



CS 597: SPECIAL TOPICS INFORMATION RETRIEVAL

Overview

Definition

2

- “**Information Retrieval** (IR) is the activity of obtaining information *resources relevant* to an information *need* from a collection of information resources”



Information Overload

Why care about IR

3

- We need to handle different information *needs* and different *collections of resources*



Information is Everywhere



New Information Created Every Minute

Areas to focus

4

□ What we will learn this semester



Web Search



Query Suggestions



Question Answering



Recommendation
Systems



IR & Big Data



Other possible topics
of interest in IR?

Web Search

5

- Why not just database (DB)?
 - ▣ Not all text resources will be properly structured
 - ▣ Text resources versus DB records
 - Matches are easily found by comparisons
 - “Find accounts with balance > \$100”
 - Unstructured text require more involved comparisons
 - “Information Retrieval Conferences”

Web Search

6

- IR goal of web search
 - Retrieve *all* the documents that are relevant to a query, while retrieving as *few* non-relevant documents as possible”
- Main IR issues related to web search
 - Comparing a query to text resources and determining what is a **good match**
 - A *relevant* document/resource contains the information a person was looking for when he submitted a query to the search engine
 - Many factors *influence* a person’s decision about what is relevant: e.g., task, context, novelty, background

Web Search

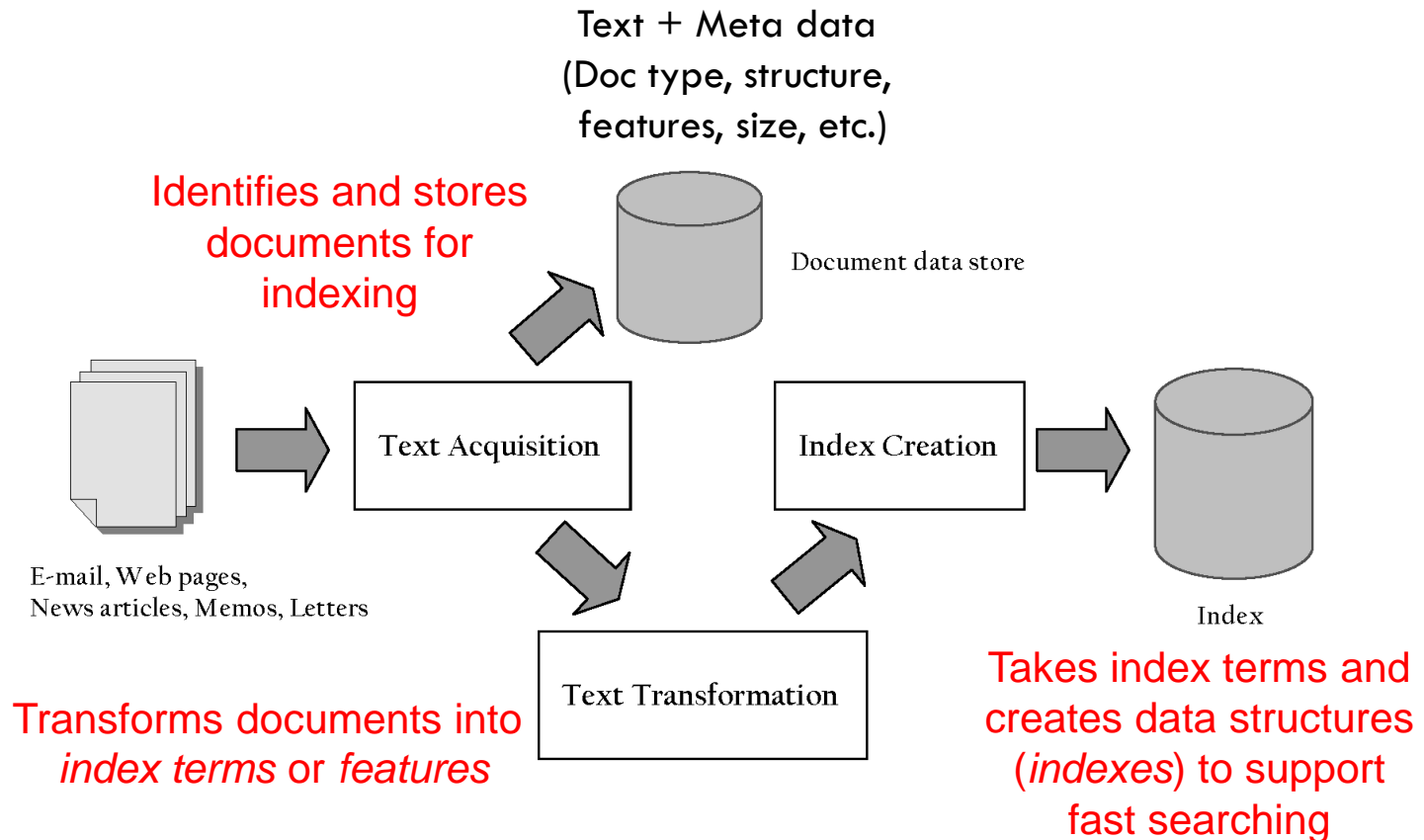
7

- Retrieval models, based on which *ranking algorithms* used in search engines are developed, define a view of relevance
 - ▣ Most models describe statistical (rather than linguistic) properties of text
 - Example: counting simple text features, such as word occurrences, instead of parsing and analyzing the sentences
 - ▣ More info can be considered for relevance ranking
 - Example: document topics, user context, etc.
- Remember: exact matching of words is not enough
 - ▣ There are many different ways to express the same thing in languages like English
 - E.g., does a news story containing the text “bank director in Hollywood steals funds” match the query “Bank scandals in western mass?”

