

Chapter IR:I

I. Introduction

- Examples of Retrieval Tasks
- Terminology
- Delineation
- Historical Background
- Web Search

Examples of Retrieval Tasks

Task: *Learn everything there is to learn about information retrieval.*

Examples of Retrieval Tasks

Task: *Learn everything there is to learn about information retrieval.*

Search for texts that contain ‘information’ and ‘retrieval’.

Google search results for "information retrieval". The search bar shows the query. Below it, a navigation bar includes "All", "Books", "Images", "News", "Videos", "More", "Settings", and "Tools". The results section starts with a summary: "About 15,100,000 results (0.38 seconds)". The first result is a link to the Wikipedia page on Information retrieval.

Information retrieval - Wikipedia

https://en.wikipedia.org/wiki/Information_retrieval ▾

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.
Overview · History · Model types · Performance and ...

[PDF] Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/pdf/01book.pdf> ▾

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

Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/> ▾

The book aims to provide a modern approach to information retrieval from a computer science perspective. It is based on a course we have been teaching in ...
Introduction to Information ... · Information Retrieval and Web ... · Boolean retrieval

Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html> ▾

Introduction to Information Retrieval. By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze. Website: <http://informationretrieval.org/>. Cambridge ...

Information Retrieval and Web Search: CS 276

<https://cs276.stanford.edu/> ▾

Information retrieval is the process through which a computer system can respond to a user's query for text-based information on a specific topic. IR was one of ...

[PDF] Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Aug 1, 2006 - Information. Retrieval. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. Cambridge University Press. Cambridge, England ...

Information Retrieval Journal - Springer

<https://link.springer.com/journal/10791>

The journal provides an international forum for the publication of theory, algorithms, and experiments across the broad area of information retrieval. Topics of ...

Bing search results for "information retrieval". The search bar shows the query. Below it, a navigation bar includes "Web", "Images", "Videos", "Maps", "News", and "My saves". The results section starts with a summary: "67,200,000 RESULTS Any time ▾".

Information retrieval - Wikipedia

https://en.wikipedia.org/wiki/Information_retrieval ▾

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Introduction to Information Retrieval

nlp.stanford.edu/IR-book ▾

Introduction to Information Retrieval. This is the companion website for the following book. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze ...

Bib · Errata

Information Retrieval | Definition of Information ...

https://www.merriam-webster.com/dictionary/information_retrieval ▾

Define **information retrieval**: the techniques of storing and recovering and often disseminating recorded data especially through the use of a...

Information Retrieval | Article about Information ...

encyclopedia2.thefreedictionary.com/information+retrieval ▾

information retrieval[in-fər'mā-shən rē-tē-väl] (computer science) The technique and process of searching, recovering, and interpreting **information** ...

CS 276: Information Retrieval and Web Search

<https://web.stanford.edu/class/cs276> ▾

Information retrieval is the process through which a computer system can respond to a user's query for text-based information on a specific topic.

[PDF] Information Retrieval - Stanford University

<https://web.stanford.edu/class/cs276/handouts/lecture2-dictionary.pdf>

3 Introduction to Information Retrieval Introduction to Information Retrieval Terms The things indexed in an IR system Introduction to Information Retrieval

Information retrieval - Britannica.com

<https://www.britannica.com/topic/information-retrieval> ▾

information retrieval: Recovery of information, especially in a database stored in a computer. Two main approaches are matching words in the query against the ...

Remarks:

- ❑ Here, the search engine is treated like a database of documents, looking up those that contain specific words a user expects to be contained in a relevant document. Unlike a database, the search engine ranks the retrieved documents with respect to its estimation which of them the user will most likely find useful.
- ❑ Compare the different total numbers of search results. The discrepancy may be due to different amounts of documents being indexed, indicating that Bing indexes many more documents than Google. However, search engines have long stopped sharing the numbers of documents they index for competitive reasons, and these numbers are only estimations based on a partial search. Usually, these estimations very much overestimate the actual number of documents that can be delivered.
- ❑ How can the number of search results be reduced without loosing many useful documents?

Using the phrase search operator (i.e., enclosing “information retrieval” in quotes) ensures that the words are contained in order in all documents retrieved, severely reducing the number of results. Conceivably, all documents about information retrieval will contain this particular phrase at least once, whereas most documents talking about something else involving “information” and “retrieval” will not. Interestingly, only about 1.12% of all search engine users know about this or other search operators [[White and Morris 2007](#)].

- ❑ Results 2, 3, 4, and 6 of the Google result are from the same website. Similarly, results 3, 4, and 7 of the Bing result are dictionary sites. What’s wrong with that?
- ❑ The preview text (so-called snippet) of Bing’s result 6 is flawed.

Examples of Retrieval Tasks

Task: *Plan a trip from San Francisco to Paris, France.*

Examples of Retrieval Tasks

Task: *Plan a trip from San Francisco to Paris, France.*

Search for flights from San Francisco to Paris, and for a hotel.

 flight sf paris 

Web Images Videos Maps News | My saves

68,900,000 RESULTS Any time ▾

Flights from Paris - Fly from Paris at less
Ad · www.CheapOair.com/Paris
Fly from Paris at less! Grab the Travel Deals & Save Maximum.
Cheap Price Worldwide · Millions of Cheap Flights · 24/7 Customer Care

Flights to Flights Paris - eDreams.com
Ad · www.eDreams.com/Flights-Paris
Book Today Cheap Flights to Flight Paris. Book Now!
Cheap Flights from £19 · More than 750 Airlines · Mobile Friendly · Service 7/7
Save on Flight+Hotel Deal **Flights to London 75% Off**
Best Offers to/from Paris **Cheap City Breaks Europe**

Showing results for **flights paris**.
No results found for flight sf paris.



Round trip ▾ Economy ▾ 1 Passenger ▾

CMF - Chambery - Chambery CDG - Paris - Paris Charles De Gaulle

Sun, Aug 27 Tue, Sep 19

We couldn't find any flights. Try changing your dates or airports.

 paris hilton 

WEB IMAGES VIDEO NEWS TRANSLATE DISK MAIL ALL

Hilton Paris Opera 4* – Онлайн бронирование номеров
hotelhunter.com ad
Эксклюзивные тарифы напрямую со скидкой до 70%
Онлайн бронирование · Бесплатный Wi-Fi · Фото и отзывы · Специальные скидки

Official website
parishilton.com ▾

Paris Hilton - Wikipedia
en.wikipedia.org · Paris Hilton ▾
Paris Whitney Hilton (born February 17, 1981) is an American businesswoman, socialite, television and media personality, model, actress, singer, and DJ.

paris hilton – browse images
yandex.com/images > paris hilton



Complain

Paris Hilton - YouTube
youtube.com > user/ParisHilton ▾
Paris Hilton - Official YouTube Stars Are Blind - Duration: 4 minutes, 33 seconds ... Paris Hilton - Good Time (Explicit) ft. Lil Wayne - Duration: 3 minutes, 47 seconds.

Paris Hilton - IMDb
imdb.com > name/nm0385296/ ▾
Socialite Paris Whitney Hilton was born on February 17, 1981 in New York City, into the Hilton family, and has three younger siblings, Nicky Hilton Rothschild...

