The Art and Science behind Search Engines

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

Information is the resolution of uncertainty. **- Claude Shannon.**

Agenda

Introduce "Search" as a potential career option (engineering/research).

Google
Google Search
I'm Feeling Lucky
Google offered in: हिन्दी বাংলা తెలుగు मराठी தமிழ் ગુજરાતી ಕನ್ನಡ മലയാളം ਪੰਜਾਬੀ

Life without search engines is difficult to imagine!

Role of Information

- We are always seeking information
  - How to be safe during COVID times?
  - Which universities provide the specialization I need?
  - How to get into a top college?
  - Which course to register for?
  - What are people talking about that new movie?
  - How to prepare for interviews?
  - …
- If only you had the information, you could rule this world!

What happens when all the information is deprived from you?

Solitary Confinement is Cruel

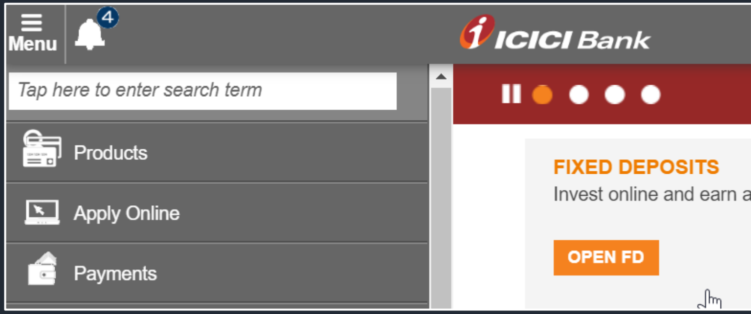
Picture Source: https://jjustice.org/resources/solitary-confinement

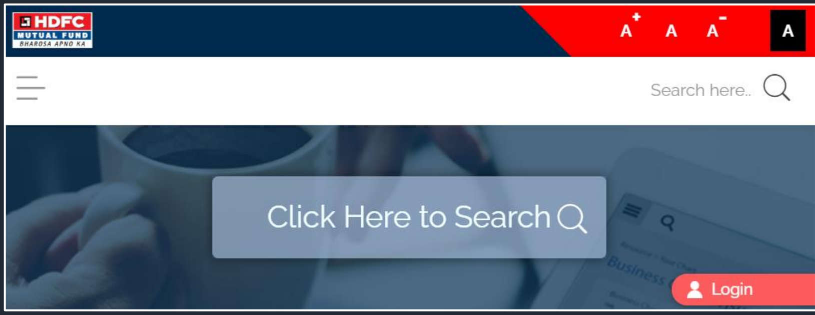
There is search beyond the web search engines.

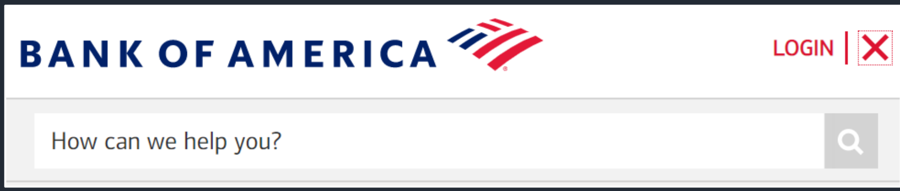
Search in Banking and Finance

Lots of products to sell

Reach a part of documentation faster

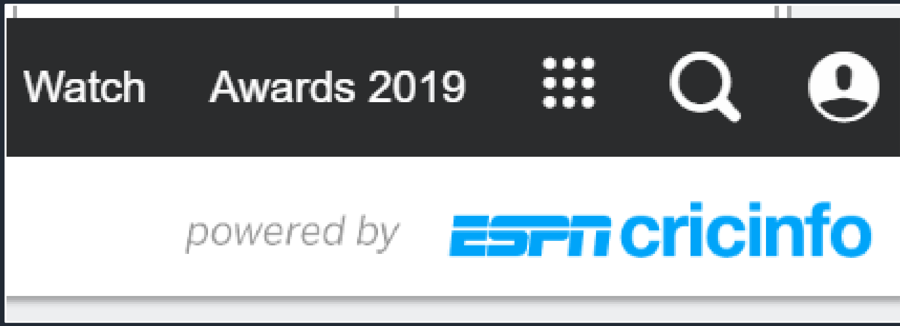


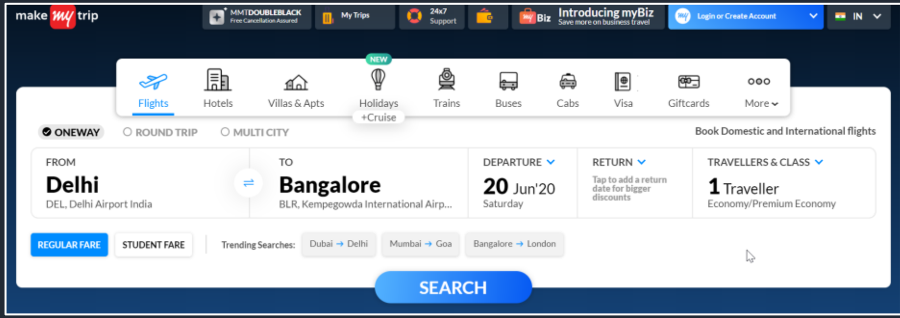




Search in Sports, Travel & Entertainment

Search events, programs, and schedules





● ONEWAY ○ ROUND TRIP ○ MULTI CITY

Book Domestic and International flights

FROM: Delhi — DEL, Delhi Airport India

TO: Bangalore — BLR, Kempegowda International Airp...

DEPARTURE: 20 Jun'20 Saturday

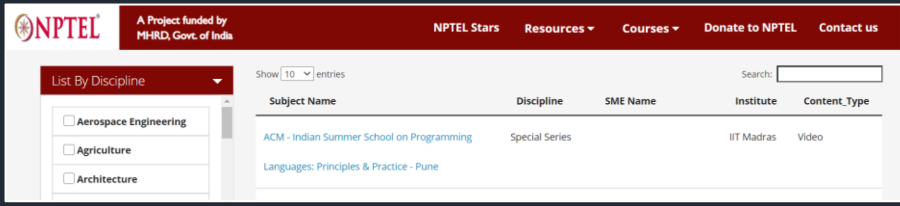
RETURN: Tap to add a return date for bigger discounts

TRAVELLERS & CLASS: 1 Traveller — Economy/Premium Economy

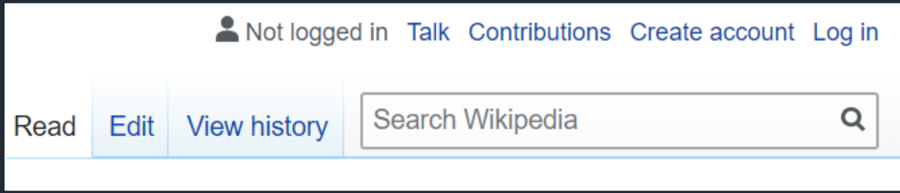
Trending Searches: Dubai → Delhi | Mumbai → Goa | Bangalore → London

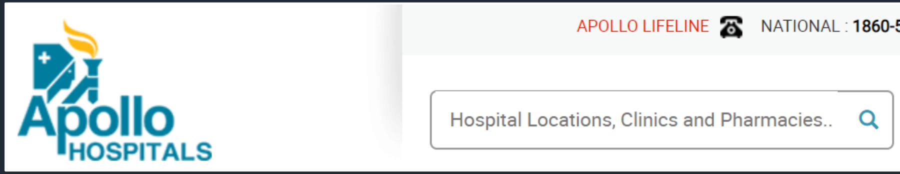
Search in Education, Ecommerce and Healthcare

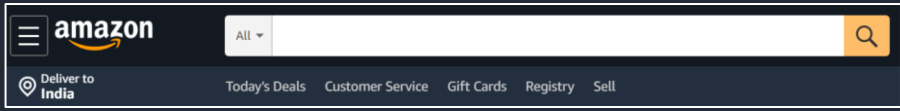
Search courses, articles, symptoms, books, etc.