Web search - Tasks

8

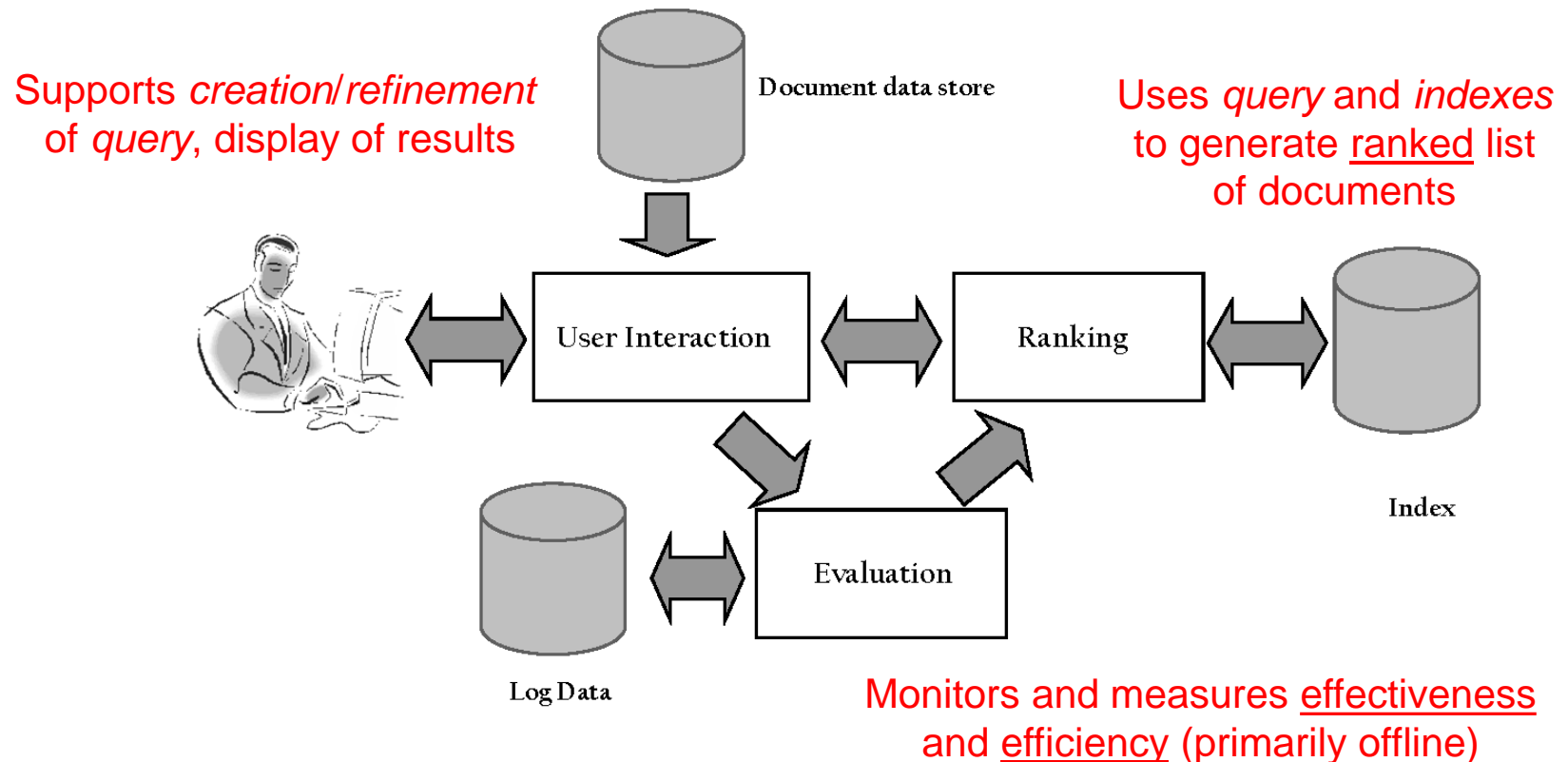
□ Indexing Processing



Web search - Tasks

9

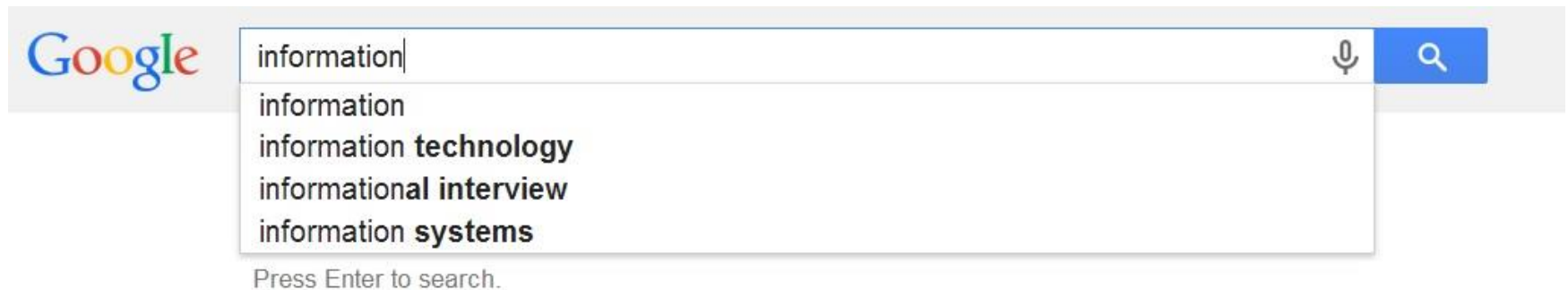
□ Query Process



Query Suggestions

10

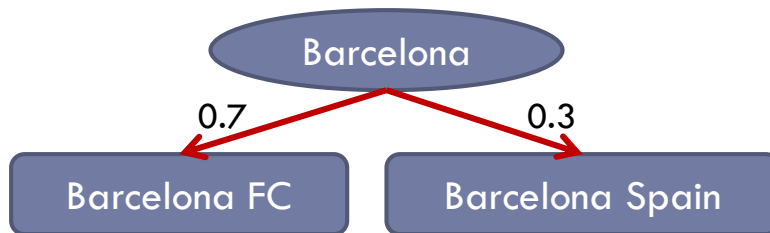
- Goal
 - ▣ Assist users by providing a list of suggested queries that could potentially capture their information needs



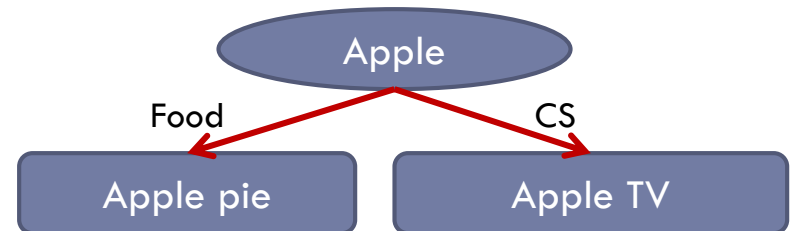
Query Suggestions

11

- Existing methodologies
 - ▣ Query-log based
 - Examine large amounts of past data to identify, given a user query Q , other queries that frequently created in users' sessions that included Q
 - ▣ Corpus-based (in the absence of query log)
 - Examine document corpus, e.g., Wikipedia, or web pages, to determine the likelihood of (co-)occurrence of pairs of words or phrases
- Regardless of the approach, QS modules



Need a **ranking strategy** to identify suggestions that *most likely* capture the intent of a user



Offer **diverse** suggestions that multiple *topical categories* to which Q belongs or *polysemy* (terms with multiple meaning)

Query Suggestions

12

□ Types of query refinement (reformulation)

Type	Goal	User Activity
