

Introduction to Information Retrieval and Recommender Systems



Francesco Ricci

Most of these slides comes from the course:
Information Retrieval and Web Search,
Christopher Manning and Prabhakar
Raghavan

What you should learn

- [J] The scientific underpinnings of the field of Information Search and Retrieval
- [J] 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)
- [J] The **pros** and **cons** of different techniques
- [J] To reason about the **benefits** and limitations of the techniques and systems for the various actors involved in the process
- [J] The capability to **decide** when (in which context, for what kind of products or services) a technique can be **useful** or not
- [J] To identify **new applications** of the techniques.

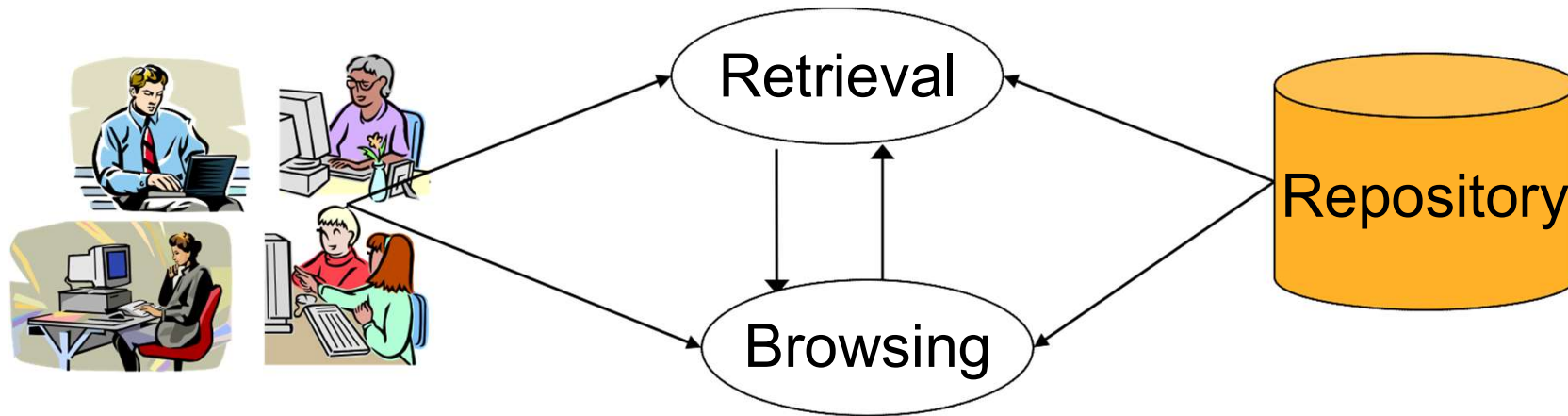
Basic Concepts in Information Retrieval

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

Retrieval and **Browse**



The User Task



[J] Retrieval

- Search for particular information
- Usually focused and purposeful

[J] Browsing

- General looking around for information
- For example: Asia-> Thailand -> Phuket -> Tsunami

Search Engines: Information Retrieval Tools

Web Images Maps News Video Gmail more ▼ Sign in

Google™ Search [Advanced Search](#)
[Preferences](#)

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

Web IR- IR on the Web

[J] **First Generation**

- Classical approach (boolean, vector, and probabilistic models)
- Informational: IR/DB techniques on page content. E.g., Lycos, Excite, AltaVista

[J] **Second Generation**

- Web as a graph
- Navigational: use off-page Web specific data – links topology. E.g., Google

