Information Search and Retrieval

Academic Year: 2010-2011

Francesco Ricci
Free University of Bozen-Bolzano
Contact Details

- Francesco Ricci
  - Room 204 (POS)
  - fricci@unibz.it
- Availability Hours: Tue 16.00-18.00
  - by prior arrangement via e-mail
- Course web site
  - http://www.inf.unibz.it/~ricci/ISR/
Course Structure

- **Lectures**: 48 hours
- **Labs**: 24 hours
- **Timetable**:
  - Lectures: Mon 14:00-16:00; Tue 14:00-16:00 - Room E411
  - Labs: Mon 16.00-18.00 - Computer Room E431
- **Assessment**:
  - final exam, written, 50 % of the grade
  - project in a small "team" (1 or 2 students) 50%.
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.**
Syllabus

- Basic information retrieval concepts
- Boolean retrieval
- Indexing
- Vector space model
- Text and vector space classification
- Evaluation in information retrieval
- Recommender systems
- Collaborative- and Content-based filtering
- Hybrid recommender systems
- Knowledge based recommenders
- Conversational recommender systems
- Evaluation of recommender systems
- Human Computer Interaction and recommender systems
- Context-dependent recommender systems
- Decision making
- Web search and link analysis
- Ranking and machine learning on documents
What you should learn

- The **scientific** underpinnings of the field of Information Search and Retrieval
- A catalogues of information search and discovery **techniques and tools** that can be exploited in the design and implementation of a specific Web site (eCommerce, eGovernment)
- The **pros** and **cons** of different techniques
- To reason about the **benefits** and limitations of the techniques and systems for the various actors involved in the process
- The capability to **decide** when (in which context, for what kind of products or services) a technique can be **useful** or not
- To identify **new applications** of the techniques.
Course Format

- **24 Lectures** on various topics in Information Search and Recommender Systems
- **12 Labs** where we shall
  - look to further details about the topics illustrated in the lectures
  - Solve some exercises (you try first!)
  - Experiment the usage of some systems (evaluate it)
  - Discuss together some papers on RS and information search
  - Work on the project that you must prepare to pass the exam.
Learning Material

- For the **information retrieval** topics

- Research on **Recommender Systems** is quite new (first works presented around '95) - a recent collection of review papers is:

- *You must learn from papers not only the slides!*
Project

- The project is conducted in **small groups** (2 students)
- **Design** an information search and advisory system in a given application scenario (e.g. bikes, courses, events, ski, eGovernment, group travel, etc.)
- **Choose** the topic and scenario you like the best
- The project results include:
  - a written **report** approximately 5.000 words
  - and optionally a system **prototype**
- The **report** must describe the proposed system and should be targeted to the **IS provider**
- It is not required to fully implement the proposed **system**, just focus on the core functionality and provide a user interface for it.
Structure of the report

- Executive summary
- Description of the application problem
- Survey of existing information search applications and studies on the chosen application domain (quote at least 3-4 specific papers)
- Critical evaluation/comparison of the pros and cons of the techniques presented in the course that could be applied to the selected problem
- Description of the proposed system functions
- Description of the core techniques used in the prototype and how have been applied to this specific example
- Advantages for the customer.
How the project will be evaluated

- The report **must follow the defined structure** (see a previous slide)
- The report must be **clearly written**
- The proposed functions and techniques must be **significant** and **sound**
- The report must show that you have **deeply investigated** the problem and you can **convince** the client
- The **system** should be enough developed to show some of the potential benefits for the users
- The presentation must be **understandable and raise the audience attention**
- The presenters must be **able to reply** to the questions of the other participants.
What a student must do to pass

- Read the papers/chapters that will be suggested for each lecture
- The slides should be enough for a general understanding of the topic
- If something is not clear during a lecture you must take a note and rise a question (especially in the labs)
- The team should start working on the project as soon as possible
  - To have enough time to prepare the report
  - To submit to me a draft that I will review
  - To be able to reason on the application of the techniques as they are presented in the course.
Exam

- The final grade is obtained evaluating the project result and the knowledge acquired about the lectures’ topics in a written exam.

- **Written exam:** questions similar to those I make during the lectures (see slides) and to the exercises in the lab.

- The project results will be presented in a seminar (20 mins for each presentation).

- The final written report and (eventually) the link to the system prototype must be sent to me the same date of the presentation.

- You will have two grades: P (project), max 16 points, and W (written exam), max 16 points.

- The final grade is \( F = \max\{P + W, 31\} \).