Paris Hilton (@ParisHilton) | Твиттер
twitter.com > parishilton ▾



Paris Hilton начан(а) читать. that time we sent @irin to mexico city with @ParisHilton and all my @marieclaire editor dreams came truehttp...
Friends 15 million subscribers, 6 thousand friends

Paris Hilton - Google+
plus.google.com > ParisHilton ▾
Mike Kerwin: I do really like cats *Paris Hilton, I was just sitting her thinking my way through petting your cat and using that as an excuse to . . .

Remarks:

- ❑ Users use informal language to describe their goals. Here, Bing does not understand ‘sf’ as an abbreviation for San Francisco. ‘SFO’, the location identifier of San Francisco Airport, would have worked. Google at least offers a spell correction for ‘sf’ to ‘sfo’.
- ❑ Users use ambiguous queries whose interpretation depends on the user’s context. For example, when searching for the Hilton hotel in Paris, Yandex returns results about the celebrity Paris Hilton. Only the ad at the top of the search results hints at the intended interpretation. Searching for “hilton paris” instead yields better results.
- ❑ Search engines introduce shortcuts and additional information as so-called oneboxes into search results. The flight search box in Bing’s results is an example.
- ❑ Search engines follow a “universal search” paradigm, offering different kinds of results. The images in Yandex’ results are an example.

Examples of Retrieval Tasks

Task: *What were the news of the day?*

Examples of Retrieval Tasks

Task: *What were the news of the day?*

Hit the news feeds.

Google News

Search

Headlines Local For You U.S. ▾

Top Stories

Bannon's departure is unlikely to calm the turmoil in Trump's White House
Washington Post · 9h ago

RELATED COVERAGE
Steve Bannon, Unrepentant
Highly Cited · The American Prospect · Aug 16, 2017

IS conflict: Iraq launches ground offensive in Tal Afar
BBC News · 4h ago

RELATED COVERAGE
Iraqi forces to commence Tal Afar operation 'in the next few days'
Local Source · Rudaw · 14h ago

Protesters Flood Streets, and Trump Offers a Measure of Praise
New York Times · 6h ago

RELATED COVERAGE
Protesters face a tricky balance on free speech
Local Source · The Boston Globe · 11h ago

Researchers find wreckage of famed Navy cruiser Indianapolis, sunk in 1945
Los Angeles Times · 3h ago

RELATED COVERAGE
Wreckage From USS Indianapolis Located In Philippine Sea | Paul Allen
Most Referenced · Paul Allen · 3h ago

#news

Home Notifications Email Twitter #news Search

Top Latest People Photos Videos News Broadcasts

KRTpro News @KRTpro_News · 1m
#BREXIT
UK to release tranche of Brexit position papers reut.rs/2wl5myy
#KRTpro #News



UK to release tranche of Brexit position papers
Britain will issue a cluster of new papers this week to outline its strategy positions in divorce talks with the European Union, ranging from regulation reuters.com

Breaking News India @Golstream · 3m
Boston March Against Hate Speech Avoids Charlottesville Chaos - NDTV
ift.tt/2wlLnv #India #News



Examples of Retrieval Tasks

Task: *What were the news of the day?*

Hit the news feeds.



LIBERAL

SHOWING POSTS ABOUT:
"HEALTH CARE"

CONSERVATIVE

\$ 1.3 Billion Health Care Fraud Just Erupts! Jef...
Attorney General Jeff Sessions said at a news conference. "We will cont...
USAPOLITICSTODAY.COM

1.3K 72 558

Allen West about a week ago
Sing it with me: And another one gone, and another one gone.
Another one bites the dust.

Implosion: ANOTHER major healthcare compa...
Epic FAIL
ALLENWEST.COM | BY ALLEN WEST

2.3K 227 1K

Breitbart about a week ago
Sinking like a stone...

Remarks:

- ❑ One cannot search for things one does not know. Instead of searching for news, they are explored. Information systems for this purpose include news aggregators and social networks, but also simply the front page of a news outlet. The former recommend news based on user preferences.
- ❑ As of 2017, Google News does not show preview snippets for the news, anymore, but only the headlines. While claiming better usability, this change coincides with increasing pressure from news publishers as well as ancillary copyright laws being passed in various jurisdictions.
- ❑ Facebook's role in providing Americans with political news has never been stronger—or more controversial. Scholars worry that the social network can create “echo chambers,” where users see posts only from like-minded friends and media sources. Facebook encourages users to “keep an open mind” by seeking out posts that don't appear in their feeds.

To demonstrate how reality may differ for different Facebook users, The Wall Street Journal created two feeds, one “blue” and the other “red.” If a source appears in the red feed, a majority of the articles shared from the source were classified as “very conservatively aligned” in a large 2015 Facebook study. For the blue feed, a majority of each source’s articles aligned “very liberal.” These aren’t intended to resemble actual individual news feeds. Instead, they are rare side-by-side looks at real conversations from different perspectives.

[\[Wall Street Journal\]](#)

Examples of Retrieval Tasks

Task: *Answer “Can Kangaroos jump higher than the Empire State Building?”*

Examples of Retrieval Tasks

Task: Answer “Can Kangaroos jump higher than the Empire State Building?”

Search for facts

Google search results for "how high can kangaroos jump":

Search term: how high can kangaroos jump

Results:

- Kangaroos Can Jump 30 Feet High. Mar 10, 2014 
- Kangaroos Can Jump 30 Feet High - YouTube <https://www.youtube.com/watch?v=1L0YNsVZYNQ>
- How can kangaroos jump so high? How high can they jump - Quora <https://www.quora.com/How-can-kangaroos-jump-so-high-How-high-can-they-jump>
About 7 metres. ... so high? How high can they jump. UpdateCancel. Answer Wiki. 1 Answer. Bruce Josephs, Runs private tours to see kangaroos in the wild.
- Secret of kangaroo's bounce - The Telegraph www.telegraph.co.uk/News/science/science-news/
Mar 10, 2011 - Kangaroos can cover 25 feet in a single leap and are known to jump obstacles How birds learned to fly: high-speed footage of chicks falling ...
- How and why do kangaroos hop? | Discover Wildlife www.discoverwildlife.com/animals/mammals/how-and-why-do-kangaroos-hop/
Sep 22, 2015 - The jumping motion drives their gut up and down, which inflates and deflates ...
Kangaroos usually hop at about 25kph, though they can reach ...
- How far can a Kangaroo Jump? - AnimalWised <https://www.animalwised.com/fun-facts-facts-about-the-animal-kingdom/>
Jump to **Do you want to know more about Kangaroos?** - The 10 Highest-Jumping Animals in the World ... Differences between Kangaroos and ...
- Red Kangaroo | National Geographic www.nationalgeographic.com/animals/mammals/r/red-kangaroo/
... the animal that can cover 25 feet in a single leap and jump as high as 6 feet. ... Red kangaroos hop along on their powerful hind legs and do so at great speed.
- How high can a kangaroo jump? | Reference.com <https://www.reference.com/Pets-Animals/Mammals/Marsupials>
Kangaroos can jump as high as 6 feet in a single jump. Kangaroo legs cannot move independently of one another, which is why they are restricted to a ...