[J] **Third Generation**

- Open research
- E.g. Mobile information search

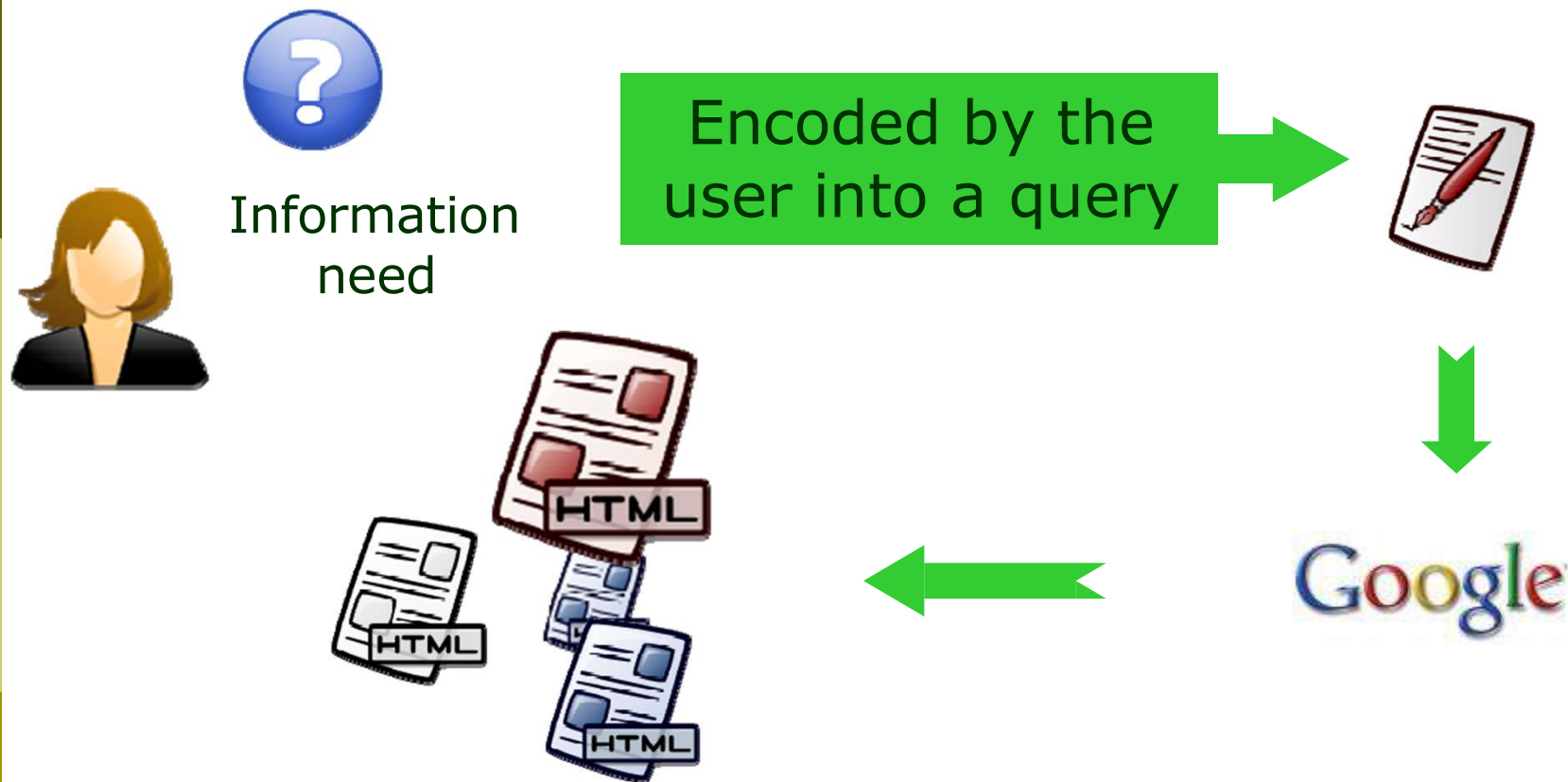
Problems with Using IR for Web

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

IR: The Basic Concepts

- [J] The user has an **information need**, that is expressed as a **free-text query**
- [J] 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]
- [J] The query **encodes** the information search need
- [J] The query **is a “document”**, to be compared to a collection of documents
- [J] **Effectiveness vs Efficiency**
- [J] How to **compare documents**? Similarity metrics needed!
- [J] How to **avoid** doing a **sequential search**?
Can we search in parallel in a set of servers?

From needs to queries



[J] Information need -> query -> search engine -> results -> browse OR query -> ...