Modification	Consider analogous, but not exactly-matching, terms	Q: "Single ladies song" QS: "Single ladies lyrics"
Expansion	Generate a more "detailed" query that captures the real interest of a user	Q: " <u>Sports Illustrated</u> " Q: " <u>Sports Illustrated 2013</u> "
Deletion	Create a more "high level", i.e., less restrictive query	Q: "Ebay <u>Auction</u> " QS: "Ebay"

Query Suggestions

13

- Challenges - Most of QS modules rely on query logs
 - ▣ Suitable for systems w/ large user base/interactions/past usage
 - ▣ Not suitable for
 - Systems with smaller user base or without large logs
 - Newly deployed systems, e.g., desktop/personal email search
 - ▣ Log-based QS modules
 - Not always can infer “unseen” queries
- (Long) Tail queries (i.e., rare queries)
- Difficult queries (i.e., queries referring to topics users are not familiar with)

Question Answering

14

- Goal
 - ▣ Automatically answer questions submitted by humans in a natural language form
- Approaches
 - ▣ Rely on techniques from diverse areas of study, such as IR, NLP, Onto, and ML, to identify users' information needs & textual phrases potentially suitable answers for users
- Exploit

YAHOO!
ANSWERS

 stackoverflow

Data from Community Question
Answering Systems (CQA)