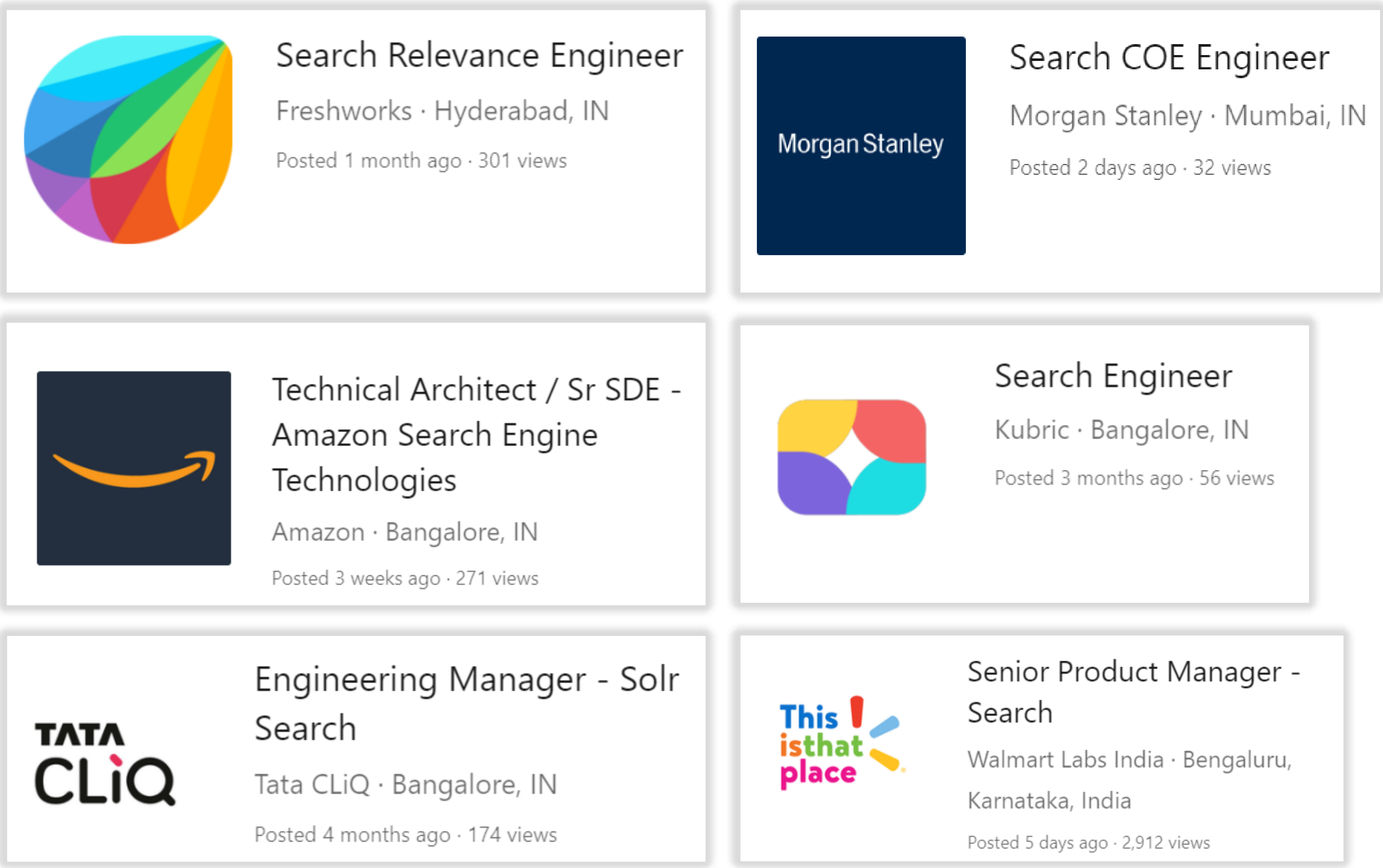


Subject Name	Discipline	SME Name	Institute	Content_Type
ACM - Indian Summer School on Programming	Special Series		IIT Madras	Video
Languages: Principles & Practice - Pune				




Apollo HOSPITALS
APOLLO LIFELINE ☎ NATIONAL : 1860-5
Hospital Locations, Clinics and Pharmacies..





Search Relevance Engineer

Freshworks · Hyderabad, IN

Posted 1 month ago · 301 views

Search COE Engineer

Morgan Stanley · Mumbai, IN

Posted 2 days ago · 32 views

Technical Architect / Sr SDE - Amazon Search Engine Technologies

Amazon · Bangalore, IN

Posted 3 weeks ago · 271 views

Search Engineer

Kubric · Bangalore, IN

Posted 3 months ago · 56 views

Engineering Manager - Solr Search

Tata CLiQ · Bangalore, IN

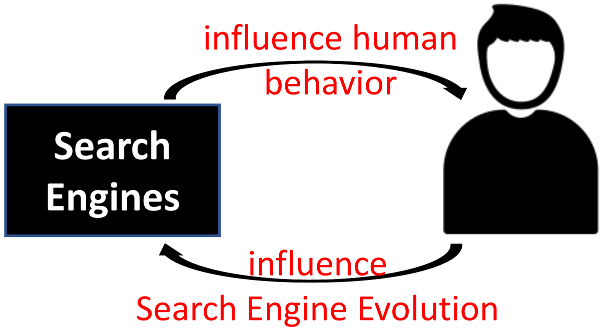
Posted 4 months ago · 174 views

Senior Product Manager - Search

Walmart Labs India · Bengaluru, Karnataka, India

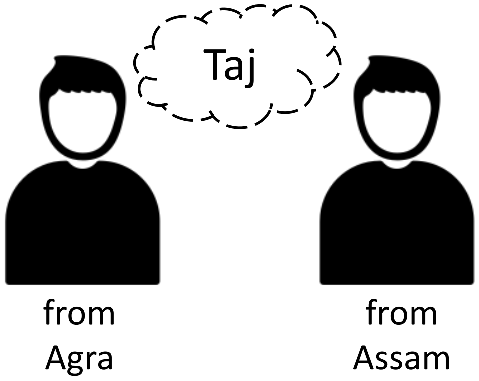
Posted 5 days ago · 2,912 views

Results of job search conducted on 18th June 2020 on https://www.linkedin.com/jobs for solr/information retrieval/search

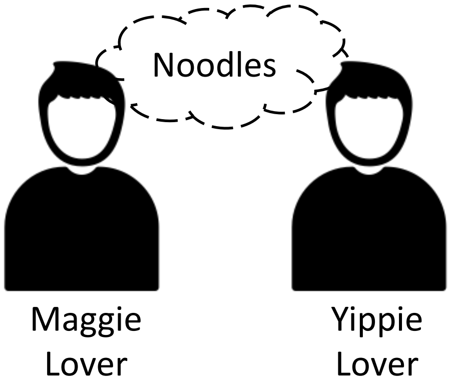


We have amazing search products.
Yet, an effective search is still a work of art!

