cmi | CHENNAI MATHEMATICAL INSTITUTE

Presented (remotely) at
Qualcomm
on 04th August 2021

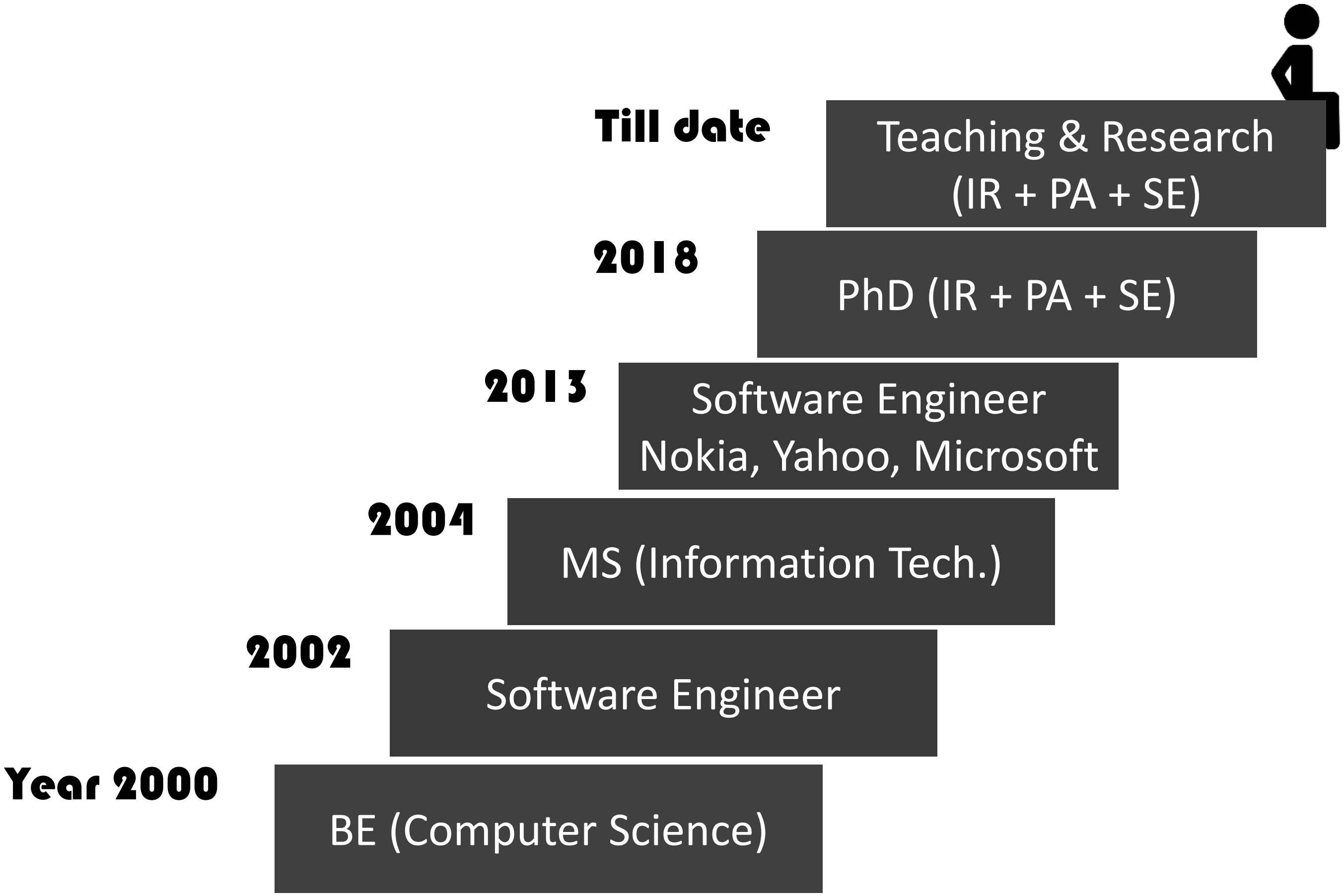
The Mathematics behind Search Engines

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

Chennai Mathematical Institute

Mathematics, the science of structure, order, and relation that has evolved from elemental practices of counting, measuring, and describing the shapes of objects.
Britannica, https://www.britannica.com/science/mathematics.

About Me



Till date

Teaching & Research
(IR + PA + SE)

2018

PhD (IR + PA + SE)

