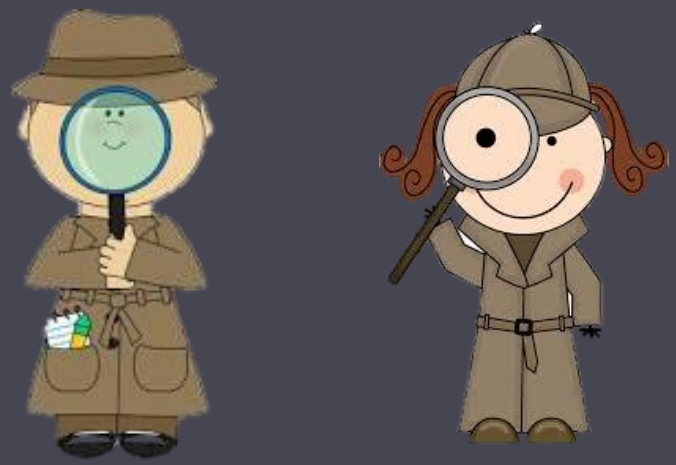


CS 597: SPECIAL TOPICS
INFORMATION RETRIEVAL

Overview

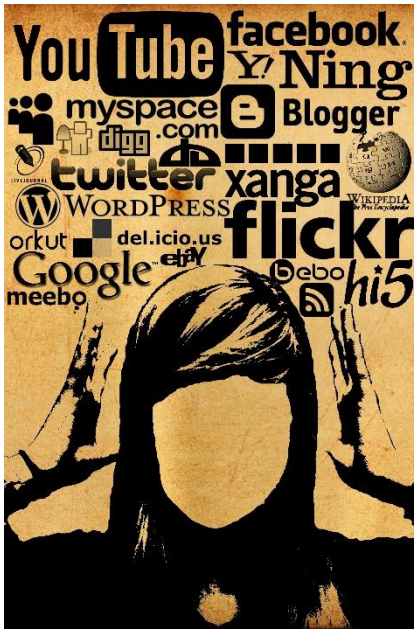
Definition

- "**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

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



Information is Everywhere

New Information Created Every Minute

Areas to focus

- 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

- 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

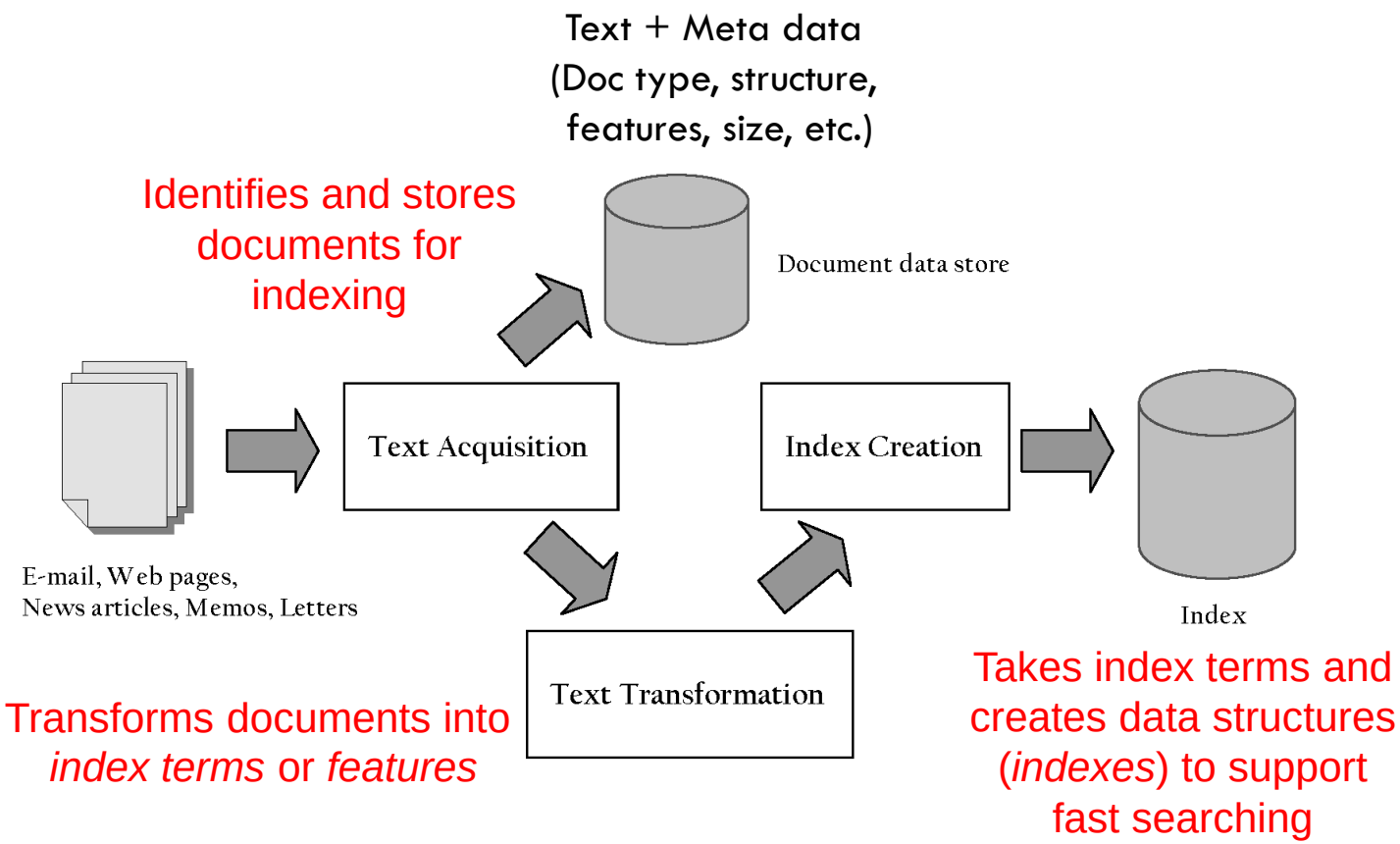
- IR goal of web search
  - Retrieve *all* the documents that are relevant to a query, while retrieving as *few* non-relevant documents as possible"
- Main IR issues related to web search
  - Comparing a query to text resources and determining what is a **good match**
    - A *relevant* document/resource 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