WolframAlpha search results for "height of empire state building in feet":

Search term: height of empire state building in feet

Results:

Input interpretation:
convert Empire State Building total height to feet
[Open code](#)

Result:
1250 feet [Show details](#)

Additional conversions:
0.2367 miles
417 yards
0.2057 nmi (nautical miles)
381 meters
0.381 km (kilometers)

Comparisons as height:
≈ 0.69 × height of the CN Tower (≈ 553 m)
≈ 0.7 × architectural height of One World Trade Center (1776 ft)
≈ 1.2 × Eiffel Tower height (≈ 324 m)

Corresponding quantity:
Distance to horizon (ignoring topography and other obstructions):
70 km (kilometers)
69 716 meters
43 miles

Sources [Download page](#) POWERED BY THE WOLFRAM LANGUAGE

Examples of Retrieval Tasks

Task: Answer “Can Kangaroos jump higher than the Empire State Building?”

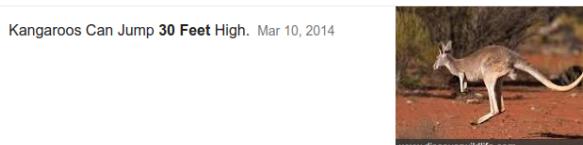
Search for facts, or ask the question outright.

Google search results for "how high can kangaroos jump".

Search bar: how high can kangaroos jump

Results:

- Kangaroos Can Jump 30 Feet High. Mar 10, 2014
- Kangaroos Can Jump 30 Feet High - YouTube
<https://www.youtube.com/watch?v=1L0YNsVzYNQ>
- About 6,340,000 results (0,32 seconds)



Kangaroos Can Jump 30 Feet High - YouTube
<https://www.youtube.com/watch?v=1L0YNsVzYNQ>

How can kangaroos jump so high? How high can they jump - Quora
<https://www.quora.com/How-can-kangaroos-jump-so-high-How-high-can-they-jump>
About 7 metres. ... so high? How high can they jump. UpdateCancel. Answer Wiki. 1 Answer. Bruce Josephs, Runs private tours to see kangaroos in the wild.

Secret of kangaroo's bounce - The Telegraph
www.telegraph.co.uk/News/science/science-news/
Mar 10, 2011 - Kangaroos can cover 25 feet in a single leap and are known to jump obstacles How birds learned to fly: high-speed footage of chicks falling ...

How and why do kangaroos hop? | Discover Wildlife
www.discoverwildlife.com/animals/mammals/how-and-why-do-kangaroos-hop
Sep 22, 2015 - The jumping motion drives their gut up and down, which inflates and deflates ...
Kangaroos usually hop at about 25kph, though they can reach ...

How far can a Kangaroo Jump? - AnimalWised
[https://www.animalwised.com/fun-facts-facts-about-the-animal-kingdom-jump-to-do-you-want-to-know-more-about-kangaroos?-the-10-highest-jumping-animals-in-the-world-differences-between-kangaroos-and...](https://www.animalwised.com/fun-facts-facts-about-the-animal-kingdom-jump-to-do-you-want-to-know-more-about-kangaroos-the-10-highest-jumping-animals-in-the-world-differences-between-kangaroos-and...)

Red Kangaroo | National Geographic
www.nationalgeographic.com/animals/mammals/r/red-kangaroo/
... the animal that can cover 25 feet in a single leap and jump as high as 6 feet. ... Red kangaroos hop along on their powerful hind legs and do so at great speed.

How high can a kangaroo jump? | Reference.com
<https://www.reference.com/pets-animals/marsupials>
Kangaroos can jump as high as 6 feet in a single jump. Kangaroo legs cannot move independently of one another, which is why they are restricted to a ...

Google search results for "can kangaroos jump higher than the empire state building".

Search bar: can kangaroos jump higher than the empire state building?

Results:

- can a kangaroo jump higher than empire state building?
- Germany Safe Search: Strict Any Time

Can a kangaroo jump higher than empire state building?
A healthy kangaroo can jump higher than any rabbit. Kangaroos also jump faster and further than rabbits. This is due, in part, to their larger size.
answers.com/Q/Can_a_kangaroo_jump_higher_than_empire...

Can a kangaroo jump higher than the Empire State Building ...
The Empire State Building can't jump. ... Can a kangaroo jump higher than the Empire State Building? ... but it can jump higher than the World Trade Center" ...
https://www.reddit.com/r/Jokes/comments/55ndr2/can_a_kangaroo_ju...

The Big Apple: "Can a kangaroo jump higher than the Empire ...
"What animal can jump higher than the Empire State building?" was cited in 1939. "What animal has eyes and can't see, legs and can't walk, ...
barrypopik.com/index.php/new_york_city/entry/can_a_kangaroo...

Can a kangaroo jump higher than the Empire State Building ...
The Empire State Building can't jump. Jump to content. my subreddits. edit subscriptions. popular-all ... Can a kangaroo jump higher than the Empire State Building?
https://www.reddit.com/r/cleanjokes/comments/6lzo2n/can_a_kangar...

Can a Kangaroo Jump Higher than the Empire State Building ...
Can a Kangaroo Jump Higher than the Empire State Building? ... Kangaroos Can Jump 30 Feet High - Duration: ... guy jumping off Empire State Building ...
youtube.com/watch?v=7M-a_s_pGW4

Can A Kangaroo Jump Higher Than The Empire State Building ...
Can A Kangaroo Jump Higher Than The Empire State Building.
increaseverticaljump.blogspot.com/download/can-a-kangaroo-jump-higher-than-...

Question: Can a kangaroo jump higher than the Empire State ...
Best Answer: Yes, but only on Tuesdays. ... Not at all, My Dear. Empire State Building can't jump. But this is very old matter. ... Yes. Buildings ...
<https://answers.yahoo.com/question/index?qid=20101223195232AAgjjqQ>

Remarks:

- ❑ Users search for facts; search engines employ various strategies to meet these requests, such as using knowledge bases like Wikidata, or by extracting factual information from web pages (e.g., Google's Knowledge Graph).
- ❑ Google's highlighted top search result appears to answer the question, but the given height is false; it mixes height with distance. Google does not necessarily validate the truth of a statement, but only returns the ones best matching the query. Other snippets mention distances inconsistent with the top one. The two bottom-most snippets claim that kangaroos can jump only about 6 feet high. In any case, a YouTube video may not be a reliable source.
- ❑ WolframAlpha allows for asking questions requiring computation, resorting to a knowledge base to fill in required facts.
- ❑ Asking the original question directly shows that it is a well-known one; some snippets of DuckDuckGo reveal that it is a riddle by giving away the answer. The odd formulation of the riddle might have revealed this already.

Examples of Retrieval Tasks

Task: *Build a fence.*

Examples of Retrieval Tasks

Task: *Build a fence.*

Search for tutorials.

