Information Retrieval: Course Introduction

Pawan Goyal

CSE, IITKGP

January 9th, 2017

Course Info

Course Website:

`http://cse.iitkgp.ac.in/~pawang/courses/IR17.html`

Shared with Prof. Animesh Mukherjee

Course Info

Course Website:

`http://cse.iitkgp.ac.in/~pawang/courses/IR17.html`

Shared with Prof. Animesh Mukherjee

Meeting Times

- Regular Hours:
  - Monday - 8:00 - 10:00 (CSE - 107)
  - Tuesday - 12:00 - 13:00 (CSE - 107)

Course Info

My Contact

- **Email:** pawang@cse.iitkgp.ernet.in
- **Office:** CSE - 308
- **Webpage:** `http://cse.iitkgp.ac.in/~pawang/`

Course Info

My Contact

- **Email:** pawang@cse.iitkgp.ernet.in
- **Office:** CSE - 308
- **Webpage:** `http://cse.iitkgp.ac.in/~pawang/`

Teaching Assistants

- Mayank Singh
- Koustav Rudra
- Suman Kalyan Maity
- Sandipan Sikdar

Books and Materials

Reference Books

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Cambridge university press.

Books and Materials

Reference Books

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Cambridge university press.

Lecture Material

- Additional Readings
- Lecture Slides

Course Evaluation Plan: Tentative

Course Evaluation Plan: Tentative

- Mid-Sem : 25%

Course Evaluation Plan: Tentative

- Mid-Sem : 25%
- End-Sem : 45%

Course Evaluation Plan: Tentative

- Mid-Sem : 25%
- End-Sem : 45%
- Term Project: 30%

What is Information Retrieval?

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.

What is Information Retrieval?

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.

What is a document?

What is Information Retrieval?

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.

What is a document?

web pages, email, books, news stories, scholarly papers, text messages, Powerpoint, PDF, forum postings, patents, IM sessions, Tweets, question answer postings etc.

Document vs. Database Records

Document vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes),

Document vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes),
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.

Document vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes),
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches

Document vs. Database Records

Example bank database query

- Find records with balance > $50,000 in branches located in Amherst, MA.

Document vs. Database Records

Example bank database query

- Find records with balance > $50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Document vs. Database Records

Example bank database query

- Find records with balance > $50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Example search engine query

- *bank scandals in western mass*

Document vs. Database Records

Example bank database query

- Find records with balance > $50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Example search engine query

- *bank scandals in western mass*
- This text must be compared to the text of entire news stories

So, what do we do in IR?

So, what do we do in IR?

- The indexing and retrieval of textual documents.

So, what do we do in IR?

- The indexing and retrieval of textual documents.
- Concerned firstly with retrieving *relevant* documents to a query.

So, what do we do in IR?

- The indexing and retrieval of textual documents.
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

So, what do we do in IR?

- The indexing and retrieval of textual documents.
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

What is the "killer" app?

So, what do we do in IR?

- The indexing and retrieval of textual documents.
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

What is the "killer" app?

Searching for the pages on WWW

Typical IR tasks

Given:

Typical IR tasks

Given:

- A corpus of textual natural-language documents.

Typical IR tasks

Given:

- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Typical IR tasks

Given:

- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Find:

Typical IR tasks

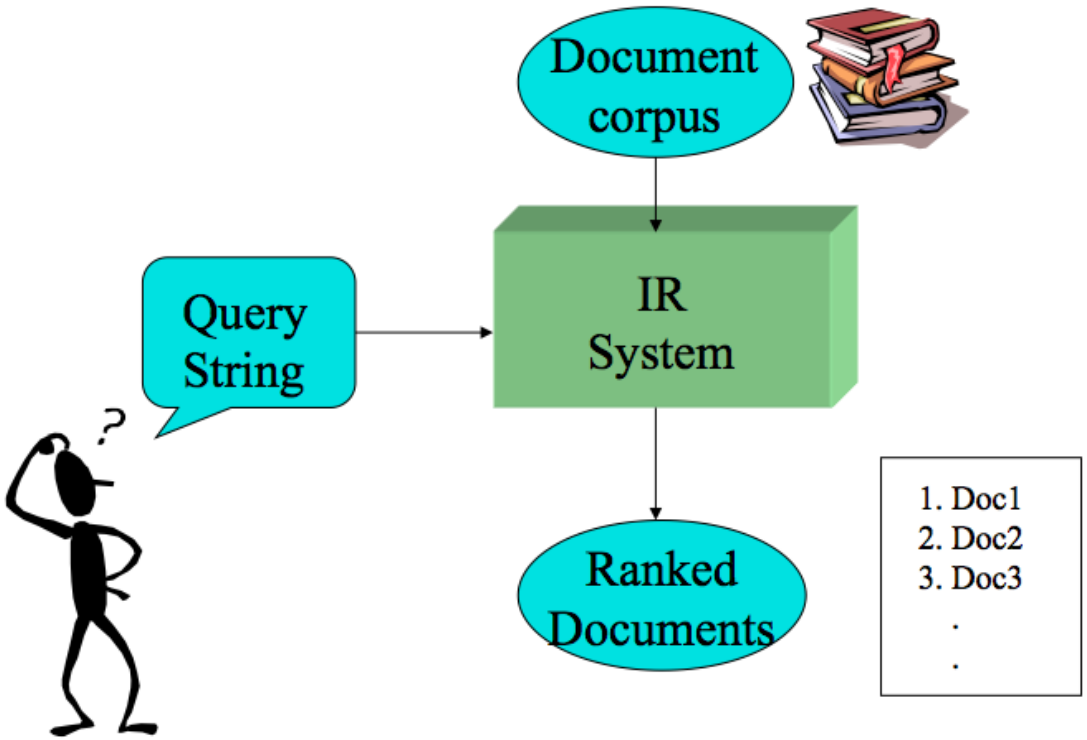
Given:

- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Find:

- A ranked set of documents that are relevant to the query.

IR System



The system should be able to retrieve the relevant docs efficiently

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.
- Being timely (recent information).

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.
- Being timely (recent information).
- Being authoritative (from a trusted source).

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.
- Being timely (recent information).
- Being authoritative (from a trusted source).
- Satisfying the goals of the user and his/her intended use of the information (information need).

Simplest notion of Relevance from Retrieval Models' Perspective

Simplest notion of Relevance from Retrieval Models' Perspective

Keyword Search

Simplest notion of Relevance from Retrieval Models' Perspective

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.

Simplest notion of Relevance from Retrieval Models' Perspective

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that (most of) the words in the query appear frequently in the document, in any order (*bag of words*).

Problems with Keywords Search

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Ambiguity

May retrieve irrelevant document that include ambiguous terms (due to polysemy)

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Ambiguity

May retrieve irrelevant document that include ambiguous terms (due to polysemy)

- 'Apple' (company vs. fruit)

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Ambiguity

May retrieve irrelevant document that include ambiguous terms (due to polysemy)

- 'Apple' (company vs. fruit)
- 'Java' (programming language vs. Island)

An Intelligent IR system will

An Intelligent IR system will

- Take into account the *meaning* of the words used.

An Intelligent IR system will

- Take into account the *meaning* of the words used.
- Adapt to the user based on *direct* or *indirect* feedback.

An Intelligent IR system will

- Take into account the *meaning* of the words used.
- Adapt to the user based on *direct* or *indirect* feedback.
- Take into account the *importance* of the page.

An Intelligent IR system will

- Take into account the *meaning* of the words used.
- Adapt to the user based on *direct* or *indirect* feedback.
- Take into account the *importance* of the page.
- ...

Active Areas of Research

Compiled based on the most recent papers at SIGIR and related conferences, just indicative, not exhaustive

What to retrieve

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.
- *Temporal Feedback for Tweet Search with Non-Parametric Density Estimation*. SIGIR 2014.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.
- *Temporal Feedback for Tweet Search with Non-Parametric Density Estimation*. SIGIR 2014.
- *Concept Embedded Convolutional Semantic Model for Question Retrieval.* WSDM 2017.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.
- *Temporal Feedback for Tweet Search with Non-Parametric Density Estimation*. SIGIR 2014.
- *Concept Embedded Convolutional Semantic Model for Question Retrieval.* WSDM 2017.
- *Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale.*. SIGIR 2016.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.
- *Temporal Feedback for Tweet Search with Non-Parametric Density Estimation*. SIGIR 2014.
- *Concept Embedded Convolutional Semantic Model for Question Retrieval.* WSDM 2017.
- *Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale.*. SIGIR 2016.
- *On Effective Personalized Music Retrieval by Exploring Online User Behaviors.*. SIGIR 2016.

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Retrieval of Relevant Opinion Sentences for New Products*. SIGIR 2015.
- *Temporal Feedback for Tweet Search with Non-Parametric Density Estimation*. SIGIR 2014.
- *Concept Embedded Convolutional Semantic Model for Question Retrieval.* WSDM 2017.
- *Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale.*. SIGIR 2016.
- *On Effective Personalized Music Retrieval by Exploring Online User Behaviors.*. SIGIR 2016.
- *ANNE: Improving Source Code Search using Entity Retrieval Approach.* WSDM 2017.