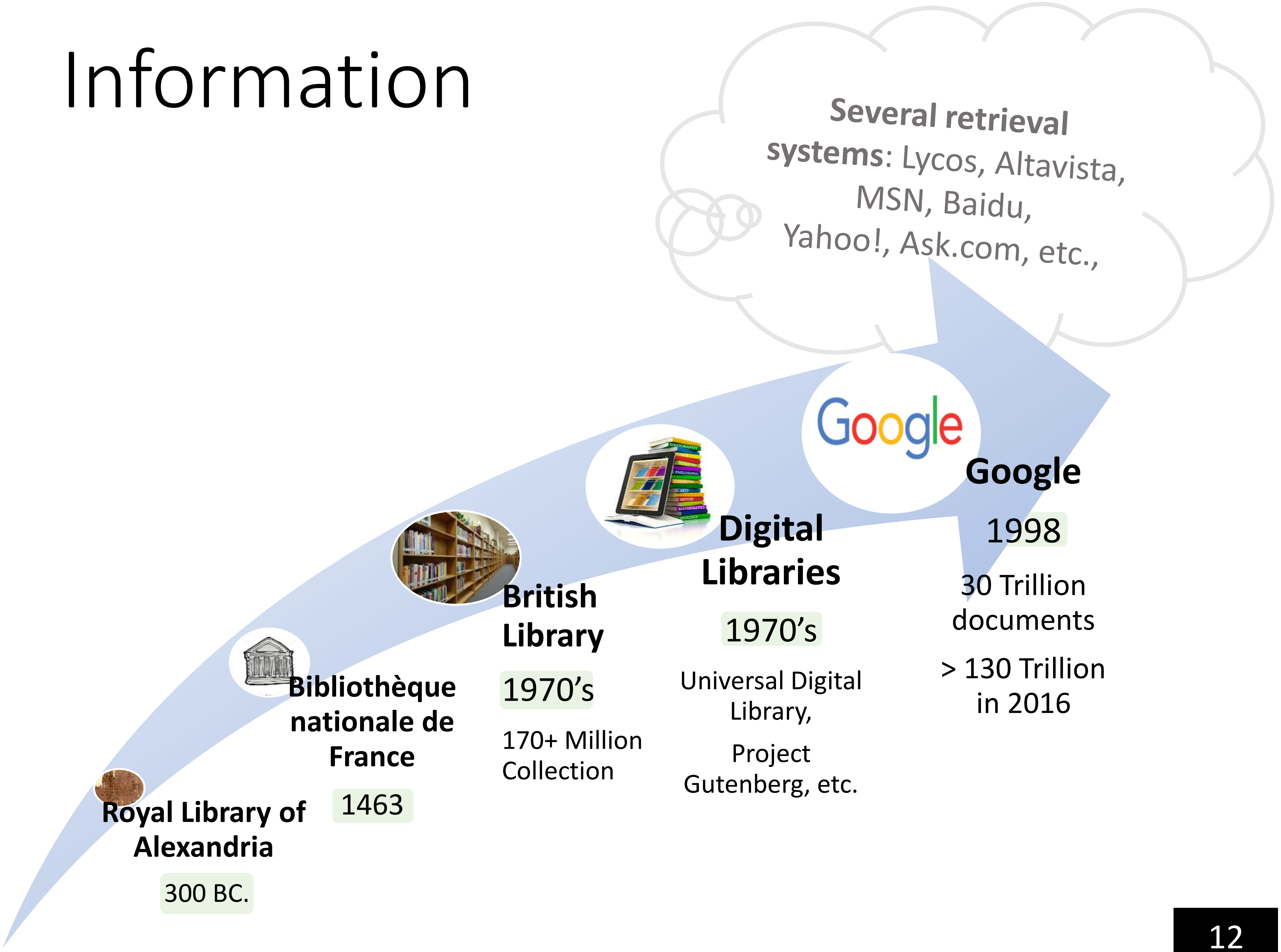
User needs to express his information need properly



Relevance may be subjective



Information



Several retrieval systems: Lycos, Altavista, MSN, Baidu, Yahoo!, Ask.com, etc.,

Royal Library of Alexandria

300 BC.

Bibliothèque nationale de France

1463

British Library

1970's

170+ Million Collection

Digital Libraries

1970's

Universal Digital Library,

Project Gutenberg, etc.

Google

1998

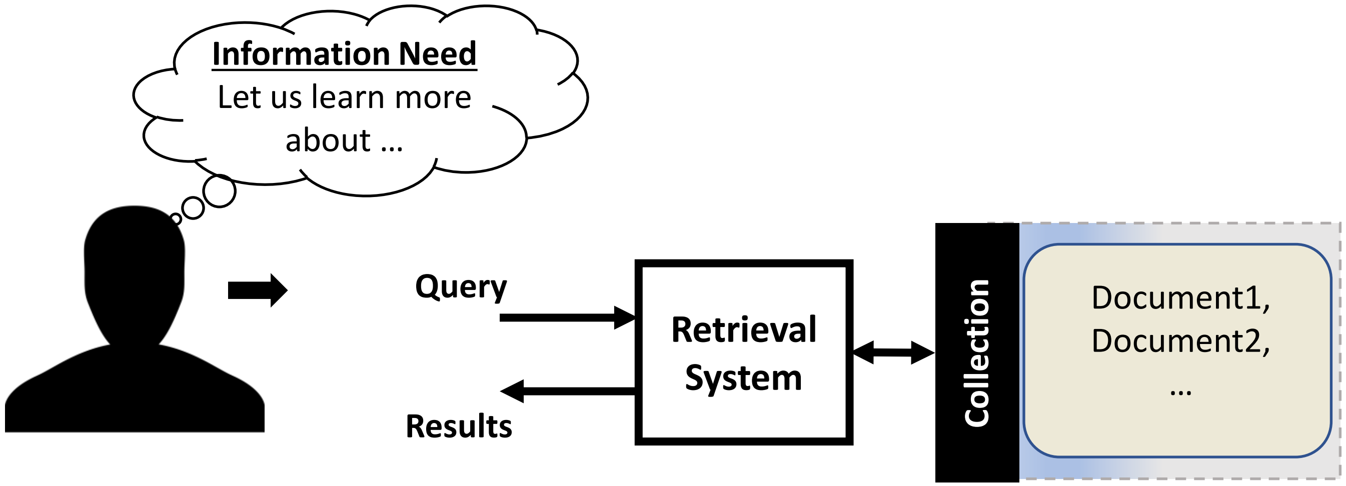
30 Trillion documents

> 130 Trillion in 2016

The Science behind Search

How do search engines work?

What is Information Retrieval?



