
ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

*Lecture 1:
Introduction*

S. M. Vahidipour

Vahidipour@kashanu.ac.ir

Outline

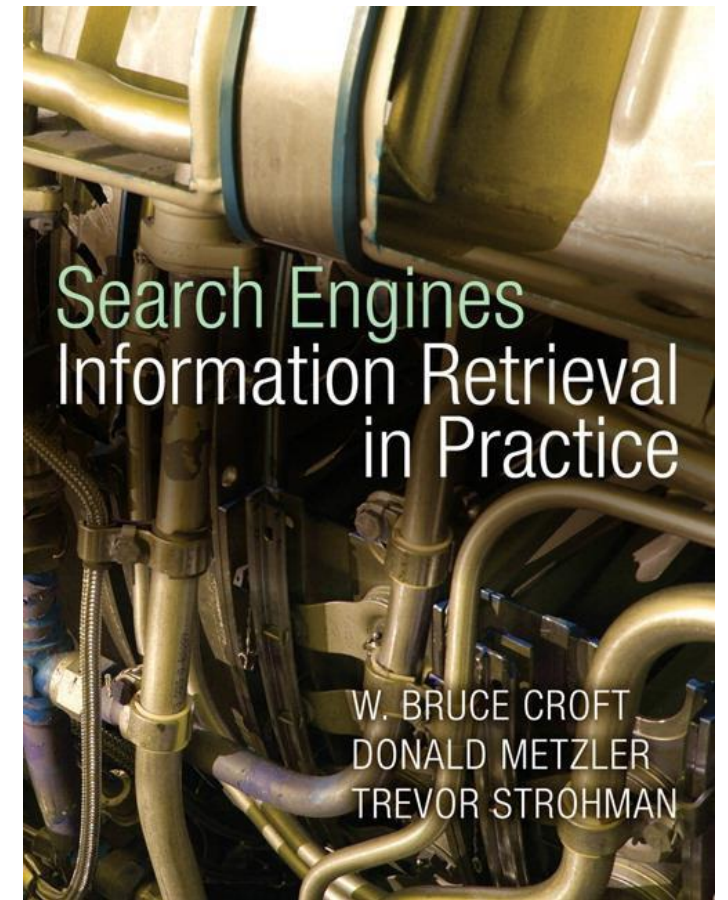
- **Introduction to the Course**
- Overview of the Semester

Text Books

Search Engines: Information Retrieval in Practice

W. Bruce Croft, Donald Metzler, Trevor Strohman

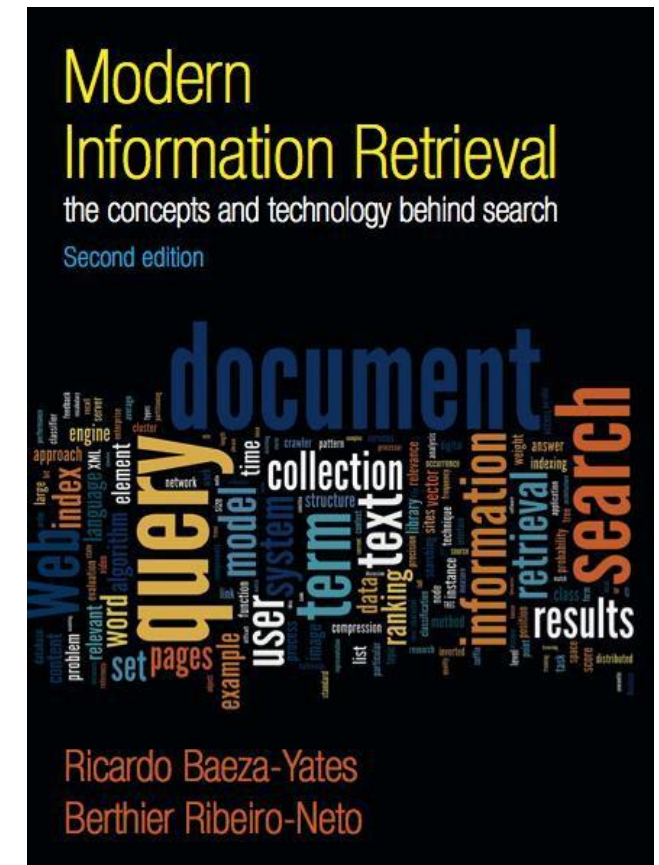
Pearson Education, 2010



Text Books

Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)