Google™



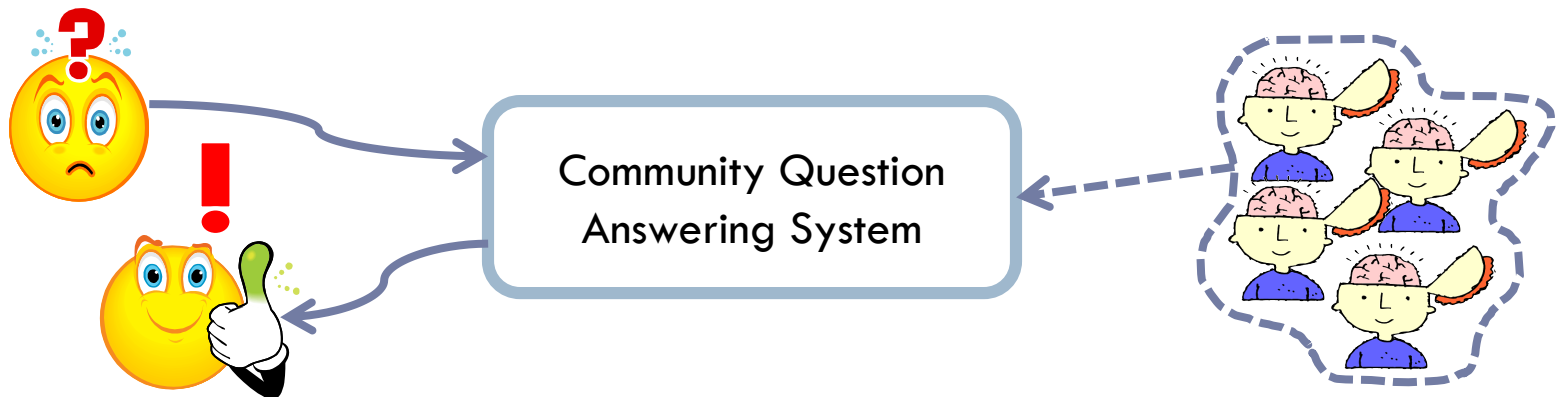
(Web) Data Sources, i.e., doc corpus

Question Answering

CQA-based

15

- CQA-based approaches
 - ▣ Analyze questions (and corresponding answers) archived at CQA sites to locate answers to a newly-formulated question
 - ▣ Exploit “wealth-of-knowledge” already provided by CQA users



- ▣ Existing popular CQA sites
 - Yahoo! Answers, WikiAnswers, and StackOverflow

Question Answering

CQA-based

16

➤ Example.

The screenshot shows the Yahoo! Answers homepage. At the top left is the "YAHOO! ANSWERS" logo with a "Japan Relief" banner. To the right is a search bar with the text "Search" and a "Web Search" button. Below the logo are four tabs: "HOME", "BROWSE CATEGORIES", "MY ACTIVITY", and "ABOUT". The main content area is green and divided into three sections: "Ask", "Answer", and "Discover". The "Ask" section has a circular icon with a question mark and a text input field containing "What would you like to ask?", with a "Continue" button below it. The "Answer" section has a circular icon with a smiley face and a text input field containing "Share your knowledge, Help others and be an Expert", with a "Browse Open Questions" button below it. The "Discover" section has a circular icon with an exclamation mark and a text input field containing "The Best Answers chosen by the Community", with a "Browse Resolved Questions" button below it. At the bottom of the green section is a large search bar with the text "What are you looking for?", a "Search Y! Answers" button, and a link to "Advanced Search".

The screenshot shows a "Best of Answers" window. At the top is the title "Best of Answers" and a window control bar. Below the title is a question: "What is the difference between a tsunami and a tidal wave?" asked by "Gregorio Tirso - Earth Sciences & Geology". The question is accompanied by an orange exclamation mark icon. Below the question is an answer by "titacabreros" with a smiley face icon. The answer text is "Most people assume that there is no difference between a tidal wave and a tsunami, and often use the words interchangeably. This is... More >". The answer has 57 thumbs up and 10 thumbs down. At the bottom of the window is a "Send to a friend" button.

Question Answering

CQA-based

17

- Challenges for finding an answer to a new question from QA pairs archived at CQA sites



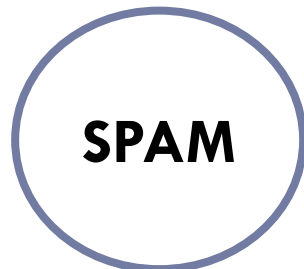
Misleading Answers



No Answers



Incorrect Answers



Spam Answers



Answerer reputation

Question Answering


CQA-based

18

□ Challenges (cont.)

Search


Do I need to apply for a canadian visa if I am taking a cruise to Alaska? [Search Y! Answers](#)

 We did not find results for: Do I need to apply for a canadian visa if I am taking a cruise to Alaska. Try the suggestions below or type a new query above.