Query Completion

- *Analyzing User's Sequential Behavior in Query Auto-Completion via Markov Processes*. Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, Hongyuan Zha, Ricardo Baeza-Yates
- *adaQAC: Adaptive Query Auto-Completion via Implicit Negative Feedback*. Aston Zhang, Amit Goyal, Weize Kong, Hongbo Deng, Anlei Dong, Yi Chang, Carl A. Gunter, Jiawei Han

Search experience contd ...

(a) Example of a satisfied(SAT) search session

(b) Example of a dissatisfied(DSAT) search sessioin

Figure 1: Examples of Users' Mouse Movement Trails on SERPs

Different users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, Xuan Zhu

Search experience contd ...

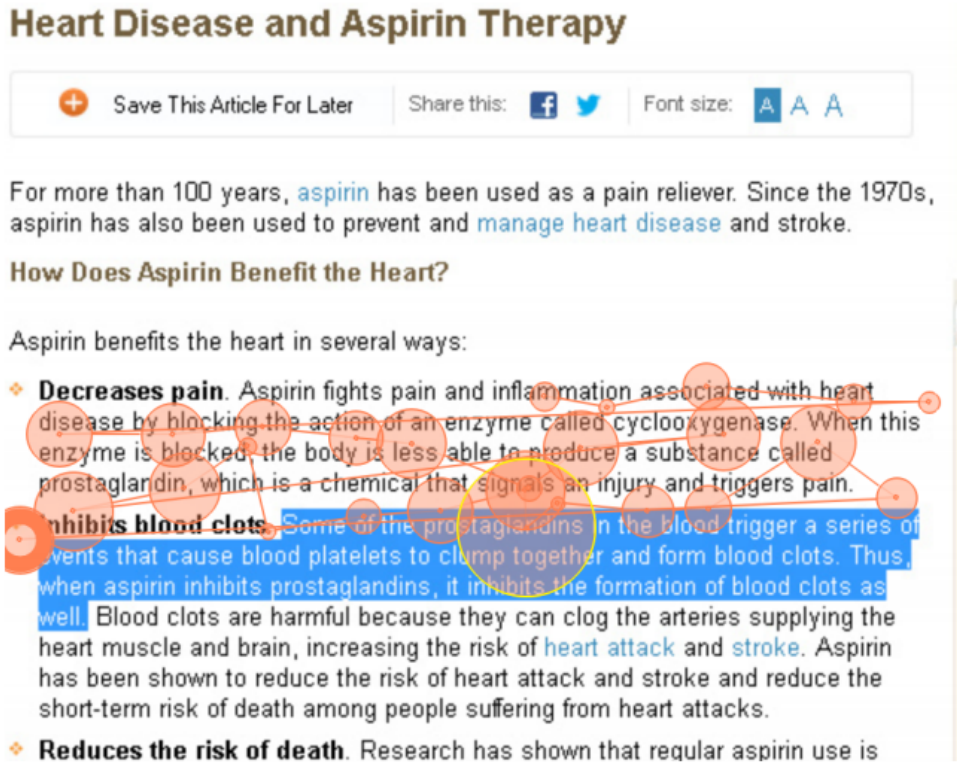


Figure 1: Eye-gaze patterns of Session 001 prior to query reformulation show strong evidence for acquisition of the medical term "prostaglandin" occurring in a paragraph of user-highlighted text.

An Eye-Tracking Study of Query Reformulation. Carsten Eickhoff, Sebastian Dungs, Vu Tran

Search Experience

- *How many results per page? A Study of SERP Size, Search Behavior and User Experience.* Diane Kelly, Leif Azzopardi
- *Influence of Vertical Result in Web Search Examination.* Liu Zeyang, Yiqun Liu, Ke Zhou, Min Zhang, Shaoping Ma
- *Unconscious Physiological Effects of Search Latency on Users and Their Click Behaviour.* Miguel Barreda-Angeles, Ioannis Arapakis, Xiao Bai, B. Barla Cambazoglu, Alexandre Pereda-Banos
- *Context-Aware Web Search Abandonment Prediction.* Yang Song, Xiaolin Shi, Ryen W. White, Ahmed Hassan

What do we cover in this course

IR Basics

- Boolean retrieval
- The term vocabulary & postings lists
- Dictionaries and tolerant retrieval
- Index construction and compression
- Scoring, term weighting & the vector space model
- Computing scores in a complete search system
- Evaluation in information retrieval
- Relevance feedback & query expansion
- Probabilistic information retrieval
- Language models for information retrieval
- Text classification & Naïve Bayes

Course Contents

Classification, clustering and Web

- Web crawling and indexes
- Vector space classification
- Flat clustering
- Hierarchical clustering
- Matrix decompositions & latent semantic indexing
- Link analysis