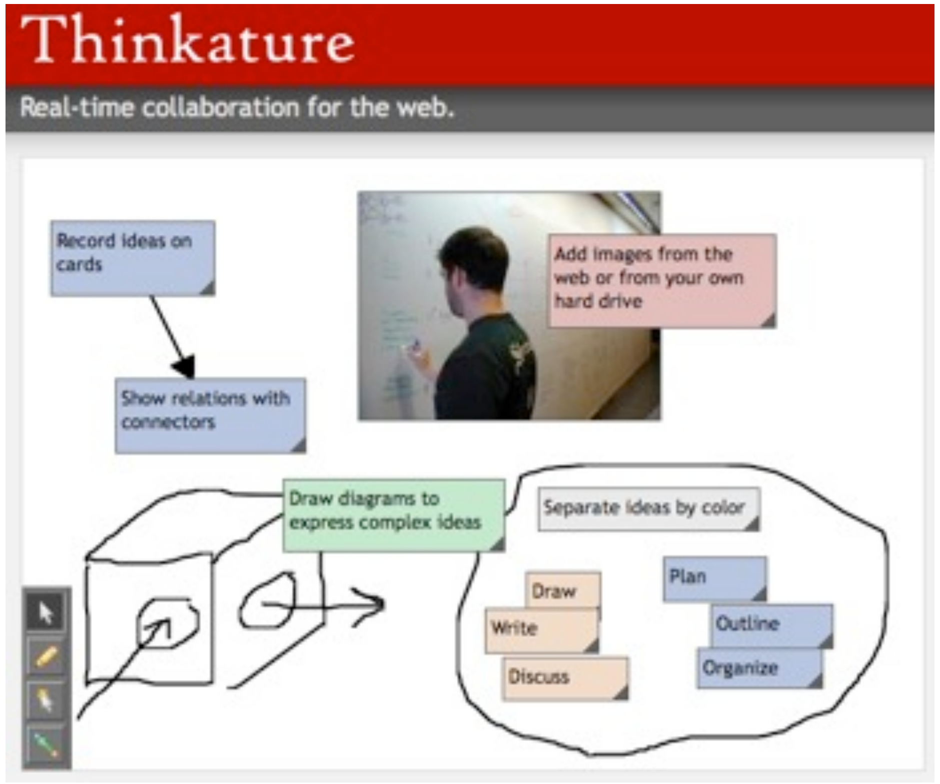
General	Educational
Websites aggregate and tag user preferences to make novel recommendations	Recommender systems designed for research and education users



