

# CS4051

Information Retrieval

Week 01

---

Muhammad Rafi

January 24, 2023

## Agenda

- Course Introduction
  - Basic Terminology
  - Structured Vs. Semi Structured Vs. Unstructured
  - Difference between Database & IR
  - Ad hoc IR setup
  - Conclusion
-

## Course Objectives

- Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources usually document collections.
- The course discusses important retrieval models (Boolean, vector space, probabilistic, inference net, language modeling, link analysis), clustering algorithms, collaborative filtering, automatic text categorization, and experimental evaluation.
- Search Engines Algorithms, Advanced IR techniques and Evaluations are discussed.

## Course Outline (1/4)

- Week -01
  - Introduction to IR course, IR Problem and its Components, Basic IR model – Boolean Information Retrieval, Extended Boolean Model, Example of Commercial Systems – WestLaw <Ch: 1 >
- Week-02
  - Document Processing, Term vocabulary and positing lists, Stemming & Lemmatization, posting list processing via skip lists, Phrase Query, positional indexing, bi-word indexing, Combining different indexing techniques. <Ch: 2>

## Course Outline (1/4)

### ■ Week -03

- Dictionary and Tolerant Retrieval, Search Structures for Dictionary, Wildcard queries, permuterm index, k-gram index, Spelling Correction, Edit Distance, Phonetic Correction <Ch: 3>

### ■ Week-04

- Index construction, single pass, in-memory, distributed indexing, dynamic indexing.
- Heaps law, Zipf's law, dictionary indexing, fixed length and variable length coding,  $\gamma$  codes <Ch:4 and 5>

## Course Outline (2/4)

### ■ Week -05

- Vector Space Model <Ch: 6>

### ■ Week-06

- Midterm Exam

### ■ Week-07

- Evaluation of IR <Ch: 8>

### ■ Week-08

- Relevance Feedback <Ch: 9>

## Course Outline (3/4)

### ■ Week-9

- Basic Web Search, Crawler and Indexing <Ch: 19 and 20>

### ■ Week-10

- Link Analysis <Ch: 21>

### ■ Week-11

- Midterm II

### ■ Week-12

- Text classification <Ch: 14>
- 

## Course Outline (4/4)

### ■ Week-13

- Text Clustering

### ■ Week-14

- Neural Information Retrieval

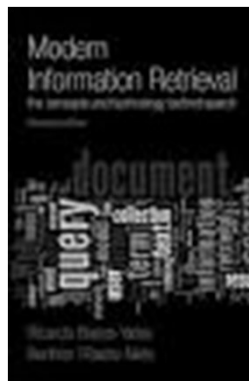
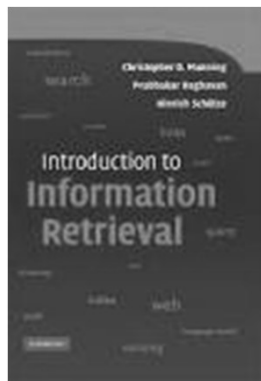
### ■ Week-15

- Topic Detection and Tracking
-

## Grading Scheme

■ Assignments	20%
■ Quizzes	10%
■ Mid-Terms (2)	20%
■ Class Project	10%
■ Final	40%

## Text Books/References



## Information Retrieval

- Information Retrieval (IR) is activity of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need (query) from within large digital collections (usually stored on computers).
- Example
  - What is your CNIC Number? CC?
  - What is De Quervain's?
  - How to get rid from kitchen bugs?
  - What is Surakav?
  - What is Sarcopenia?

## Example

Google  X 🔊 🔍

[Q](#) All [Images](#) [Videos](#) [Books](#) [News](#) [More](#) [Tools](#)

About 2,000,000 results (0.53 seconds)

De Quervain's tenosynovitis is a **painful condition that affects the tendons in your wrist**. It occurs when the 2 tendons around the base of your thumb become swollen. The swelling causes the sheaths (casings) covering the tendons to become inflamed. This puts pressure on nearby nerves, causing pain and numbness. 07-May-2020

<https://familydoctor.org/de-quervains-tenosynovitis/> ⋮


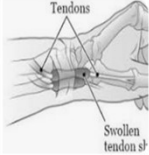
de Quervain's Tenosynovitis: Causes & Treatment

About featured snippets • Feedback

People also ask ⋮

How long does it take for de Quervain to heal? ▼

Can De Quervain's be cured? ▼

## Terminology

- Data / Information / Knowledge/ Wisdom
  - Database systems Vs. Information retrieval
  - Data Mining Vs. Text mining
  - Text Mining Vs. Information retrieval
  - Ad hoc Information Retrieval
- 

## Database Systems

- Generally data collected by some database management systems are collectively called Database.
  - Mostly structured one, to support business operations.
  - SQL is used to get the required information from it. Routines Vs. Ad hoc queries.
-

## IR Vs. Databases

- Format of data:
  - DB: Structured data. Clear semantics based on a formal model.
  - IR: Mostly Semi-Structured or unstructured. Free text.
- Queries:
  - DB: Formal (like SQL)
  - IR: often expressed in natural language (keywords search)
- Result:
  - DB: exact result
  - IR: Sometimes relevant, often not

## Data Mining

- Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD). It is an interdisciplinary subfield of computer science
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.