- Both P and W must be greater or equal to 9.
Introduction to Information Retrieval and Recommender Systems
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**
  - **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

- 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.

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.

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/Slocate/adviceengine.html

SciRus - for scientific information
SciRus is the most comprehensive science specific search engine on the Internet. ... patents and institutional repository and website information. ...
www.sciirus.com/

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

Search Tools - Enterprise Search Engines - Information, Guides and...
SearchTools reports on web sites, intranet and portal search tools, providing news about local
Brief History of Search Engines

- Bing ([www.bing.com](http://www.bing.com)) – (2009-) Microsoft rebarded search engine, was Live in 2006 and MSN search before.
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

<table>
<thead>
<tr>
<th>Domain</th>
<th>August 2010</th>
<th>September 2010</th>
<th>Month-over-Month Point Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Sites</td>
<td>65.4</td>
<td>66.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Yahoo Sites</td>
<td>17.4</td>
<td>16.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>11.1</td>
<td>11.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Ask Network</td>
<td>3.8</td>
<td>3.7</td>
<td>-0.1</td>
</tr>
<tr>
<td>AOL Network</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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

Encoded by the user into a 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%):
     - $q = \text{compaq}$ - 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%)
     - $q = \text{canon 5d mkII}$ - probable target a page reviewing canon 5d mkII

  3. **Transactional**: The intent is to perform some web-mediated activity (30%)
     - $q = \text{hotel Vienna}$ - probable target "Expedia"
Exploratory Search

- 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

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

[Church and Smyth, 2008]
Exploratory Search: Example

![Diagram with various music bands and relationships]

Search: **hot chili**
Information Discovery: Example

http://musicover.com
Exploratory Search: People

hannes werthner

Did you mean: hannes werner

Posts by Friends

Dimitrios Buhalis
Hannes Werthner
ENTER2010 @ Lugano

Sat at 2:33pm · Comment · Like · Share

Web Results

Home Page von Hannes Werthner
Prof. Hannes Werthner Electronic Commerce Group Institute for Software Technology and Interactive Systems Vienna University of Technology
www.ec.tuwien.ac.at/~werthner

Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner, Univ.Prof. | EC
Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner, Univ.Prof. E.Mail: werthner [at] ec.tuwien.ac.at; Website: http://www.ec.tuwien.ac.at/~werthner; Phone: www.ec.tuwien.ac.at/~hannes....

Hannes Werthner - Tetherless World Wiki
This page was last modified on 5 March 2008, at 02:33. This page has been accessed 93 times. Privacy policy; About Tetherless World Wiki; Disclaimers
wv.rpi.edu/wiki/Hannes_Wert...
AIDS (Acquired Immune Deficiency Syndrome) is the final and most serious stage of HIV disease, which causes severe damage to the immune system. According to the Centers for Disease Control and Prevention, AIDS begins when a person with HIV infection has a CD4 cell count below 200. (CD4 is also called “T-cell”, a type of immune cell.) AIDS is also defined by numerous opportunistic infections and cancers that occur in the presence of HIV infection.
Dynamic Search Engine

Donau Oberösterreich
A warm welcome! 170 km of the 2,860 km long Danube flow through Upper Austria. Only a short distance but nevertheless a breathtakingly beautiful one. You can cycle along the sparkling blue waters from Passau to Ybbs and explore the many fascinating aspects of this unique region where some places look back on 1000, 2000 and even 3000 years of history! In a healthy ambiance you can find a variety of possibilities for recuperation and relaxation. Whether you wish to go cycling, hiking, horseback riding, play tennis or golf, or enjoy a leisurely boat-trip you will find exactly what you are looking for. For your gastronomic treats traditional delicacies are offered in our family friendly farm-hotels, first class hotels etc. The D... Donau Oberösterreich - 27.0 - Similar resorts

Bromberg
Bromberg is inviting its visitors to days of relaxation and recreation. Simply spare-time that you can enjoy according to your mood and desires. The Lambert Church, landmark of Bromberg, is greeting from the slope, decorated with a huge “Christophorus” fresco dating back to the 14th century. Indeed, one of the most marketable and beautiful landmarks of the Bucklige Welt area. Downhill, in the centre the old parish courtyard (built in the 15th century) is welcoming you. Another sight, worth seeing is the marvellous courtyard with its arcades and the Florian chapel, nestled close to it. After the large earth quake in 1972, remarkable, magnificent frescoes dating back to the 16th century have been discovered. On widely spread hiking and walking... Bromberg - 27.0 - Similar resorts
Yotify