Search

Do I need a visa if I am taking a cruise to Alaska [Search Y! Answers](#)

Sort by: [Relevance](#) | [Newest](#) | [Most Answers](#)


 **Do I need Transit Visa or Tourist Visa for Canada if I am taking a cruise to Alaska from Seattle?**
I am taking an Alaskan cruise from Seattle and this cruise... in Victoria (Canada). I am an Indian Citizen on H1B visa in USA. Do I need to apply for a tourist visa for...
★ In Other - Canada - Asked by Imran Moin - 2 answers - 2 years ago


Account for the fact that questions referring to the same topic might be formulated using **similar**, but not the same, words


Search


Good places to visit in Alaska? [Search Y! Answers](#)

Sort by: [Relevance](#) | [Newest](#) | [Most Answers](#)

 **What are some good places to visit in Alaska?**
... about moving to Alaska but want to go and visit for a few weeks before I move there...
★ In Other - United States - Asked by jp_vou - 4 answers - 4 years ago

 **Are Alaska and Russia good places to visit on a Round the World trip?**
I want to go on a big trip in the summer, preferably around the world. I really want to see India, China and Japan which are...but I'd really love to go to Russia and Alaska. So my question is this: have you ever...
2★ In Other - Destinations - Asked by Sneha - 5 answers - 1 year ago

 **where is a good place to visit in Alaska?**
i m going to visit alaska in spring break but i dnt know where since its such a big place...i wanna see moose,bears,...
★ In Other - United States - Asked by isaac - 3 answers - 2 years ago

 **What are some good places to see in Alaska?**
... planning a trip to Alaska in August. I would like to know of some good places to visit. We will be flying in to...
★ In Other - United States - Asked by Kate E - 4 answers - 3 years ago

Identifying the most suitable answer among the **many** available

Question Answering

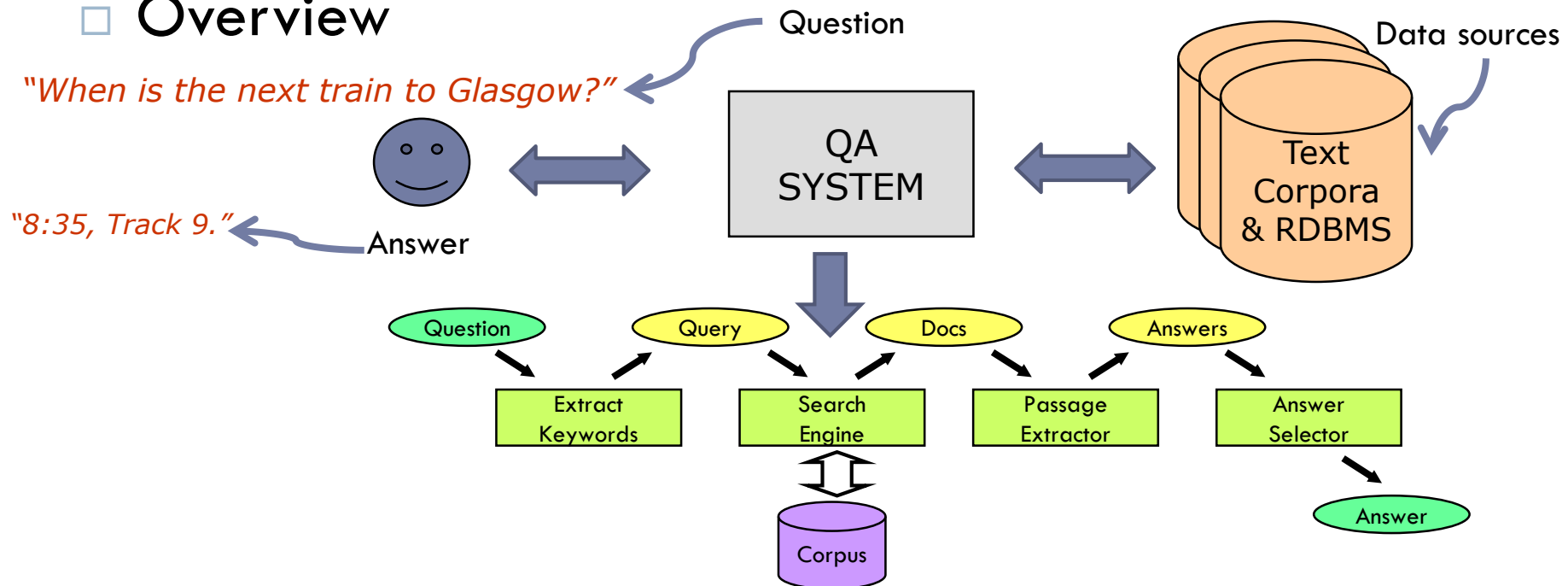
Corpus-based

19

□ Corpus-based approaches

- Analyze text documents from diverse online sources to locate answers that satisfy the info. needs expressed in a question

□ Overview



Question Answering

Corpus-based

20

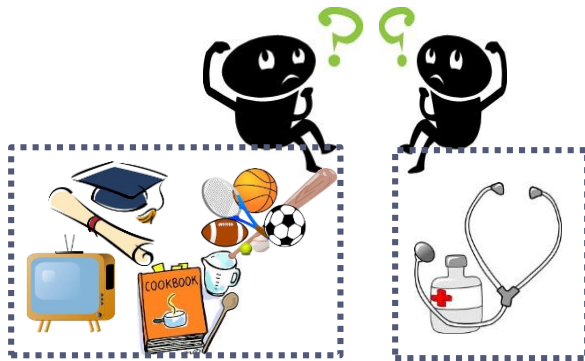
□ Classification

Factoid vs. List (of factoids) vs. Definition

“What lays blue eggs?” -- one fact

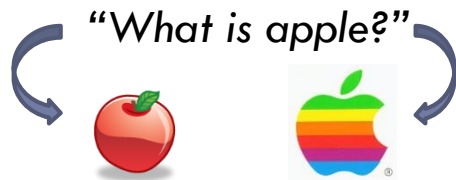
“Name 9 cities in Europe” -- multiple facts