Search results 1-10 for *how to build a fence*

Total results: 2037230 (retrieved in 1025.3ms)

[how to build a fence](#)
fence-posts.org/tag/how-to-build-a-fence ▾
Hi, I'm Alex Barnett and I'd like to welcome you to my web site, **How To Build A Fence**. All right, I know, I've heard all of the jokes before. Why on earth would I want to build a web site about fence building? Is there anything more boring? Well, the truth is that building a good, sturdy, long

[How To Build a Wooden Fence](#)
fence-posts.org/how-to-build-a-w... ▾
More results from fence-posts.org

[How to build a fence like a pro](#)
www.how-to-build-a-fence-like-a-pro.com/ ▾

you will need to build your fence. We will answer frequently asked questions and provide a photo gallery of pictures for your enjoyment. Take your time in planning out your fence. Have fun Make it an event with friends and family members. © Copyright 2008, Trigon Corporation, All Rights Reserved

[How To Build A Fence](#)
siteexpansion.com/tag/how-to-build-a-fence ▾

Gardening information structured to support backyard garden themes. Provide seasonal gardening information and largest garden store on the Web. Finally a single source for the backyard gardener. Tags: a **fence** , annual flowers , bedtime stories , bedtime stories ringtone , bulbs Do it yourself

[How to Build a Fence with Goat Panels](#)
feedlotpanels.com/how-to-build-a-fence-with-goat-panels ▾

article, you will learn how to build a fence using goat panels. The first step of building a goat fence includes finding out how large you want your goat enclosure or goat pen to be. Each goat panel is about 16 feet long and 48 inches tall. Consequently, if you want a goat fence with an area of 16

[How to Build a Fence, Garden Fencing](#)
www.beestonfencingcompany.co.uk/howtobuildafence.htm ▾

You may be wondering how to erect fencing without digging holes or mixing concrete. Here are some quick and easy solutions for building a fence with timber panels and fence posts. To erect a fence on grass or soil, drive metal spikes into the ground with a sledge hammer and insert the upright fence

[How To Build A Jackleg Fence](#)
www.mademan.com/mm/how-build-jackleg-fence.html ▾

If you are looking for a relatively inexpensive fencing option, you may want to explore how to build a jackleg fence. A jackleg fence is not an option for keeping small critters or children in or out; however, it is the perfect choice for livestock or just as a property divider. A jackleg fence has

Search results for *how to build a fence*

About 1,320,000 results

[Filters ▾](#)

How to Build a Fence | Mitre 10 Easy As
Mitre 10 New Zealand
5 years ago • 406,654 views
Knowing how to build a fence is one of the more useful DIY skills to learn. And if you learn how to do it properly, your DIY fence ...

Building a Picket Fence | Setting the Posts
The Restoration Couple
1 year ago • 181,300 views
Part 1 - Here is a look at our picket fence installation around the vegetable garden. Enjoy! CONTACT US ...

BUILDING A FENCE CONSIDERATIONS
The Home Depot
8 years ago • 1,236,065 views
Find our full guide to learn how to build a fence:
http://thd.co/2ozT1W3j Shop all the products you'll need at The

93 Building 70 Feet of Wooden Fence
Memphis Applegate
6 months ago • 92,372 views
I'm using the cool South Carolina "winter" to finish the wooden fence around my backyard. In this edition I install about 70 feet

DIY: How To Build A Fence (BYOT #12)
BYOT
1 year ago • 109,604 views
This DIY project is all about how to build a fence. With the right tools you can turn an ugly fence into a beautiful new cedar fence ...

Building a Board on Board Cedar Fence - Part 2
April Wilkerson
1 year ago • 657,037 views
If you missed Part 1, here is a link: https://youtu.be/v3BlyCf1YPM
Looking for Part 3? Here ya go: ...

Remarks:

- ❑ Users search for guidelines on complex tasks. Besides textual information, this includes instructive multimedia contents, e.g., from YouTube.
- ❑ ChatNoir is the only publicly available research search engine that operates at scale.

Examples of Retrieval Tasks

Task: *Write an essay on video surveillance.*

Examples of Retrieval Tasks

Task: Write an essay on video surveillance.

Search for other's opinions on video surveillance.

Google search results for "video surveillance". The search bar shows "video surveillance". Below it, a row of filters: All, Images, Shopping, News, Videos, More, Settings, Tools. The "All" filter is selected. Below the filters, it says "About 14,800,000 results (0,59 seconds)".

Surveillance Data Storage - Free White Paper - quantum.com

<https://landing.quantum.com/Download-Now> ▾

Free White Paper: Benefits of Tiered Storage for Surveillance Data.

Highlights: Surveillance Beyond Security, Intelligent Storage,...

IHS Industry Report

Video Surveillance

Performance Matters

Surveillance Infographic

Surveillance - Wikipedia

<https://en.wikipedia.org/wiki/Surveillance> ▾

Surveillance is the monitoring of behavior, activities, or other changing information for the

Surveillance cameras are video cameras used for the purpose of observing an area. They are often connected to a recording device or IP network. ...

PRISM (surveillance program) · Foreign Intelligence · Terrorist Surveillance Program

People also ask

What is the video surveillance system? ▾

What is the best CCTV system for home? ▾

What is a security camera? ▾

What is closed circuit television? ▾

Feedback

Closed-circuit television - Wikipedia

https://en.wikipedia.org/wiki/Closed-circuit_television ▾

Closed-circuit television (CCTV), also known as **video surveillance**, is the use of video cameras to transmit a signal to a specific place, on a limited set of ...

Video Surveillance Systems Integrator | VideoSurveillance.com

<https://wwwvideosurveillance.com/> ▾

VideoSurveillance.com is a trusted virtual security integrator for enterprises and organizations alike, specializing in customized surveillance systems.

Video surveillance from ABUS for private houses and apartments

<https://www.abus.com> · ABUS - August Bremicker Söhne KG · Home Security ▾

Video surveillance from ABUS enables fast localization of threats and protects persons and material assets. If worst comes to worst, detailed recordings help ...

Video surveillance and security cameras for your home | Ivideon

<https://www.iveideon.com/video-surveillance-for-home/> ▾

An affordable video surveillance system that's easy to install and use to protect your family and home. Keep an eye on what's important to you. See for ...

args search results for "video surveillance". The search bar shows "video surveillance". Below it is a navigation bar with "Page 1 of 40 arguments (retrieved in 21.1ms)", "Pro vs. Con View", and "Overall Ranking View".

[con] Often surveillance camera images are not clear and police...

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Often **surveillance** camera images are not clear and police cannot identify the criminal. ... ▾

[con] Surveillance cameras cannot physically protect the public....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Surveillance cameras cannot physically protect the public, only film what is happening. ... ▾

[pro] Surveillance cameras are not closely monitored and are only....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Surveillance cameras are not closely monitored and are only usually viewed if a crime has taken place. ... ▾

[con] Crime camera evidence is very rarely used in court cases....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Crime camera evidence is very rarely used in court cases. ... ▾

[pro] There is not much privacy in public places....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

There is not much privacy in public places. ... ▾

[con] Filming without consent is actually illegal....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Filming without consent is actually illegal. ... ▾

[pro] It is no different to police monitoring a dangerous area....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

It is no different to police monitoring a dangerous area. ... ▾

[pro] Crime cameras offer conclusive, unbiased evidence in court....