Taxonomy of Web search

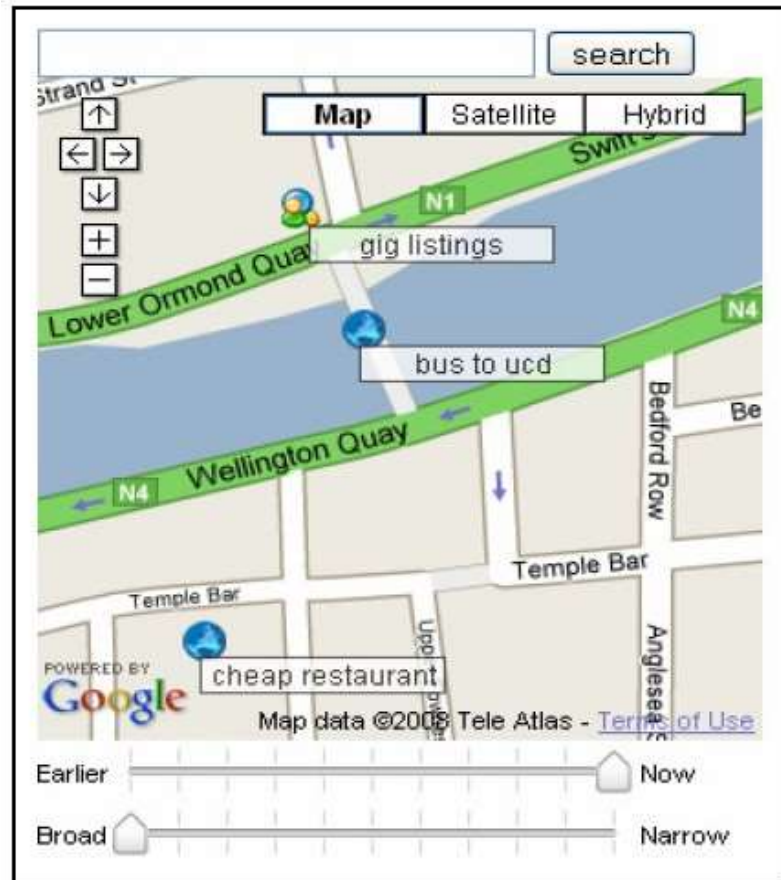
- [J] In the web context the "need behind the query" is often not informational in nature
- [J] [Broder, 2002] classifies web queries according to their intent into 3 classes:
 - 1. Navigational:** The immediate intent is to reach a particular site (20%):
 - [J] $q = compaq$ - probable target <http://www.compaq.com>
 - 2. Informational:** The intent is to acquire some information assumed to be present on one or more web pages (50%)
 - [J] $q = canon\ 5d\ mkII$ - probable target a [page](#) reviewing canon 5d mkII
 - 3. Transactional:** The intent is to perform some web-mediated activity (30%)
 - [J] $q = hotel\ Vienna$ - probable target "Expedia"





Strategies and Tools

- [J] A search engine is just a tool, among others, that can be exploited, within a strategy, to achieve a goal (perform a task)
- [J] New tools have emerged, and will be developed, to combine work in Human Computer Interaction and Information Retrieval
- [J] Exploratory search is the area where new tools will be developed mostly