- 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

- Indexing Processing

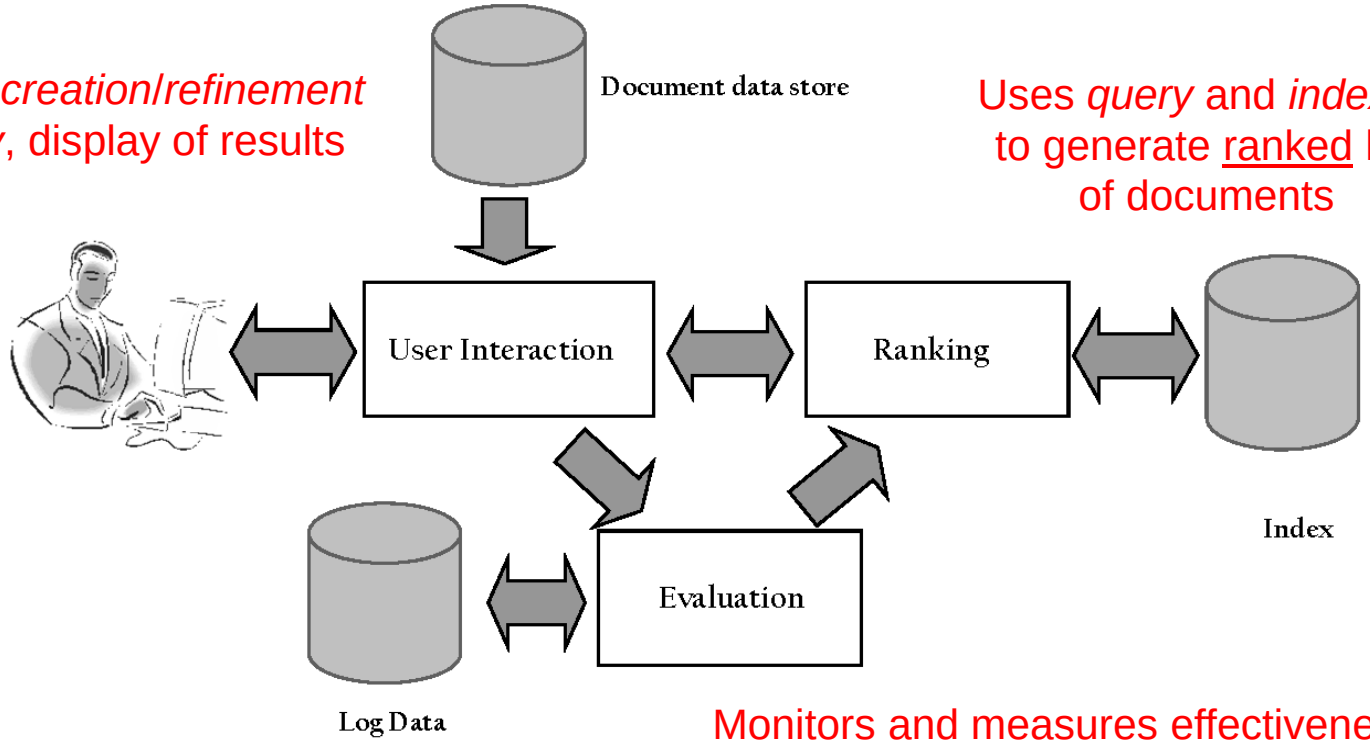


Web search - Tasks

- Query Process

Supports *creation/refinement* of *query*, display of results

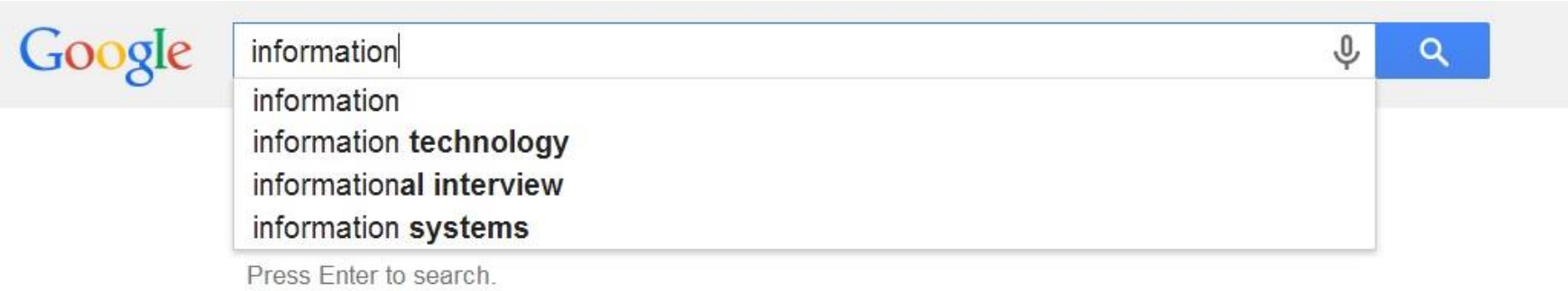
Uses *query* and *indexes* to generate ranked list of documents



Monitors and measures effectiveness and efficiency (primarily offline)