- Yotify is designed to make a shopping search (e.g., for an apartment) **persistent**
- Search runs at regular intervals (e.g. daily) with results sent back to the user via e-mail
- Yotify asks partner sites (e.g. craigslist, or shopping.com) to integrate its software into their systems

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

- Information Retrieval
- Recommender System
- Product Search
- Decision Support

Item Complexity

News, Article, webpage
Music, DVD, Book
Laptop, Camera, Travel
Investment, Real Estate, Politics

User involvement increases

Risk (Price)

low

high

low

high

Constraints
CP-Nets
MAUT
Decision Strategies
Critiquing
Preference Elicitation
Collaborative Filtering
Data Mining
Keyword-based search
PageRank

Preference Elicitation
Collaborative Filtering
Data Mining
Keyword-based search
PageRank

Collaborative Filtering
Critiquing
Decision Strategies
Input
Output
Constraints
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

A search engine is not a recommender system

Querying a SE for a recommendation will return a list of recommender systems
Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to see all recommendations.

- Artificial Intelligence: A Modern Approach (Hardcover) by Stuart Russell
  - Rating: 4.5 out of 5 stars (18) - $113.77

- Introduction to Algorithms (Hardcover) by Thomas H. Cormen
  - Rating: 4.5 out of 5 stars (22) - $98.00

- Operating Systems: Internals and Design Principles (Hardcover) by William S. Stevens
  - Rating: 4.3 out of 5 stars (64) - $118.05

- Database System Concepts (Hardcover) by Abraham Silberschatz
  - Rating: 4.3 out of 5 stars (25) - $111.22

New For You

- Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition) (Hardcover) by Ian H. Witten
  - Rating: 4.4 out of 5 stars (6) - $44.37

- Speedlight's Handbook: Learn... (Paperback) by Syl Arena
  - Rating: 4.3 out of 5 stars (99) - $30.85

- Strobist Photo Tricks & Secrets... (Paperback) by Zeke Kamm
  - Rating: 4.3 out of 5 stars (9) - $14.33

- The Photoshop Elements 9 Book (Paperback) by Scott Kelby
  - Rating: 4.3 out of 5 stars (19) - $28.15

From Your Wish List

- Consumer Behavior
  - Rating: 4.5 out of 5 stars (17) - $20.00

- Advanced Photoshop Techniques
  - Rating: 4.5 out of 5 stars (13) - $30.00

- Mobile Media and Applications
  - Rating: 4.5 out of 5 stars (9) - $25.00
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.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Movie Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>⭐⭐⭐⭐⭐ 3.0 stars</td>
<td>Austin Powers: International Man of Mystery (1997) Action, Adventure, Comedy</td>
</tr>
<tr>
<td>⭐⭐⭐⭐⭐ 4.0 stars</td>
<td>Contact (1997) Drama, Sci-Fi</td>
</tr>
<tr>
<td>??? Not seen</td>
<td>Demolition Man (1993) Action, Comedy, Sci-Fi</td>
</tr>
<tr>
<td>??? Not seen</td>
<td>Eraser (1996) Action, Drama, Thriller</td>
</tr>
<tr>
<td>??? Not seen</td>
<td>Maverick (1994) Action, Comedy, Western</td>
</tr>
<tr>
<td>⭐⭐⭐⭐ 4.5 stars</td>
<td>Philadelphia (1993) Drama</td>
</tr>
<tr>
<td>⭐⭐⭐⭐ 3.5 stars</td>
<td>Piano, The (1993) Drama, Romance</td>
</tr>
<tr>
<td>??? Not seen</td>
<td>Toy Story 2 (1999) Adventure, Animation, Children, Comedy, Fantasy</td>
</tr>
</tbody>
</table>

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

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

Users:

- a
- b
- c
- d
- e
- f
- g
- h
- i
- j
- k
- l
- m
- n
- o
- p
- q
- r
- s
- t
- u
- v
- w
- x
- y

Items:

- 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
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
  \[
  u_{ik} = \frac{\sum_j (v_{ij} - v_i)(v_{kj} - v_k)}{\sqrt{\sum_j (v_{ij} - v_i)^2 \sum_j (v_{kj} - v_k)^2}}
  \]
- Where the sum (and averages) is over $j$ s.t. $v_{ij}$ and $v_{kj}$ are not “?”. [Breese et al., 1998]
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
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.
Google