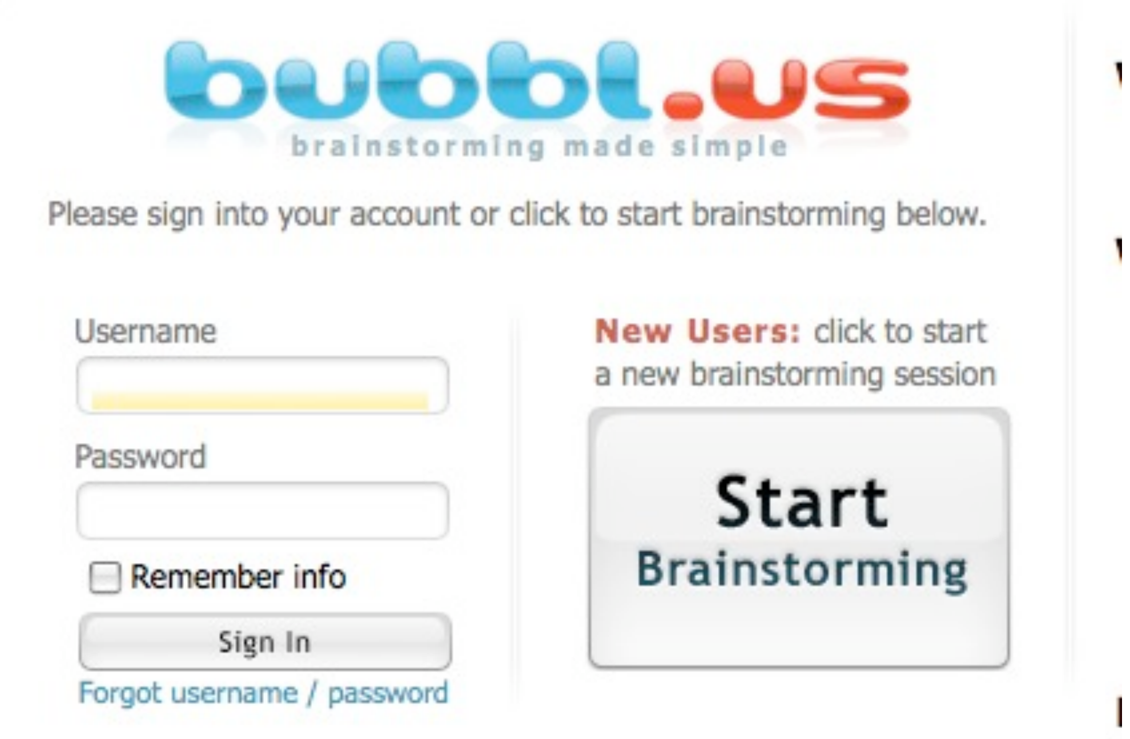
Ratemyteachers: An (infamous) example of recommendation technology in education involves user evaluation of teachers.

Collaborative Editing

General	Educational
Web tools used collaboratively to design, construct and distribute digital product	Text, spreadsheets and other documents can be stored centrally and permit collaborative editing



Thinknature: Websites incorporate more visual tools for collaborative pages



Bubbl.us: Some emphasizing mind-maps for brainstorming

Syndication

General	Educational
Users can 'subscribe' to RSS feed enable websites so that they are automatically notified of any changes or updates in content via aggregator	Websites from which students can take advantage of syndicated content