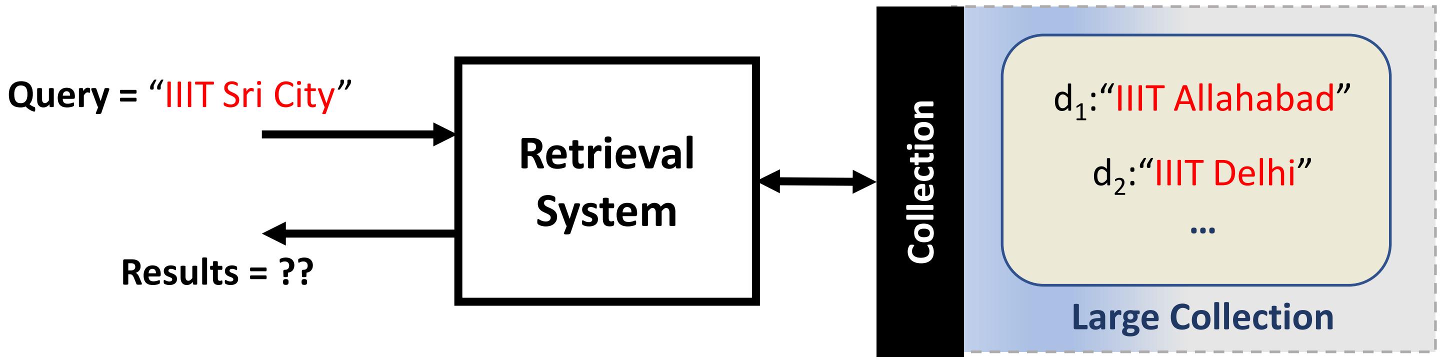
Information Retrieval (IR) is finding material (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections**.

– From the Manning et al. IR Book.

Simple Retrieval Problem

- Say, we have a **collection** with 5 **documents**, each having the following contents
  - d1: IIIT ALLAHABAD
  - d2: IIIT DELHI
  - d3: IIIT GUWAHATI
  - d4: IIIT KANCHIPURAM
  - d5: IIIT SRI CITY
- Assume, the **Query** is
  - IIIT SRI CITY
- Which **document** will you match and why?

The Problem: How to Build a Retrieval System?



One (bad) Approach

- First match the **term** IIIT.
  - Filter out documents that contain this term.
- Next match the **term** Sri.
  - Filter out documents that contain this term.
- Next match the **term** City.
  - Filter out documents that contain this term.

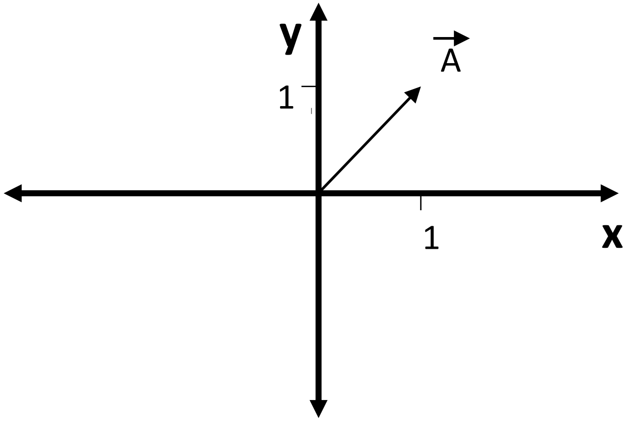
Three iterations!
Quiz: Can we do better?

A Better Approach

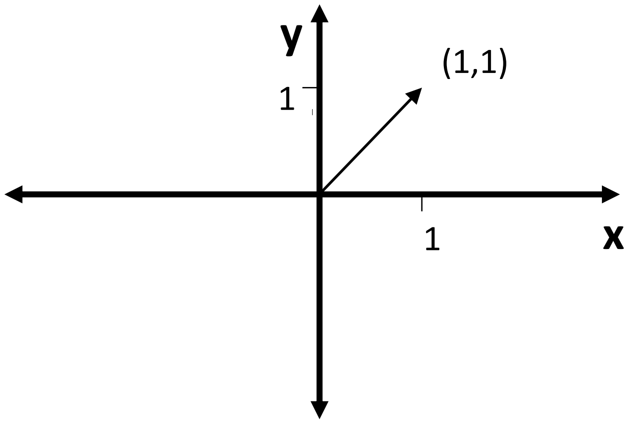
Revisiting
Linear Algebra

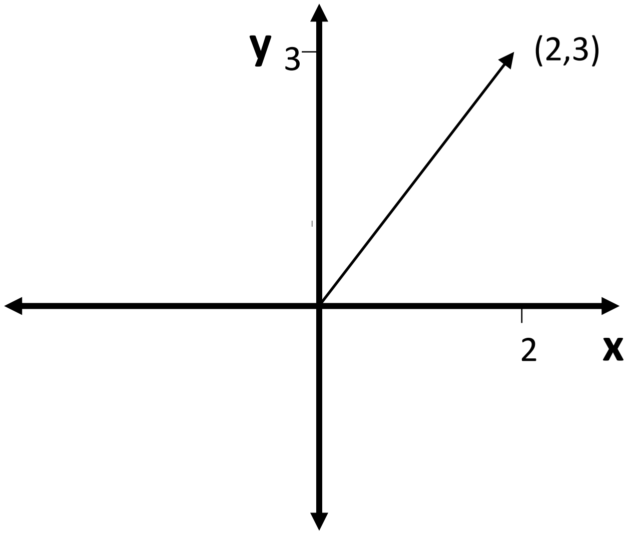
Vectors

- Geometric entity which has magnitude and direction

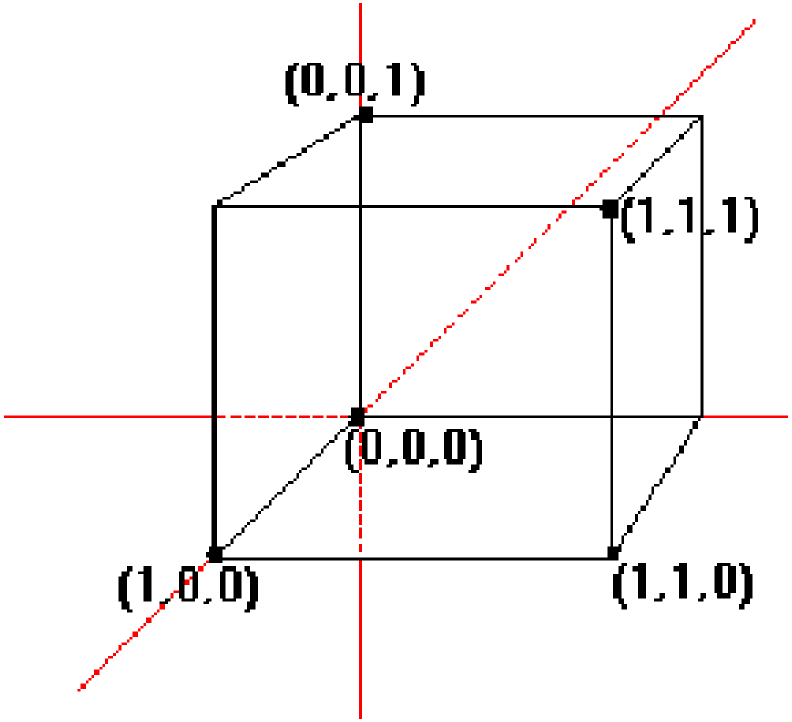


How is (2,3) Different?





What is (1,1,1) ?