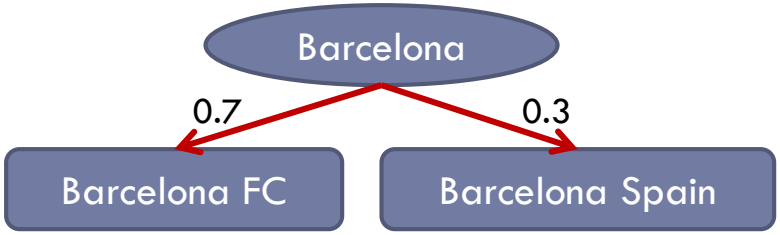
Query Suggestions

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

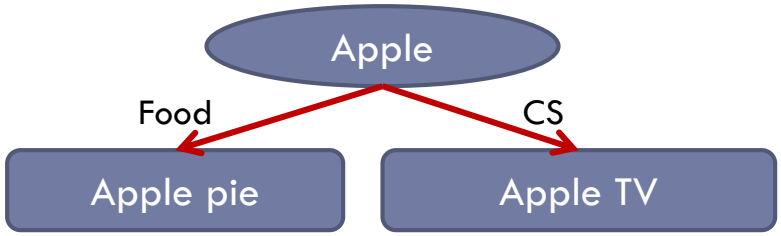


Query Suggestions

- 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

- 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: "Sports Illustrated" Q: "Sports Illustrated 2013"
Deletion	Create a more "high level", i.e., less restrictive query	Q:"Ebay Auction" QS: "Ebay"