Ricardo Baeza-Yates, Berthier Ribeiro-Neto
ACM Press Books, 2010

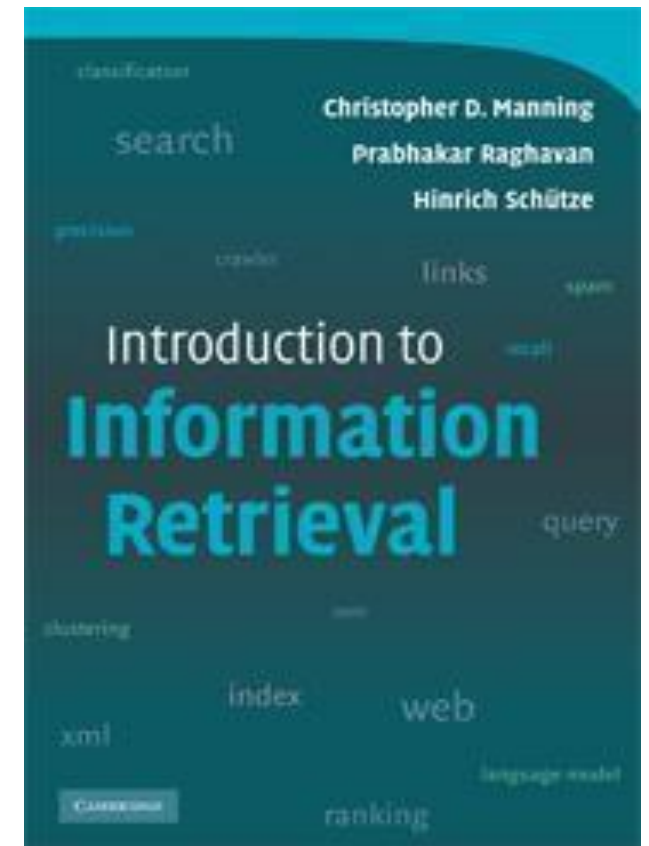


Text Books

Introduction to Information Retrieval

C. Manning, P. Raghavan, and H. Schütze

Cambridge University Press, 2008



Search and Information Retrieval

- Search on the Web is a daily activity for many people throughout the world
 - Google: 40,000 searches per second (3.5 billion per day; 1.2 trillion per year)
 - Yahoo: 3,200 searches per second (280 million per day; 8.4 billion per month)
 - Bing: 927 searches per second (80 million per day; 2.4 billion per month)



10^6 : Million, 10^9 : billion, 10^{12} : Trillion, 10^{15} : Quadrillion, 10^{18} : Quintillion, ...

Search and Information Retrieval

- Search and communication are most popular uses of the computer.
- Applications involving search are everywhere.
- The field of computer science that is most involved with R&D for search is information retrieval (IR).

Information Retrieval

“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”
(Salton, 1968)

- General definition that can be applied to many types of information and search applications
 - Still appropriate after 40 years.
- Primary focus of IR since the 50s has been on text and documents



Data/Information

□ Storage

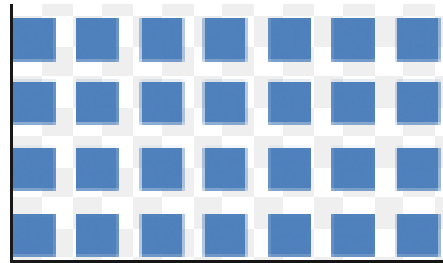


□ Search



Data/Information

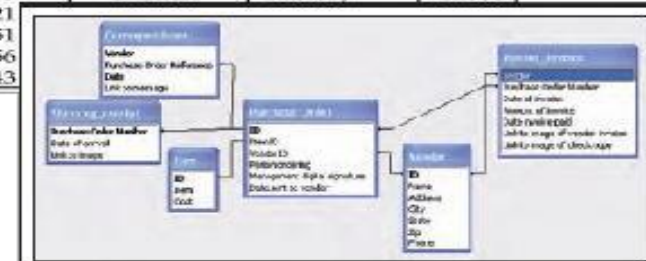
□ Structured



□ Unstructured



ID	name	dept_name	salary
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821			
15151			
33456			
76543			



Techniques such as data mining, Natural Language Processing(NLP), and text analytics provide different means to interpret this information. Common techniques for structuring text usually involve manual tagging with metadata, followed by further text mining-based structuring. Unstructured Information Management Architecture (UIMA) provides a framework for information to extract meaning and create structured data about the information.^[5]

Software that creates machine-processable structure exploits the linguistic, auditory, and visual structure in communication.^[6] Algorithms can infer this inherent structure from text, for instance, by examining word frequency and small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguity and then used to facilitate search and discovery. Examples of "unstructured data" may include books, journals, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still unstructured. For example, an HTML web page is tagged, but HTML mark-up typically serves solely for rendering. It does not support automated processing of the information content of the page. XHTML tagging of elements, although it typically does not capture or convey the semantic meaning of tagged terms.

Since unstructured data commonly occurs in electronic documents, the use of a content or document management system for entire documents is often preferred over data transfer and manipulation from within the documents. Document management systems means to convey structure onto document collections.

Search engines have become popular tools for indexing and searching through such data, especially text.

What is a Document?

□ Examples:

- ❖ Web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM (Instant Messages) sessions, etc.

□ Common properties

- ❖ Significant text content
- ❖ Some structure (\approx attributes in DB)
 - Papers: title, author, date
 - Email: subject, sender, destination, date

Comparing Text

- ❑ Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval.
- ❑ Exact matching of words is not enough
 - ❖ Many different ways to write the same thing in a “natural language” like English
 - Does a news story containing the text “karl benz built the first automobile in 1886” match the query “car inverter”?
- ❑ Defining the meaning of a word, a sentence, a paragraph, or a story is more difficult than defining the meaning of a database field.

Dimensions of IR

□ IR is more than just text, and more than just web search

❖ although these are central

□ People doing IR work with different media, different types of search applications, and different tasks

□ Three dimensions of IR

□ Content

□ Applications

□ Tasks

The Content Dimension

- ❑ Textual data, but...
- ❑ New applications increasingly involve new media
 - ❑ Video, photos, music, speech
 - ❑ Scanned documents (for legal purposes)
- ❑ Like text, content is difficult to describe and compare
 - ❑ Text may be used to represent them (e.g., tags)
- ❑ IR approaches to search and evaluation are appropriate

The Application Dimension

Web search

- Most common

Vertical search

- Restricted domain/topic
- Books, movies, suppliers

Enterprise search

- Corporate intranet
- Databases, emails, web pages, documentation, code, wikis, tags, directories, presentations, spreadsheets

Desktop search

- Personal enterprise search
- See above plus recent web pages

P2P search

- No centralized control
- File sharing, shared locality

Literature search

Forum search

...