“What is information retrieval?” -- textual answer



Open vs. Closed domain

□ Challenges



Identifying actual user's
information needs



*“Magic mirror in my
hand, who is the
fairest in the land?”*

Converting to
quantifiable measures

[Abraham Lincoln - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Abraham_Lincoln)
en.wikipedia.org/wiki/Abraham_Lincoln

Abraham Lincoln was born February 12, 1809, the second child of Thomas Lincoln and Nancy Hanks (nee Hanks), in a one-room log cabin on the Sinking ...
Sexuality - William Wallace Lincoln - Assassination of Abraham - Mary Todd Lincoln

[Who Was Abraham Lincoln?: Janet Pascal, Nancy Harrison, John O ...](https://www.amazon.com/books?pf_rd_p=1a1b1c1d-1e1f-1g1h-1i1j-1k1l1m1n1o1p1q1r1s1t1u1v1w1x1y1z1)
[www.amazon.com](https://www.amazon.com/books?pf_rd_p=1a1b1c1d-1e1f-1g1h-1i1j-1k1l1m1n1o1p1q1r1s1t1u1v1w1x1y1z1) > Books > Children's Books > Biographies > Political
Fulfillment by Amazon (FBA) is a service we offer sellers that lets them store their products in Amazon's fulfillment centers, and we directly pack, ship, and ...

[Abraham Lincoln | The White House](https://www.whitehouse.gov/about/presidents/abrahamlincoln)
www.whitehouse.gov/about/presidents/abrahamlincoln
A short biography from the official White House Web Site.

Answer ranking

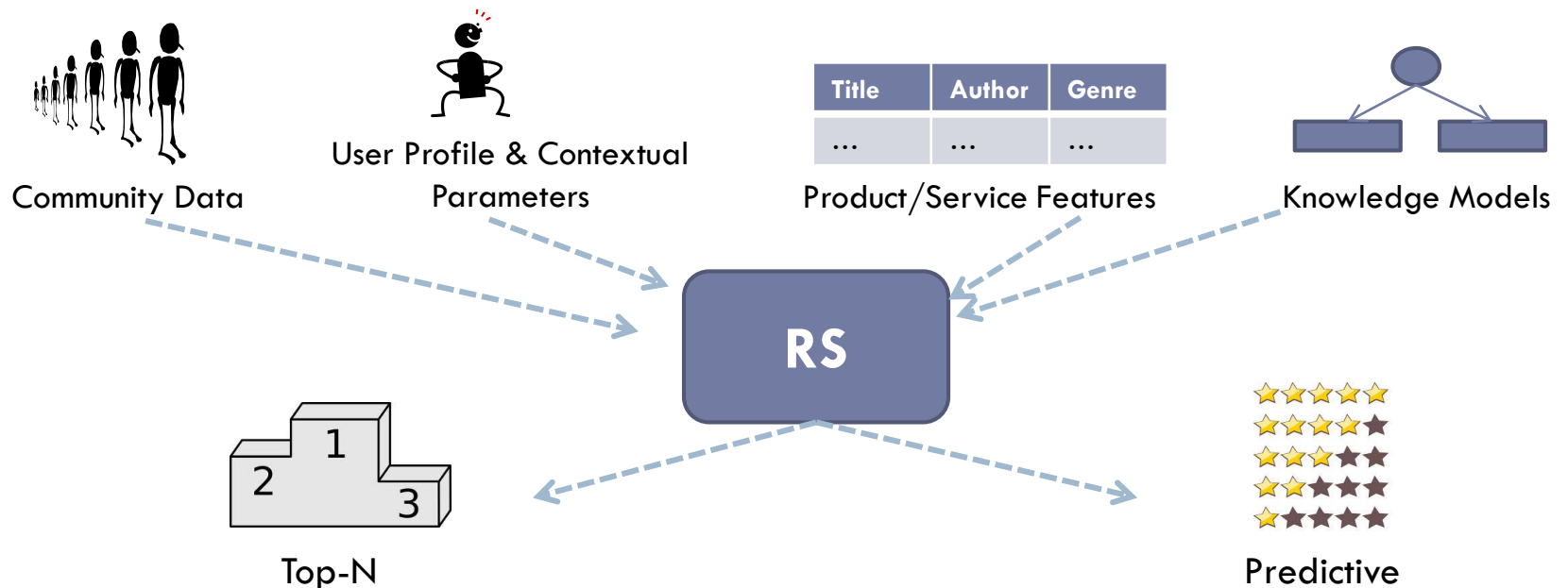
Recommendation Systems (RS)

21

□ Goal

- Enhance users' experience by assisting them in finding information (due to the information overload problem) and reduce search and navigation time