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Crime cameras offer conclusive, unbiased evidence in court. ... ▾

[pro] Crime cameras help catch criminals and get them off the...

http://www.debatepedia.org/en/index.php/Debate:_Video_surveillance

Crime cameras help catch criminals and get them off the street ... ▾

Remarks:

- ❑ Users search for other peoples' opinions and reasoning on controversial topics or when deciding what product to buy.
- ❑ Google's results include related questions from question answering platforms, concisely presented in a onebox. Yet, there's no hint at the controversiality of the topic. The results referring to Wikipedia are the only way for a user to learn the topic's background, all other results are related to shopping.
- ❑ Search engines specialized to argument retrieval, such as Args, retrieve argumentative information alongside stance (pro or con).

Examples of Retrieval Tasks

Task: *Given an example image, find more like it.*

Examples of Retrieval Tasks

Task: *Given an example image, find more like it.*

The image is an example of the information sought after.



Google image.jpg new york city

All Images Maps Shopping More Settings Tools

About 25,270,000,000 results (0.90 seconds)



Image size:
756 × 1014

Find other sizes of this image:
[All sizes](#) - [Medium](#) - [Large](#)

Best guess for this image: [new york city](#)

The Official Guide to New York City | nycgo.com

<https://www.nycgo.com/> ▾

Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.

New York City - Wikipedia

https://en.wikipedia.org/wiki/New_York_City ▾

The City of New York, often called New York City or simply New York, is the most populous city in the United States. With an estimated 2017 population of 8,622,698 distributed over a land area of about 302.6 square miles (784 km²), New York City is also the most densely populated major city in the United States. Located ...

Visually similar images



[Report images](#)

Pages that include matching images

Fifth Avenue - Wikipedia

https://en.wikipedia.org/wiki/Fifth_Avenue ▾

250 × 335 · Fifth Avenue is a major thoroughfare in the borough of Manhattan in New York City, United States. It stretches from West 143rd Street in Harlem to Washington Square North at Washington Square Park in Greenwich Village. It is considered one of the most expensive and elegant streets in the world.

Examples of Retrieval Tasks

Task: *Given an example **text**, find more like it.*

Examples of Retrieval Tasks

Task: *Given an example text, find more like it.*

The text is an example of the information sought after.



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikibooks
Wikiquote
Wikiversity
Wikivoyage

Languages

Article Talk
Read View source View history
Search Wikipedia

Mars

From Wikipedia, the free encyclopedia

This article is about the planet. For the deity, see [Mars \(mythology\)](#). For other uses, see [Mars \(disambiguation\)](#).

Mars is the fourth planet from the Sun and the second-smallest planet in the Solar System after Mercury. In English, Mars carries a name of the Roman god of war, and is often referred to as the "Red Planet"^{[14][15]} because the reddish iron oxide prevalent on its surface gives it a reddish appearance that is distinctive among the astronomical bodies visible to the naked eye.^[16] Mars is a terrestrial planet with a thin atmosphere, having surface features reminiscent both of the impact craters of the Moon and the valleys, deserts, and polar ice caps of Earth.

The rotational period and seasonal cycles of Mars are likewise similar to those of Earth, as is the tilt that produces the seasons. Mars is the site of Olympus Mons, the largest volcano and second-highest known mountain in the Solar System, and of Valles Marineris, one of the largest canyons in the Solar System. The smooth Borealis basin in the northern hemisphere covers 40% of the planet and may be a giant impact feature.^{[17][18]} Mars has two moons, Phobos and Deimos, which are small and irregularly shaped. These may be captured asteroids;^{[19][20]} similar to 5261 Eureka, a Mars trojan.

There are ongoing investigations assessing the past habitability potential of Mars, as well as the possibility of extant life. Future astrobiology missions are planned, including the Mars 2020 and ExoMars rovers.^{[21][22][23][24]} Liquid water cannot exist on the surface of Mars due to low atmospheric pressure, which is less than 1% of Earth's,^[25] except at the lowest elevations for short periods.^{[26][27]} The two polar ice caps appear to be made largely of water.^{[28][29]} The volume of water ice in the south polar ice cap, if melted, would be sufficient to cover



Mars in natural color in 2007^[2]

Designations

Pronunciation UK English: /maʊz/
US English: /maʊz/ (ⓘ listen)

Adjectives Martian

Orbital characteristics^[2]

Aphelion 249 200 000 km
(154 800 000 mi; 1.666 AU)

Perihelion 206 700 000 km
(128 400 000 mi; 1.382 AU)

Semi-major axis 227 939 200 km
(141 634 900 mi;
1.523 679 AU)

Eccentricity 0.0934

Orbital period 686.971 d
(1.880 82 yr; 668.5991 sols)

Synodic period 779.96 d
(2.1354 yr)

Average orbital 24 007 km/s



New analysis
Text alignment

Found 6 reused passages.

Detailed comparison of your submitted documents: 6 reused passages, length 600 words, 327 shared words

https://en.wikipedia.org/wiki/Mars

Although better remembered for mapping the Moon, Johann Heinrich Mädler and Wilhelm Beer were the first "areographers". They began by establishing that most of Mars's surface features were permanent and by more precisely determining the planet's rotation period. In 1840, Mädler combined ten years of observations and drew the first map of Mars. Rather than giving names to the various markings, Beer and Mädler simply designated them with letters; Meridian Bay (Sinus Meridiani) was thus feature "a"

https://en.wikipedia.org/wiki/Mars

Olympus Mons (Mount Olympus).^[114] The surface of Mars as seen from Earth is divided into two kinds of areas, with differing albedo. The paler plains covered with dust and sand rich in reddish iron oxides were once thought of as Martian "continents" and given names like Arabia Terra (land of Arabia) or

Although better remembered for mapping the Moon starting in 1830, Johann Heinrich Mädler and Wilhelm Beer were the first "areographers". They started off by establishing once and for all that most of the surface features were permanent, and pinned down Mars' rotation period. In 1840, Mädler combined ten years of observations and drew the first map of Mars ever made. Rather than giving names to the various markings, Beer and Mädler simply designated them with letters; Meridian Bay (Sinus Meridiani) was thus feature "a"

https://en.wikipedia.org/wiki/Mars

The surface of Mars as seen from Earth is consequently divided into two kinds of areas, with differing albedo. The paler plains covered with dust and sand rich in reddish iron oxides were once thought of as Martian 'continents' and given names like Arabia Terra (land of Arabia) or

Remarks:

- ❑ Users sometimes cannot express their need as a textual query, but provide an object that best exemplifies the information sought.
- ❑ Some search engines are tailored to searching for specific multimedia examples, such as images or audio.
- ❑ Using a text as an example, there can be two goals: finding other texts talking about the same subject, or finding other texts which share reused text passages with the text in question. For example, Google Scholar offers the search facet “Related Articles” to search for articles related to one designated from a prior search. Picapica is a search engine for text reuse.

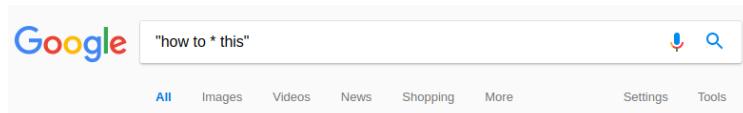
Examples of Retrieval Tasks

Task: *Find out what people commonly write in the phrase how to ? this.*

Examples of Retrieval Tasks

Task: *Find out what people commonly write in the phrase how to ? this.*

Use wildcard search operators to find matching phrases.



