Information Search and Recommender Systems

concepts and issues

Francesco Ricci

Free University of Bozen-Bolzano

Motivations

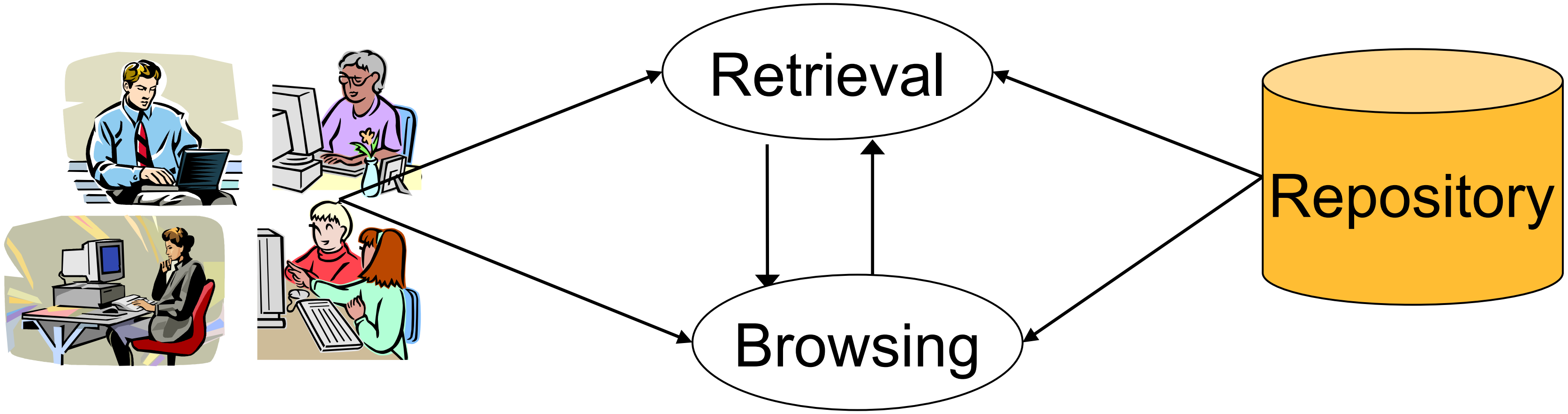
- The **World Wide Web** has become the primary source of information for leisure and work activities
- WWW huge content would be wasted if that information could not be **found**, **analyzed**, and **exploited**
- Each user should be able to **quickly find information** that is both relevant and comprehensive for their needs
- WWW has become a **principal driver of innovation** and a range of new techniques have been introduced to tame and exploit its information content
- **Recommender systems** are (web, mobile, …) tools that are becoming more and more popular for supporting the user in **finding** and **selecting products**, **services**, or **information.**

Basic Concepts in Information Retrieval

- **Information Retrieval** (IR) **deals** with the representation, storage and organization of unstructured data
- **Information retrieval** is the process of searching within a document collection for a particular information need (a **query**)
- Its mission is to assist in **information search**
- Two main search paradigms:

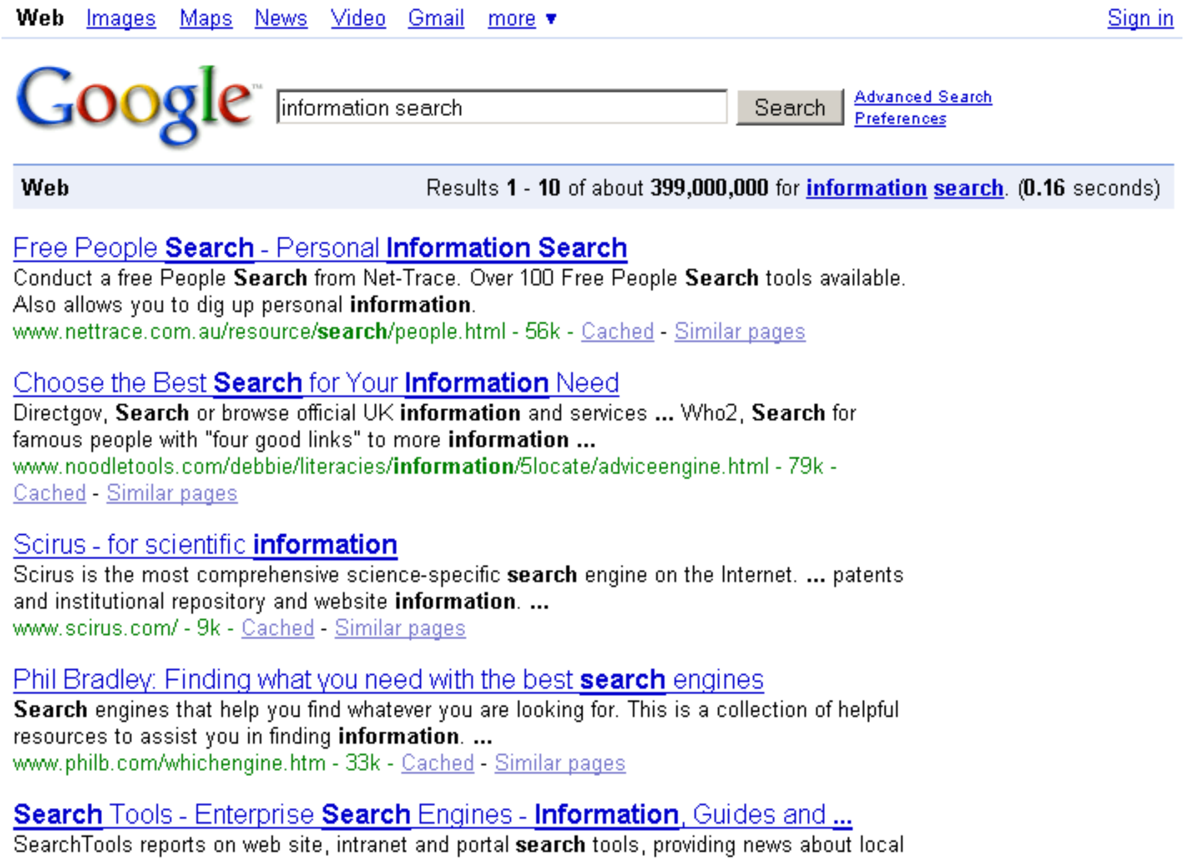
Retrieval and **Browse**

The User Task



- **Retrieval**
  - Search for particular information
  - Usually focused and purposeful
- **Browsing**
  - General looking around for information
  - For example: Asia-> Thailand -> Phuket -> Tsunami

Search Engines: Information Retrieval Tools



Web Results **1** - **10** of about **399,000,000** for **information search**. (**0.16** seconds)

Free People **Search** - Personal **Information Search**
Conduct a free People **Search** from Net-Trace. Over 100 Free People **Search** tools available. Also allows you to dig up personal **information**.
www.nettrace.com.au/resource/**search**/people.html - 56k - Cached - Similar pages

Choose the Best **Search** for Your **Information** Need
Directgov, **Search** or browse official UK **information** and services **...** Who2, **Search** for famous people with "four good links" to more **information ...**
www.noodletools.com/debbie/literacies/**information**/5locate/adviceengine.html - 79k - Cached - Similar pages

Scirus - for scientific **information**
Scirus is the most comprehensive science-specific **search** engine on the Internet. **...** patents and institutional repository and website **information**. **...**
www.scirus.com/ - 9k - Cached - Similar pages

Phil Bradley: Finding what you need with the best **search** engines
Search engines that help you find whatever you are looking for. This is a collection of helpful resources to assist you in finding **information**. **...**
www.philb.com/whichengine.htm - 33k - Cached - Similar pages

Search Tools - Enterprise **Search** Engines - **Information**, Guides and **...**
SearchTools reports on web site, intranet and portal **search** tools, providing news about local

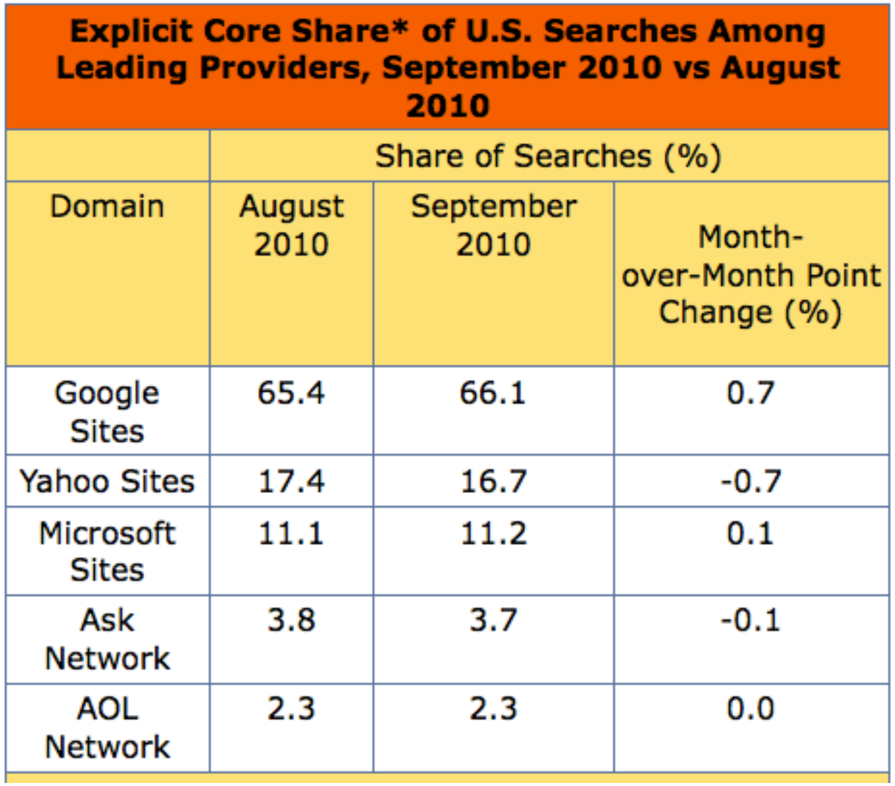
- Search engines are the primary tools people use to find information on the web
- Exclusion of a site from search engines will cut off the site from its intended audience.