2013

Software Engineer
Nokia, Yahoo, Microsoft

2004

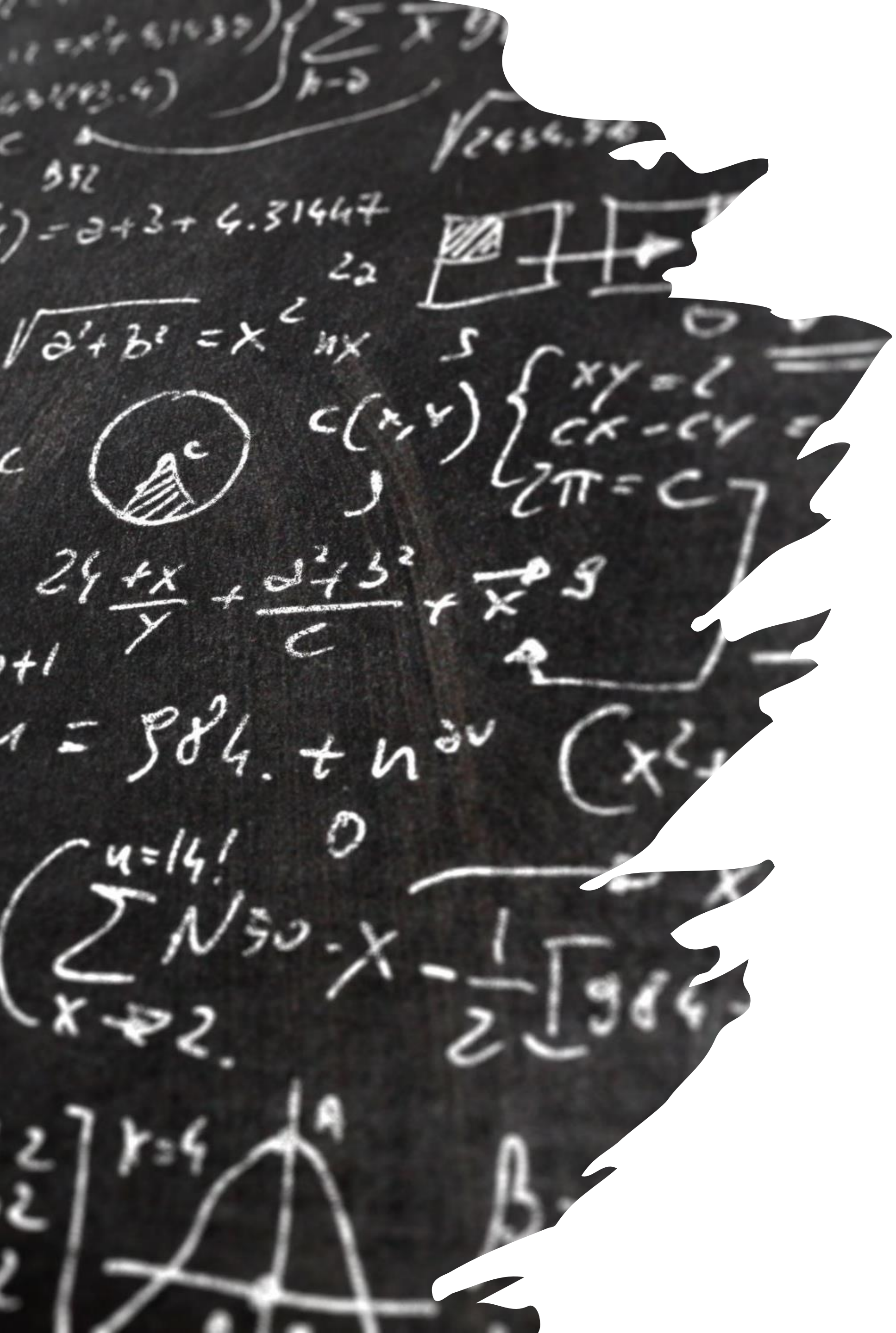
MS (Information Tech.)

2002

Software Engineer

Year 2000

BE (Computer Science)



Agenda

- Background on Search Engines
- Vector Space Models
- Probabilistic Models
- Current Trends and Research Problems