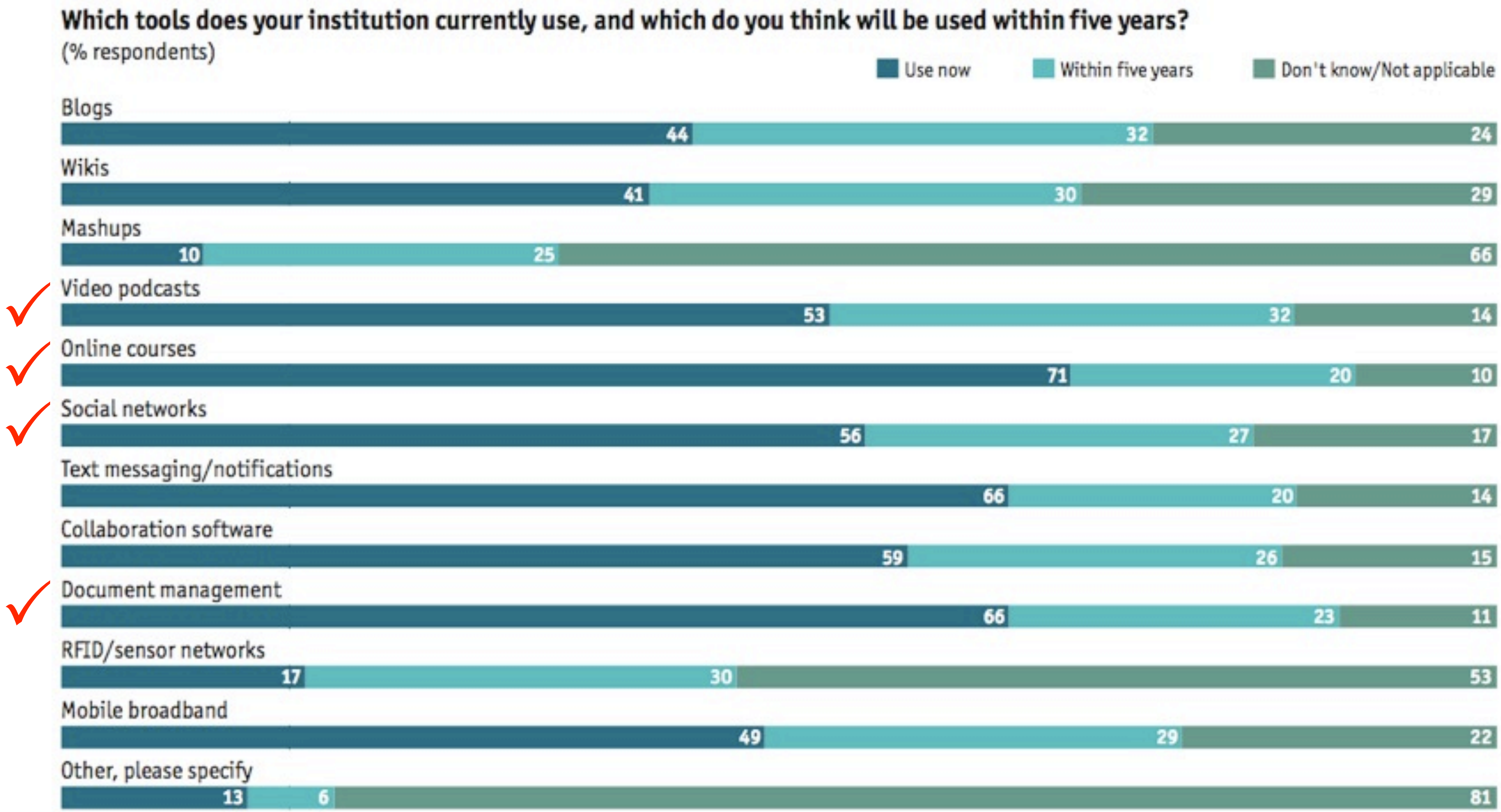
Podcastschool: A website contains podcasts for school students

Stanford: A website contains syndicated material sponsored by Stanford

Tensions and Areas for Further Research

- Teaching vs. learning
- Walled garden vs. open arena
- Private learning vs. collaborative learning
- Digital native vs. digital immigrant
- Social networking vs. anti-social networking
- Rip-mix-burn vs. cut-tweak-paste
- Transitory marks vs. persistent marks
- Print literacy vs. digital literacy
- Serial processing vs. parallel processing

Economist Intelligent Unit 2008



Summary

- New availability of resources for learning
  - Easy access to free and a variety of information resources
  - Education providers pressured to open up their resources to show their quality
- New learner empowerment and networks
  - New empowerment in choosing the learning provider
  - New means to express and show one's skills
- New participation in learning processes
  - Digital natives expect to use participative approaches

Concluding Remarks

- Social Computing is here to stay!
- Relations are important!
- Discovering new paradigms by blending different social media and interactions
- Be concerned about computational techniques to search, rank, and mine data and information to achieve collective intelligence/wisdom

"On the Internet, nobody knows you're a dog."

Acknowledgments

- Prof. Michael R. Lyu
- Prof. Jimmy Lee
- Jessie Li
- Dr. Kaizhu Huang
- Dr. Haixuan Yang
- Thomas Chan (M.Phil)
- Hongbo Deng (Ph.D.)
- Zhenjiang Lin (Ph.D.)
- Hao Ma (Ph.D.)
- Haiqin Yang (Ph.D.)
- Xin Xin (Ph.D.)
- Zenglin Xu (Ph.D.)
- Chao Zhou (Ph.D.)