The Task Dimension

- ❑ User queries / ad-hoc search
 - ❑ Range of query enormous, not pre-specified
- ❑ Filtering
 - ❑ Given a profile (interests), notify about interesting news stories
 - ❑ Identify relevant user profiles for a new document
- ❑ Classification / categorization
 - ❑ Automatically assign text to one or more classes of a given set
 - ❑ Identify relevant labels for documents
- ❑ Question answering
 - ❑ Similar to search
 - ❑ Automatically answer a question posed in natural language
 - ❑ Provide concrete answer, not list of documents.

Main Issues in IR

□ Relevance

- A relevant document contains the information a user was looking for when he/she submitted the query

□ Evaluation

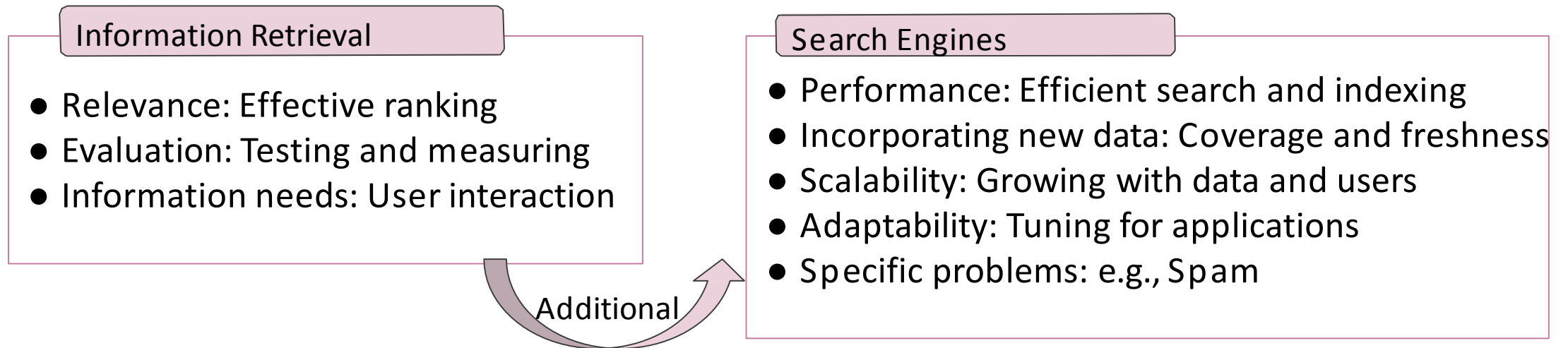
- How well does the ranking meet the expectation of the user

□ Users and information needs

- Users of a search engine are the ultimate judges of quality

IR and Search Engines

- ❑ A search engine is the practical application of information retrieval techniques to large scale text collections
- ❑ Big issues include main IR issues but also some others...