Scope

The **Mathematics** behind **Search Engines**

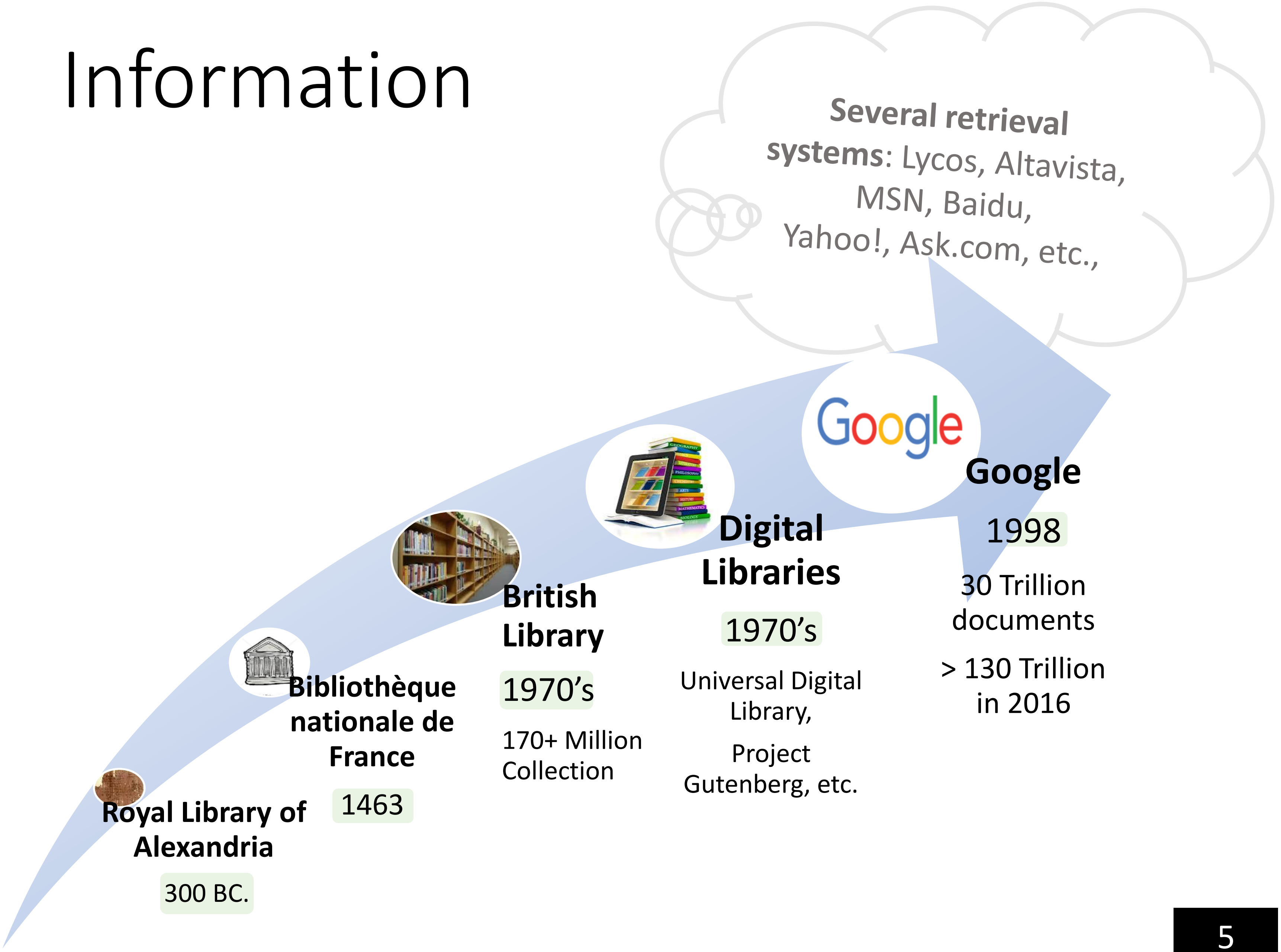
Will Discuss

- ✓ Concepts
- ✓ Illustrations
- ✓ Intuitions
- ✓ Purpose
- ✓ Properties

Will not Discuss

- 🛇 Details
- 🛇 Definitions
- 🛇 Formalism
- 🛇 Derivations
- 🛇 Proofs

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

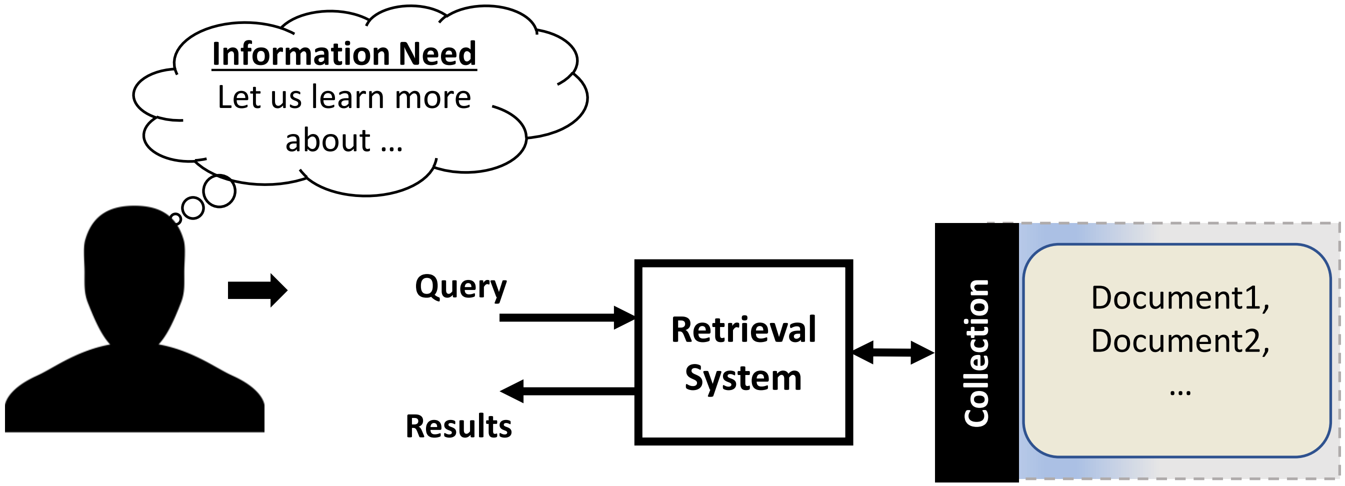
Solitary Confinement is Cruel

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

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!

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.

Vector Space Model

Let us build a search engine!

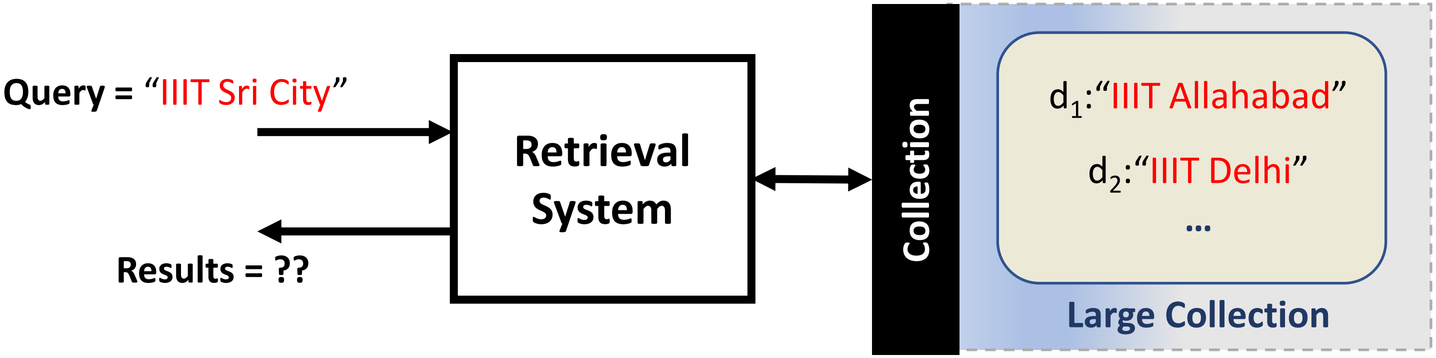
Stumpedia
zapmeta
Ask
dmoz open directory project
altavista
exalead
Clusty
inSuggest
BingBo
HOTBOT
Mahalo Beta
RedZee
accoona
spokeo
aftervote search socially
sproose User Improved Results
WEBBRAIN 2.0
Mooter The power of relevance
quintura see & find
INFOMI
Scholarly Internet Resource
Search the Deep Net
Trexy
all+
dogpile
Alster find the pearls
KAZAZZ! SEARCH ENGINE
trovando.it search different!
SortFix Start Dragging Stop Typing
GIGABLAST
CompletePlanet The Deep Web Directory
Dogpile
juxt

Image Source https://philbradley.typepad.com/phil_bradleys_weblog/2016/02/100-search-engines-logos-image-for-you-to-use.html

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.

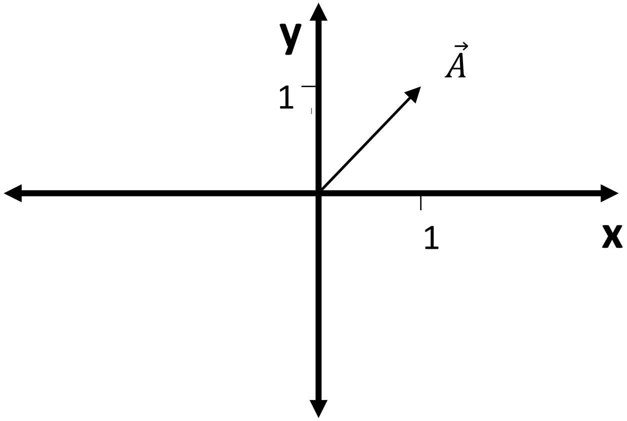
Three iterations!
Quiz: Can we do better?