□ Overview

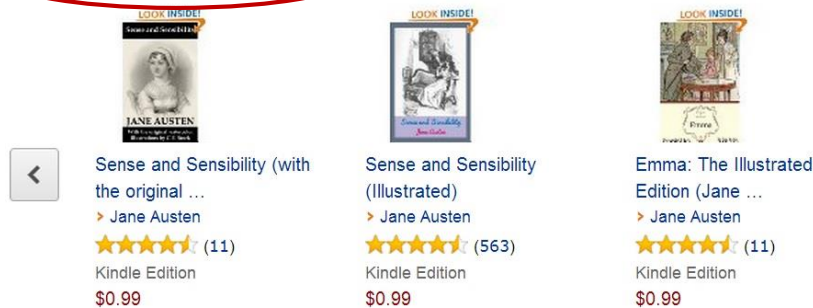


Recommendation Systems

22

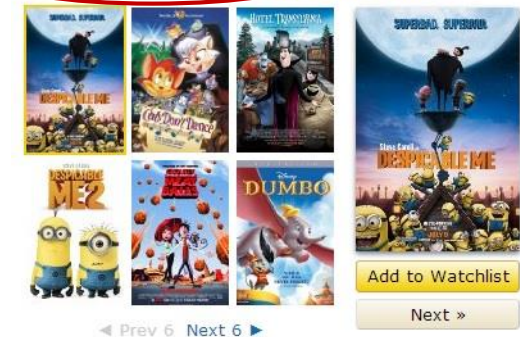
□ Examples

Customers Who Bought This Item Also Bought

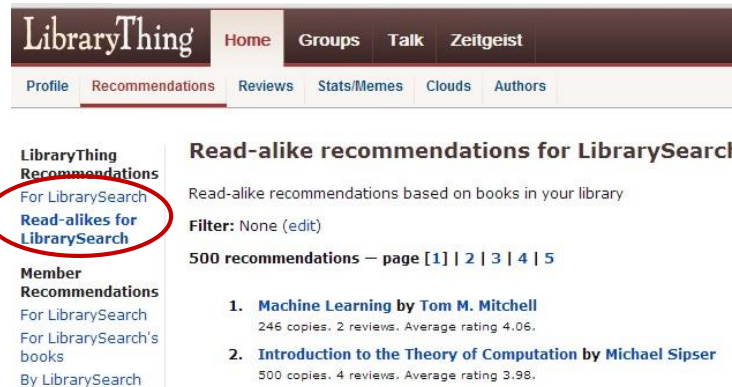


Amazon.com

People who liked this also liked...