Brief History of Search Engines

- Yahoo! (www.yahoo.com) - (1994-) directory service and search engine.
- Infoseek – (1994-2001) search engine.
- Inktomi – (1995-) search engine infrastructure, acquired by Yahoo! 2003.
- AltaVista – (1995-) search engine, acquired by Overture in 2003.
- AlltheWeb – (1999-) search engine, acquired by Overture in 2003 .
- Ask Jeeves (www.ask.com) - (1996-) Q&A and search engine, acquired by IAC/InterActiveCorp in 2005.
- Overture – (1997-) pay-per-click search engine, acquired by Yahoo! 2003.
- Bing (www.bing.com) – (2009-) Microsoft rebarded search engine, was Live in 2006 and MSN search before.
- Google (www.google.com) – (1998-) – search engine.

Search Engine Statistics

- Google has over 40,000 searches a second.
- In 2005 Google has 36.5% searches but as of 2010 Google dominates with Bing and Yahoo far behind.
- In China and Korea local engines are more popular.
- Users are spending more time on the web (over 34 hours a month, Feb. 2009).



Explicit Core Share* of U.S. Searches Among Leading Providers, September 2010 vs August 2010

	Share of Searches (%)		
Domain	August 2010	September 2010	Month-over-Month Point Change (%)
Google Sites	65.4	66.1	0.7
Yahoo Sites	17.4	16.7	-0.7
Microsoft Sites	11.1	11.2	0.1
Ask Network	3.8	3.7	-0.1
AOL Network	2.3	2.3	0.0

Source: ComScore

Web IR- IR on the Web

- **First Generation**
  - Classical approach (boolean, vector, and probabilistic models)
  - Informational: IR/DB techniques on page content. E.g., Lycos, Excite, AltaVista
- **Second Generation**
  - Web as a graph
  - Navigational: use off-page Web specific data – links topology. E.g., Google
- **Third Generation**
  - Open research
  - Mobile information search
  - A lot of business potential, “monetarization of infomediary role”, matching services

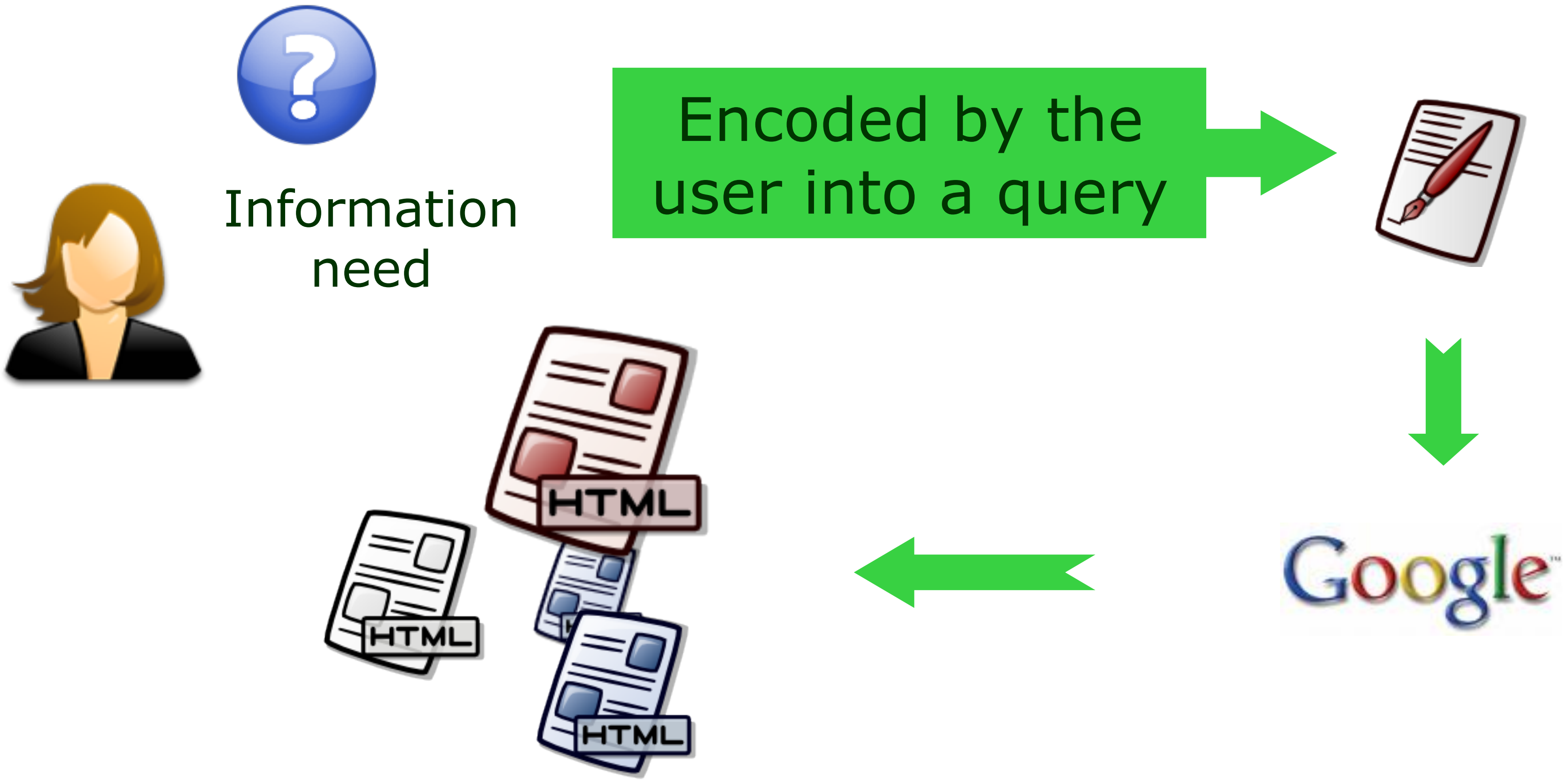
Problems with Using IR for Web

- Very **large** and **heterogeneous** collection
  - Dynamic
  - Self-organized
  - Hyperlinked
- Very **short queries**
- **Unsophisticated** users
- **Difficult to judge relevance** and to rank results
- **Synonymy** and **ambiguity**
- Authorship styles (in content writing and query formulation)
- Search engine **persuasion**, keyword *stuffing* (a web page is loaded with keywords in the meta tags or in content).

IR: The Basic Concepts

- The user has an **information need**, that is expressed as a **free-text query**
- Information need: *the perceived need for information that leads to someone using an information retrieval system in the first place* [Schneiderman, Byrd, and Croft. 1997]
- The query **encodes** the information search need
- The query **is a** “**document**”, to be compared to a collection of documents
- **Effectiveness vs Efficiency**
- How to **compare documents**? Similarity metrics needed!
- How to **avoid** doing a **sequential search**? Can we search in parallel in a set of servers?

From needs to queries



Information need

Encoded by the user into a query

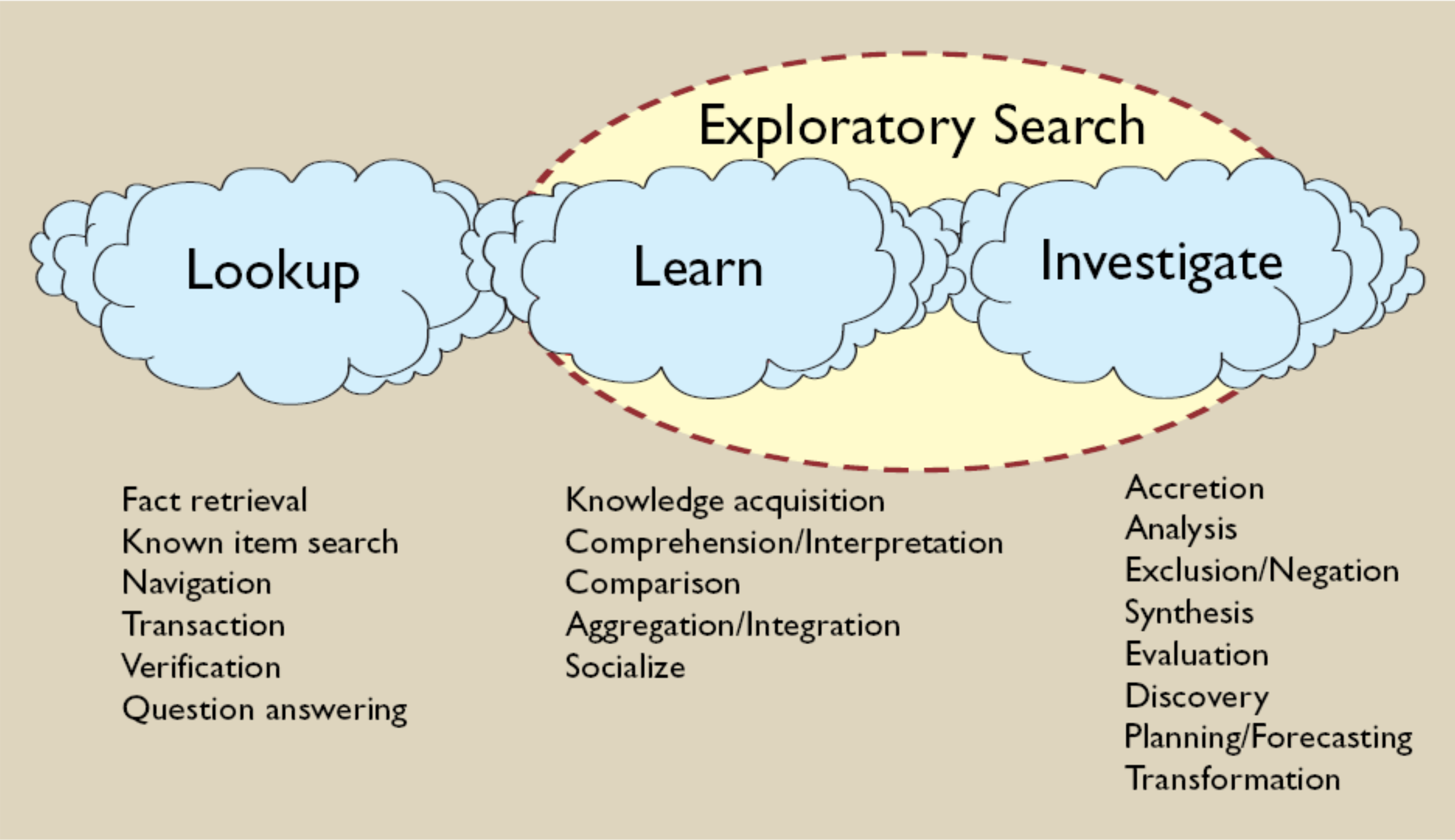
- Information need -> query -> search engine -> results -> browse OR query -> ...