A Better Approach

Revisiting
Linear Algebra

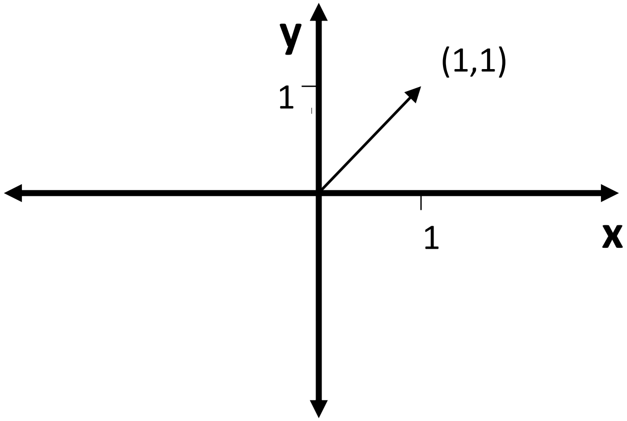
Vector

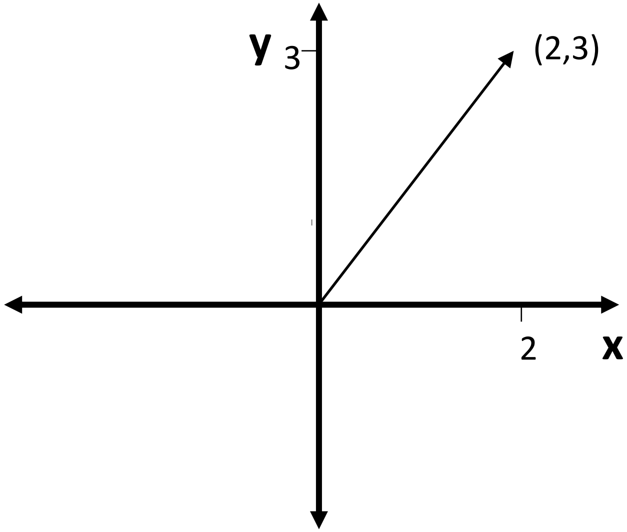
- Geometric entity which has magnitude and direction



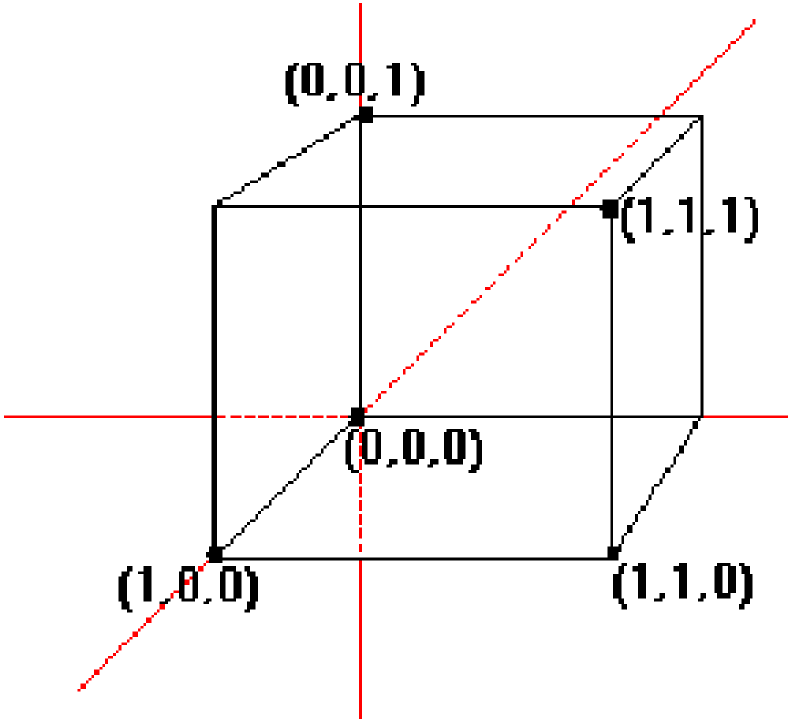
$\vec{A}$ is fixed at (0,0)

How is (2,3) Different?





What is (1,1,1) ?

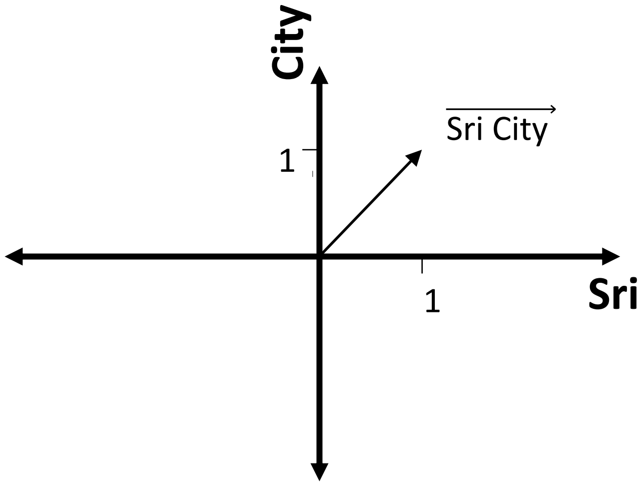


Remember!