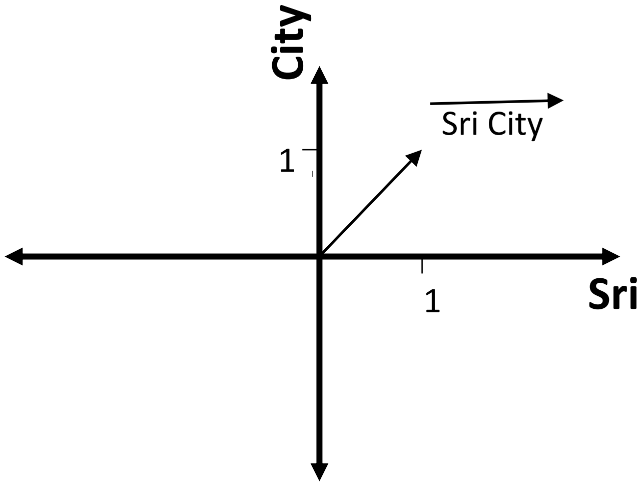


Remember!

A number is just a mathematical object. We give meaning to it!

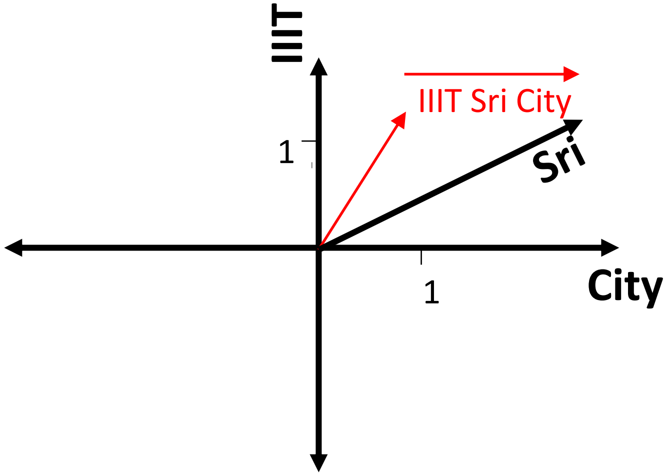
Sentences are Vectors

- “Sri City” as a vector



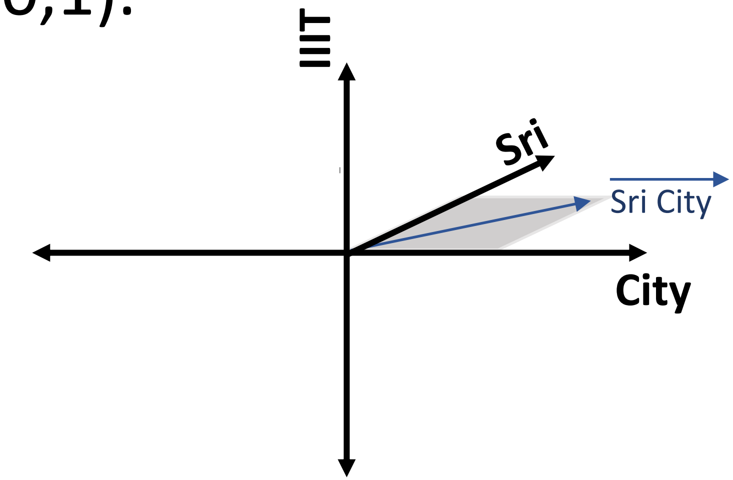
Sentences are Vectors

- “IIIT Sri City” is a 3-dimensional vector



Sentences are Vectors

- On this 3D space, “Sri City” vector will lie on the x (City) and z (Sri) plane. If (x,y,z) denotes the vector, “Sri City” is (1,0,1).



Natural Language Phrases as Vectors

Let query q = “IIIT Sri City”.

Let document, d_1 = “IIIT Sri City” and d_2 = “IIIT Delhi”.

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

q = (1,1,1,0), d_1= (1,1,1,0) and d_2 = (1,0,0,1)

Quiz

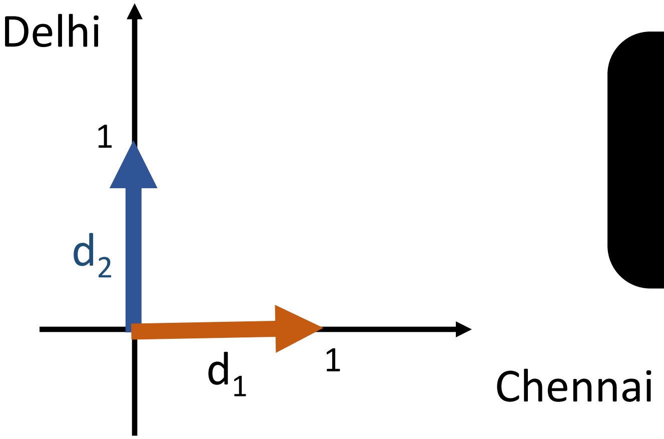
- Considering the following vectors:

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

- What is the Natural Language (NL) equivalent of (0,1,1,0) ?
- What is the NL equivalent of (1,0,0,1) ?
- What is the vector for Delhi?
- What is the NL equivalent of q here?

Similarity Score

- Assume, we have the following two documents:
    - d_1 = “Chennai”
    - d_2 = “Delhi”
- On a scale of 0 – 1, how similar are d_1 and d_2?
- What is the angle between d_1 and d_2 vectors?



Can we express similarity as a function of the angles?

0 – 90 to 1 – 0: How?

	0°	30°	45°	60°	90°
sin	0	1/2	1/√2	√3/2	1
cos	1	√3/2	1/√2	1/2	0
tan	0	1/√3	1	√3	Not defined

Back to Trigonometry: Dot Product

- If x and y are non-unit vectors, what is the cosine of angle between them (cos Ɵ)?

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos(\theta)$$

(or)

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}$$

Matching Documents to Queries

- Document as a vector of term-occurrence

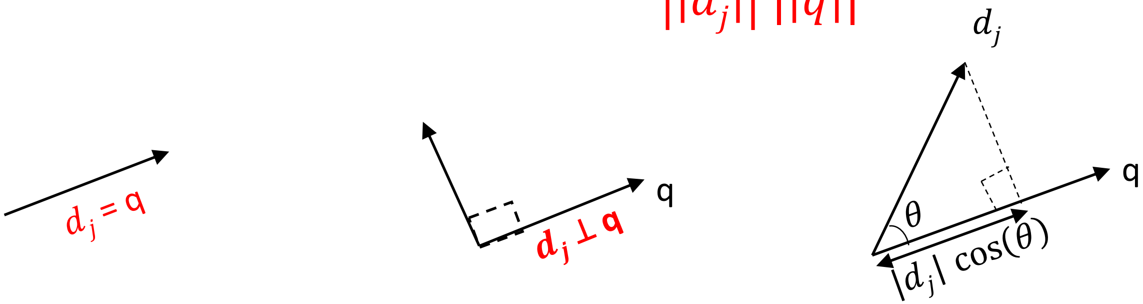
$$d_j = (w_{1j}, w_{2j}, \ldots w_{nj})$$

- Query as a vector of term-occurrence

$$q = (w_{1q}, w_{2q}, \ldots w_{mq})$$

- Similarity between these vectors can be represented as

$$Cosine\ Similarity = \cos(\theta) = \frac{d_j . q}{||d_j||\ ||q||}$$



Example

Let query q = “BITS Pilani”.