Query Suggestions

- 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

- 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

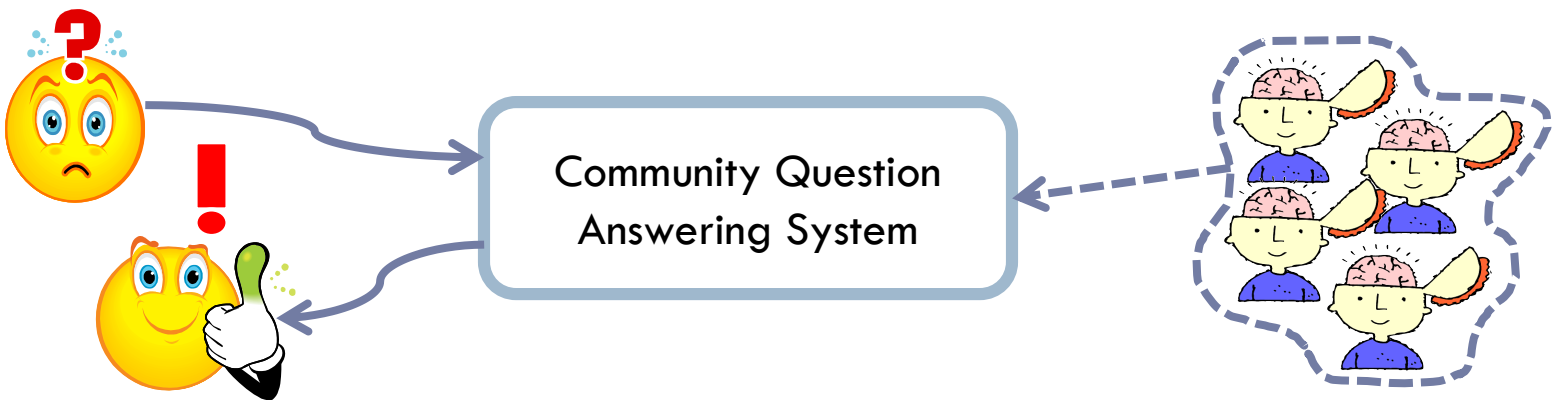
Data from Community Question Answering Systems (CQA)