Exploratory Search: Mobile Search



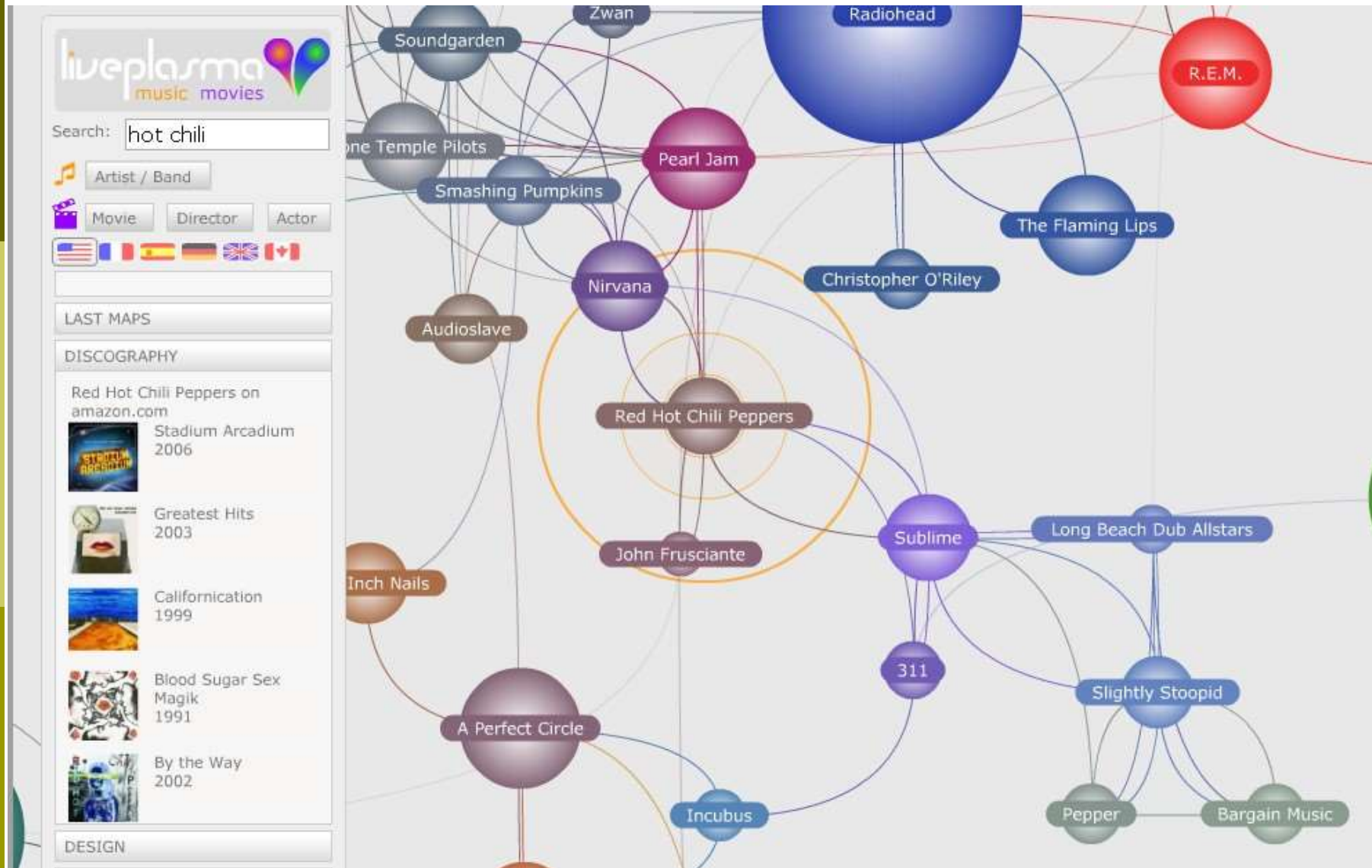
 1. Query	 2. Query with result- selections
 3. Query with comments	 4. Query with comments & result-selections

(b) Icons used to identify queries

[Church and Smyth, 2008]

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

Exploratory Search: Example



Information Search Features

- [J] There is **no single best strategy** or tool for finding information
- [J] 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

- [J] Information Search (IS) and Decision Making (DM) are strictly connected
- [J] **IS for DM:** we search information (external and internal) before taking decisions
 - Classical in DM and Consumer Behavior
- [J] **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.