Outline

- Introduction to the Course
- **Overview of the Semester**

Search Engine

❑ Basic architecture

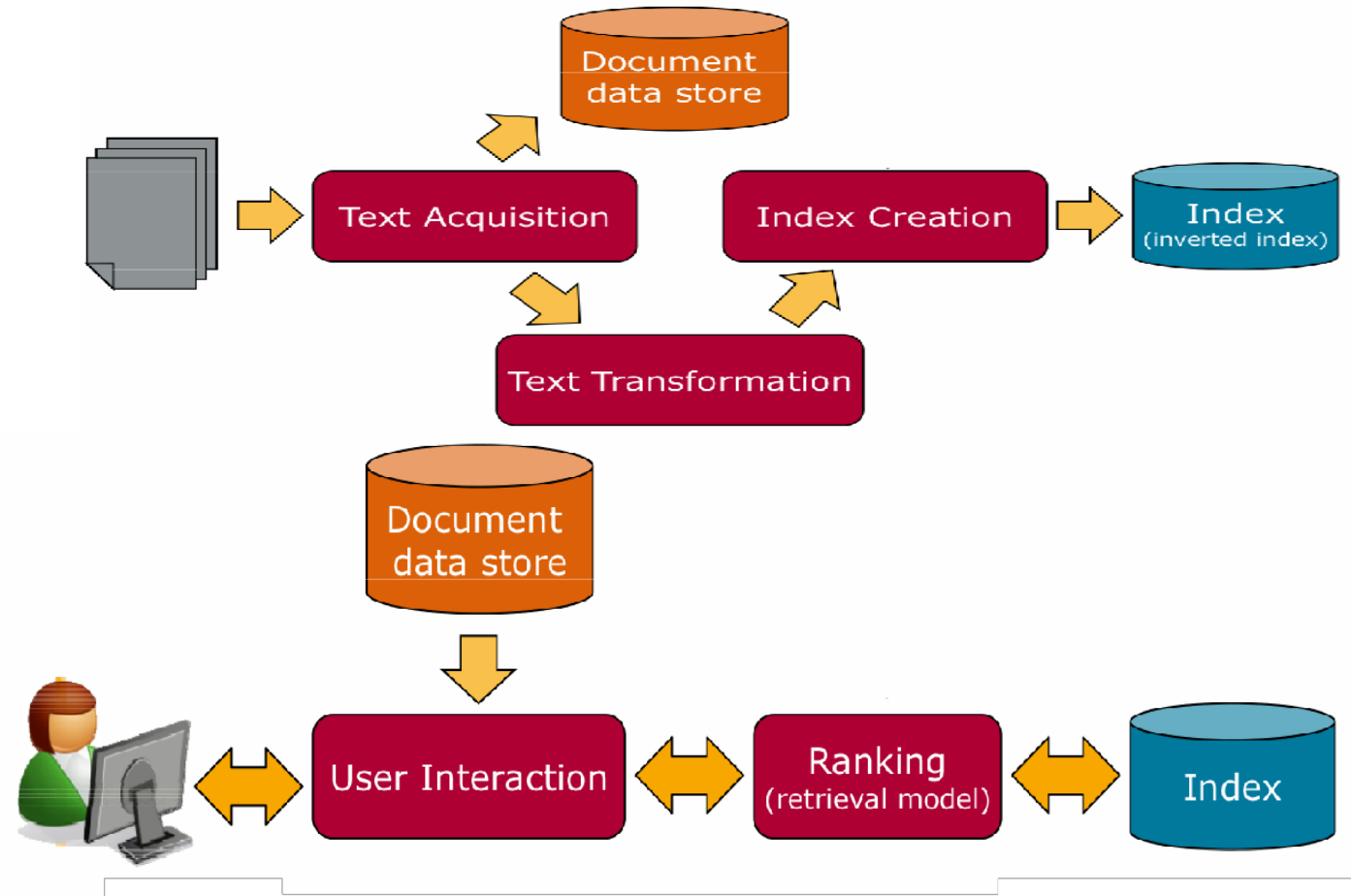
Main issues

❑ Indexing

- ❖ Text acquisition
- ❖ Text transformation
- ❖ Index creation

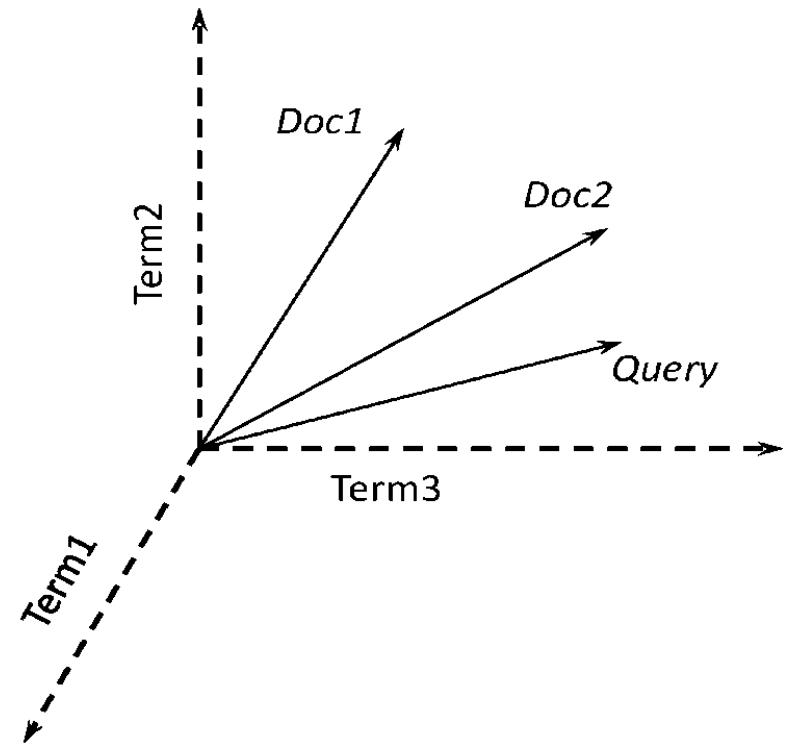
❑ Querying

- ❖ User interaction
- ❖ Ranking
- ❖ Evaluation



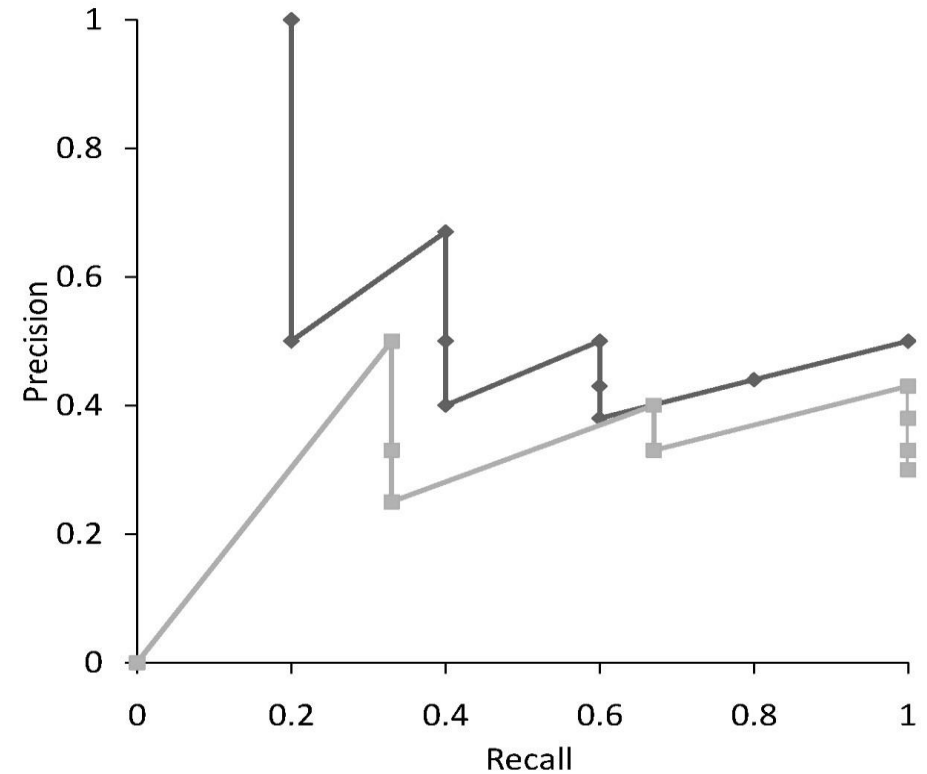
Overview of Traditional Retrieval Models