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

Question Answering

CQA-based

- 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

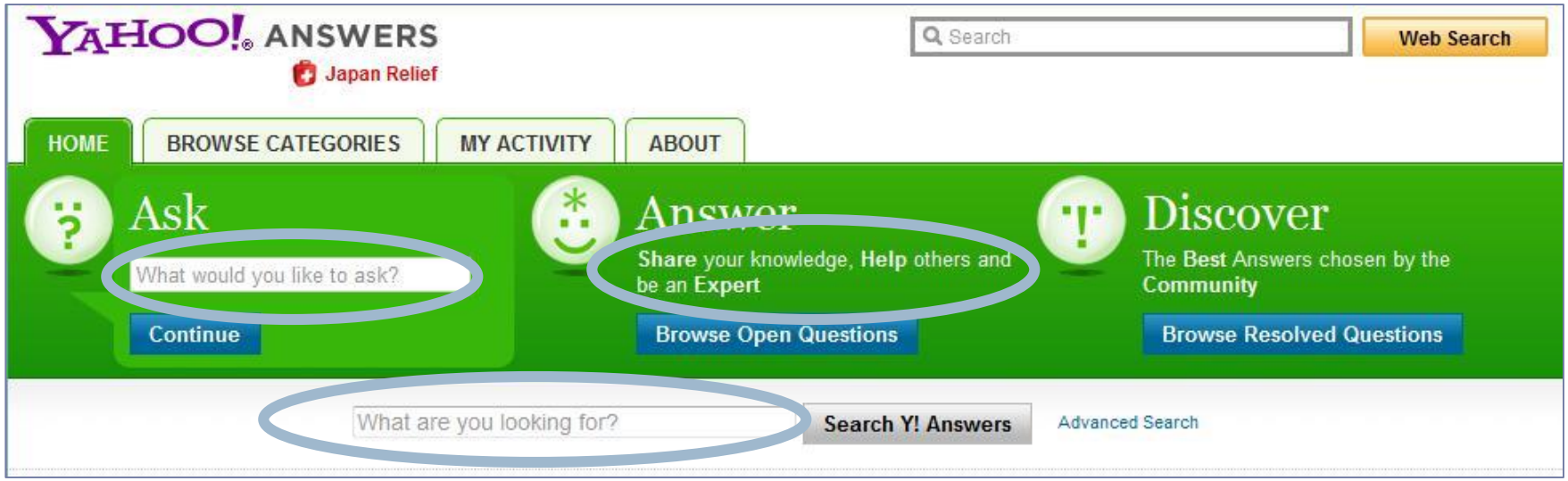


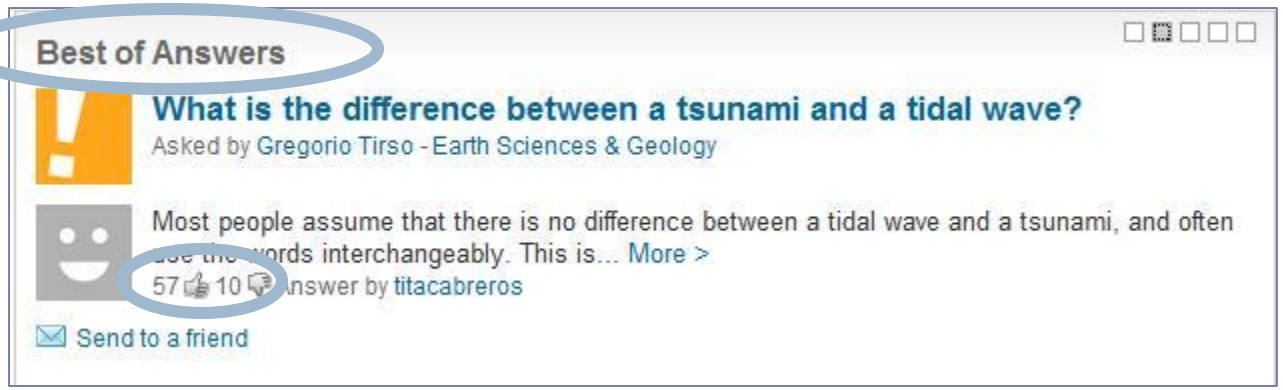
  - Existing popular CQA sites
    - Yahoo! Answers, WikiAnswers, and StackOverflow

Question Answering

CQA-based

- Example.



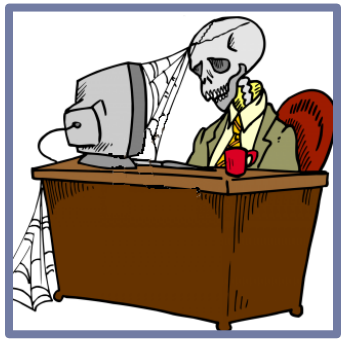


Question Answering

CQA-based

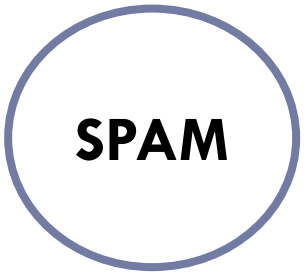
- Challenges for finding an answer to a new question from QA pairs archived at CQA sites