Let document, d_1 = “BITS Pilani Goa Campus” and d_2 = “IIIT Delhi”.

	BITS	Pilani	Goa	Campus	IIIT	Delhi
q	1	1	0	0	0	0
d_1	1	1	1	1	0	0
d_2	0	0	0	0	1	1

In our VSM, q = (1,1,0,0,0,0), d_1= (1,1,1,1,0,0) and d_2 = (0,0,0,0,1,1)

similarity(d_1, q) $= \frac{d_1.q}{||d_1||\ ||q||} = \frac{1.1 + 1.1}{\sqrt{1^2+1^2+1^2+1^2}\ \sqrt{1^2+1^2}} = 0.71.$

similarity(d_2, q) $= \frac{d_2.q}{||d_2||\ ||q||} = 0.$

Which of the Following are Sets?

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Bag

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Set of Words Representation

- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → {IIIT, Sri, City}

(Assuming, we ignore the punctuations)

	IIIT	Sri	City
q	1	1	1

Bag of Words Representation

- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → [IIIT, Sri, Sri, City, City]

IIIT Sri City

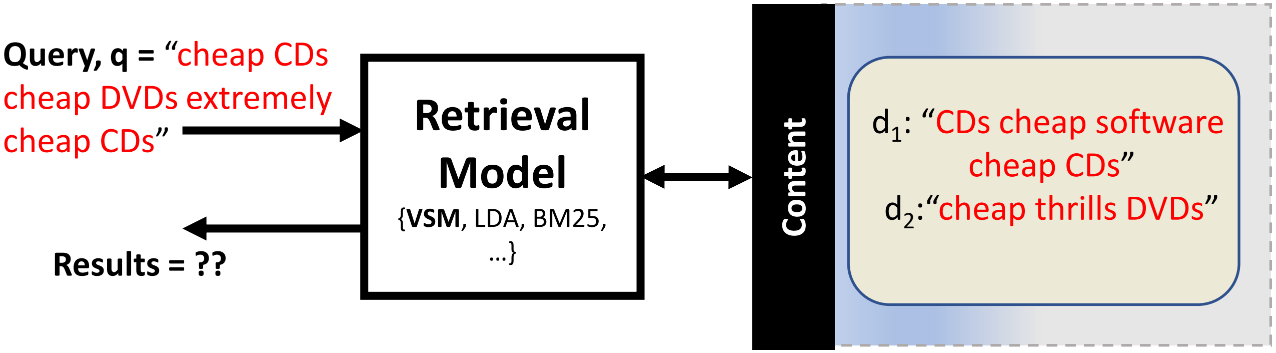
	IIIT	Sri	City
q	1	1	1

IIIT Sri City, Sri City

	IIIT	Sri	City
q	1	2	2

Leads to different vectors

Which Document to Retrieve?



Query, q = “cheap CDs cheap DVDs extremely cheap CDs”

Results = ??

d_1: “CDs cheap software cheap CDs”

d_2:“cheap thrills DVDs”

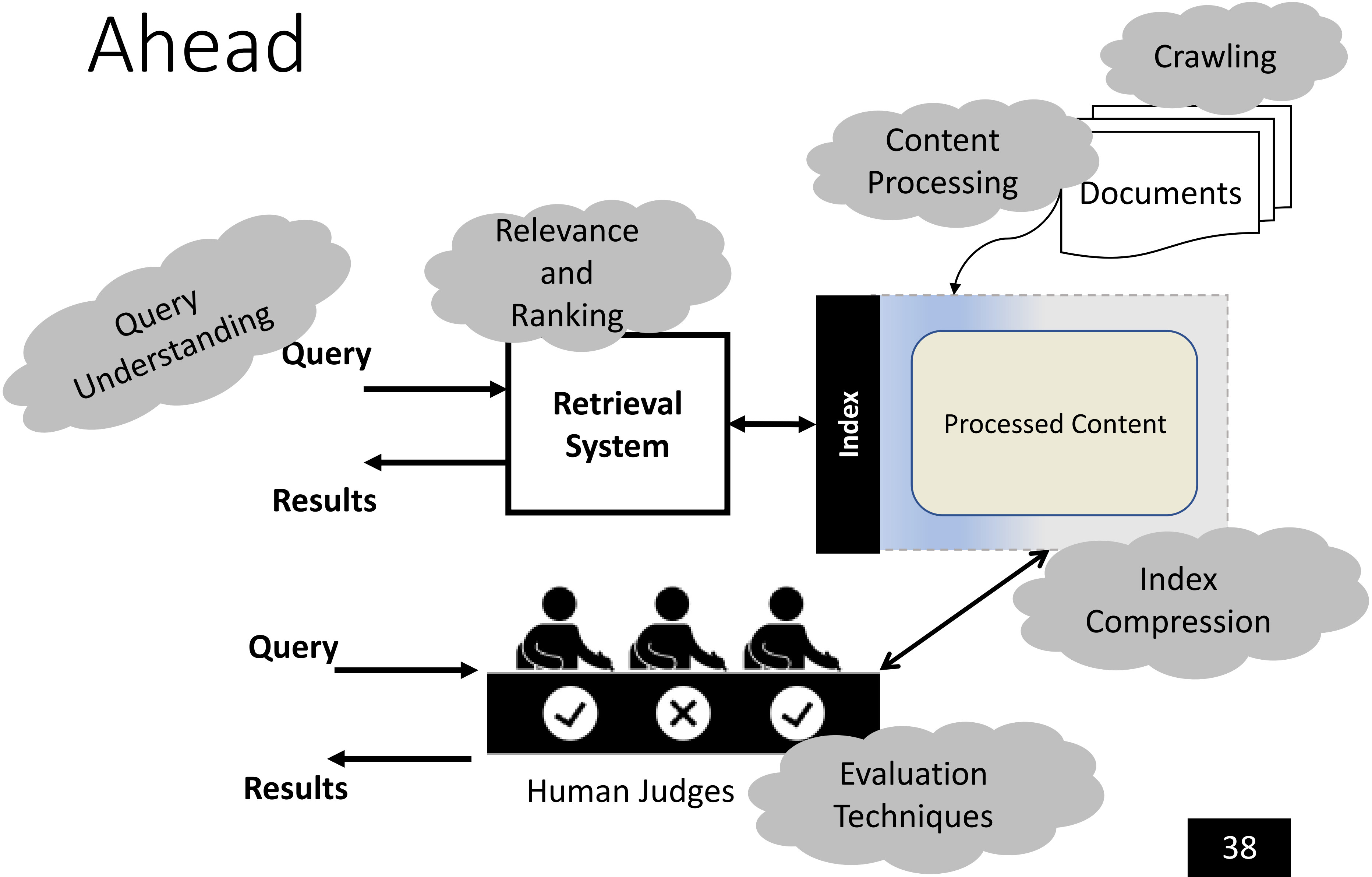
	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

$\text{sim}(q,d_1) = 0.86$

$\text{sim}(q,d_2) = 0.59$

Example taken from Manning’s IR Book (https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf)