Taxonomy of Web search

- In the web context the "need behind the query" is often not informational in nature
- [Broder, 2002] classifies web queries according to their intent into 3 classes:

1. **Navigational:** The immediate intent is to reach a particular site (20%):
   - *q = compaq* - probable target http://www.compaq.com
2. **Informational:** The intent is to acquire some information assumed to be present on one or more web pages (50%)
   - q= canon 5d mkII - probable target a page reviewing canon 5d mkII
3. **Transactional:** The intent is to perform some web-mediated activity (30%)
   - q = hotel Vienna - probable target "Expedia"

Exploratory Search



Exploratory Search

Lookup

Learn

Investigate

Fact retrieval
Known item search
Navigation
Transaction
Verification
Question answering

Knowledge acquisition
Comprehension/Interpretation
Comparison
Aggregation/Integration
Socialize

Accretion
Analysis
Exclusion/Negation
Synthesis
Evaluation
Discovery
Planning/Forecasting
Transformation

[Marchionini, 2006]

Strategies and Tools

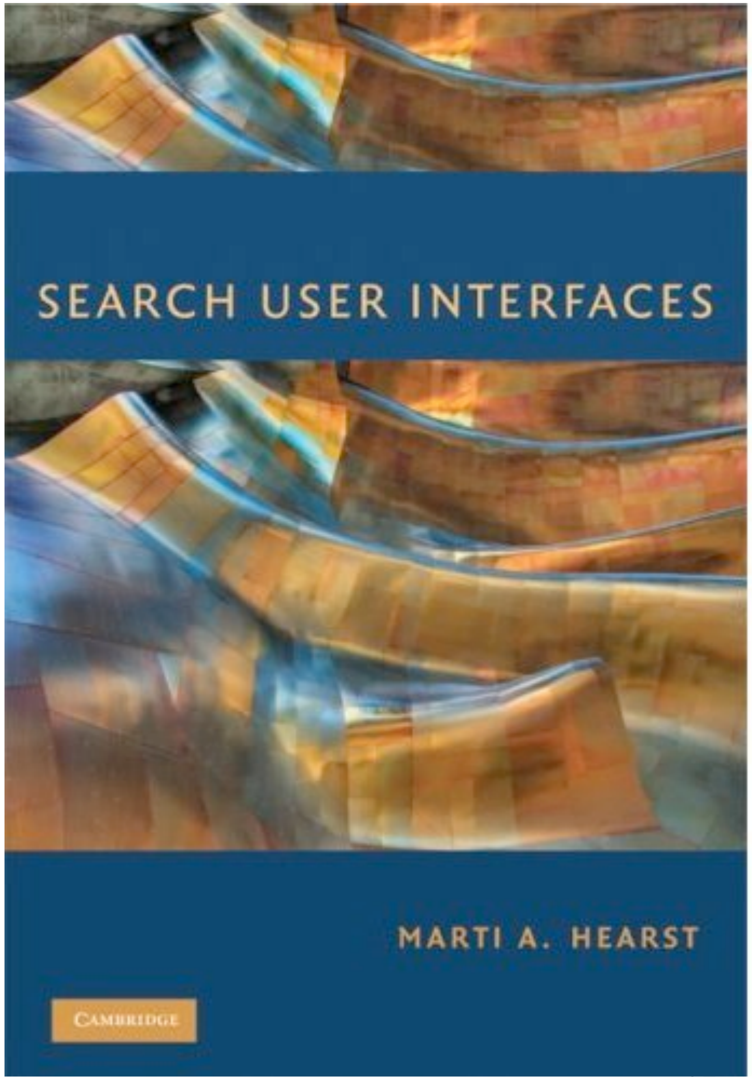
- A search engine is just a tool, among others, that can be exploited, within a strategy, to achieve a goal (perform a task)

- New tools have emerged, and will be developed, to combine work in Human Computer Interaction and Information Retrieval

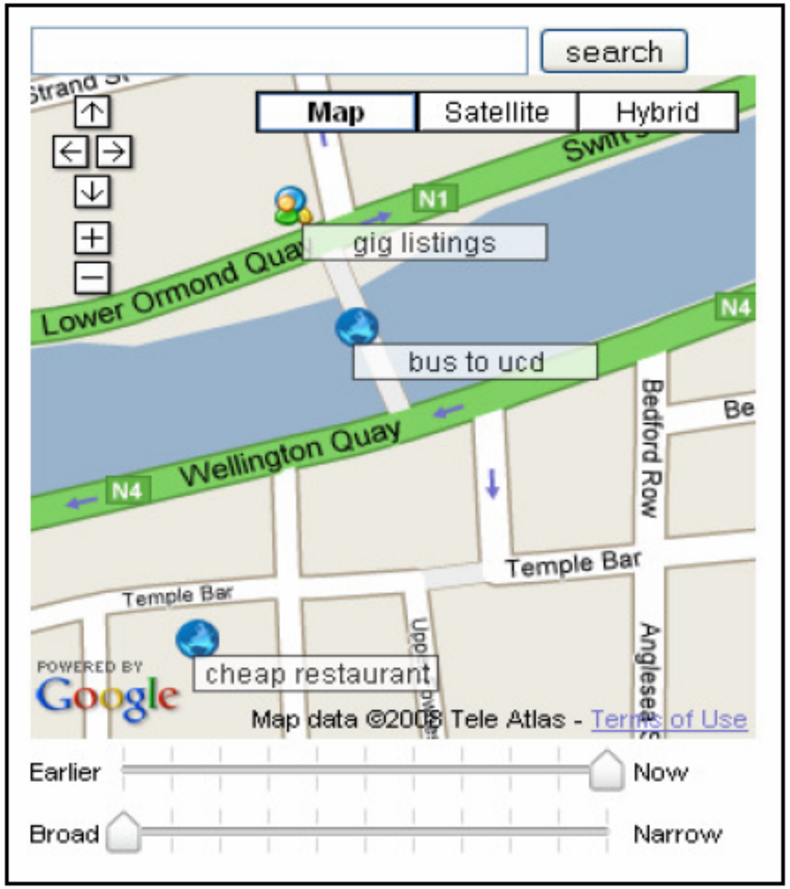
- Exploratory search is the area where new tools will be developed mostly

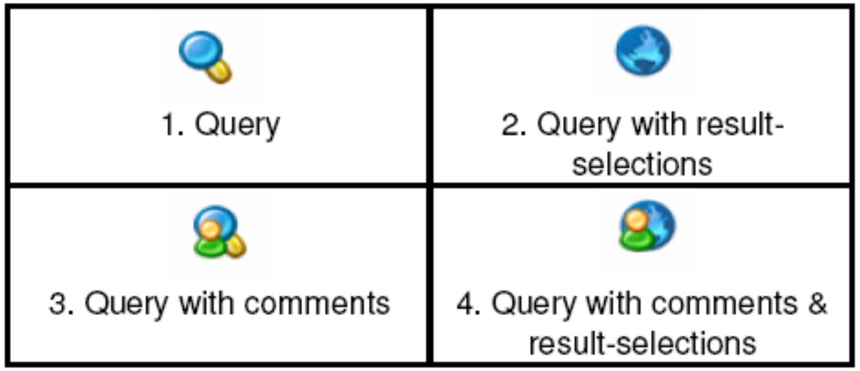
Information Search Interfaces

- Design Search User Interfaces
- Evaluate Search User Interfaces
- Models of the Information Seeking Process
- Search Interfaces Fundamentals:
  - Query Specification
  - Presentation of Search Results
  - Query Reformulation
- Advanced Topics, including:
  - Integrating Navigation with Search
  - Personalization in Search
  - Information Visualization
  - Mobile Search
  - Social Search
  - Multimedia Search