Misleading Answers



No Answers

Incorrect Answers



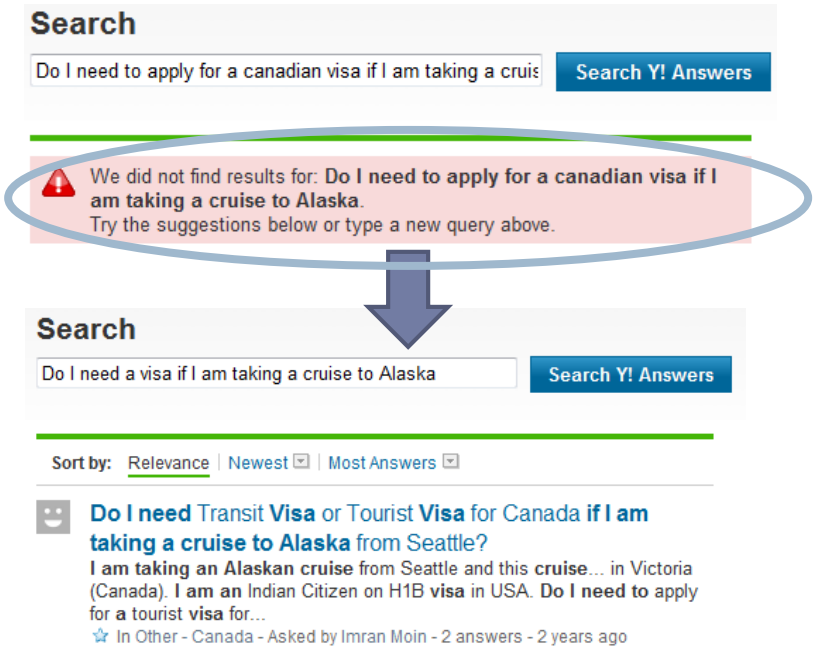
Spam Answers

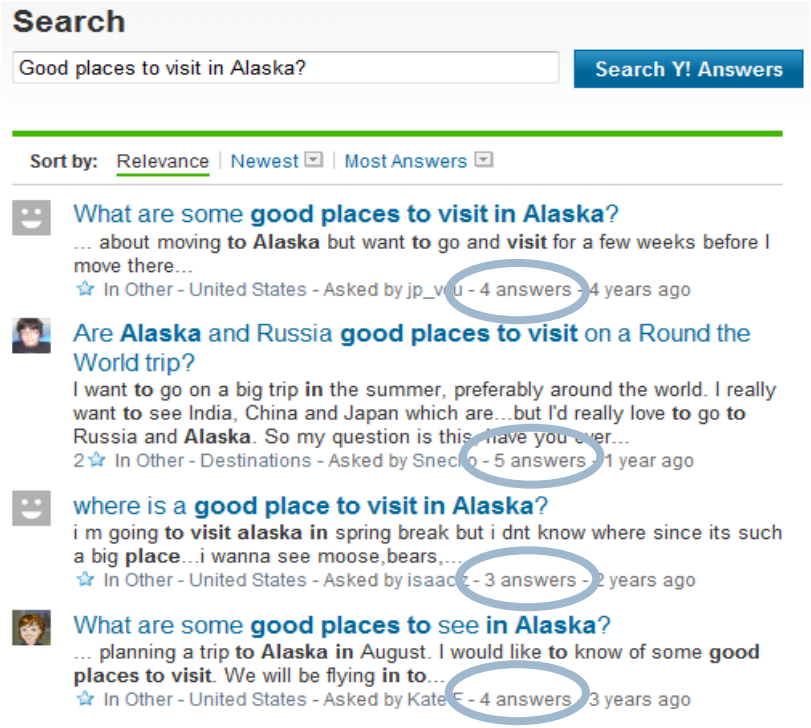
Answerer reputation

Question Answering

CQA-based

- Challenges (cont.)