On-Going Social Computing Research

Machine Learning

- Direct Zero-norm Optimization for Feature Selection (ICDM'08)
- Semi-supervised Learning from General Unlabeled Data (ICDM'08)
- Learning with Consistency between Inductive Functions and Kernels (NIPS'08)
- An Extended Level Method for Efficient Multiple Kernel Learning (NIPS'08)
- Semi-supervised Text Categorization by Active Search (CIKM'08)
- Transductive Support Vector Machine (NIPS'07)
- Global and local learning (ICML'04, JMLR'04)

Web Intelligence

- Effective Latent Space Graph-based Re-ranking Model with Global Consistency (WSDM'09)
- Formal Models for Expert Finding on DBLP Bibliography Data (ICDM'08)
- Learning Latent Semantic Relations from Query Logs for Query Suggestion (CIKM'08)
- RATE: a Review of Reviewers in a Manuscript Review Process (WI'08)
- MatchSim: link-based web page similarity measurements (WI'07)
- Diffusion rank: Ranking web pages based on heat diffusion equations (SIGIR'07)
- Web text classification (WWW'07)

Collaborative Filtering

- Recommender system: accurate recommendation based on sparse matrix (SIGIR'07)
- SoRec: Social Recommendation Using Probabilistic Matrix Factorization (CIKM'08)

Human Computation

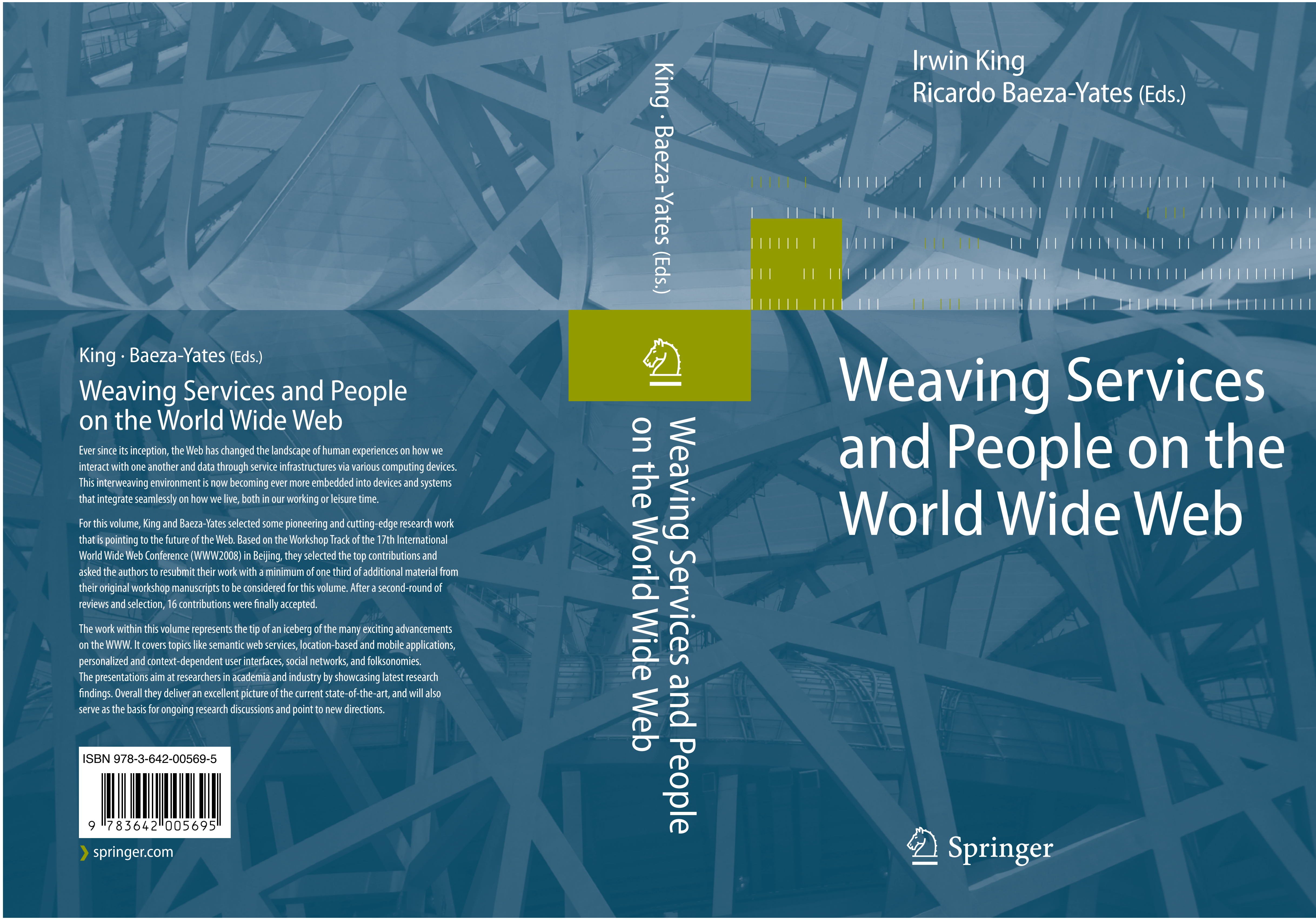
- An Analytical Study of Puzzle Selection Strategies for the ESP Game (WI'08)
- An Analytical Approach to Optimizing The Utility of ESP Games (WI'08)