Exploratory Search: Mobile Search





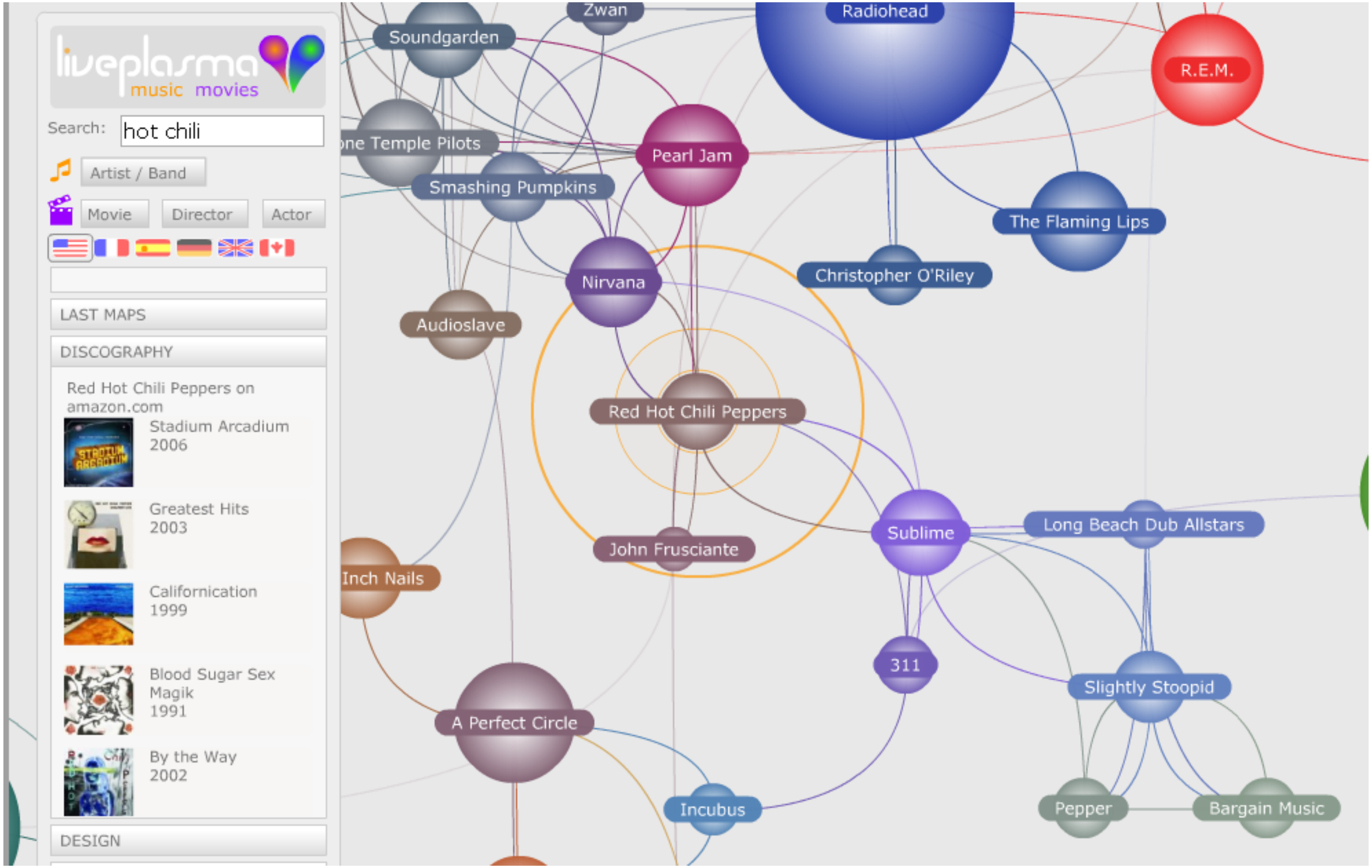
1. Query	2. Query with result-selections
3. Query with comments	4. Query with comments & result-selections

(b) Icons used to identify queries

[Church and Smyth, 2008]

- User can browse searches (query and results) performed by other users in a location.

Exploratory Search: Example



Red Hot Chili Peppers on amazon.com

Stadium Arcadium 2006

Greatest Hits 2003

Californication 1999

Blood Sugar Sex Magik 1991

By the Way 2002

http://www.visualcomplexity.com/vc/

Information Search Features

- There is **no single best strategy** or tool for finding information
- The strategy depends on:
  - the **nature** of the **information** the user is seeking,
  - the nature and the **structure** of the **content repository**,
  - the **search tools** available,
  - the user **familiarity** with the **information** and the **terminology** used in the repository,
  - and the **ability** of the user to **use the search tools** competently.

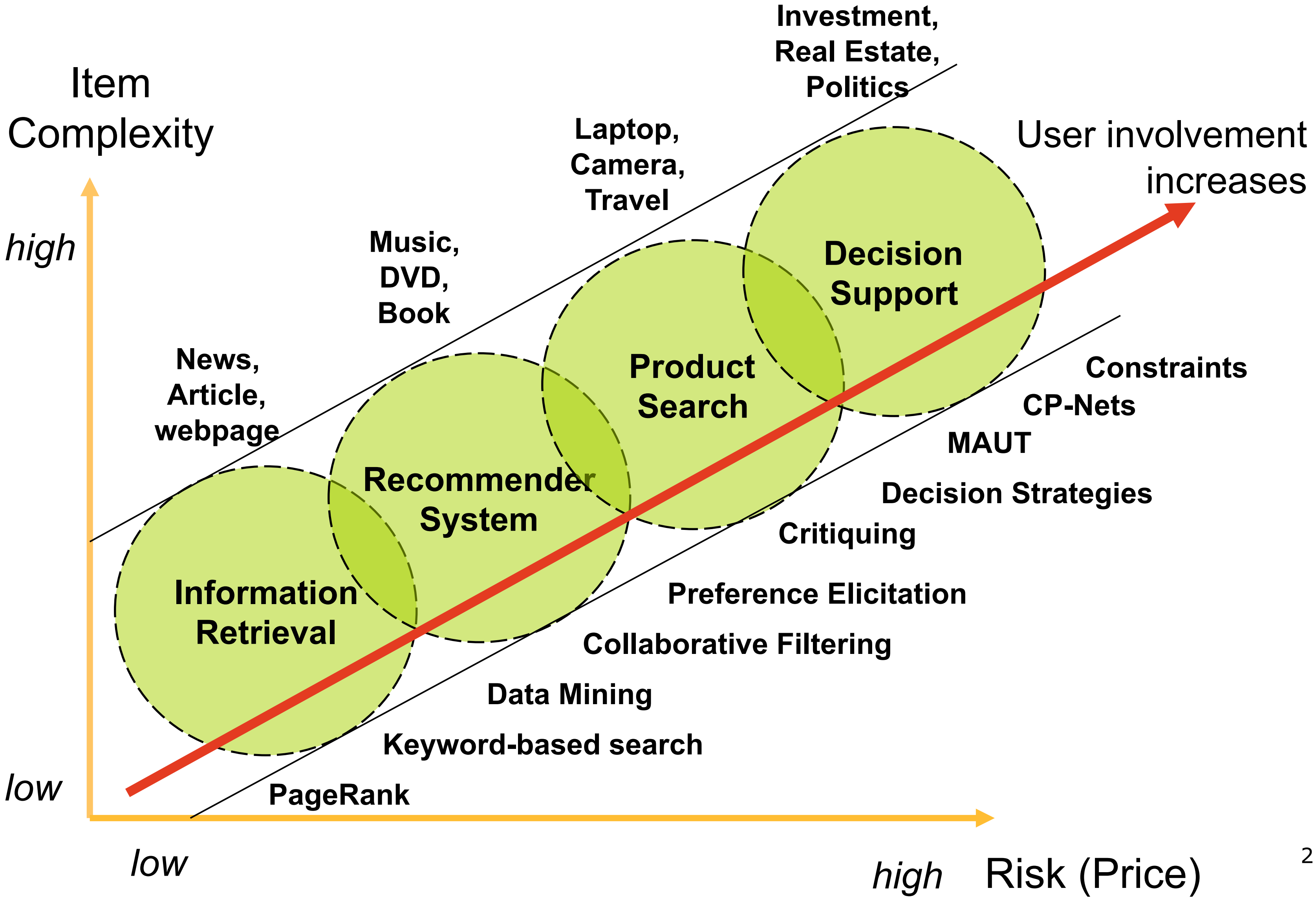
Information Search and Decision Making

- Information Search (IS) and Decision Making (DM) are strictly connected
- **IS for DM:** we search information (external and internal) before taking decisions
  - Classical in DM and Consumer Behavior
- **DM for IS:** we must take decisions about what information to consider, or when to stop searching
  - New feature of the Web, caused by Information Overload.