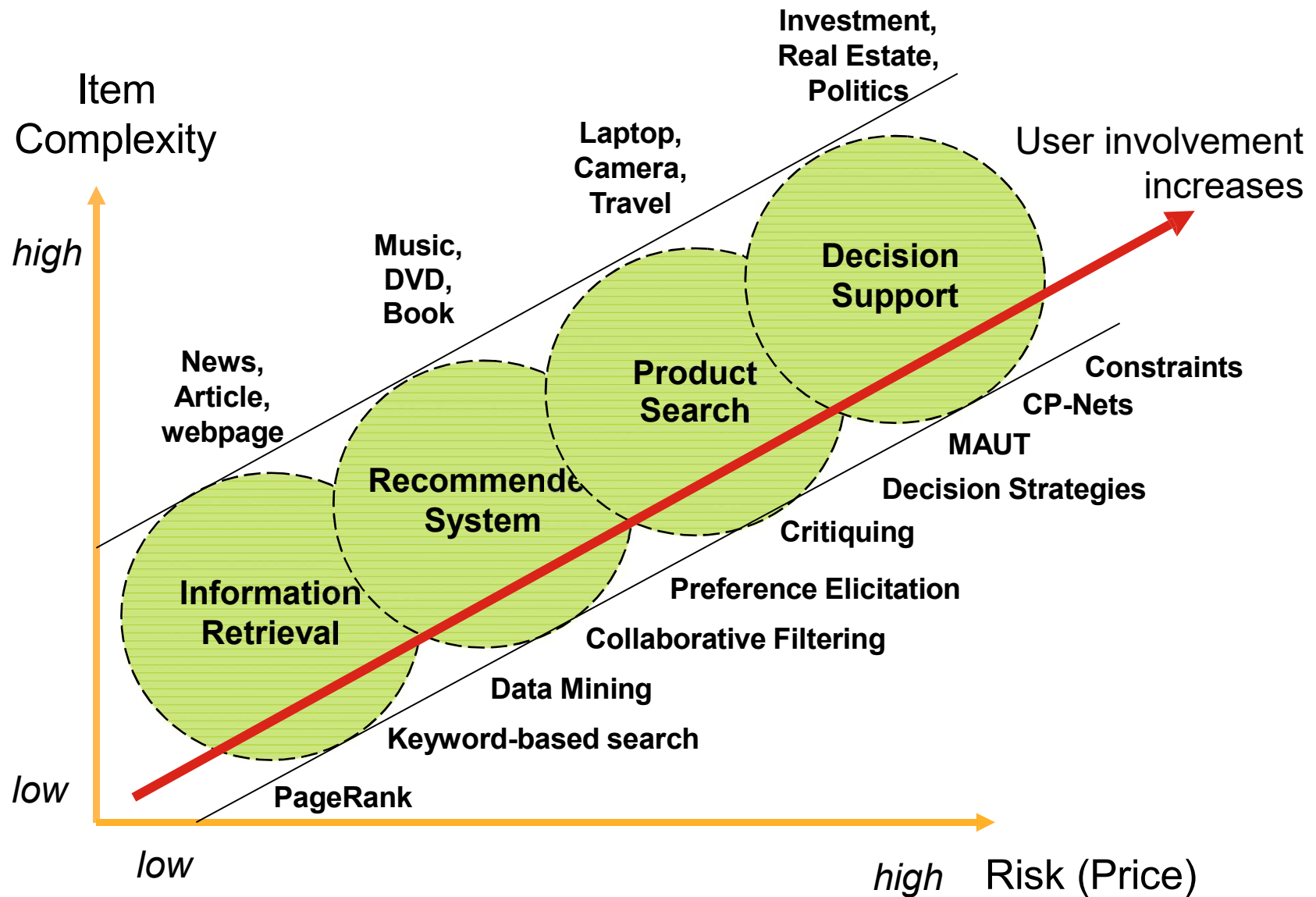


Information Overload



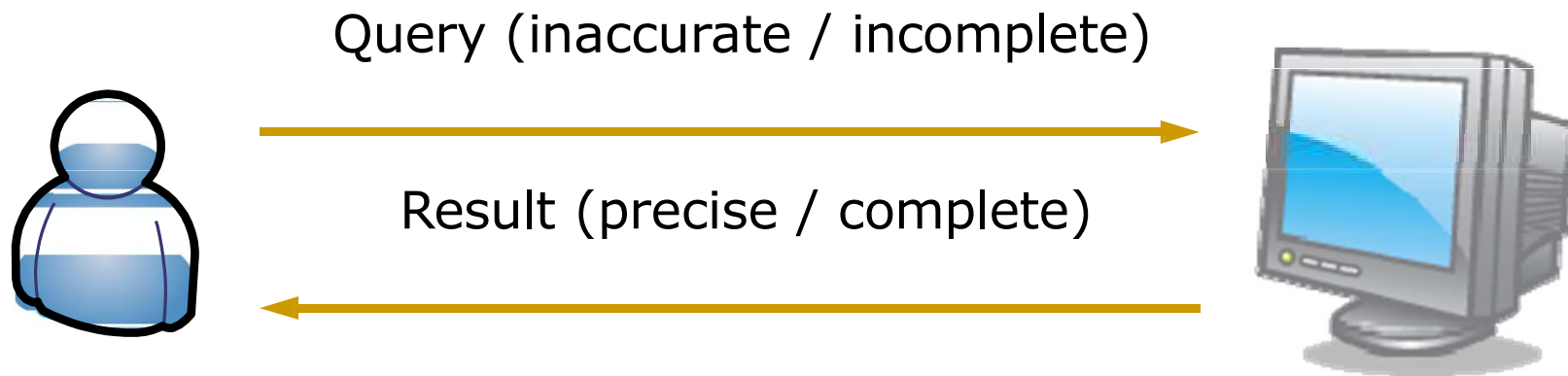
- [J] **Internet = information overload**, i.e., the state of having too much information to **make a decision** or **remain informed about a topic**
- [J] Information retrieval technologies can assist a user to **look up** content if the user knows exactly what he is looking for (i.e. for lookup)
- [J] 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

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



Recommender Systems

- [J] In everyday life **we rely on recommendations** from other people either by word of mouth, recommendation letters, movie and book reviews printed in newspapers ...
- [J] 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

Top 10 Digital Cameras

[Cameras.PCWWorld.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-cameras/ - 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-cameras/ - 106k - [Cached](#) - [Similar pages](#)

Digital Camera reviews - Best Reflex Camera

Digital photography BLOG, full reviews and articles about the **digital camera** world.

www.bestreflex.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/blcameraquiz.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**

Core Computations of Recommender Systems

- [J] **Rating Prediction:** a model must be built to predict ratings for items not currently rated by the user
 - **Numeric ratings:** regression
 - **Discrete ratings:** classification
- [J] **Ranking:** compute a score for each item and then rank the items with respect to the score (e.g. search engine)
- [J] **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