- ❑ Boolean retrieval
- ❑ Vector space model
- ❑ Probabilistic models



Overview of Evaluation Metrics

- ❑ Effectiveness metrics
- ❑ Efficiency metrics
- ❑ Training, testing, and statistics



Advanced Retrieval Models

- ❑ Language model-based retrieval
- ❑ Learning to rank

Word Mismatch Problem

□ Language model-based approaches

- Translation model
- Topic model
- Word cluster model
- Wordnet
- Dependency model

□ Query expansion approaches

Advanced/Specific IR Tasks

- Query log and query suggestion
- Personalized search
- Information extraction
- Cross-language IR
- Question answering
- Recommendation systems
- Enterprise search
- Digital library
- Structured text retrieval
- Multimedia retrieval

Personalized Search



[Windows Phone \(United States\)](#)

www.windowsphone.com/

“I love my **Windows Phone**. The stuff that's important to me is pinned to my Start screen, like the Live Tile for my All Recipes app which updates regularly with ...

My Phone - Windows Phone 7 - Canada - India

[Windows Phone Apps+Games Store \(United ... - WindowsPhone.com\)](#)

www.windowsphone.com/en-us/store

Browse Apps+Games for Windows Phone. ... Popular and Famous. This collection features the most famous and innovative apps in the Store. First Row ...

[Windows Phone - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Windows_Phone

Windows Phone is a family of mobile operating systems developed by Microsoft, and is the successor to its Windows Mobile platform, although incompatible with ...

[Phones | Windows Phone \(United States\)](#)

www.windowsphone.com/en-us/phones

Features · Phones · Apps+Games · News · How-to. Explore. My Phone. Have a Windows Phone? Sign in · My Phone · Find My Phone · Purchase history ...

[Windows Phone 8 - New Windows 8 Phones, Windows 7 ... - AT&T](#)

www.att.com · [Wireless](#) · [Devices](#)

AT&T has the newest **Windows Phone 8** & Windows 7 Phones, including smartphones and devices from HTC, Samsung, and Nokia. Choose the best Windows 8 ...

[Windows Phone \(windows phone\) on Twitter](#)

<https://twitter.com/windowsphone>

The Official **Windows Phone** Twitter Channel - keeping you updated with the latest ...
“If iOS bores you and Android intimidates you, #WindowsPhone will likely ...

[Windows Phone \(United States\)](#)

www.windowsphone.com/

“I love my **Windows Phone**. The stuff that's important to me is pinned to my Start screen, like the Live Tile for my All Recipes app which updates regularly with ...

My Phone - Windows Phone 7 - Canada - India

[Windows Phone Apps+Games Store \(United ... - WindowsPhone.com\)](#)

www.windowsphone.com/en-us/store

Browse Apps+Games for Windows Phone. ... Popular and Famous. This collection features the most famous and innovative apps in the Store. First Row ...

You've visited this page 2 times. Last visit: 10/29/12

[Yes, you can use Google+ through Internet Explorer on Windows ...](#)

<https://plus.google.com/.../posts/bCS7m7wedyF>



Danny Sullivan – 46 minutes ago – Yes, you can use Google+ through Internet Explorer on **Windows Phone 8**, but it's not a pleasant experience, I've found. I can haz app please

[Windows Phone - Wikipedia, the free encyclopedia](#)



en.wikipedia.org/wiki/Windows_Phone

Windows Phone is a family of mobile operating systems developed by Microsoft, and is the successor to its Windows Mobile platform, although incompatible with ...

[Images for windows phone - Report images](#)



Information Extraction

Google  

[All](#) [Images](#) [Maps](#) [Videos](#) [News](#) [More ▾](#) [Search tools](#)

About 17,700,000 results (0,53 seconds)

Microsoft Corporation / Headquarters



Redmond, Washington, United States

Points of interest and overview

[Feedback](#)



People also ask

- Who is Microsoft? ▾
- Who is the owner of Microsoft? ▾
- When did Microsoft go public with their stocks? ▾
- How much is the Microsoft company worth? ▾

Microsoft - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Microsoft> ▾

Jump to **Headquarters** - The corporate headquarters, informally known as the Microsoft Redmond campus, is located at One Microsoft Way in Redmond, Washington. Microsoft initially moved onto the grounds of the campus on February 26, 1986, weeks before the company went public on March 13.

Headquarters: Microsoft Redmond campus, R... Founders: Bill Gates; Paul Allen
Owner: Bill Gates (3%) Founded: April 4, 1975; 41 years ago; Albuque...

Google  

[All](#) [Images](#) [Maps](#) [News](#) [Videos](#) [More ▾](#) [Search tools](#)

About 911,000 results (0,55 seconds)



See photos

Milad Hospital ★ [Website](#) [Directions](#)

4.0 ★★★★★ 56 Google reviews
Hospital in Tehran, Iran

Milad Hospital is the largest specialized and subspecialized hospital in Iran. This hospital is a complementary health service provider in Iran's Social Security organization chain of hospitals. [Wikipedia](#)

Address: Tehran Province, Tehran, Hemmat Expy, Iran
Height: 54 m
Phone: +98 919 500 7207
Number of beds: 1,000
Founded: 2001
People also search for: Milad Tower, Roudaki Hall, more

[Suggest an edit](#) - [Own this business?](#)

Cross-language Retrieval



"دانشگاه کاشان"

All Images Maps Books Videos More Settings Tools

Kashan > Colleges and Universities

University of Kashan



Islamic Azad University of Kashan



Kashan University of Medical Sciences



دانشگاه کاشان

<https://kashanu.ac.ir/> Translate this page

وب سایت رسمی دانشگاه دولتی کاشان شامل معرفی دانشگاه ، دانشکده ها و کلیه موسسات و سازمانهای زیرمجموعه دانشگاه.