Information Overload

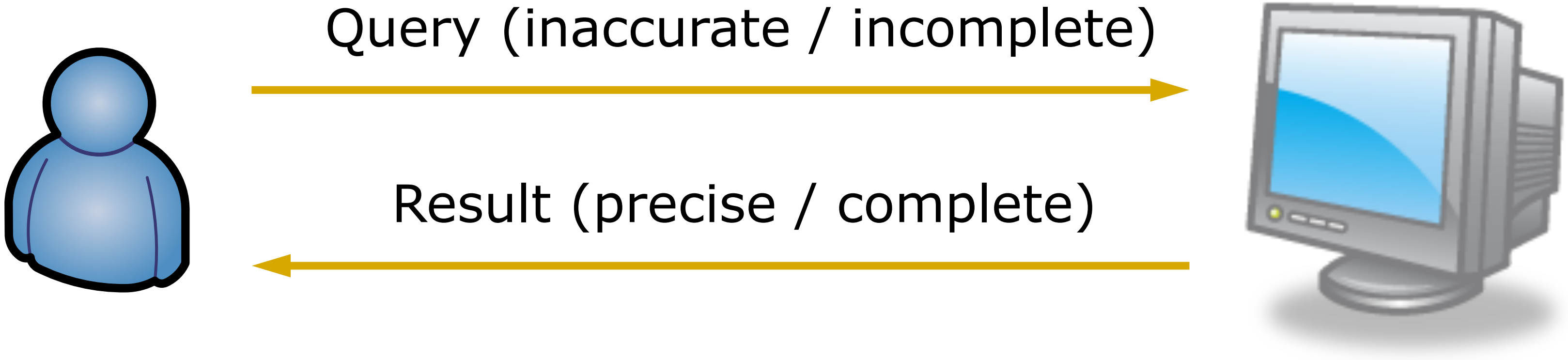
- **Internet = information overload**, i.e., the state of having too much information to **make a decision** or **remain informed about a topic**
- Information retrieval technologies can assist a user to **look up** content if the user knows exactly what he is looking for (i.e. for lookup)
- But to **make a decision** or **remain informed about a topic you must perform an exploratory search** (e.g., comparison, knowledge acquisition, product selection, etc.)
  - not aware of the range of available options
  - may not know what to search
  - if presented with some results may not be able to choose.

Type of Techniques



Min input vs. Max output

- Most users are impatient to get results providing just minimal input
- Users' preferences are constructive and context dependent
- Users want to make accurate choices, i.e., get relevant information items



Query (inaccurate / incomplete)

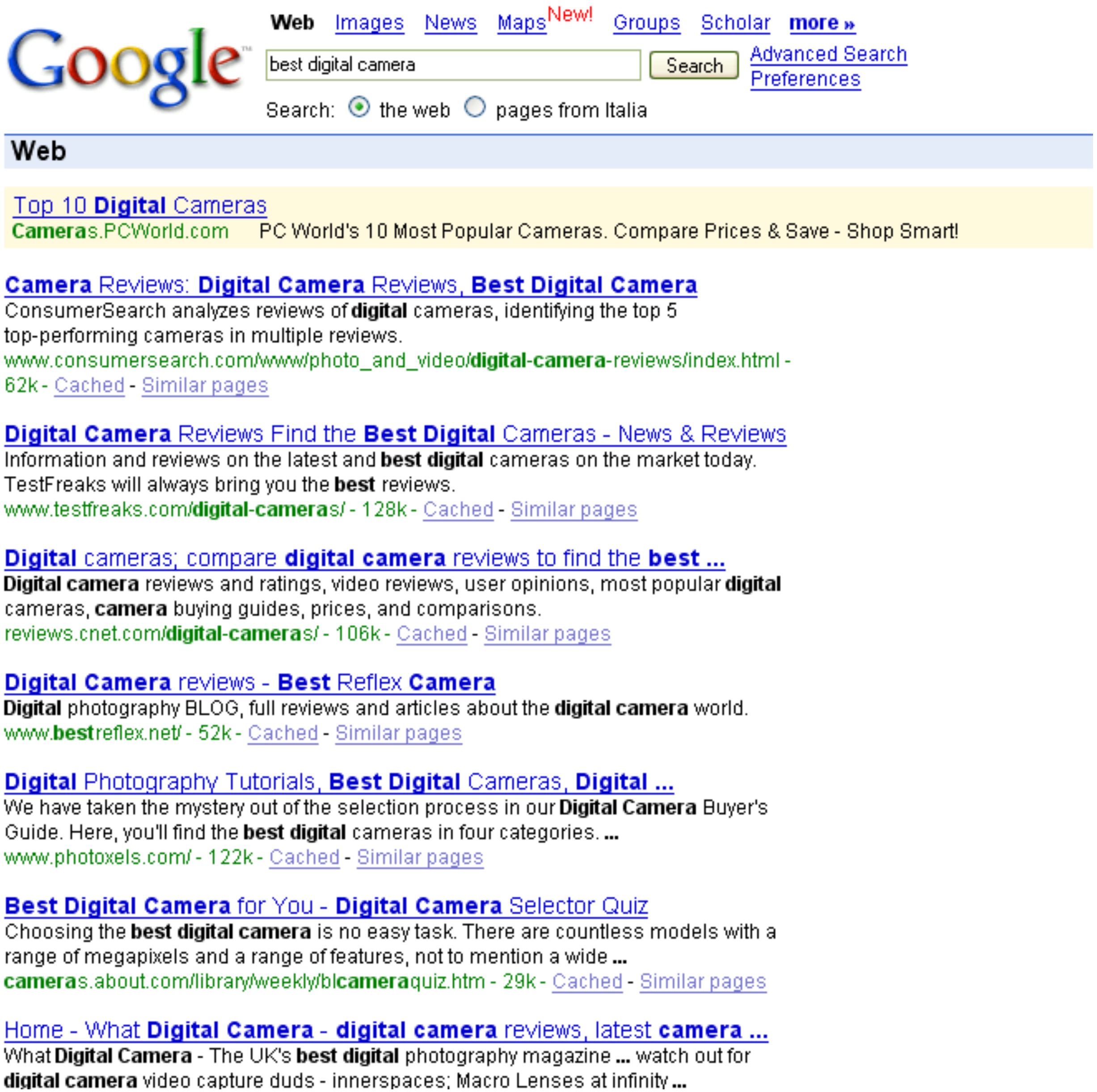
Result (precise / complete)

Recommender Systems

- In everyday life **we rely on recommendations** from other people either by word of mouth, recommendation letters, movie and book reviews printed in newspapers …
- In a typical recommender system **people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients**
  - Aggregation of recommendations
  - Match the recommendations with those searching for recommendations

[Resnick and Varian, 1997]

Recommenders and Search Engines



Web Images News Maps New! Groups Scholar more »

best digital camera | Search | Advanced Search | Preferences

Search: ● the web ○ pages from Italia

Web

Top 10 **Digital** Cameras
Cameras.PCWorld.com PC World's 10 Most Popular Cameras. Compare Prices & Save - Shop Smart!

Camera Reviews: **Digital Camera** Reviews, **Best Digital Camera**
ConsumerSearch analyzes reviews of **digital** cameras, identifying the top 5 top-performing cameras in multiple reviews.
www.consumersearch.com/www/photo_and_video/**digital-camera**-reviews/index.html - 62k - Cached - Similar pages

Digital Camera Reviews Find the **Best Digital** Cameras - News & Reviews
Information and reviews on the latest and **best digital** cameras on the market today. TestFreaks will always bring you the **best** reviews.
www.testfreaks.com/**digital-camera**s/ - 128k - Cached - Similar pages

Digital cameras; compare **digital camera** reviews to find the **best** ...
Digital camera reviews and ratings, video reviews, user opinions, most popular **digital** cameras, **camera** buying guides, prices, and comparisons.
reviews.cnet.com/**digital-camera**s/ - 106k - Cached - Similar pages

Digital Camera reviews - **Best** Reflex **Camera**
Digital photography BLOG, full reviews and articles about the **digital camera** world.
www.**best**reflex.net/ - 52k - Cached - Similar pages