- **Google** is the leading search and online advertising company - founded by Larry Page and Sergey Brin (Ph.D. students at Stanford University)
- “googol” or $10^{100}$ is the mathematical term Google was named after
- Google’s success in search is largely based on its **PageRank™** algorithm
- Gartner reckons that Google now make use of more than 1 million servers, spitting out search results, images, videos, emails and ads
- Google reports that it spends some 200 to 250 million US dollars a year on IT equipment.
Ranking web pages

- To count *inlinks*: http://siteexplorer.search.yahoo.com
- Web pages are not equally “important”
  - www.unibz.it vs. www.stanford.edu
  - Inlinks as votes
    - www.stanford.edu has 743,482 inlinks
    - www.unibz.it has 4,989 inlink (Feb 2010)
- Are all *inlinks* equal?
  - Recursive question!
Simple recursive formulation

- Each link’s vote is **proportional** to the importance of its **source** page
- If page $P$ with importance $x$ has $n$ outlinks, each link gets $x/n$ votes
Simple “flow” model

The web in 1839

\[
y = \frac{y}{2} + \frac{a}{2}
\]
\[
a = \frac{y}{2} + m
\]
\[
m = \frac{a}{2}
\]
Solving the flow equations

- 3 equations, 3 unknowns, no constants
  - No unique solution
  - All solutions equivalent modulo scale factor
- Additional constraint forces uniqueness
  - \( y + a + m = 1 \) (normalization)
  - \( y = 2/5, a = 2/5, m = 1/5 \)
- Gaussian elimination method works for small examples, but we need a better method for large graphs.
Matrix formulation

- Matrix $\mathbf{M}$ has one row and one column for each web page (square matrix).
- Suppose page $i$ has $n$ outlinks:
  - If $i$ links to $j$, then $M_{ij} = 1/n$
  - Else $M_{ij} = 0$
- $\mathbf{M}$ is a row stochastic matrix:
  - Rows sum to 1
- Suppose $\mathbf{r}$ is a vector with one entry per web page:
  - $r_i$ is the importance score of page $i$
  - Call it the rank vector
Example

\[
\begin{align*}
  y & = y/2 + a/2 \\
  a & = y/2 + m \\
  m & = a/2
\end{align*}
\]

\[
\begin{bmatrix}
  y & a & m \\
  y & 1/2 & 1/2 & 0 \\
  a & 1/2 & 0 & 1/2 \\
  m & 0 & 1 & 0
\end{bmatrix}
\]

\[(y \ a \ m) = (y \ a \ m)M\]
Power Iteration Example

\[
\begin{align*}
\text{Yahoo} & \rightarrow \text{Amazon} & \rightarrow \text{Microsoft} \\
& & \\
\begin{pmatrix}
\begin{array}{c}
y \\
a \\
m
\end{array}
\end{pmatrix} &= 
\begin{pmatrix}
\begin{array}{ccc}
\frac{5}{12} & \frac{3}{8} & 2/5 \\
\frac{1}{2} & 0 & \frac{1}{2} \\
0 & 1 & 0
\end{array}
\end{pmatrix}
\begin{pmatrix}
\begin{array}{c}
y \\
a \\
m
\end{array}
\end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
&= \begin{pmatrix}
\frac{1}{3} & \frac{1}{3} & 1/6 & 1/4 & 1/6 & \ldots & 1/5
\end{pmatrix}
\begin{pmatrix}
\begin{array}{c}
y \\
a \\
m
\end{array}
\end{pmatrix}
\end{align*}
\]

\[(y \ a \ m) = (y \ a \ m)M \ M \ M \ M \ M \ \ldots \ M\]
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 ...

Real time  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.
Vertical search engines and LBS

- **Vertical search engines** are specialists (focusing on specific topics) in comparison to generalists (e.g., Google and Yahoo!)
  - Health and medicine: medstory.com
  - Travel sites: Kayak.com or Expedia.com
  - Real-estate: Zillow.com or Trulia.com (exploit location based search)
  - Job search: Indeed.com or Monster.com
  - Shopping search engines: Shopzilla.com and MySimon.com

- **Location-based search** uses geographic information about the searcher to provide more relevant search results.
Dynamic Search Engine

Donau Oberösterreich

A warm welcome! 170 km of the 2,860 km long Danube flow through Upper Austria. Only a short distance but nevertheless a breathtakingly beautiful one. You can cycle along the sparkling blue waters from Passau to Ybbs and explore the many fascinating aspects of this unique region where some places look back on 1000, 2000 and even 3000 years of history! In a healthy ambiance you can find a variety of possibilities for recuperation and relaxation. Whether you wish to go cycling, hiking, horseback riding, play tennis or golf, or enjoy a leisurely boat-trip you will find exactly what you are looking for. For your gastronomic treats traditional delicacies are offered in our family friendly farm-hotels, first class hotels etc. The D... Donau Oberösterreich - 27.0 - Similar resorts