Information Retrieval – Road Ahead



Technologies & Frameworks

Lucene™

Apache

Solr

Apache

nutch™

Apache

Terrier

Univ. of Glasgow

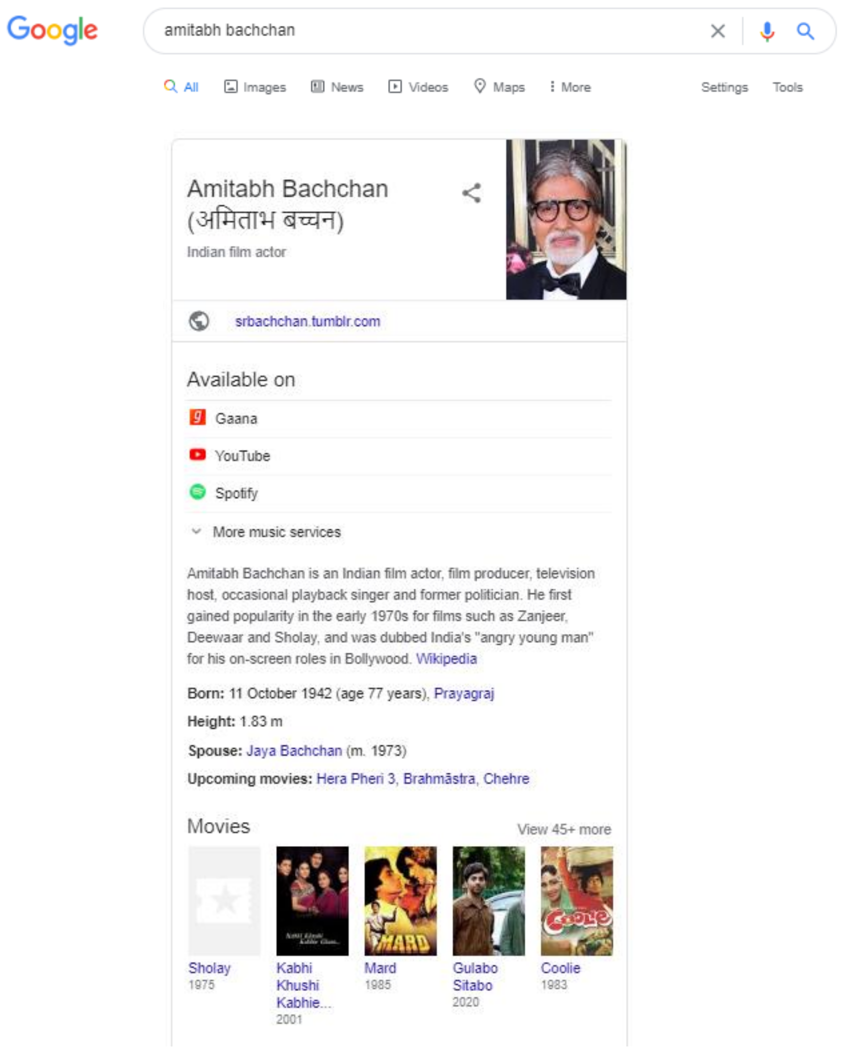
Lemur

Galago, Indri

UMass & CMU

Thanks to these… We can now focus on more complex problems.

Entity Search



amitabh bachchan

Amitabh Bachchan (अमिताभ बच्चन)

Indian film actor

srbachchan.tumblr.com

Available on

- Gaana
- YouTube
- Spotify

Amitabh Bachchan is an Indian film actor, film producer, television host, occasional playback singer and former politician. He first gained popularity in the early 1970s for films such as Zanjeer, Deewaar and Sholay, and was dubbed India's "angry young man" for his on-screen roles in Bollywood. Wikipedia

Born: 11 October 1942 (age 77 years), Prayagraj

Height: 1.83 m

Spouse: Jaya Bachchan (m. 1973)

Upcoming movies: Hera Pheri 3, Brahmāstra, Chehre

Movies

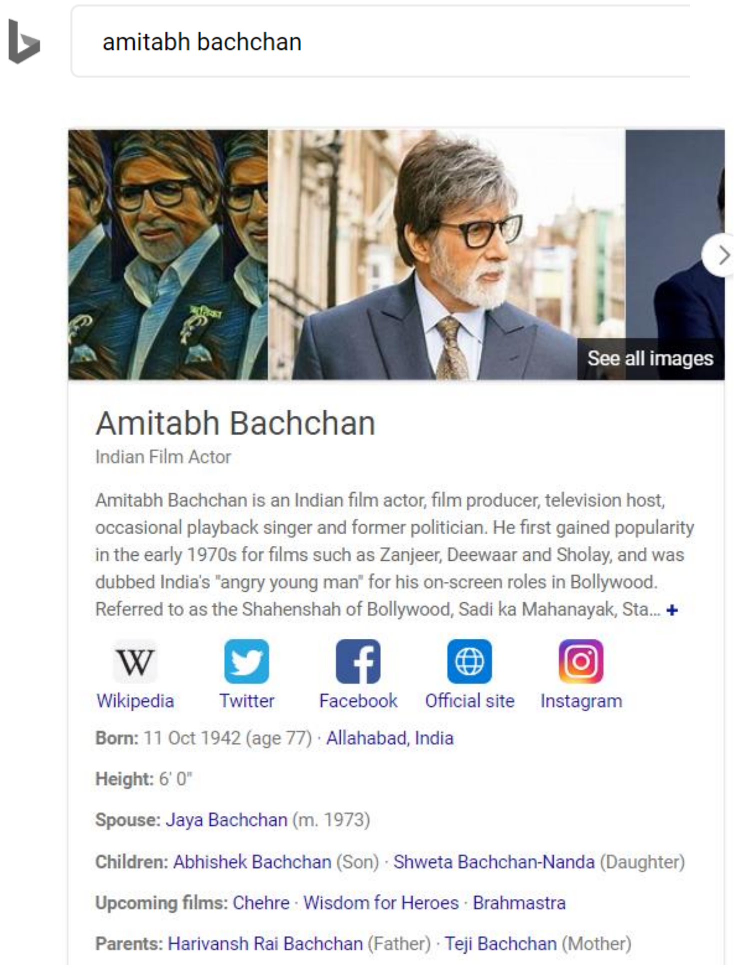
Sholay 1975

Kabhi Khushi Kabhie... 2001

Mard 1985

Gulabo Sitabo 2020

Coolie 1983



amitabh bachchan

Amitabh Bachchan

Indian Film Actor

Amitabh Bachchan is an Indian film actor, film producer, television host, occasional playback singer and former politician. He first gained popularity in the early 1970s for films such as Zanjeer, Deewaar and Sholay, and was dubbed India's "angry young man" for his on-screen roles in Bollywood. Referred to as the Shahenshah of Bollywood, Sadi ka Mahanayak, Sta... +