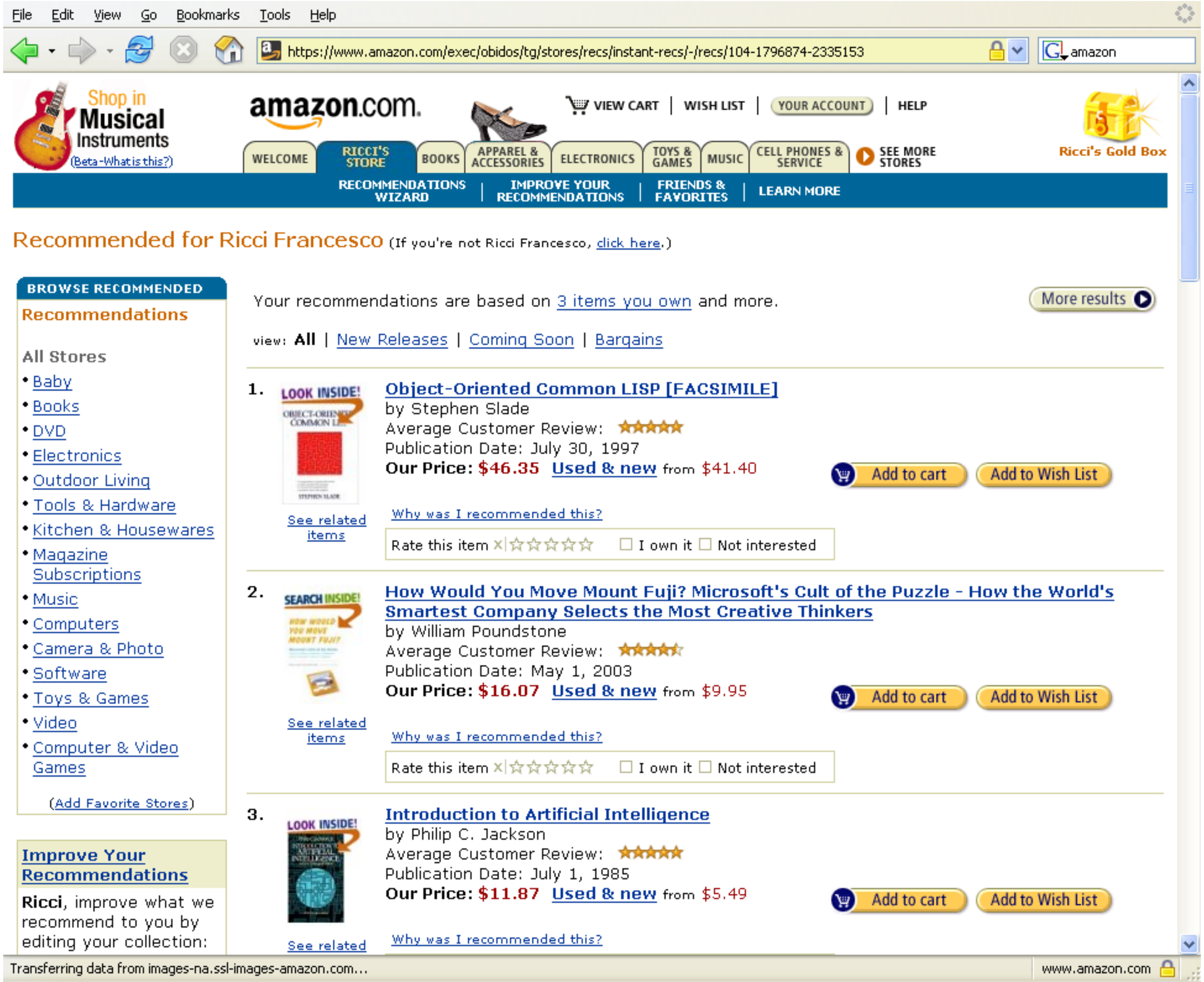
Digital Photography Tutorials, **Best Digital** Cameras, **Digital** ...
We have taken the mystery out of the selection process in our **Digital Camera** Buyer's Guide. Here, you'll find the **best digital** cameras in four categories. ...
www.photoxels.com/ - 122k - Cached - Similar pages

Best Digital Camera for You - **Digital Camera** Selector Quiz
Choosing the **best digital camera** is no easy task. There are countless models with a range of megapixels and a range of features, not to mention a wide ...
cameras.about.com/library/weekly/bl**camera**quiz.htm - 29k - Cached - Similar pages

Home - What **Digital Camera** - **digital camera** reviews, latest **camera** ...
What **Digital Camera** - The UK's **best digital** photography magazine ... watch out for **digital camera** video capture duds - innerspaces; Macro Lenses at infinity ...

A search engine is not a recommender system

Querying a SE for a recommendation will return a list of **recommender systems**


File Edit View Go Bookmarks Tools Help
https://www.amazon.com/exec/obidos/tg/stores/recs/instant-recs/-/recs/104-1796874-2335153
amazon
Shop in Musical Instruments
(Beta-What is this?)
amazon.com.
VIEW CART | WISH LIST | YOUR ACCOUNT | HELP
WELCOME
RICCI'S STORE
BOOKS
APPAREL & ACCESSORIES
ELECTRONICS
TOYS & GAMES
MUSIC
CELL PHONES & SERVICE
SEE MORE STORES
Ricci's Gold Box
RECOMMENDATIONS WIZARD | IMPROVE YOUR RECOMMENDATIONS | FRIENDS & FAVORITES | LEARN MORE

Recommended for Ricci Francesco (If you're not Ricci Francesco, click here.)

BROWSE RECOMMENDED

Recommendations

All Stores

- Baby
- Books
- DVD
- Electronics
- Outdoor Living
- Tools & Hardware
- Kitchen & Housewares
- Magazine Subscriptions
- Music
- Computers
- Camera & Photo
- Software
- Toys & Games
- Video
- Computer & Video Games

(Add Favorite Stores)

Improve Your Recommendations

Ricci, improve what we recommend to you by editing your collection:

Your recommendations are based on 3 items you own and more.

view: **All** | New Releases | Coming Soon | Bargains

1.
Object-Oriented Common LISP [FACSIMILE]
by Stephen Slade
Average Customer Review: ★★★★★
Publication Date: July 30, 1997
Our Price: $46.35 Used & new from $41.40

See related items
Why was I recommended this?
Rate this item ✕ ☆☆☆☆☆ □ I own it □ Not interested

2.
How Would You Move Mount Fuji? Microsoft's Cult of the Puzzle - How the World's Smartest Company Selects the Most Creative Thinkers
by William Poundstone
Average Customer Review: ★★★★½
Publication Date: May 1, 2003
Our Price: $16.07 Used & new from $9.95

See related items
Why was I recommended this?
Rate this item ✕ ☆☆☆☆☆ □ I own it □ Not interested

3.
Introduction to Artificial Intelligence
by Philip C. Jackson
Average Customer Review: ★★★★★
Publication Date: July 1, 1985
Our Price: $11.87 Used & new from $5.49

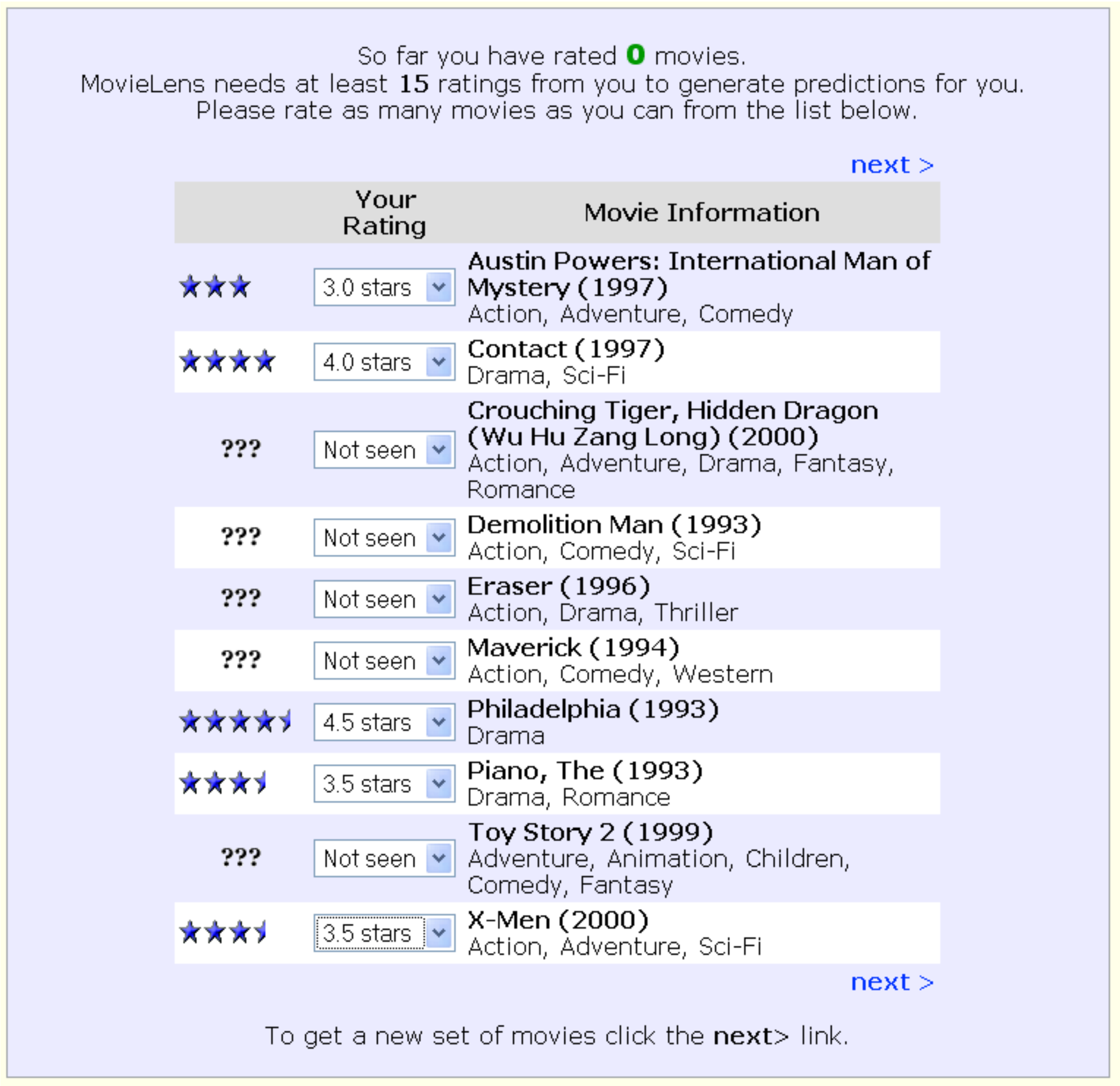
See related
Why was I recommended this?

Core Computations of Recommender Systems

- **Rating Prediction:** a model must be built to predict ratings for items not currently rated by the user
  - **Numeric ratings:** regression
  - **Discrete ratings:** classification
- **Ranking:** compute a score for each item and then rank the items with respect to the score (e.g. search engine)
  - Simpler than rating prediction - just the order matter
- **Selection task:** a model must be built that selects the N most relevant items – new for the user
  - Can be thought to be a post-process of rating prediction or ranking – but different evaluation strategies are applied.