## Text Mining

- “The objective of Text Mining is to exploit information contained in textual documents in various ways, including ...discovery of patterns and trends in data, associations among entities, predictive rules, etc.” (Grobelnik et al., 2001)
- “Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known.” (Hearst, 1999)

## Text Vs. Data

	<b>Data Mining</b>	<b>Text Mining</b>
Data Object	Numerical & categorical data	Textual data
Data structure	Structured	Unstructured & semi-structure
Data representation	Straightforward	Complex
Space dimension	<tens of thousands	> tens of thousands
Methods	Data analysis, machine learning statistic, neural networks	Data mining, information retrieval, NLP,...
Maturity	Broad implementation since 1994	Broad implementation starting 2000
Market	10 <sup>5</sup> analysts at large and mid size companies	10 <sup>8</sup> analysts corporate workers and individual users

## Information Retrieval Vs. Text Mining

- Information Retrieval is more concerned with getting information related to a user query from unstructured collections of generally textual data. Activity like:
    - Crawling, parsing, indexing and evaluating the retrieved information by using generalized methods
  - Text Mining generally utilizes the NLP based methods for doing activity like:
    - Clustering, classification, summarization and QA
- 

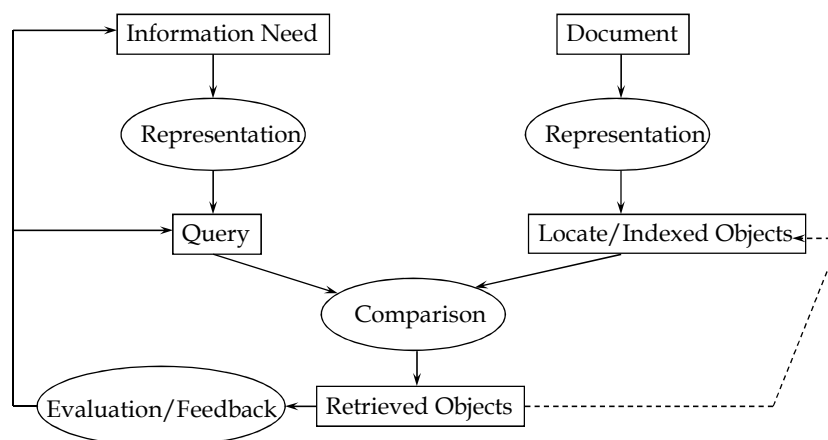
## Information Retrieval Vs. Information Extraction

- Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.
    - Searches can be based on metadata or on full-text indexing.
  - Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.
-

## Adhoc Retrieval Systems

- It is a system that aims to provide documents from within the collection that are relevant to an arbitrary user information need, communicated to the system by means of a one-off, user-initiated query.
- Example
  - Google
  - Bing (later Decision engine)

## Basic Information Retrieval Process



## Content wise IR Systems

### ■ Contents

- Size
  - Personal / Desktop (Spotlight, Instant Search, anywhere)
  - Web Scale (Google, DuckDuckgo)
  - Enterprise Search (WestLaw, PolicyBazar)
- Static (Offline) or Dynamic (Online)
- Type
  - Text
  - Multimedia
  - Mixed
- Exact Match vs. Best Match

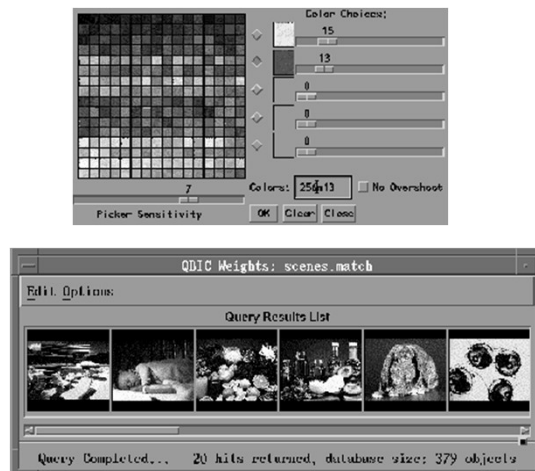
## Dimension of IR

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	Clustering
Music	P2P search	
	Literature search	

## IBM's QBIC

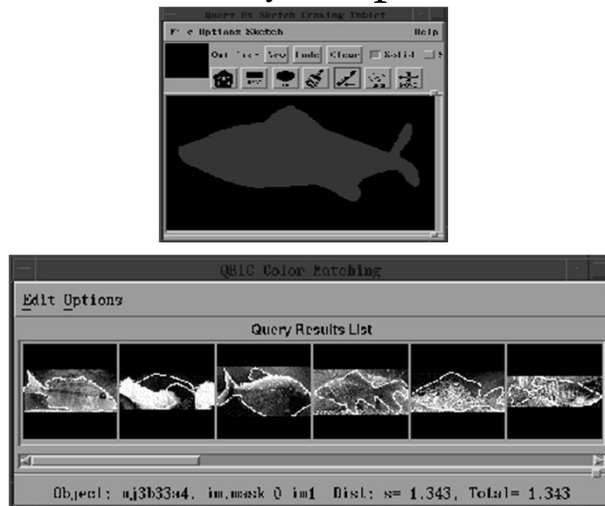
- QBIC – Query by Image Content
- First commercial CBIR system.
- Model system – influenced many others.
- Uses color, texture, shape features
- Text-based search can also be combined.
- Uses R\*-trees for indexing