IMDB.com

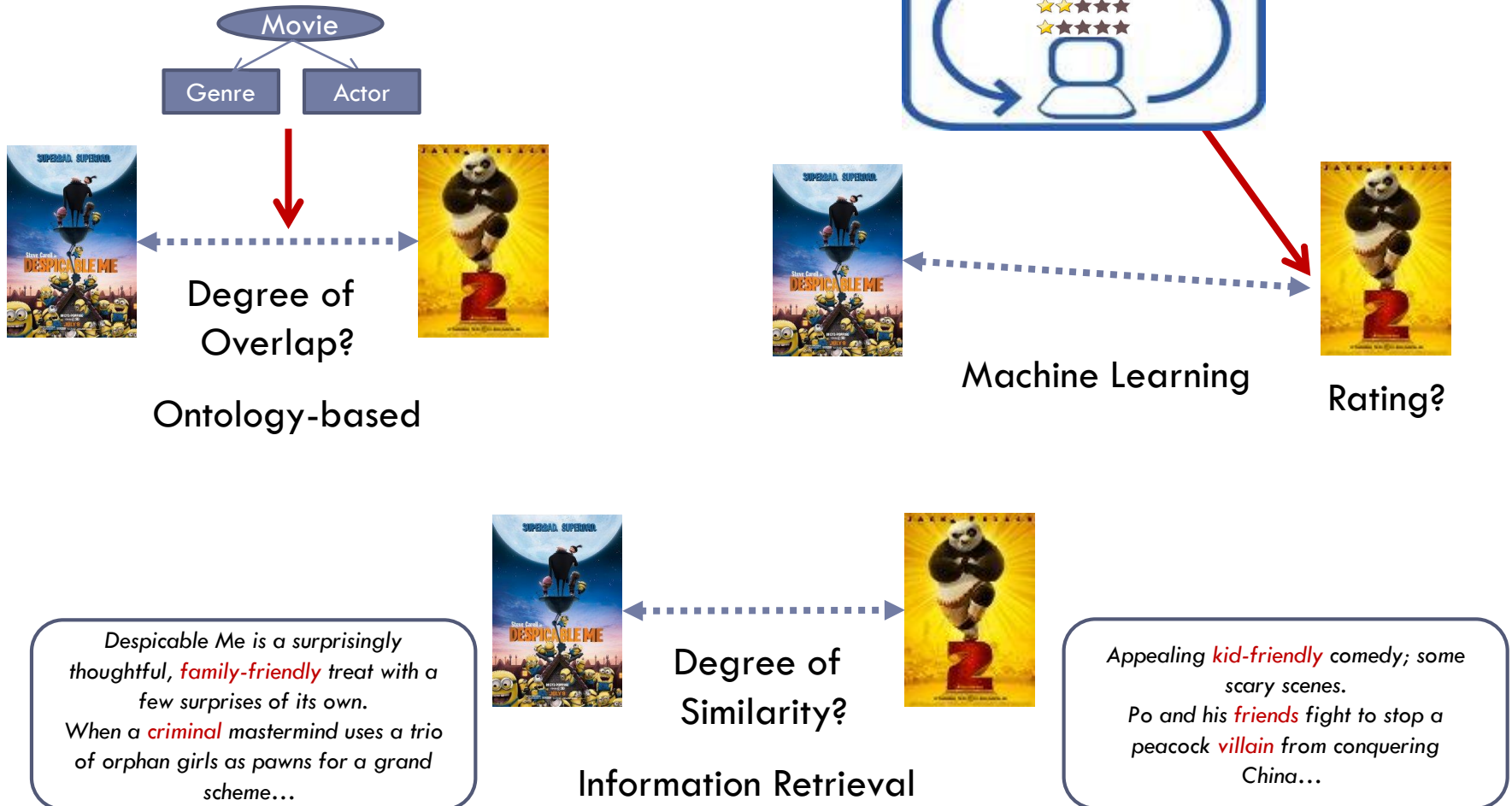


LibraryThing.com

Recommendation Systems

23

□ Approaches



Recommendation Systems

24

- Categorization
 - ▣ Content-based: examine textual descriptions of items
 - ▣ Collaborative filtering: Examine historical data in the form of user and item ratings
 - ▣ Hybrid: Examine content, ratings, and other features to make suggestions
- Other considerations
 - ▣ Target of recommendations, e.g., books suggested for an individual vs. groups of people
 - ▣ Purpose of recommendations, e.g., movies for family vs. friends
 - ▣ Trust-based recommendation, e.g., considering the opinion/suggestions of the social network of a user

Recommendation Systems

25

□ Challenges

▣ Capture users' preferences/interests

- What type of information should be included in users' profiles?
How is this information collected?

▣ Finding the relevant data for describing items

- What metadata should be considered to best capture an item?

▣ Introduce “novelty” and “serendipity” to recommendations

- Provide variety among suggestions
- E.g., suggesting “Kung-Fu Panda 2” to someone who has viewed “Kung-Fu Panda” is not unexpected

Recommendation Systems

26

□ Challenges (continued)

▣ Personalization

- Avoid “one-size-fits-all”, like Amazon’s recommender that provides to every user the same suggestion

▣ Cold start

- No information on new items/users

▣ Sparsity

- Very few items are assigned a high number of ratings

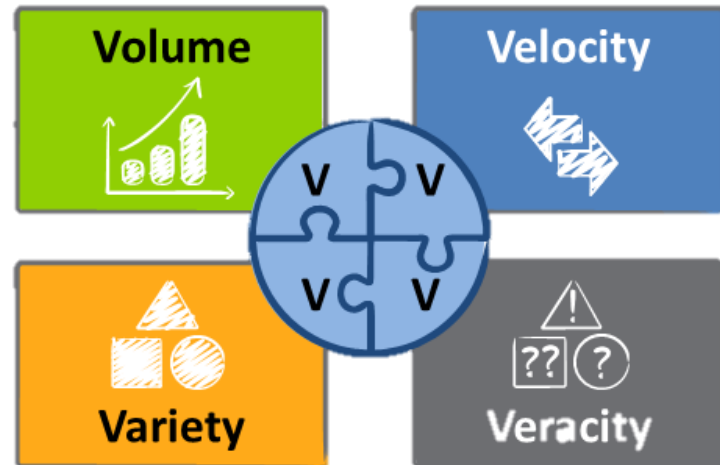
▣ Popularity bias

- Well-known items are favored at the time of providing ratings

Big Data

27

- Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured
 - Deals with extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions
 - The 4 V's in Big Data



IR & Big Data

28

□ Challenges

- Analyze huge amount of data distributed on several servers on the internet
 - Handle search queries that need to go to several servers in parallel to identify relevant resources
- Personalization based on individual (e.g., user click) and social (e.g., Facebook) data
- Efficiency in all IR tasks
- Scaling
 - 50 billion indexed webpages
 - 3 billion google search requests per day

Information Retrieval Topics

29

- Other topics pertaining to information retrieval
 - ▣ NLP for IR
 - ▣ Cross- and multi-lingual IR
 - ▣ Query intent (for QS and QA)
 - ▣ Spoken queries
 - ▣ Ranking in databases

Information Retrieval Topics

30

- Other topics pertaining to IR (continued)
 - ▣ Multimedia IR
 - Examples: image search, video search, speech/audio search, music search
 - ▣ IR Applications
 - Examples: digital libraries, enterprise search, genomics IR, legal IR, patent search, text reuse
 - ▣ Evaluation
 - Examples: test collections, experimental design, effectiveness measures, simulated work task evaluation as opposed to benchmark-based evaluation