The Collaborative Filtering Idea

- Trying to **predict** the opinion the user will have on the different items and be able to recommend the "best" items to each user based on: **the user's previous likings** and the **opinions of other like minded users**
- From an historical point of view CF came after content-based (we'll see this later) but it is the most famous method
- CF is a typical **Internet application** – it must be supported by a networking infrastructure
  - But we are thinking of using many servers
  - At least many users and one server
- There is no stand alone CF application.



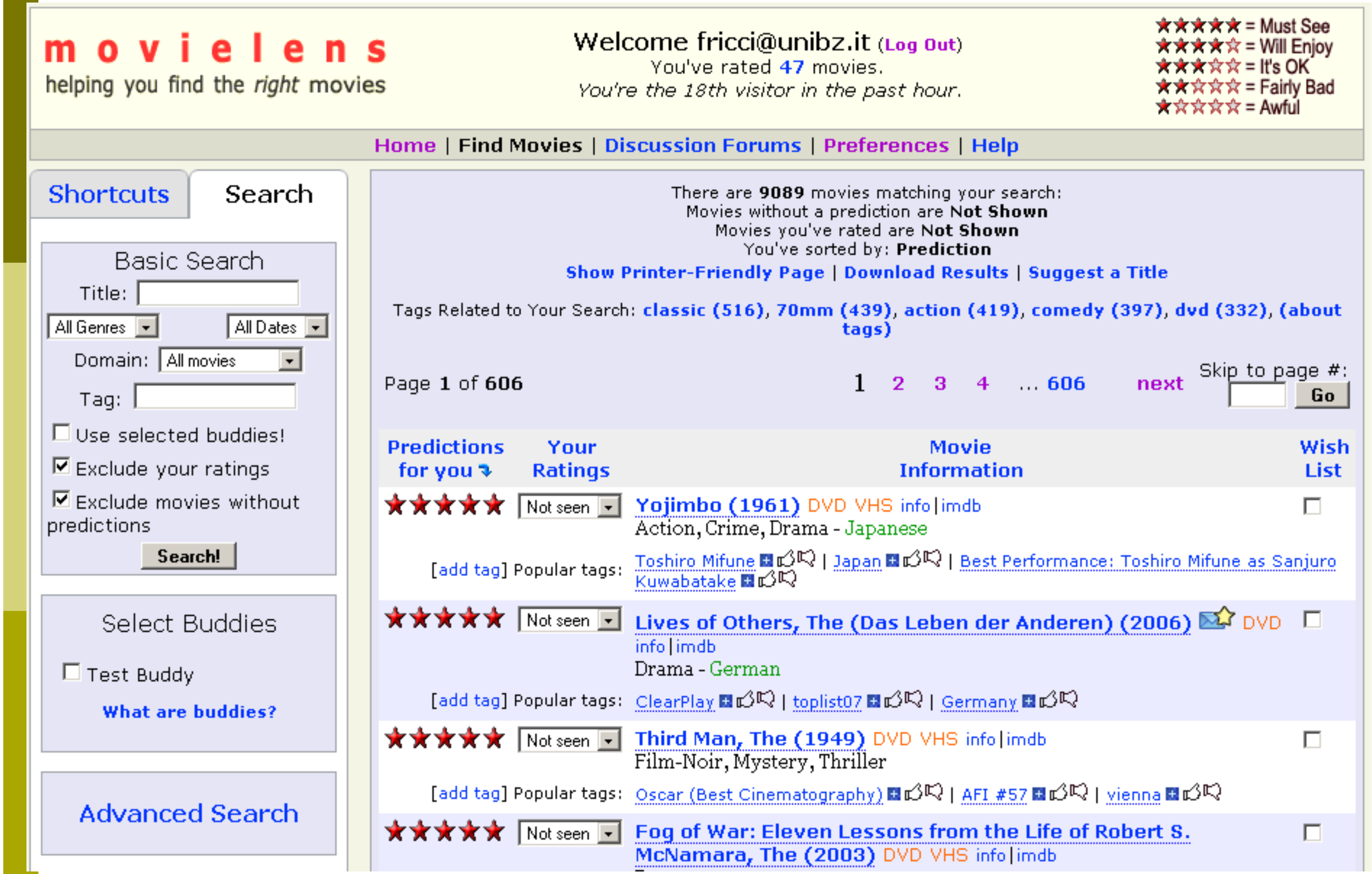
So far you have rated **0** movies.
MovieLens needs at least **15** ratings from you to generate predictions for you.
Please rate as many movies as you can from the list below.

next >

	Your Rating	Movie Information
★★★	3.0 stars	**Austin Powers: International Man of Mystery (1997)** Action, Adventure, Comedy
★★★★	4.0 stars	**Contact (1997)** Drama, Sci-Fi
???	Not seen	**Crouching Tiger, Hidden Dragon (Wu Hu Zang Long) (2000)** Action, Adventure, Drama, Fantasy, Romance
???	Not seen	**Demolition Man (1993)** Action, Comedy, Sci-Fi
???	Not seen	**Eraser (1996)** Action, Drama, Thriller
???	Not seen	**Maverick (1994)** Action, Comedy, Western
★★★★⯨	4.5 stars	**Philadelphia (1993)** Drama
★★★⯨	3.5 stars	**Piano, The (1993)** Drama, Romance
???	Not seen	**Toy Story 2 (1999)** Adventure, Animation, Children, Comedy, Fantasy
★★★⯨	3.5 stars	**X-Men (2000)** Action, Adventure, Sci-Fi

next >

To get a new set of movies click the **next>** link.



m o v i e l e n s
helping you find the *right* movies

Welcome fricci@unibz.it (Log Out)
You've rated 47 movies.
You're the 18th visitor in the past hour.

★★★★★ = Must See
★★★★☆ = Will Enjoy
★★★☆☆ = It's OK
★★☆☆☆ = Fairly Bad
★☆☆☆☆ = Awful

Home | Find Movies | Discussion Forums | Preferences | Help

Shortcuts | Search

Basic Search

Title:

All Genres | All Dates

Domain: All movies

Tag:

☐ Use selected buddies!

☑ Exclude your ratings

☑ Exclude movies without predictions

Search!

Select Buddies

☐ Test Buddy

What are buddies?

Advanced Search

There are **9089** movies matching your search:
Movies without a prediction are **Not Shown**
Movies you've rated are **Not Shown**
You've sorted by: **Prediction**

Show Printer-Friendly Page | Download Results | Suggest a Title

Tags Related to Your Search: **classic (516), 70mm (439), action (419), comedy (397), dvd (332), (about tags)**

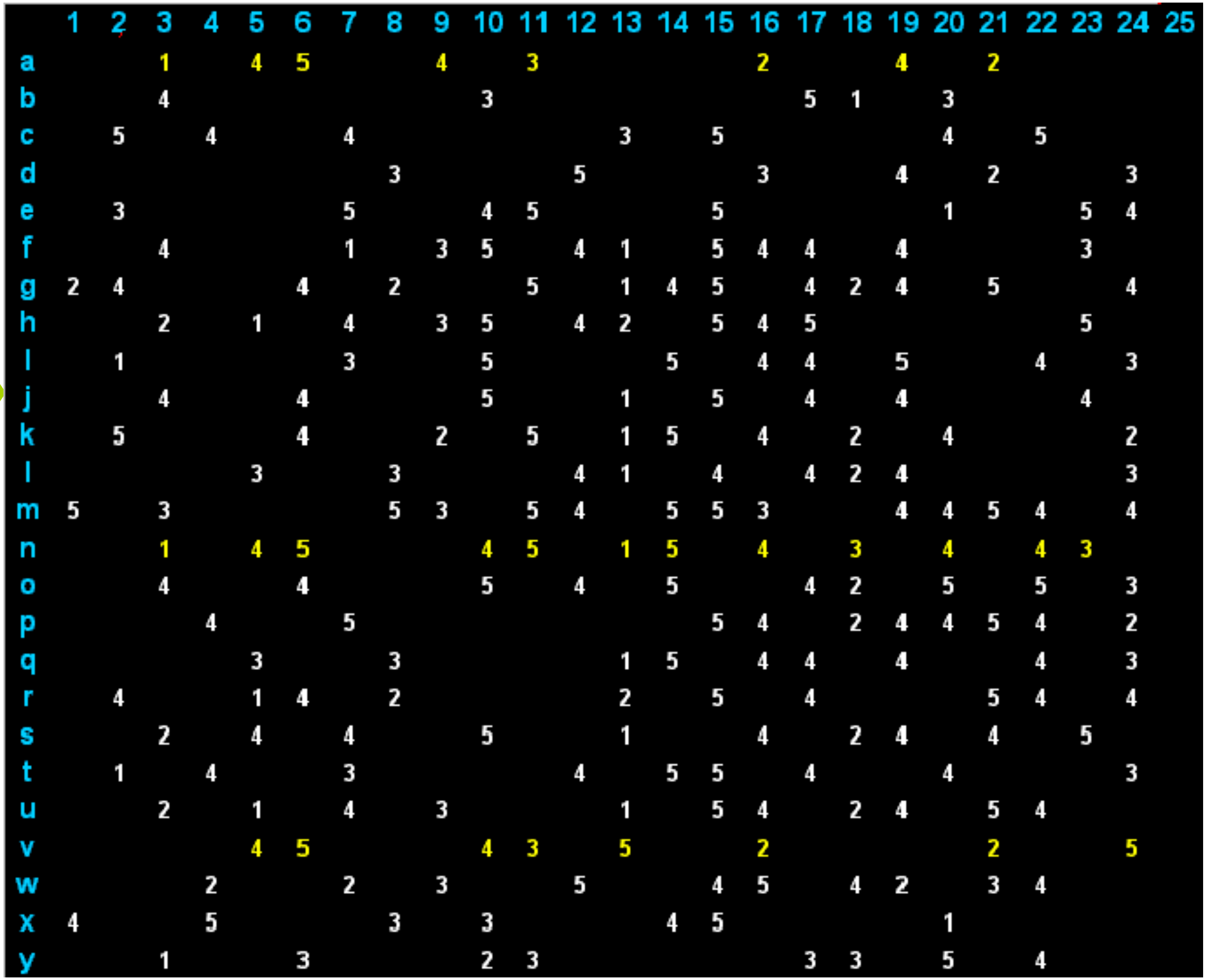
Page **1** of **606** | 1 2 3 4 ... 606 | next | Skip to page #: Go