## QBIC – Search by color



**\*\* Images courtesy : Yong Rao**

## QBIC – Search by shape



**\*\* Images courtesy : Yong Rao**

## QBIC – Query by sketch



**\*\* Images courtesy : Yong Rao**

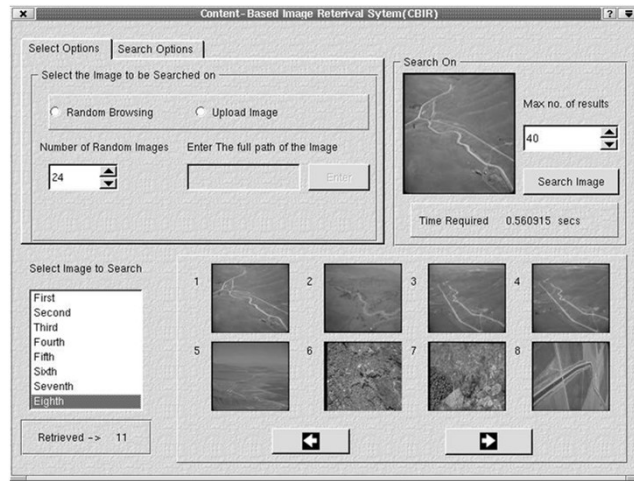
## Virage

- Developed by Virage inc.
  - Like QBIC, supports queries based on color, layout, texture
  - Supports arbitrary combinations of these features with weights attached to each
  - This gives users more control over the search process
- 

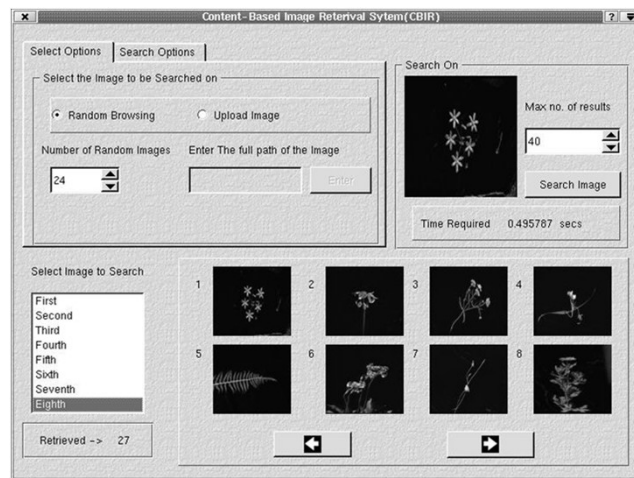
## VisualSEEk

- Research prototype – University of Columbia
  - Mainly different because it considers spatial relationships between objects.
  - Global features like mean color, color histogram can give many false positives
  - Matching spatial relationships between objects and visual features together result in a powerful search.
-

# ISearch



# ISearch





# ISearch



## Example: Content-based Image Retrieval:

- <http://wang.ist.psu.edu/IMAGE>



## Music Search Engine

- Search by metadata likes: artists biography, music reviews, new releases, concerts dates
- Search for Lyrics: Lyrics.com. Allmusic.com musicpedia.com and SearchLyrics.com
- Recommend Similar Music: based on seed elements (artist, track) users are recommended similar tracks or artists.
- User query by Humming
- Generate Playlists: automatic generation of playlists, that satisfy user constraints

## Mathematical Search Engines

- The mathematical contents on the web is continuously on rise. The IR perspective of contents are quite low.
- We need to redesign a specialized search engine for it.
- There are few very good attempts:
  - [www.wolframalpha.com](http://www.wolframalpha.com)
  - [symbolab.com](http://symbolab.com)
  - [searchenginewatch.com](http://searchenginewatch.com)

## Question & Answering Systems

- System that enables the extraction of an answer (or several) to a request (a question) based on a corpus
  - An answer or several, possibly a list from one or several documents, an answer of the type Yes/No...,
    - Askjeeves.com
    - Easyask.com
    - Answerlogic.com
- 

## General Search Engines

- [www.google.com](http://www.google.com)
  - [www.visimo.com](http://www.visimo.com)
  - [www.clusty.com](http://www.clusty.com)
  - [www.bing.com](http://www.bing.com)
  - [www.yahoo.com](http://www.yahoo.com)
-

## Bing

- MSN Search
  - Microsoft Search
  - Live Search
  - Bing
    - ASP.NET Launched June 01, 2009 supported 40 languages now
    - Bing and Decide was the slogan in 2009
    - Bing a decision engine - 2010
    - “Bing is for doing” is in 2012
-