سیستم اتوماسیون تغذیه

سامانه اتوماسیون تغذیه دانشگاه کاشان ... همچنین لازم به نگر است ...

دانشگاه کاشان

تهیه شده توسط مرکز فناوری اطلاعات و ارتباطات دانشگاه فردوسی مشهد ...

آدرس

آدرس: کاشان، کیلومتر 8 بلوار قطب روانی، دانشگاه کاشان، دانشکده ...

[More results from kashanu.ac.ir »](#)

پورتال کتابخانه دیجیتال

پورتال کتابخانه دیجیتال. فهرست ها. صفحه اصلی . کتابخانه ...

ایمیل

جهت دسترسی به صفحه ورود ایمیل تان بر روی پیوند سرویس ...

مدیریت تربیت بدنی

نشانی: کاشان - کیلومتر 5 بلوار قطب روانی - دانشگاه کاشان ...

دانشگاه کاشان - ویکی‌پدیا، دانشنامه آزاد



Kashan

City in Iran

Kashan is a city of Isfahan, Iran. At the 2006 census, its population was 248,789, in 67,464 families. [Wikipedia](#)

Weather: 23°C, Wind NW at 19 km/h, 22% Humidity

Population: 275,325 (2011) UNdata

Question Answering



Vind Antwoorden

Recommendation Systems

amazon **Books** modern information retrieval **Prime student** 5
Departments - Your Amazon.com Today's Deals Gift Cards & Registry Sell Help Hello, Sign in Your Account - Try Pri

Books Advanced Search New Releases Best Sellers The New York Times Best Sellers Children's Books Textbooks Textbook Rentals Sell Us Your Books Best Books of the Month Deals in Books

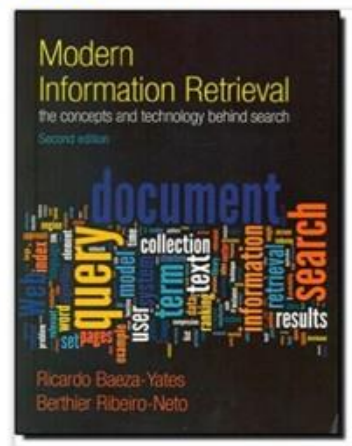
Prime student **FREE Two-Day Shipping** for college students [Learn more](#)

Back to search results for "modern information retrieval"

Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books)

by Ricardo Baeza-Yates (Author), Berthier Ribeiro-Neto (Author)

★★★★☆ 4 customer reviews



ISBN-13: 978-0321416919
ISBN-10: 0321416910

Paperback
\$63.74

Other Sellers
from \$53.95

Buy new
Usually ships v
Ships from and

Customers Who Bought This Item Also Bought



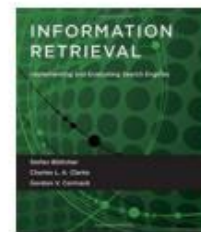
Introduction to Information Retrieval
Christopher D. Manning
★★★★☆ 25
Hardcover
\$65.51 **Prime**



Information Retrieval: Implementing and Evaluating Search Engines (MIT Press)
Stefan Büttcher
★★★★☆ 5
Hardcover



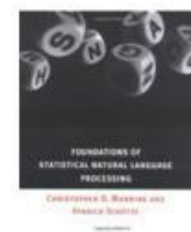
Search Engines: Information Retrieval in Practice
Bruce Croft
★★★★☆ 17
Hardcover
\$123.22 **Prime**



Information Retrieval: Implementing and Evaluating Search...
Stefan Büttcher
★★★★☆ 5
Paperback
\$41.42 **Prime**



Managing Gigabytes: Compressing and Indexing Documents and Images...
Ian H. Witten
★★★★☆ 11
Hardcover
\$93.50 **Prime**