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

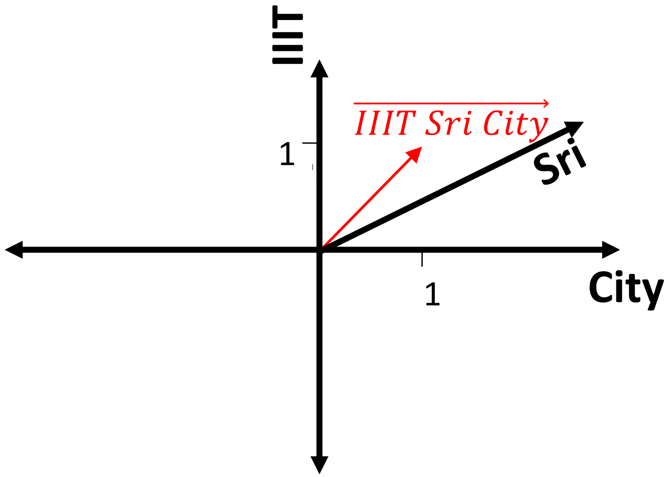
Sentences are Vectors

- "Sri City" as a vector



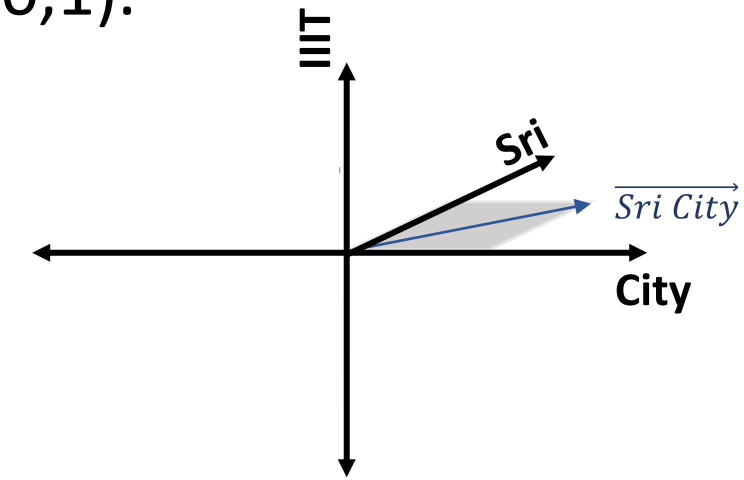
Sentences are Vectors

- "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

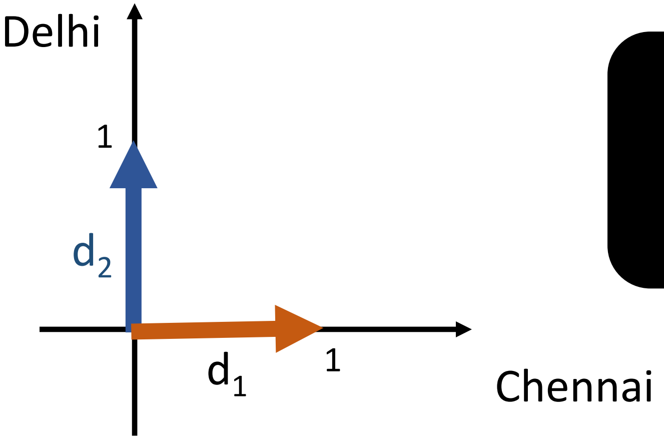
- Considering the following vectors:

	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 vector for Delhi?

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	$1/2$	$1/\sqrt{2}$	$\sqrt{3}/2$	1
cos	1	$\sqrt{3}/2$	$1/\sqrt{2}$	$1/2$	0
tan	0	$1/\sqrt{3}$	1	$\sqrt{3}$	Not defined

Back to Trigonometry: Dot Product

- If a and b are non-unit vectors, what is the cosine of angle between them (cos ϴ)?

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

(or)

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

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

An Assumption

More frequent appearance of a query term implies higher document relevance.

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}

Vadivelu, a popular tamil actor

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

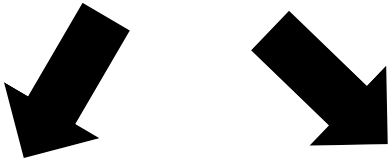
- “IIIT Sri City” → {IIIT, Sri, City}
- “IIIT Sri City, Sri City” → {IIIT, Sri, City}

(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

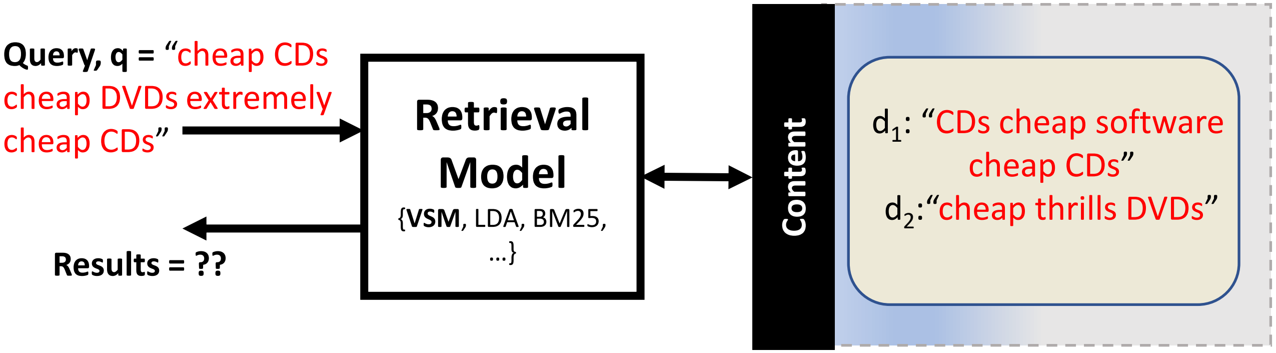
	IIIT	Sri	City
q	1	1	1

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 = ??

Retrieval Model

{**VSM**, LDA, BM25, …}

Content

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

$sim(q,d_1) = 0.86$

$sim(q,d_2) = 0.59$

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

Probabilistic Model

Can we do it another way?

Photo by Scott Graham on Unsplash

The Law – Robert M. Coates

From the book, “The World of Mathematics – Volume IV”.

Triborough Bridge, NY, USA.
(aka Robert F. Kennedy Bridge)

Late 1940s, NY: *No other bridge or main highway was affected, and though the two preceding nights had been equally balmy and moonlit, on both of these the bridge traffic had run close to **normal**.*

And then, one day…

PROGRESS
IF THIS MESSAGE IS PRESENT THIS IMAGE IS BEING USED WITHOUT PERMISSION
CHRIS MADDEN.
www.chrismadden.co.uk

*It just looked as if **everybody** in Manhattan who owned a car had decided to drive out to Long Island that evening.*

No Reason!

Sergeant: "I kept askin' them" he said, "Is there night football somewhere that we don't know about? Is it the races you're goin' to?"

But the funny thing was half the time they'd be askin' me. "What's the crowd for, Mac?" they would say. And I'd just look at them.

If normal things stop happening, if we lose regularities in life, our planet could become unlivable!

Time for Action

- At this juncture, it was inevitable that Congress should be called on for action.

- Senator said, “*You can control it*”. Re-education and reforms were decided upon. He said, (we need to lead people back to) “*the basic regularities, the homely **averageness** of the American way of life*”.

The Law of Large Numbers

Known as the **Fundamental Theorem of Probability**

The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

Probabilistic Retrieval

- Information Need: Taj Mahal
- Let a query q be “Taj”
- Let the results be:
  - d1: Taj
  - d2: Taj Mahal
  - d3: Taj Tea
- Two judges were asked to provide relevance judgments:

Document	Judge 1	Judge 2
Taj	R	N
Taj Mahal	R	R
Taj Tea	N	N

Probability of Relevance

- Documents **can** have probability of being relevant and of being non-relevant at the same time.
- Example:
  - Documents in our collection :

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	?	?
Taj Tea	?	?