A screenshot of a Google search results page. The search query is "how to * this". The results include:

- How to couple two different loop elements - MATLAB Answers ...**
<https://www.mathworks.com/.../393724-how-to-couple-two-different-loop-elements> ▾
16 hours ago - ... on i and j since I am summing over them. My a and b are just certain shifts. Mind that i+a and j+b cannot exceed number N=20, so whenever i+a and j+b goes beyond N=20, I just need to start numbering from the very beginning (cyclic ordering). I would appreciate if anyone could help me how to do this.
- python - Would a time series model be appropriate for this problem ...**
<https://stats.stackexchange.com/.../would-a-time-series-model-be-appropriate-for-this...> ▾
10 hours ago - Time Series forecasting with explanatory variables · 0 · Selecting the best time lagged moving average for time series analysis · 0 · cross validating time series model · 0 · A time-series analysis problem · 0 · How to formulate a classification problem with time series element · 0 · Not sure how to approach this ...
- variance - How to calculate covariance matrix with batch data ...**
<https://stats.stackexchange.com/.../how-to-calculate-covariance-matrix-with-batch-data> ▾
11 hours ago - I am reading the following paper. In the paper, they define the CORAL loss for using the feature covariance matrices (see 3.1). And, in the end of this chapter, they said: "We use batch covariances and ...". I am not sure how to calculate this batch-covariance matrix. Deep CORAL: Correlation Alignment for ...
- How to Win -- and Why You Fail - Entrepreneur**
<https://www.entrepreneur.com/article/311467> ▾
2 days ago - This one shift in my life has made a huge difference, so I wanted to record an episode where I break down how to do this. Most of the powerful advice I've been given in my life has been really simple. And this is no different. But it makes the biggest difference day in and day out in my results. I'm giving you ...
- 15 Easy Solutions To Your Data Frame Problems In R (article ...**
<https://www.datcamp.com/community/.../15-easy-solutions-data-frame-problems-r> ▾
Since you choose to keep all values from all corresponding variables and to add columns to the result, you set the all argument to TRUE : It could be that the fields for rows that don't occur in both data structures result in NA-values. You can easily solve this by removing them. How to do this will be discussed below.
- javascript - Unable to add google scripts library to a project ...**
<https://stackoverflow.com/.../unable-to-add-google-scripts-library-to-a-project-owned-...> ▾
13 hours ago - Google Apps Script: how to make this autocomplete library work in UI? 0 · Deploy and use Google Sheets add-on with Google Apps Script · 0 · How to publish a chart to the web with google apps script · 1 · Sending Emails from different accounts with Google App Script · 4 · Google Script Library: Project Key ...



A screenshot of a Netspeak search results page. The search query is "how to ? this". The results are presented as a table:

Search Query	Count	Percentage	Is Popular
how to ? this			
how to use this	1.1 million	34.8%	+
how to do this	0.6 million	18.2%	+
how to cite this	238,000	7.5%	+
how to replace this	107,000	3.4%	+
how to fix this	92,000	2.9%	+
how to make this	87,000	2.8%	+
how to read this	74,000	2.4%	+
how to buy this	68,000	2.2%	+
how to get this	64,000	2.1%	+
how to solve this	52,000	1.7%	+
how to handle this	37,000	1.2%	+
how to purchase this	34,000	1.1%	+
how to play this	27,000	0.9%	+
how to book this	27,000	0.9%	+
how to accomplish this	25,000	0.8%	+
how to achieve this	24,000	0.8%	+
how to implement this	21,000	0.7%	+
how to improve this	21,000	0.7%	+
how to take this	21,000	0.7%	+
how to resolve this	21,000	0.7%	+
how to put this	20,000	0.7%	+
how to set this	19,000	0.6%	+
how to order this	18,000	0.6%	+
how to access this	18,000	0.6%	+
how to avoid this	15,000	0.5%	+
how to apply this	14,000	0.5%	+
how to add this	13,000	0.4%	+
how to obtain this	12,000	0.4%	+
how to say this	12,000	0.4%	+
how to about this	12,000	0.4%	+
how to approach this	12,000	0.4%	+
how to manage this	12,000	0.4%	+
how to join this	11,000	0.4%	+
how to complete this	11,000	0.4%	+
how to install this	11,000	0.4%	+
how to change this	10,000	0.3%	+

Remarks:

- ❑ Users “misuse” search engines and all other kinds of tools to achieve goals beside their originally intended purpose due to lack of specialized tools, or lack of knowledge of their existence.
- ❑ Many web search engines support some wildcard search operators, but they cannot be used to solve this kind of retrieval task. The search engine still interprets the query in terms of its contents, ranking documents according to their relevance to the query. Moreover, only few search results fit on a single page, while many more alternatives may be in practical use.
- ❑ Netspeak indexes only short phrases alongside their web frequencies, offering a wildcard search interface tailored to searching for usage commonness.

Terminology

Information science distinguishes the concepts data, information, and knowledge.

Definition 1 (Data)

A sequence of symbols recorded on a storage medium.

Definition 2 (Information)

A useful portion of data.

useful: the data in question serves as a means to an end for a person

Definition 3 (Knowledge)

Product of reflection upon information combined with experience to the extent of recognizing acquaintance, reaching belief, or justified understanding.

Examples: One might claim “I know $2+2=4$,” but also “I know God exists.”

Remarks:

- ❑ Data is typically organized into documents, each containing purposefully chosen partitions of data. Examples: a book, a video tape. A digital document corresponds to specific sequences of bits on a digital storage medium. Example: files on a hard drive formatted with a file system.
- ❑ Definitions of the three concepts, but especially those of information and knowledge, differ wildly among scholars of information science. [\[Zins 2007\]](#) collects 44 different attempts.
- ❑ Epistemology studies the nature of knowledge, justification, and the rationality of belief. In mathematics, it is known that $2+2=4$, but there is also knowing how to add two numbers, and knowing a person (e.g., oneself), place (e.g., one's hometown), thing (e.g., cars), or activity (e.g., addition). Some philosophers think there is an important distinction between “knowing that” (know a concept), “knowing how” (understand an operation), and “acquaintance-knowledge” (know by relation), with epistemology being primarily concerned with the first of these. Plato’s definition of knowledge as *justified true belief* was widely accepted until the 1960s. At this time, a paper written by the American philosopher Edmund Gettier entitled “Is Justified True Belief Knowledge?” called into question the theory of knowledge that had been dominant among philosophers for thousands of years. [\[Wikipedia\]](#)
- ❑ The fundamental question of rationality: “Why do you believe what you believe?”, or alternatively, “What do you think you know, and how do you think you know it?” [\[LessWrong\]](#)

Terminology

Definition 4 (Information System)

An organized system for collecting, creating, storing, processing, and distributing information, typically including hardware, software, users, and the data itself.

Definition 5 (Information Need)

A user's desire to locate and obtain information to satisfy a conscious or unconscious goal.