- [J] 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**
- [J] From an historical point of view CF came after content-based (we’ll see this later) but it is the most famous method
- [J] 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
- [J] 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.

Shortcuts

Search

Basic Search

Title:

All Genres All Dates

Domain:

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
★★★★★	<input type="text"/>	Yojimbo (1961) DVD VHS info imdb Action, Crime, Drama - Japanese	<input type="checkbox"/>
[add tag] Popular tags: Toshiro Mifune Japan Best Performance: Toshiro Mifune as Sanjuro Kuwabatake			
★★★★★	<input type="text"/>	Lives of Others, The (Das Leben der Anderen) (2006) DVD info imdb Drama - German	<input type="checkbox"/>
[add tag] Popular tags: ClearPlay toplist07 Germany			
★★★★★	<input type="text"/>	Third Man, The (1949) DVD VHS info imdb Film-Noir, Mystery, Thriller	<input type="checkbox"/>
[add tag] Popular tags: Oscar (Best Cinematography) AFI #57 vienna			
★★★★★	<input type="text"/>	Fog of War: Eleven Lessons from the Life of Robert S. McNamara, The (2003) DVD VHS info imdb	<input type="checkbox"/>

Matrix of ratings

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			

Items

Collaborative Filtering and Google

- [J] Search engines are not recommender systems, BUT
- [J] Actually Google and Collaborative Filtering have **many similarities**
 - They both **rank** items
 - The ranking is based on **opinion of their users**
 - [J] Collaborative Filtering: ratings on items
 - [J] Google: links to pages
 - Both are expressions of the Web 2.0
- [J] **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