Born: 11 Oct 1942 (age 77) · Allahabad, India

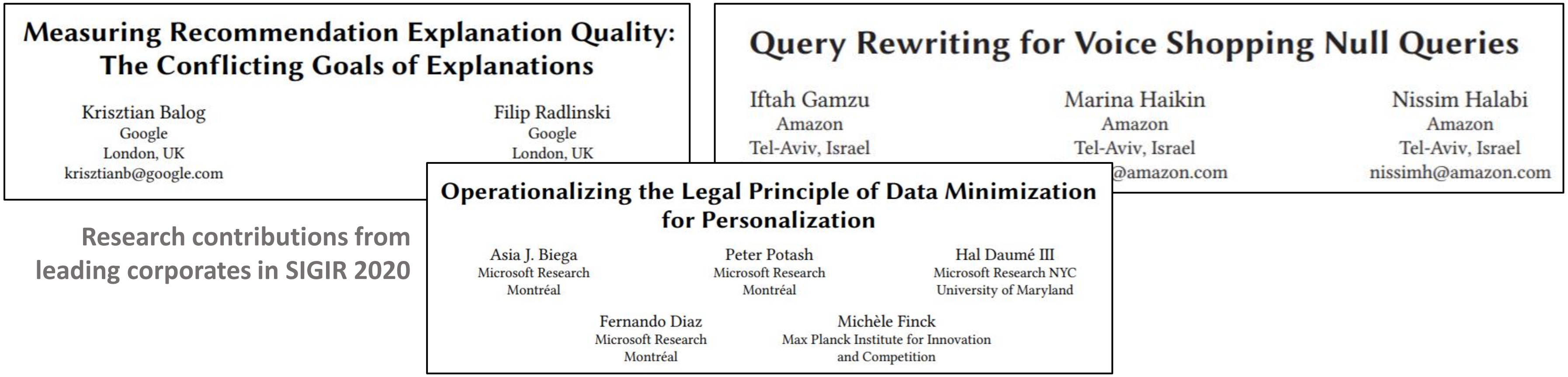
Height: 6' 0"

Spouse: Jaya Bachchan (m. 1973)

Children: Abhishek Bachchan (Son) · Shweta Bachchan-Nanda (Daughter)

Upcoming films: Chehre · Wisdom for Heroes · Brahmastra

Parents: Harivansh Rai Bachchan (Father) · Teji Bachchan (Mother)



Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations

Krisztian Balog
Google
London, UK
krisztianb@google.com

Filip Radlinski
Google
London, UK

Query Rewriting for Voice Shopping Null Queries

Iftah Gamzu
Amazon
Tel-Aviv, Israel

Marina Haikin
Amazon
Tel-Aviv, Israel
@amazon.com

Nissim Halabi
Amazon
Tel-Aviv, Israel
nissimh@amazon.com

Operationalizing the Legal Principle of Data Minimization for Personalization

Asia J. Biega
Microsoft Research
Montréal

Peter Potash
Microsoft Research
Montréal

Hal Daumé III
Microsoft Research NYC
University of Maryland

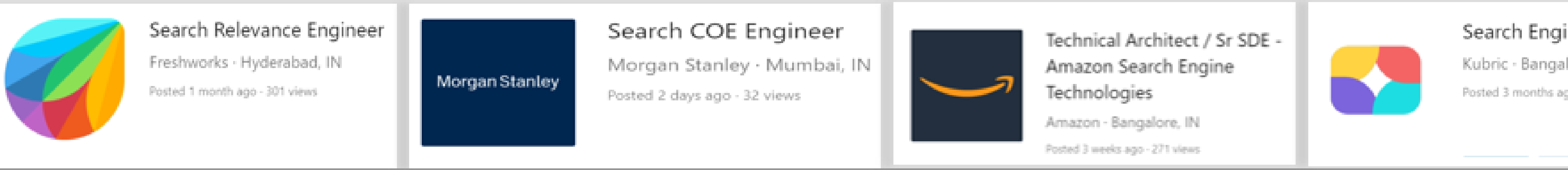
Fernando Diaz
Microsoft Research
Montréal

Michèle Finck
Max Planck Institute for Innovation
and Competition

Research contributions from leading corporates in SIGIR 2020

In spite of all these developments, “search”ing effectively has been an art.

(This is why) The academia, research labs, and the software industry needs you.
Let us strive to build better search experiences.

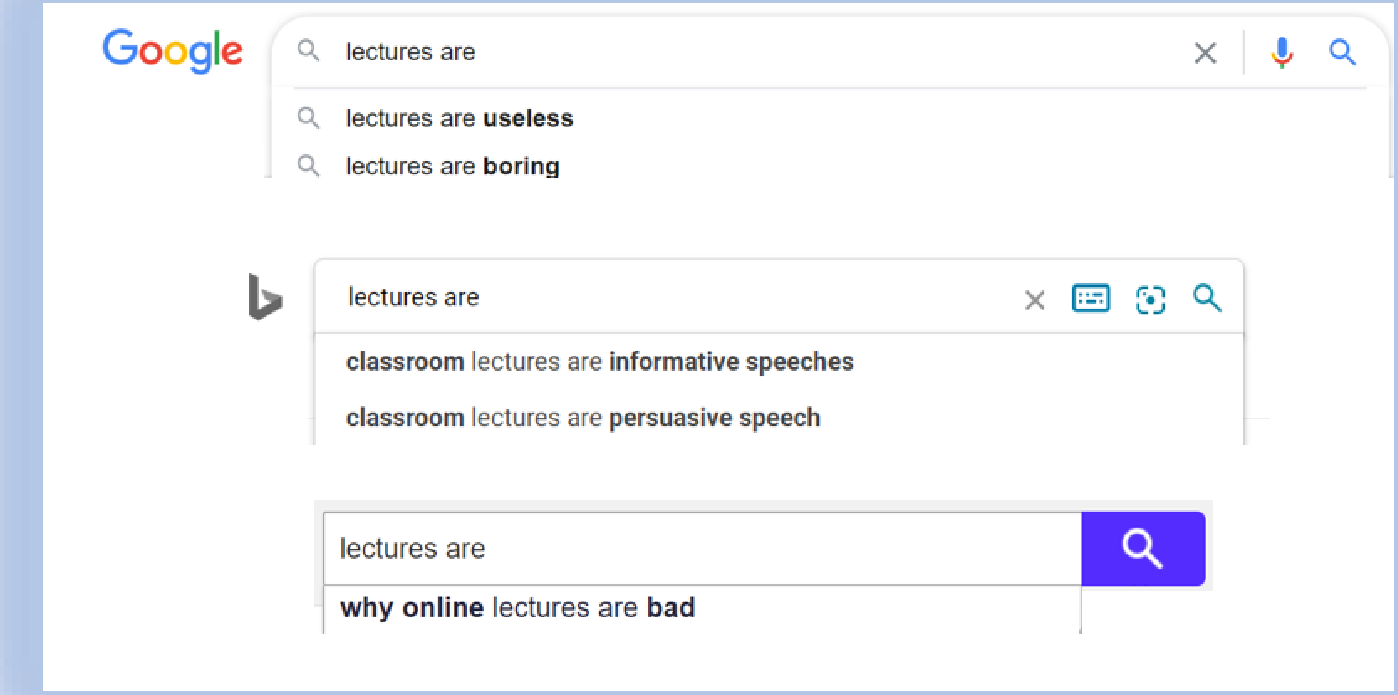


Search Relevance Engineer
Freshworks · Hyderabad, IN
Posted 1 month ago · 301 views

Search COE Engineer
Morgan Stanley · Mumbai, IN
Posted 2 days ago · 32 views

Technical Architect / Sr SDE - Amazon Search Engine Technologies
Amazon · Bangalore, IN
Posted 3 weeks ago · 271 views

Search Engi
Kubric · Bangal
Posted 3 months ag



lectures are

lectures are **useless**

lectures are **boring**

lectures are

classroom lectures are **informative speeches**

classroom lectures are **persuasive speech**

lectures are

why online lectures are **bad**

Thank You

venkateshv@cmi.ac.in