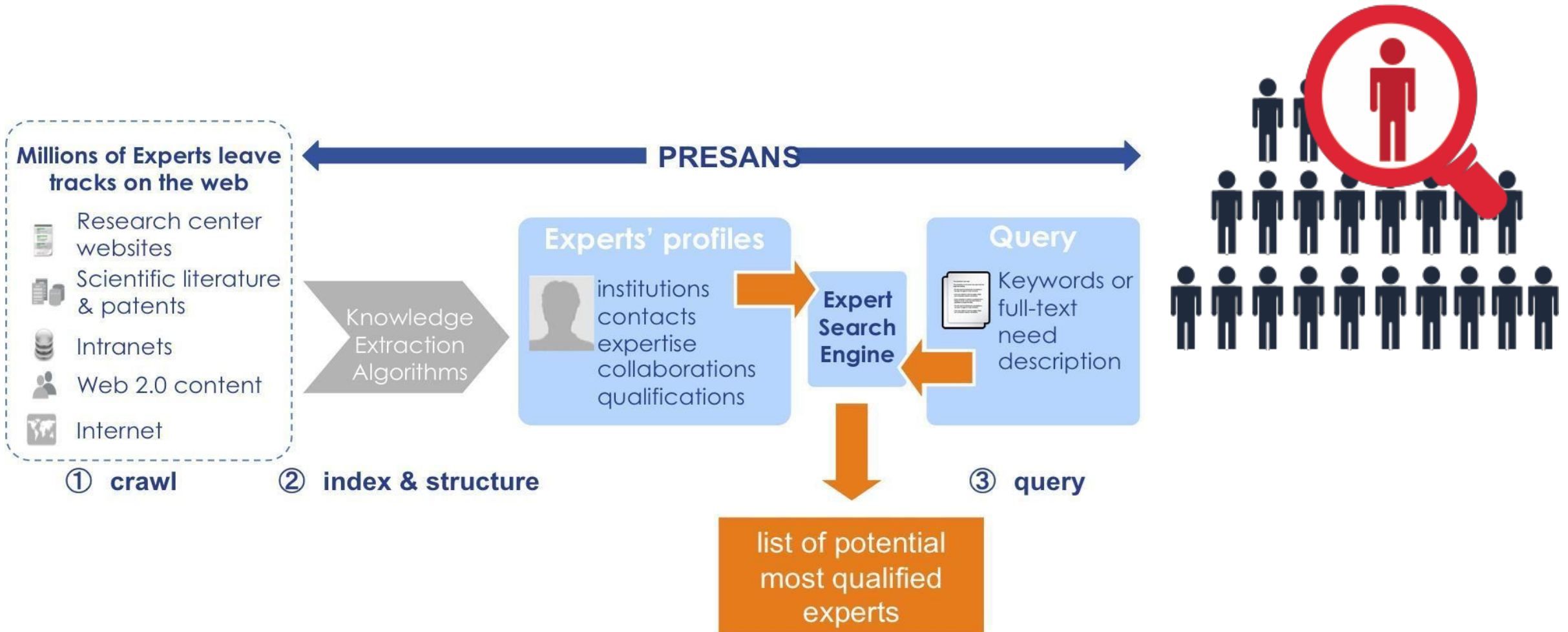
Foundations of Statistical Natural Language Processing
Christopher D. Manning
★★★★☆ 23
Hardcover
\$99.00 **Prime**



Speech and Language Processing, 2nd Edition
Daniel Jurafsky
★★★★☆ 20
Hardcover
\$174.53 **Prime**