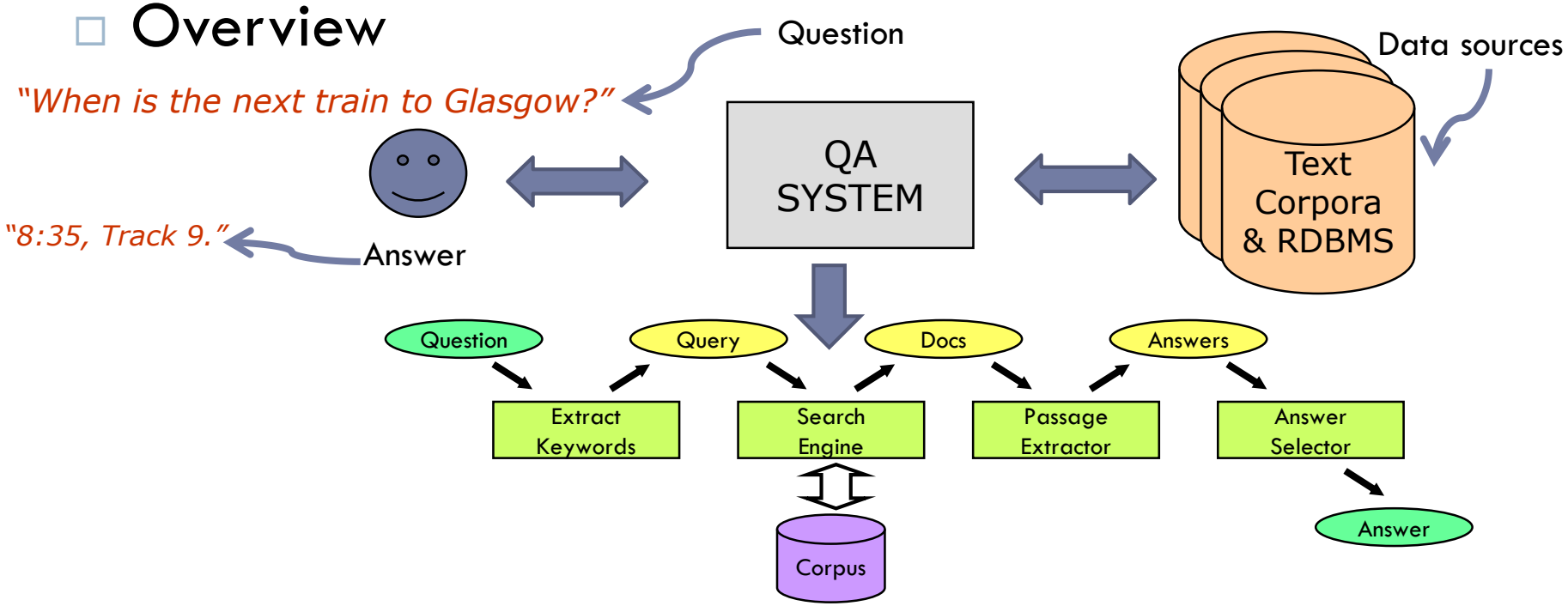
Account for the fact that questions referring to the same topic might be formulated using similar, but not the same, words

Identifying the most suitable answer among the many available

Question Answering

Corpus-based

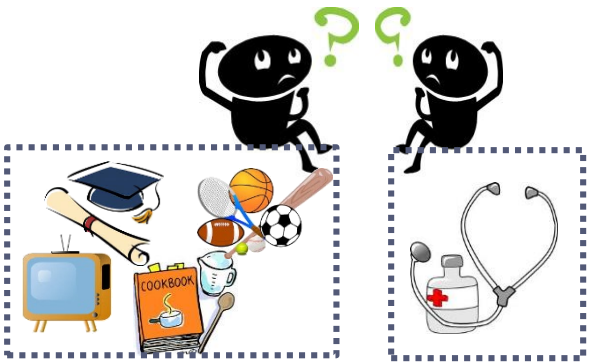
- Corpus-based approaches
  - Analyze text documents from diverse online sources to locate answers that satisfy the info. needs expressed in a question
- Overview



Question Answering

Corpus-based

- Classification



Open vs. Closed domain

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*

- Challenges

"What is apple?"