R = 0 → Non-Relevant

R = 1 → Relevant

Probability of Relevance

- Documents **can** have probability of being relevant and of being non-relevant at the same time.
- Example:
  - Documents in our collection :

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

R = 0 → Non-Relevant
R = 1 → Relevant

Probability Ranking Principle

Rank documents by the probability of relevance, $P(R=1|q,d) \quad R \in \{0,1\}$

Probability Ranking Principle

Rank documents by the probability of relevance, $P(R=1|q,d)$ $R \in \{0,1\}$

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

R = 0 → Non-Relevant
R = 1 → Relevant

Search Result:

1. Taj Mahal
2. Taj
3. Taj Tea

Bayes Optimal Decision Rule

d is relevant if
$P(R=1 \mid d,q) > P(R=0 \mid d,q)$

Document	$P(R=0 \mid d,q)$	$P(R=1 \mid d,q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

Search Result:

1. Taj Mahal

Predicting Relevance

Document	$P(R=0\|d,q)$	$P(R=1\|d,q)$
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

This is user given relevance.

Can we estimate/predict relevance based on term occurrence ?

Predicting Relevance

- You may use labeled set from judges (or mined from clicklogs)
- You may assume query and document as set of words.

Query	Document	Relevance
q1 = (x1,x2,…)	d1 = (..xi, xj,…)	1
q1	d2	1
q1	d3	0
q2	d1	0
q2	d2	0
q2	d3	1

This is user given relevance.

Can we estimate/predict relevance ?

Binary Independence Model (BIM)

- Each document is a binary vector of terms.
- Occurrence of terms is mutually independent.

$$P(R=1|d,q)=\frac{P(d|R=1,q)P(R=1|q)}{P(d|q)}$$

Bayes Rule

Quiz

$$P(R = 1|d, q) = \frac{P(d|R = 1, q)P(R = 1|q)}{P(d|q)}$$

Bayes Rule

- P(d=Taj|R=1,q) = ?
- P(R=1|q) = ?
- P(d|q) = ?

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0

Quiz

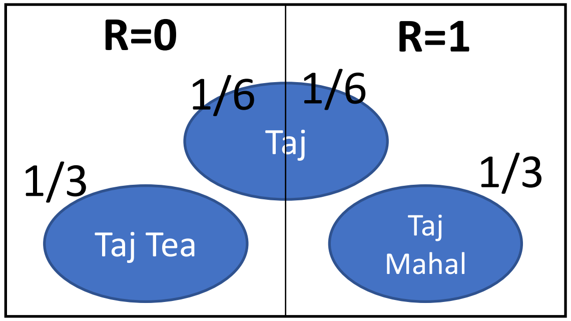
$$P(R = 1|d,q) = \frac{P(d|R = 1,q)P(R = 1|q)}{P(d|q)}$$