More Buyi
20 New from \$

Enterprise Search



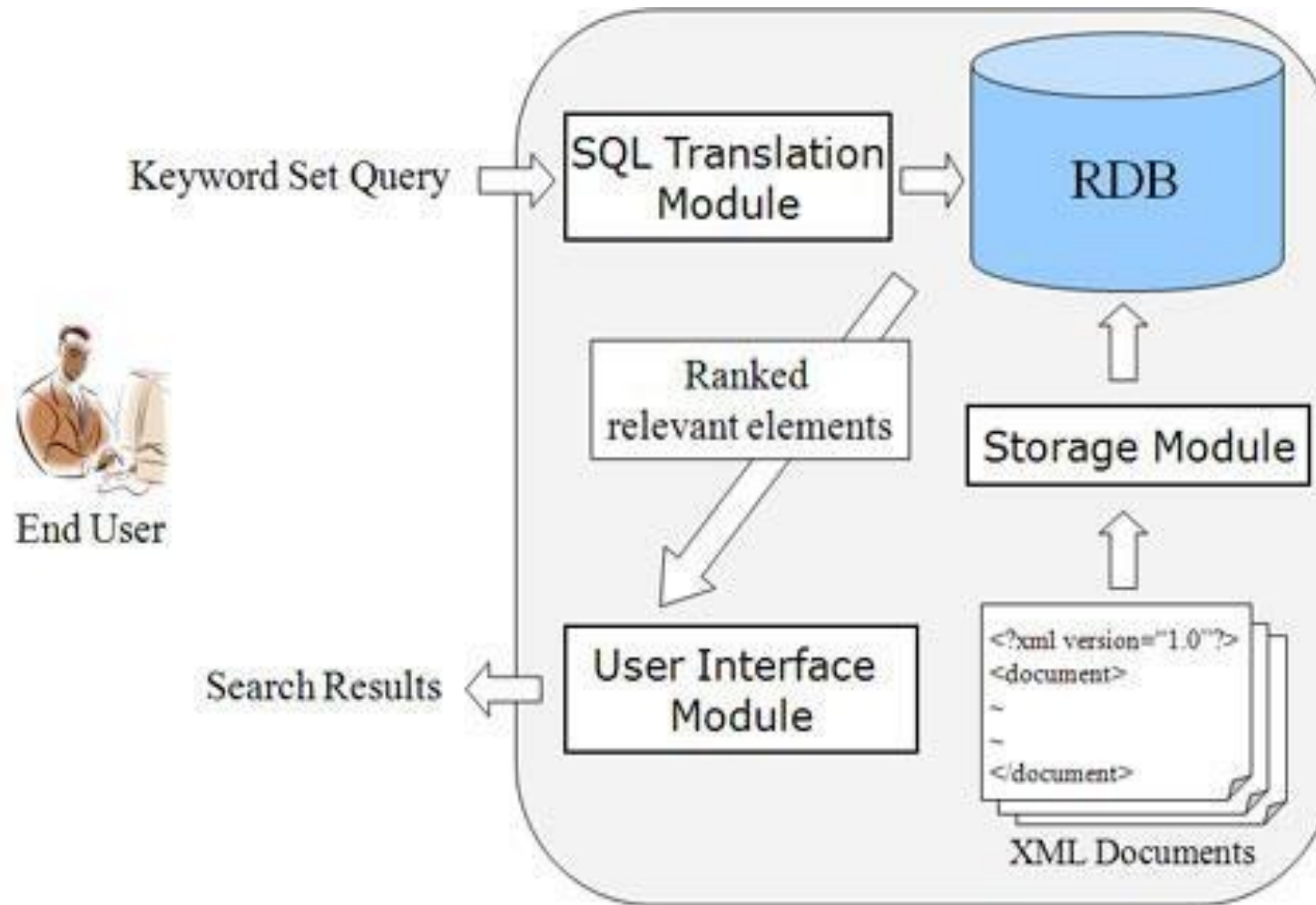
Digital Library



ACM **DL** DIGITAL LIBRARY



Structured Text Retrieval

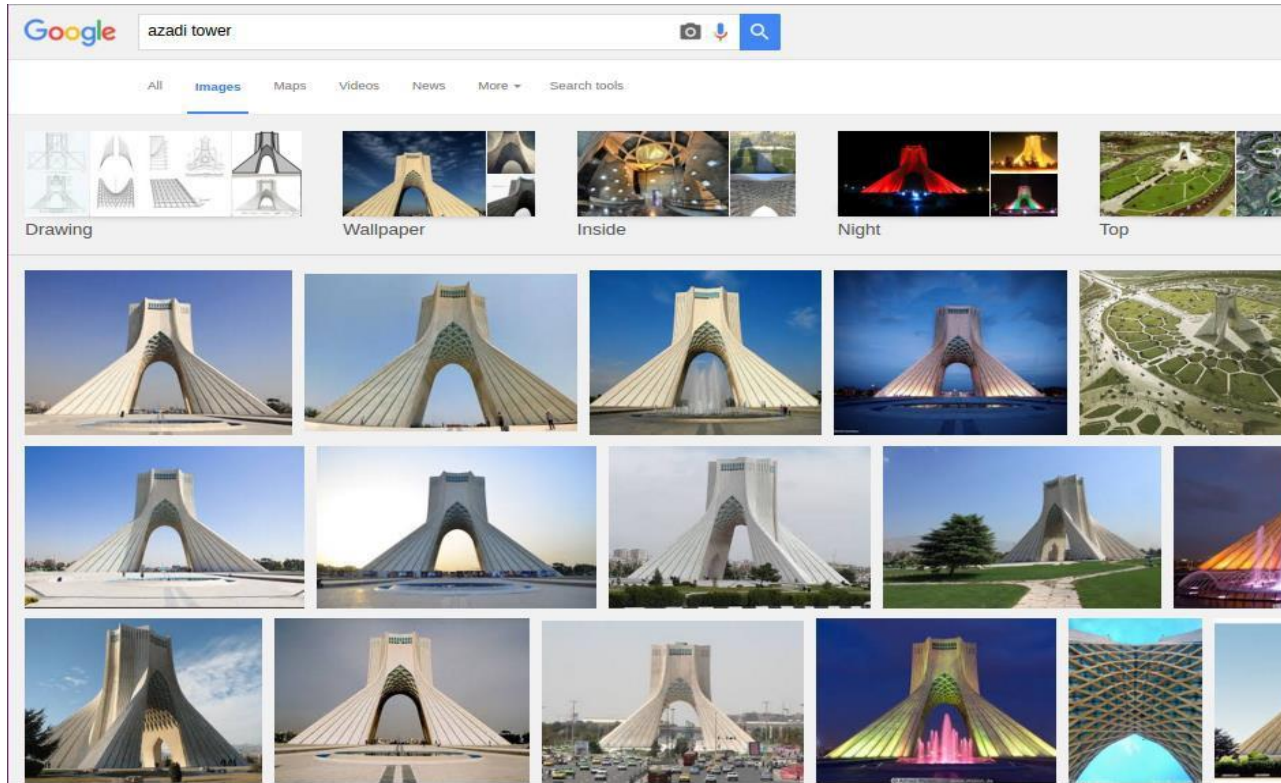


Multimedia Retrieval

Google azadi tower

All Images Maps Videos News More Search tools

Drawing Wallpaper Inside Night Top



Google azadi tower

All Images Maps Videos News More Search tools

About 5,180 results (0.14 seconds)

Azadi Tower, Tehran, Iran - YouTube
<https://www.youtube.com/watch?v=0GabRvNcK4c>
Nov 23, 2013 - Uploaded by vidd.it
Azadi Tower, Tehran, Iran Image Credits Azadi Tower, Tehran, Iran ...

Iran Téhéran Tour Azadi / Iran Tehran Azadi tower - YouTube
<https://www.youtube.com/watch?v=O4zcGhZMPJU>
Jan 15, 2016 - Uploaded by hors frontieres
Iran Téhéran Tour Azadi / Iran Tehran Azadi tower. hors frontieres. Subscribe
SubscribedUnsubscribe 1,0441K ...

Light festival on Tehran's Azadi tower - YouTube
<https://www.youtube.com/watch?v=vkjY3mhSa7A>
Oct 6, 2015 - Uploaded by Ahmad Mousavizadeh
First light installation in Iran was organized in Azadi Tower by Philipp Geist, the German artist, on Saturday ...

Azadi Tower - YouTube
<https://www.youtube.com/watch?v=JIWpg-GmRhY>
Jan 13, 2015 - Uploaded by Iran Program Presstv
Azadi Tower. Iran Program Presstv. SubscribeSubscribedUnsubscribe 5,2985K.
Loading... Loading ...

Light display underway at Tehran Azadi tower - YouTube
<https://www.youtube.com/watch?v=oLt7lh9-oZg>
Oct 4, 2015 - Uploaded by PresstV News Videos
Peace and freedom are concepts which everyone strives to realize. And that is what a German artist is also ...



Questions?