Identifying actual *user's information needs*

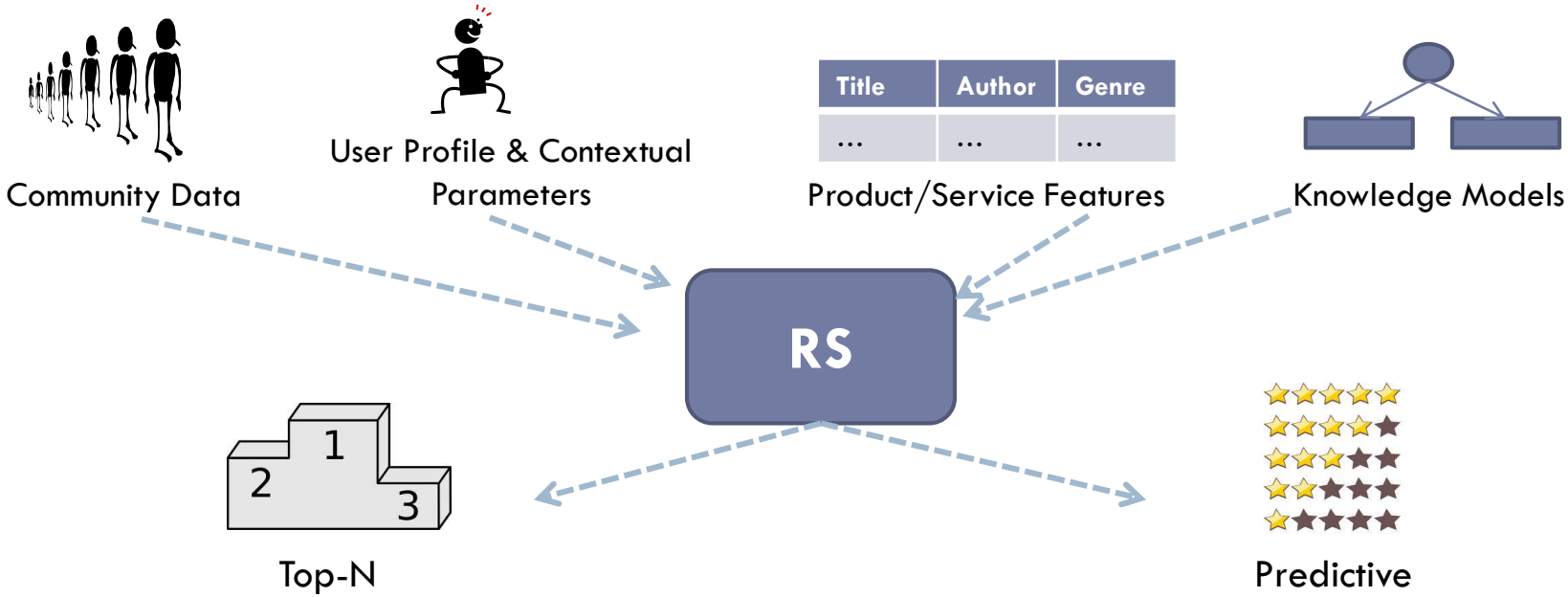
"Magic mirror in my hand, who is the fairest in the land?"

Converting to *quantifiable* measures

Answer *ranking*

Recommendation Systems (RS)

- 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

- Examples

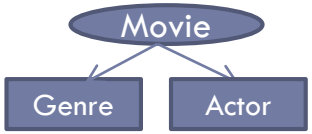
Amazon.com

IMDB.com

LibraryThing.com

Recommendation Systems

- Approaches



Degree of Overlap?

Ontology-based

Machine Learning

Rating?

Despicable Me is a surprisingly thoughtful, family-friendly treat with a few surprises of its own.
When a criminal mastermind uses a trio of orphan girls as pawns for a grand scheme…

Degree of Similarity?

Information Retrieval

Appealing kid-friendly comedy; some scary scenes.
Po and his friends fight to stop a peacock villain from conquering China…

Recommendation Systems

- 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

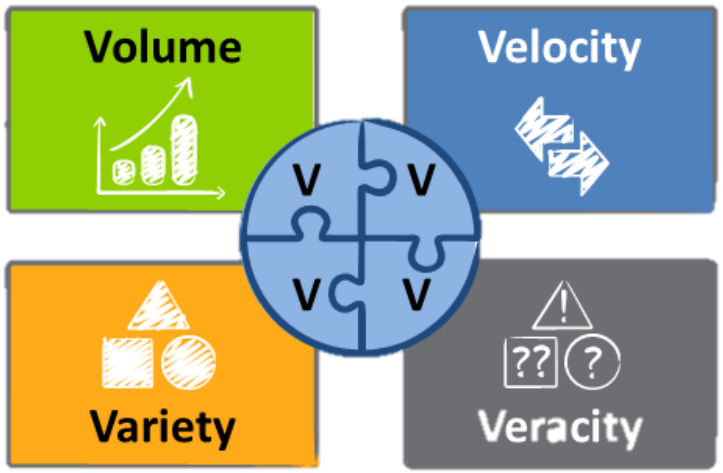
- Challenges
  - Capture users' preferences/interests
    - What type of information should be included in users' profiles? How is this information collected?
  - Finding the relevant data for describing items
    - What metadata should be considered to best capture an item?
  - Introduce "novelty" and "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

- 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

- 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

- 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

- 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

- 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