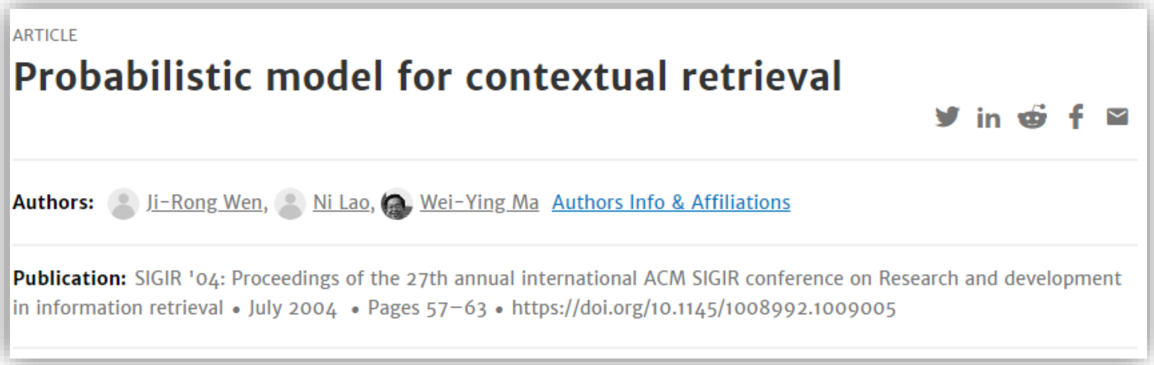
Bayes Rule

- P(d=Taj|R=1,q) = 1/3
- P(R=1|q) = 1/2
- P(d=Taj|q) = 1/3

P(R=1|d=Taj,q) = (1/3)(1/2)/(1/3) = ½

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0



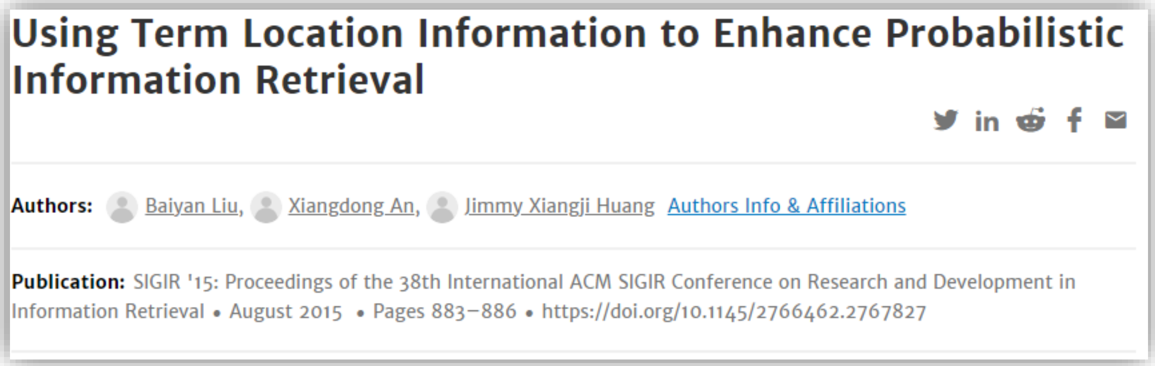


ARTICLE

Probabilistic model for contextual retrieval

Authors: Ji-Rong Wen, Ni Lao, Wei-Ying Ma Authors Info & Affiliations

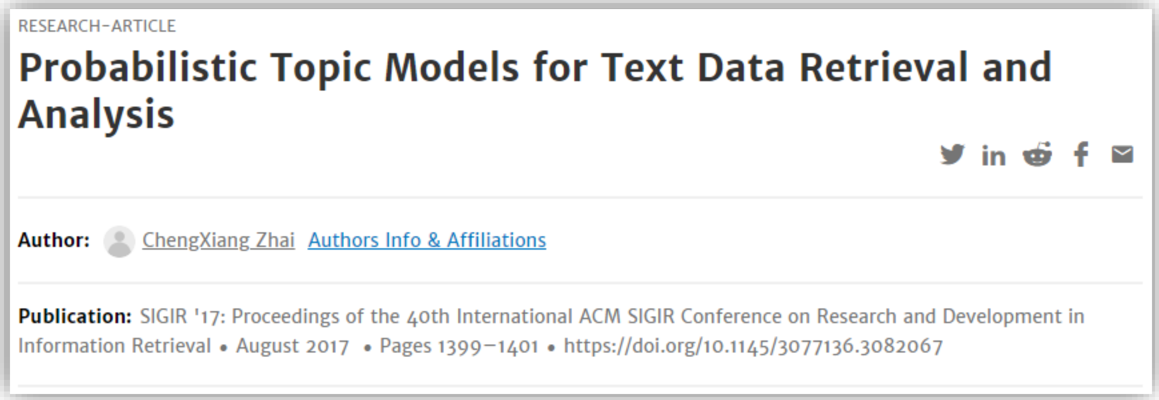
Publication: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval • July 2004 • Pages 57–63 • https://doi.org/10.1145/1008992.1009005



Using Term Location Information to Enhance Probabilistic Information Retrieval

Authors: Baiyan Liu, Xiangdong An, Jimmy Xiangji Huang Authors Info & Affiliations

Publication: SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval • August 2015 • Pages 883–886 • https://doi.org/10.1145/2766462.2767827



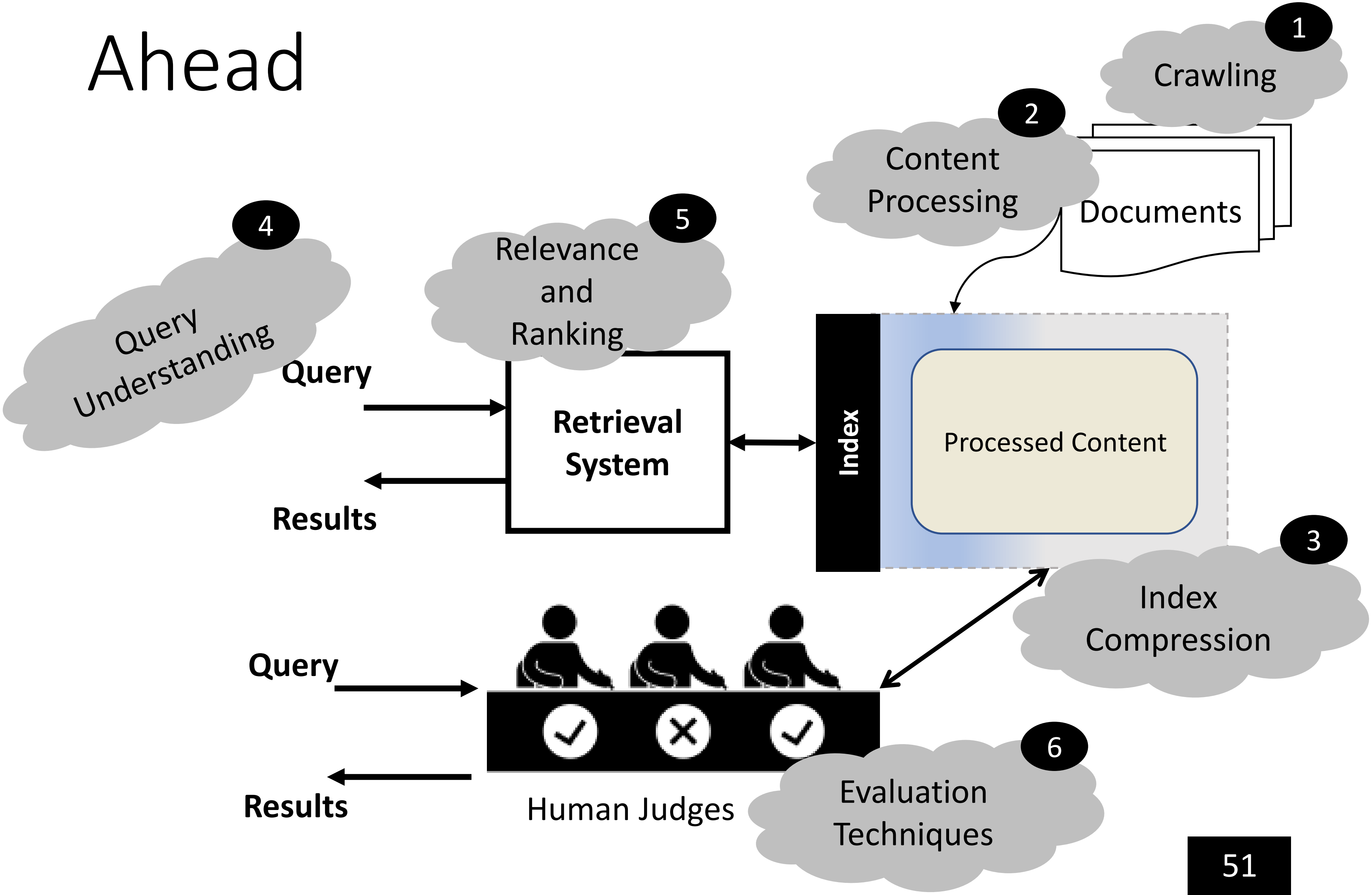
RESEARCH-ARTICLE

Probabilistic Topic Models for Text Data Retrieval and Analysis

Author: ChengXiang Zhai Authors Info & Affiliations

Publication: SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval • August 2017 • Pages 1399–1401 • https://doi.org/10.1145/3077136.3082067