1 googol = 1.0×10^{100}

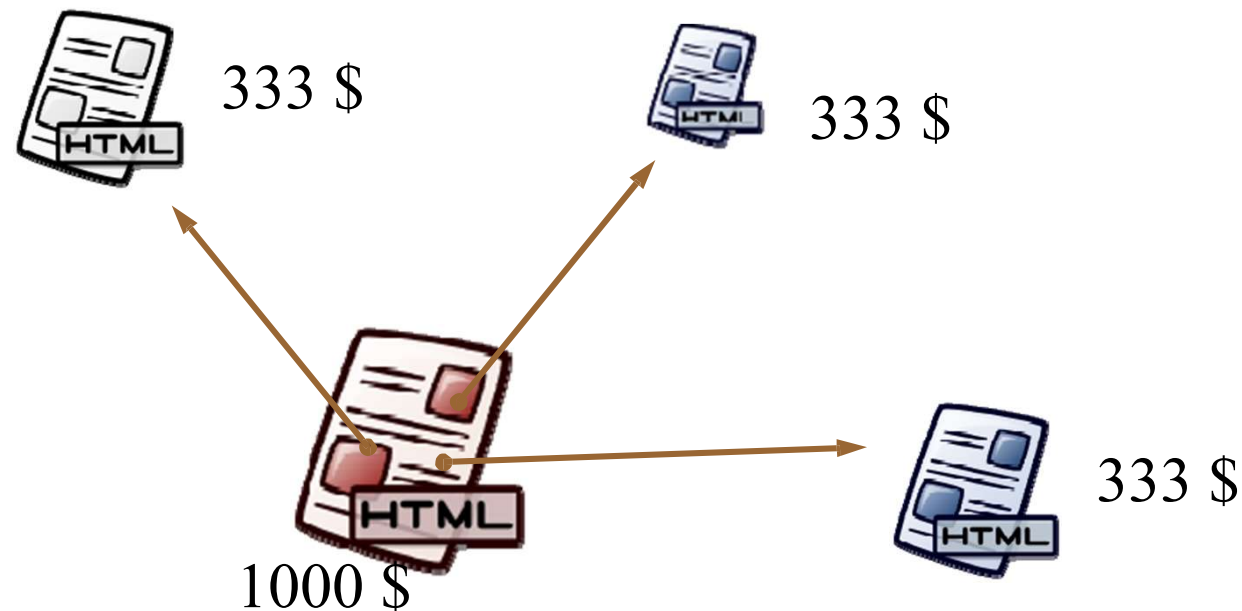
- [J] **Google** is the leading search and online advertising company - founded by Larry Page and Sergey Brin (Ph.D. students at Stanford University)
- [J] “googol” or 10^{100} is the mathematical term Google was named after
- [J] Google’s success in search is largely based on its **PageRank™** algorithm