Workshop on Social Computing in Education 2009

Home New since last time: 1 file

Welcome to the workshop on Social Computing in Education (SCE2009). The workshop is held in conjunction with the SocialComp-09, Vancouver, Canada from August 29-31, 2009.

With the advent of Web 2.0 and related technologies, Social Computing has become a new paradigm in ways we communicate, learn, and educate. Social platforms such as wikis, blogs, twitters, forums, groups, podcasts, mashups, virtual worlds, and sites for social networking, recommender systems, social bookmarking, social news, knowledge sharing, etc. are generating novel ways we acquire, access, manipulate, process, retrieve, present, and visualize information in the teaching and learning space. The social media for education has become dynamic, ubiquitous, distributed, real-time, collaborative, bottom-up, many-to-many, value-based, and personalized. This workshop solicits contributions on using Social Computing and related technologies for education, the emerging applications of Web 2.0 as an educational platform, as well as privacy, risk, security, and policy issues associated in Social Computing for Education 2.0.


King · Baeza-Yates (Eds.)
Weaving Services and People on the World Wide Web
Ever since its inception, the Web has changed the landscape of human experiences on how we interact with one another and data through service infrastructures via various computing devices. This interweaving environment is now becoming ever more embedded into devices and systems that integrate seamlessly on how we live, both in our working or leisure time.
For this volume, King and Baeza-Yates selected some pioneering and cutting-edge research work that is pointing to the future of the Web. Based on the Workshop Track of the 17th International World Wide Web Conference (WWW2008) in Beijing, they selected the top contributions and asked the authors to resubmit their work with a minimum of one third of additional material from their original workshop manuscripts to be considered for this volume. After a second-round of reviews and selection, 16 contributions were finally accepted.
The work within this volume represents the tip of an iceberg of the many exciting advancements on the WWW. It covers topics like semantic web services, location-based and mobile applications, personalized and context-dependent user interfaces, social networks, and folksonomies. The presentations aim at researchers in academia and industry by showcasing latest research findings. Overall they deliver an excellent picture of the current state-of-the-art, and will also serve as the basis for ongoing research discussions and point to new directions.
ISBN 978-3-642-00569-5
9 783642 005695
springer.com
King · Baeza-Yates (Eds.)
Weaving Services and People on the World Wide Web
Irwin King
Ricardo Baeza-Yates (Eds.)
Weaving Services and People on the World Wide Web
Springer

Economist Intelligent Unit 2008

Potential increase in student plagiarism

VeriGuide

- Similarity text detection system
- Developed at CUHK
- Promote and uphold academic honesty, integrity, and quality
- Support English, Traditional and Simplified Chinese
- Handle .doc, .txt, .pdf, .html, etc. file formats
- Generate detailed originality report including readability
- Use "WWW2009MD" for the **10-20-30** service at www.veriguide.org