Information Retrieval – Road Ahead



Technologies & Frameworks

Lucene™

Apache

Solr

Apache

nutch™

Apache

Terrier

Univ. of Glasgow

Lemur

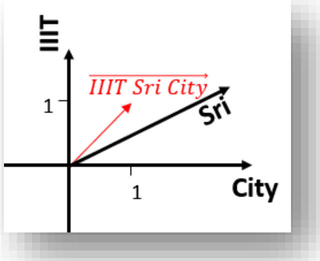
Galago, Indri

UMass & CMU

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

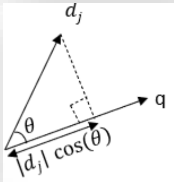
Summary: Vector Space Model

Model Documents as Vectors

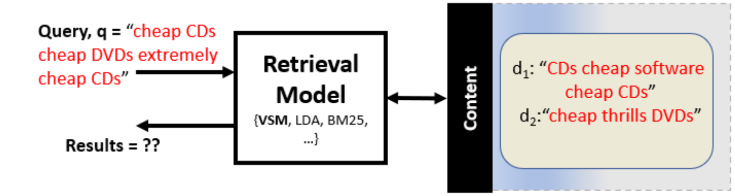


Compute Similarity

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



A Worked-Out Example



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

Retrieval Model {VSM, LDA, BM25, …}

Results = ??

Content

d_1: “CDs cheap software cheap CDs”

d_2:“cheap thrills DVDs”

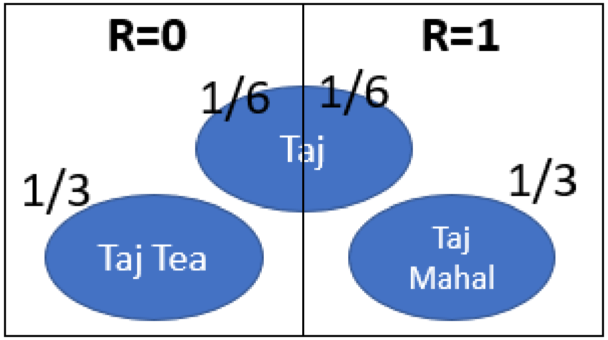
	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

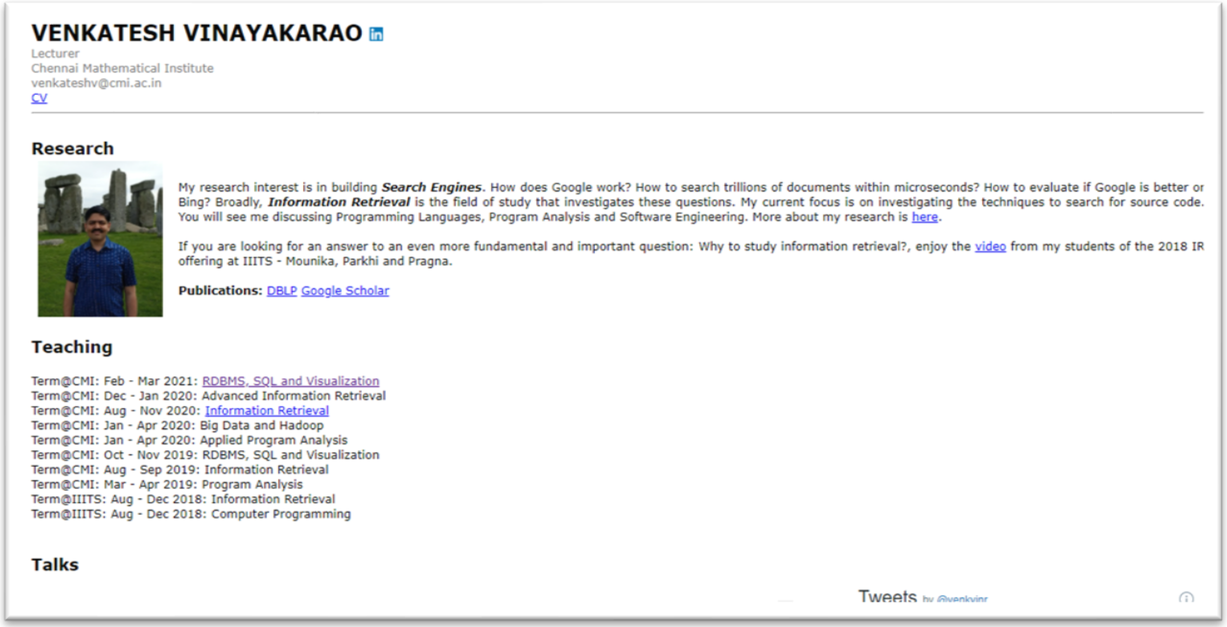
sim(q,d_1) = 0.86

sim(q,d_2) = 0.59

Summary: Probabilistic Model

Document	P(R=0|d,q)	P(R=1|d,q)
Taj	0.5	0.5
Taj Mahal	0	1
Taj Tea	1	0





VENKATESH VINAYAKARAO

Lecturer
Chennai Mathematical Institute
venkateshv@cmi.ac.in
CV

Research

My research interest is in building ***Search Engines***. How does Google work? How to search trillions of documents within microseconds? How to evaluate if Google is better or Bing? Broadly, ***Information Retrieval*** is the field of study that investigates these questions. My current focus is on investigating the techniques to search for source code. You will see me discussing Programming Languages, Program Analysis and Software Engineering. More about my research is here.

If you are looking for an answer to an even more fundamental and important question: Why to study information retrieval?, enjoy the video from my students of the 2018 IR offering at IIITS - Mounika, Parkhi and Pragna.

Publications: DBLP Google Scholar

Teaching

Term@CMI: Feb - Mar 2021: RDBMS, SQL and Visualization
Term@CMI: Dec - Jan 2020: Advanced Information Retrieval
Term@CMI: Aug - Nov 2020: Information Retrieval
Term@CMI: Jan - Apr 2020: Big Data and Hadoop
Term@CMI: Jan - Apr 2020: Applied Program Analysis
Term@CMI: Oct - Nov 2019: RDBMS, SQL and Visualization
Term@CMI: Aug - Sep 2019: Information Retrieval
Term@CMI: Mar - Apr 2019: Program Analysis
Term@IIITS: Aug - Dec 2018: Information Retrieval
Term@IIITS: Aug - Dec 2018: Computer Programming

Talks

Thank You

venkateshv@cmi.ac.in

This slide deck is available at http://vvtesh.co.in/.