Bromberg

Bromberg is inviting its visitors to days of relaxation and recreation. Simply spare-time that you can enjoy according to your mood and desires. The Lambert Church, landmark of Bromberg, is greeting from the slope, decorated with a huge “Christophorus” fresco dating back to the 14th century. Indeed, one of the most marketable and beautiful landmarks of the Bucklige Welt area. Downhill, in the centre the old parish courtyard (built in the 15th century) is welcoming you. Another sight, worth seeing is the marvellous courtyard with its arcades and the Florian chapel, nestled close to it. After the large earth quake in 1972, remarkable, magnificent frescos dating back to the 16th century have been discovered. On widely spread hiking and walking...

Bromberg - 27.0 - Similar resorts
Results for **retinoblastoma**

**Information that Matters™: click below to refine your search** | View More...

**Drugs & Substances**
- Cev Protocol
- Oncovin
- Paraplatin
- Topophos
- Cisplatin

**Conditions**
- **Retinoblastoma**
- Eys Cancer
- Squamous Cell Skin
- Lung Cancer
- Leukemia

**Procedures**
- Chemotherapy
- Radiation Therapy
- Tumor Markers
- External Beam Rad...
- Cryotherapy

**Personal Health**
- Genetic Predispos...
- Smoking
- Family History
- Aging

**People**
- Rodriguez-Galindo...
- Dunkel, Jra
- Kuchar, Brian H
- Perentesis, John P
- Villablanca, Judi...

---

**The Web 1 to 10 of about 1,030,000**

1. **Retinoblastoma International: Homepage**
   Information about the disease aimed at parents and professionals. Lobbies for early eye exams in newborns.
   http://www.retinoblastoma.net/

2. **Retinoblastoma Treatment - National Cancer Institute**
   Retinoblastoma is a disease in which malignant (cancer) cells form in the tissues of the retina. ...
   After diagnosis of retinoblastoma in one eye, regular follow-up exams of the ...
Same query in Google

Retinoblastoma Resource
www.Retinoblastoma.net Find Information, Research, Videos, Fundraising, Counseling & More

Refine results for retinoblastoma:
- Treatment
- Tests/diagnosis
- Symptoms
- For patients
- Causes/risk factors
- From medical authorities
- For health professionals
- Alternative medicine

Retinoblastoma - Wikipedia, the free encyclopedia
Retinoblastoma is a cancer of the retina. Development of this tumor is initiated by mutations [1] that inactivate both copies of the RB1 gene, ...
en.wikipedia.org/wiki/Retinoblastoma - 40k - Cached - Similar pages - Note this

Retinoblastoma International Homepage
Retinoblastoma International (RBI) is a public charitable organization dedicated to funding research, clinical treatment and international awareness for ...
www.retinoblastoma.net - 18k - Cached - Similar pages - Note this

Retinoblastoma International: What Is Retinoblastoma?
Retinoblastoma (re-hi-nob-ble-sto-ma) is a childhood cancer arising from immature retinal cells in one or both eyes and can strike from the time a child is ...
www.retinoblastoma.net/whatist.rb.html - 21k - Cached - Similar pages - Note this

retinoblastoma
Click here for the Parent's Guide to Understanding Retinoblastoma - Para obtener una copia en Español de Entendiendo el Retinoblastoma, Una Guia para Padres ...
www.retinoblastoma.com - 1k - Cached - Similar pages - Note this
Same query in Ask

Retinoblastoma Treatment, Symptoms, Causes, and Recovery [Traduci questa pagina]
A set of questions and answers about retinoblastoma. What is retinoblastoma? Retinoblastoma is a cancer of the eye in children.
www.victtech.org/retinoblastoma.html • Copia cache • Salva

Retinoblastoma Innovations Finger’s Slotted Eye Plaque for Treatment of
Salva

Retinoblastoma (NRRSF) Information for parents by Peter Kaiser, MD, Ingrid Scott, MD, Joan O’Brien, MD and Timothy G. Murray, MD...
www.cancerindex.ccc/guide2r.htm • Copia cache • Salva

Childhood Eye Cancer Trust (CEC) [Fighting Retinoblastoma] [Traduci questa pagina]
Please update your bookmarks/favorites... The Retinoblastoma Society has changed its name...
www.rbsociety.org.uk • Copia cache • Salva
Reading Material