Ranking web pages

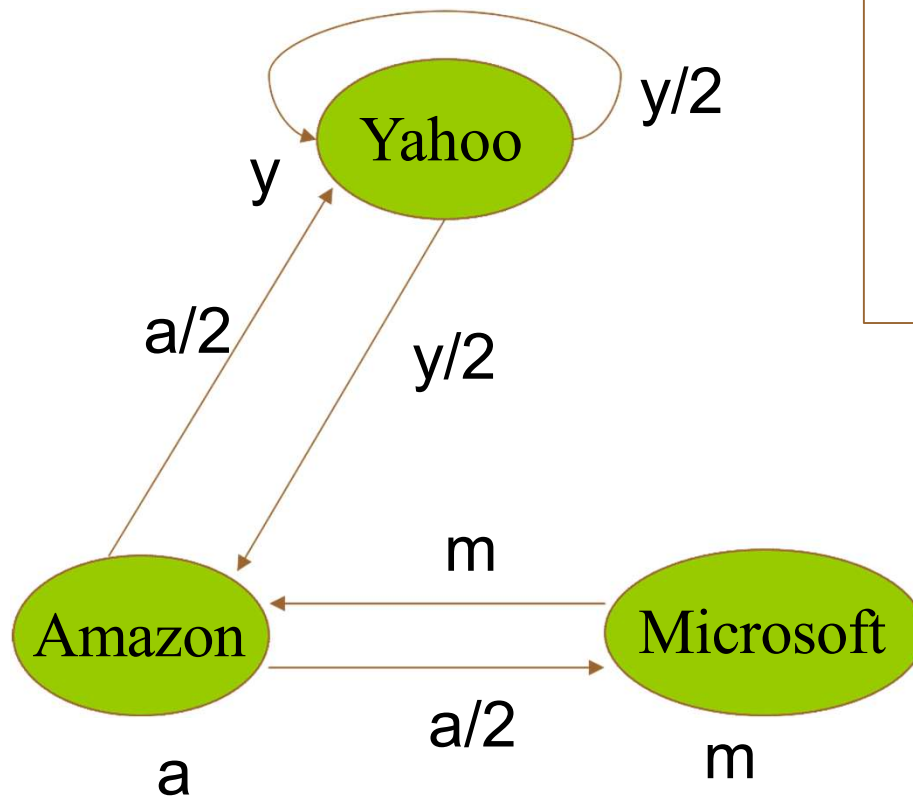
- [J] To count *inlinks*:
- [J] Web pages are not equally “important”
 - www.unibz.it vs. www.stanford.edu
 - Inlinks as votes
 - [J] www.stanford.edu has 743,482 inlinks
 - [J] www.unibz.it has 4,989 inlink
- [J] Are all *inlinks* equal?

Simple recursive formulation

- [J] Each link's vote is **proportional** to the importance of its **source** page
- [J] If page **P** with importance **x** has **n** outlinks, each link gets x/n votes



Simple “flow” model



$$\begin{aligned} y &= y/2 + a/2 \\ a &= y/2 + m \\ m &= a/2 \end{aligned}$$

Solving the flow equations

- [J] 3 equations, 3 unknowns, no constants
 - No unique solution
- [J] Additional constraint forces uniqueness
 - $y + a + m = 1$
 - $y = 2/5, a = 2/5, m = 1/5$
- [J] Gaussian elimination method works for small examples, but we need a better method for large graphs.

Recommender Systems vs Search Engines I

- [J] Recommender system research has taken techniques from IR (e.g. content-based filtering)
- [J] Search engines have used idea coming from recommender systems (a page is important is linked/endorsed by another)
- [J] **IR** deals with **large repositories of unstructured content about a large variety of topics**
- [J] **RSs** focus on **smaller** content repositories on a **single topic**
- [J] **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

- [J] IR deals with “**locating relevant content**” – the user should be able to evaluate the relevance of the retrieved set
- [J] 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
- [J] IR and RS supports different stages of the information search/discovery process
- [J] An effective information system must blend techniques coming from the two areas.

Vertical search engines and LBS


- [J] **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
- [J] **Location-based search** uses geographic information about the searcher to provide more relevant search results.

Results for **retinoblastoma**
 e-mail  del.icio.us


Health

Research






Information that Matters™: click below to refine your search | [View More...](#)
Drugs & Substances

Cev Protocol 
Oncovin 
Paraplatin 
Etopophos 
Cisplatin 

Conditions

Retinoblastoma 
Eye Cancer 
Squamous Cell Ski... 
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, Ira 
Kushner, Brian H 
Perentesis, John P 
Villablanca, Judi... 


The Web


News Media



Audio Video



Clinical Trials



Research Articles

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 ...

<http://www.cancer.gov/cancerinfo/pdq/treatment/retinoblastoma/patient/>

Same query in Google

 [Advanced Search](#)
[Preferences](#)
Search: ☒ the web ☐ pages from Italia

Web Results **1 - 10** of about **1,060,000** for **retinoblastoma**

[Retinoblastoma](#) Resource

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

Refine results for **retinoblastoma**:

[Treatment](#) [Tests/diagnosis](#) [For patients](#) [From medical authorities](#)
[Symptoms](#) [Causes/risk factors](#) [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 (reh-tin-oh-blast-oma) 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/whatisrb.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 Guía para Padres ...

www.retinoblastoma.com/ - 1k - [Cached](#) - [Similar pages](#) - [Note this](#)