Predictions for you	Your Ratings	Movie Information	Wish List
★★★★★	Not seen	**Yojimbo (1961)** DVD VHS info | imdb Action, Crime, Drama - Japanese [add tag] Popular tags: Toshiro Mifune | Japan | Best Performance: Toshiro Mifune as Sanjuro Kuwabatake	☐
★★★★★	Not seen	**Lives of Others, The (Das Leben der Anderen) (2006)** DVD info | imdb Drama - German [add tag] Popular tags: ClearPlay | toplist07 | Germany	☐
★★★★★	Not seen	**Third Man, The (1949)** DVD VHS info | imdb Film-Noir, Mystery, Thriller [add tag] Popular tags: Oscar (Best Cinematography) | AFI #57 | vienna	☐
★★★★★	Not seen	**Fog of War: Eleven Lessons from the Life of Robert S. McNamara, The (2003)** DVD VHS info | imdb	☐

Matrix of ratings

Items

Users



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
a			1		4	5			4		3					2			4		2				
b			4							3							5	1		3					
c		5		4			4						3		5					4		5			
d								3				5				3			4		2			3	
e		3					5			4	5				5					1			5	4	
f			4				1		3	5		4	1		5	4	4		4				3		
g	2	4				4		2			5		1	4	5		4	2	4		5			4	
h			2		1		4		3	5		4	2		5	4	5						5		
I		1					3			5				5		4	4		5			4		3	
j			4			4				5			1		5		4		4				4		
k		5				4			2		5		1	5		4		2		4				2	
l					3			3				4	1		4		4	2	4					3	
m	5		3					5	3		5	4		5	5	3			4	4	5	4		4	
n			1		4	5				4	5		1	5		4		3		4		4	3		
o			4			4				5		4		5			4	2		5		5		3	
p				4			5								5	4		2	4	4	5	4		2	
q					3			3					1	5		4	4		4			4		3	
r		4			1	4		2					2		5		4				5	4		4	
s			2		4		4			5			1			4		2	4		4		5		
t		1		4			3					4		5	5		4			4				3	
u			2		1		4		3				1		5	4		2	4		5	4			
v					4	5				4	3		5			2					2			5	
w				2			2		3			5			4	5		4	2		3	4			
x	4			5				3		3				4	5					1					
y			1			3				2	3						3	3		5		4			

Collaborative-Based Filtering

- A collection of n user u_i and a collection of m products p_j
- A $n \times m$ matrix of ratings v_{ij}, with $v_{ij} = ?$ if user i did not rate product j
- Prediction for user *i* and product *j* is computed as

$$v_{ij}^{*} = v_i + K \sum_{v_{kj} \neq ?} u_{ik} (v_{kj} - v_k)$$

- Where, v_i is the average rating of user i, K is a normalization factor such that the sum of u_{ik} is 1, and

$$u_{ik} = \frac{\sum_j (v_{ij} - v_i)(v_{kj} - v_k)}{\sqrt{\sum_j (v_{ij} - v_i)^2 \sum_j (v_{kj} - v_k)^2}}$$

Similarity of users i and k

- Where the sum (and averages) is over j s.t. v_{ij} and v_{kj} are not "?".

[Breese et al., 1998]

Collaborative Filtering and Google

- Search engines are not recommender systems, BUT
- Actually Google and Collaborative Filtering have **many similarities**
  - They both **rank** items
  - The ranking is based on **opinion of their users**
    - Collaborative Filtering: ratings on items
    - Google: links to pages
  - Both are expressions of the Web 2.0
- **Web 2.0:** involves the user
  - the content is created by users
  - users help organize it, share it, remix it, critique it, update it.

How Google Ranks Tweets

- Tweets: 140-character microblog posts sent out by Twitter members
- The key is to identify "reputed followers," -Twitterers "follow" the comments of other Twitterers they've selected, and are themselves "followed."
- You earn reputation, and then you give reputation
- If lots of people follow you, and then you follow someone-- then even though this [new person] does not have lots of followers, his tweet is deemed valuable
- One user following another in social media is analogous to one page linking to another on the Web. Both are a form of recommendation ...

example http://www.technologyreview.com/web/24353/

Recommender Systems vs Search Engines I

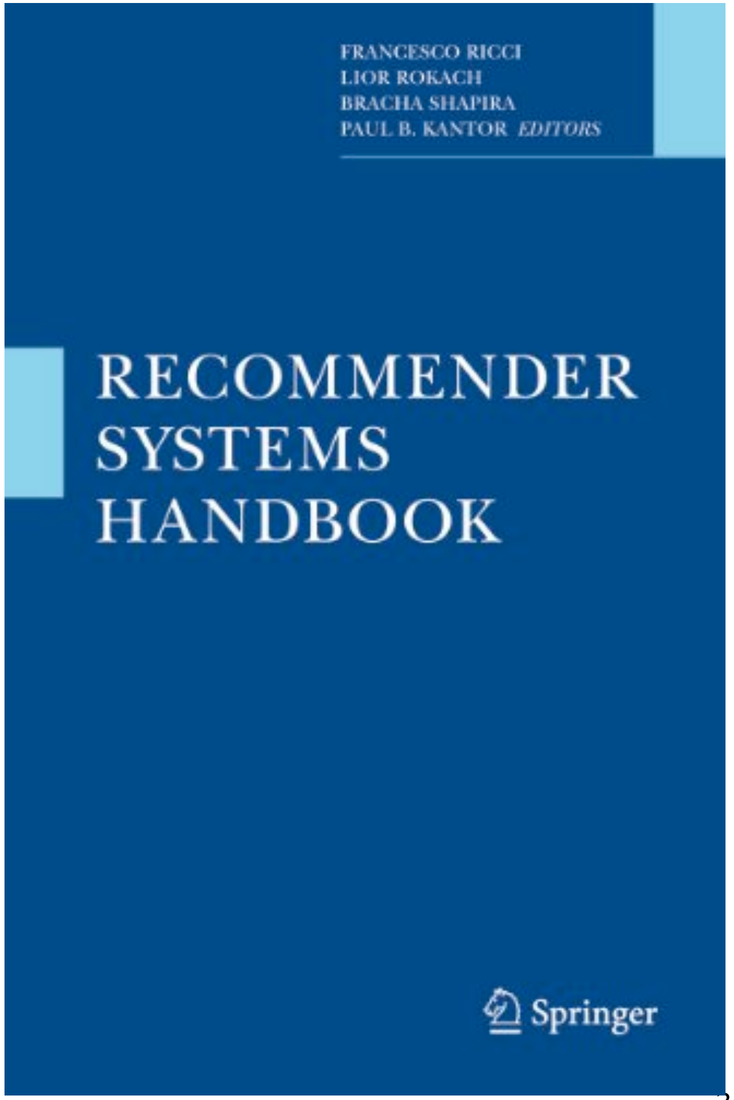
- Recommender system research has taken techniques from IR (e.g. content-based filtering)
- Search engines have used idea coming from recommender systems (a page is important is linked/endorsed by another)
- **IR** deals with **large repositories of unstructured content about a large variety of topics**
- **RSs** focus on **smaller** content repositories on a **single topic**
- **Personalization** in IR (personalized search engines) did not received much interests (e.g. personalized google) – but now could revamp because of recent research on **learning to rank.**

Recommender Systems vs Search Engines II

- IR deals with "**locating relevant content**" – the user should be able to evaluate the relevance of the retrieved set
- RS deals with "**differentiating relevant content**" – the user has not enough knowledge to evaluate relevance
  - E.g. imagine to select a camera with google and with dpreview.com
- IR and RS supports different stages of the information search/discovery process
- An effective information system must blend techniques coming from the two areas.

Topics in Recommender Systems

- Prediction Algorithms
- Evaluation methodologies
- System deployment and integration
- Method selection
- Conversational systems
- Persuasion
- Recommendation presentation and explanations
- Social computing
- Trust
- Preference elicitation and active learning
- Robustness and security



Challenges in Recommender Systems

- Scalability of the algorithms with large and real-word data sets
- Proactive recommenders
- Privacy preserving recommenders
- Diversity of the recommendations
- Integration of short- and long-term preferences
- Generic user models and cross domain solutions
- Distributed models
- Recommending a sequence of items (e.g. a playlist)
- Recommender for mobile users
- Recommendations for groups
- Context-Aware Recommendations

References

- C. D. Manning, P. Raghavan and H. Schutze. Introduction to Information Retrieval, Cambridge University Press, 2008.
- M. Levene. An Introduction to Search Engines and Web Navigation. Wiley, 2010.
- S. Buttcher, C. Clarke, G. Cormack. Information Retrieval, MIT Press, 2010.
- M. Hearst, Search User Interfaces, Cambridge Univ. Press, 2009.
- Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. Recommender Systems Handbook, Springer, 2011.