Definition 6 (Relevance)

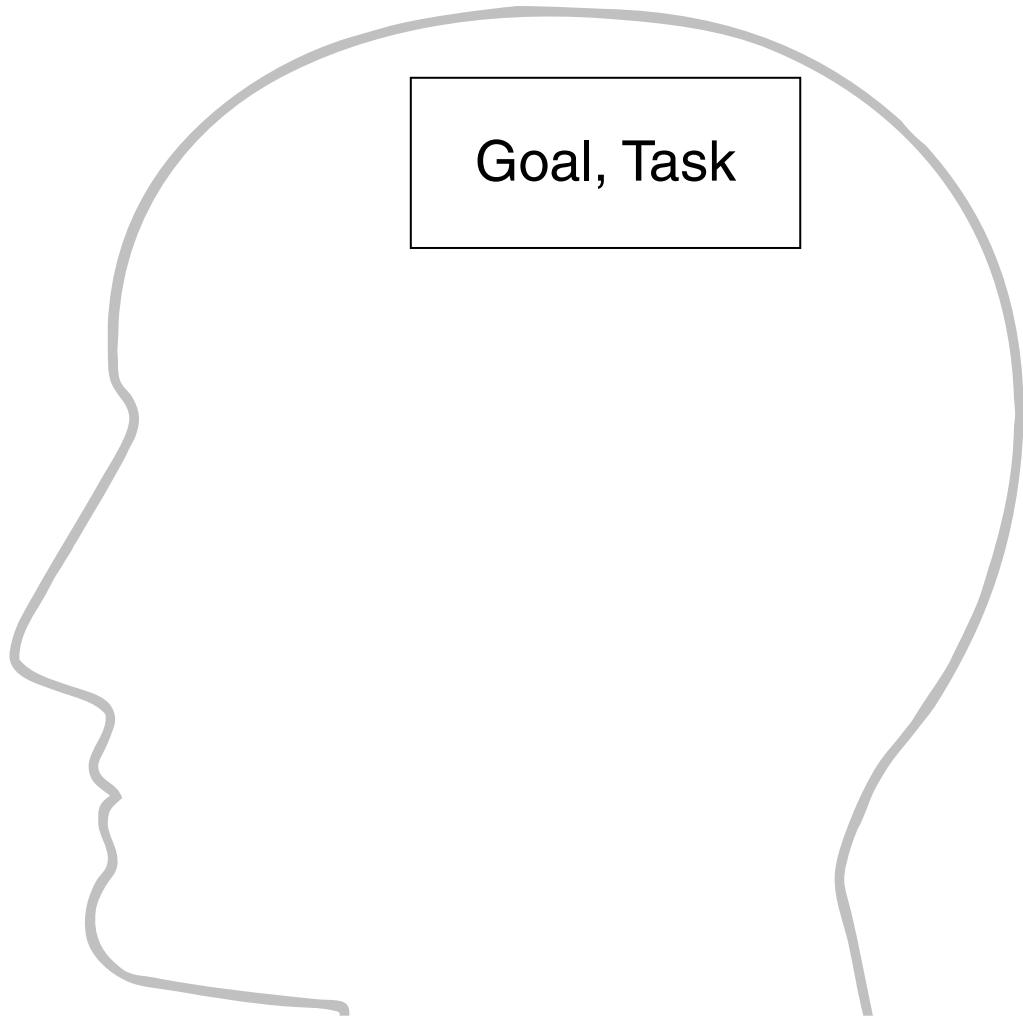
The degree to which a portion of data meets the information need of a user.

A portion of data is said to be relevant to a user's information need, if it informs the user. The closer it brings the user to satisfy their goal, the more relevant it is.

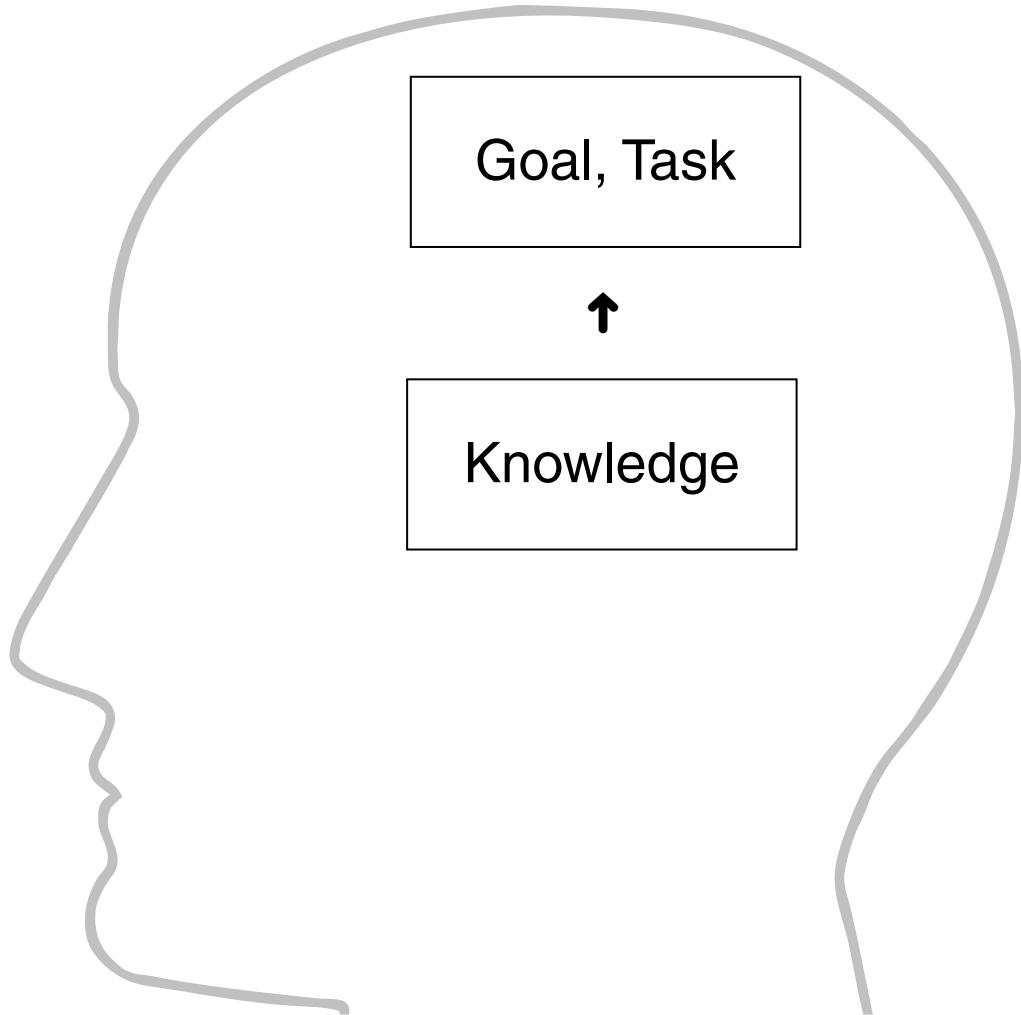
Remarks:

- ❑ Information need refers to a cognitive need that is perceived when a gap of knowledge is encountered in the pursuit of a goal.
- ❑ The study of information needs has been generalized to the study of information behavior, i.e. “the totality of human behavior in relation to sources and channels of information, including both active and passive information-seeking, and information use.” [\[Wilson 2000\]](#)

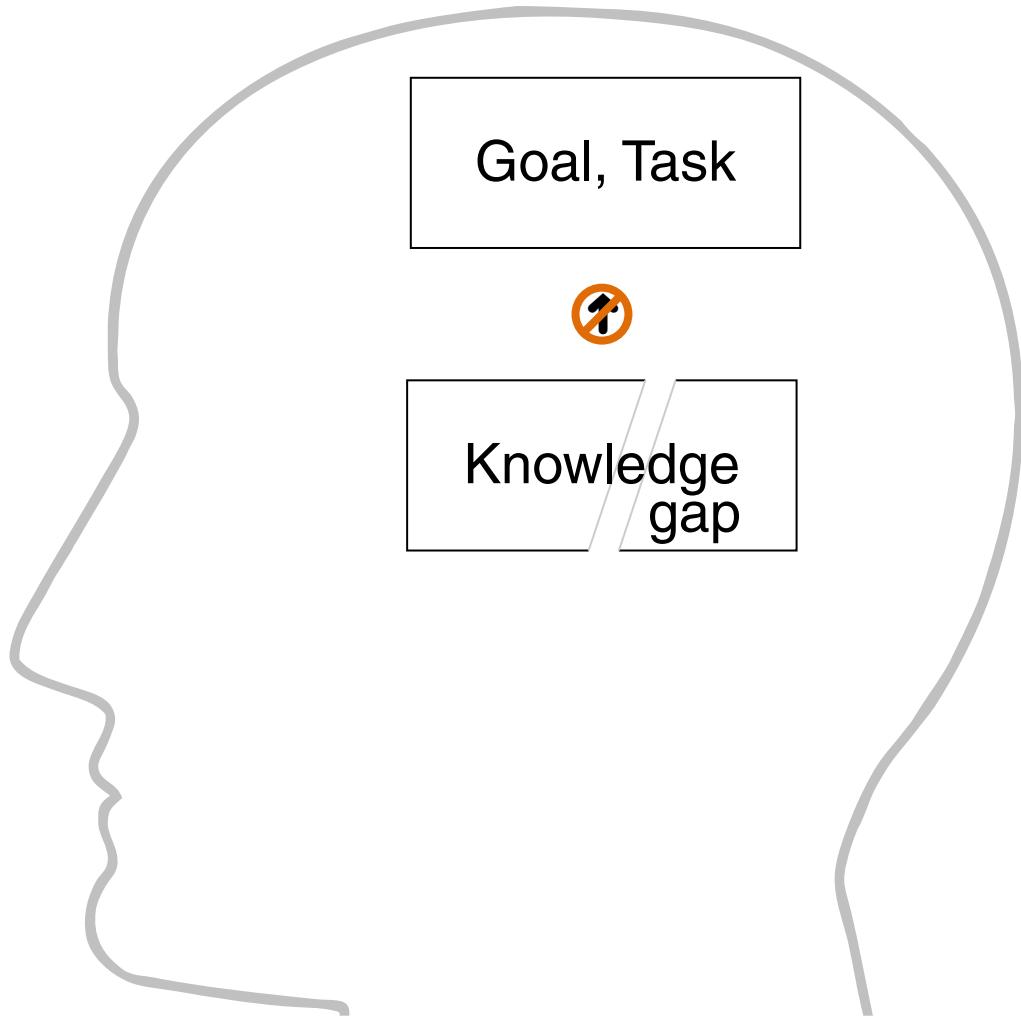
Terminology



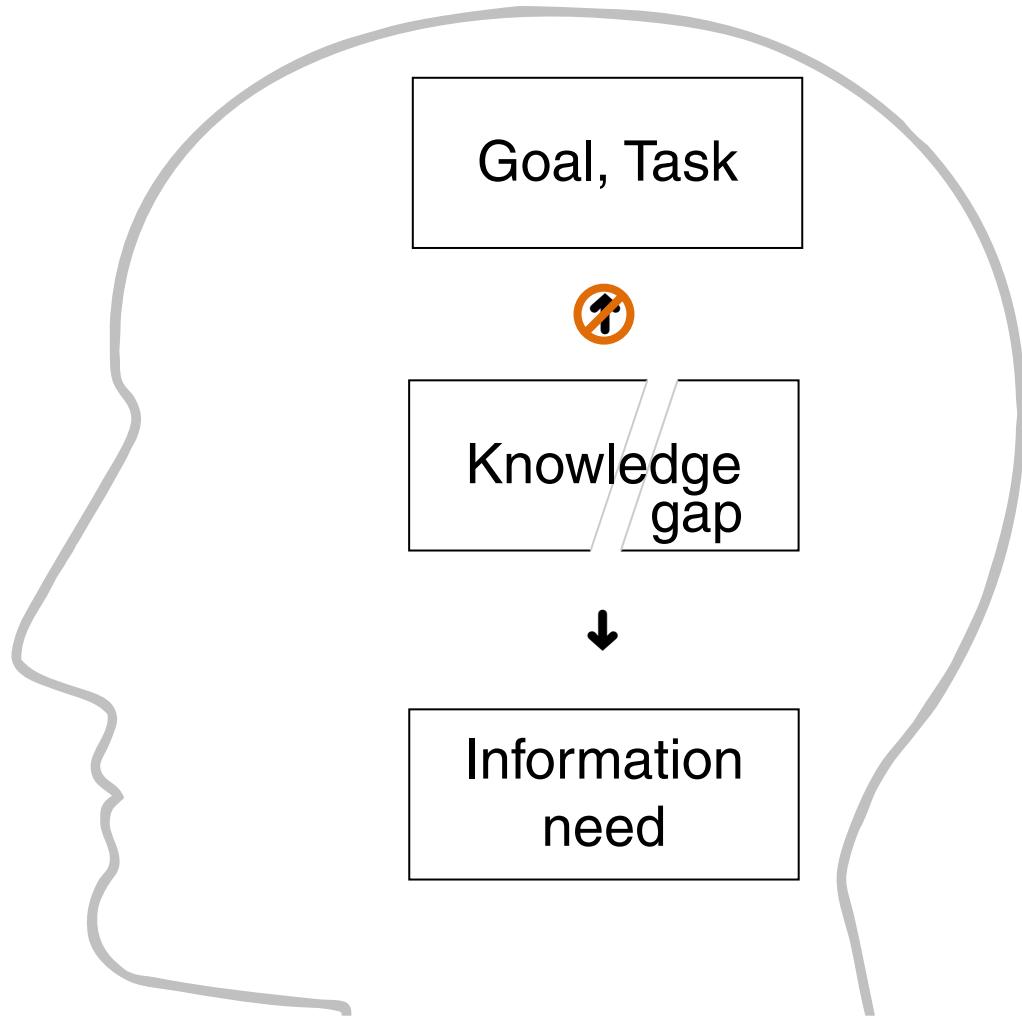
Terminology



Terminology



Terminology

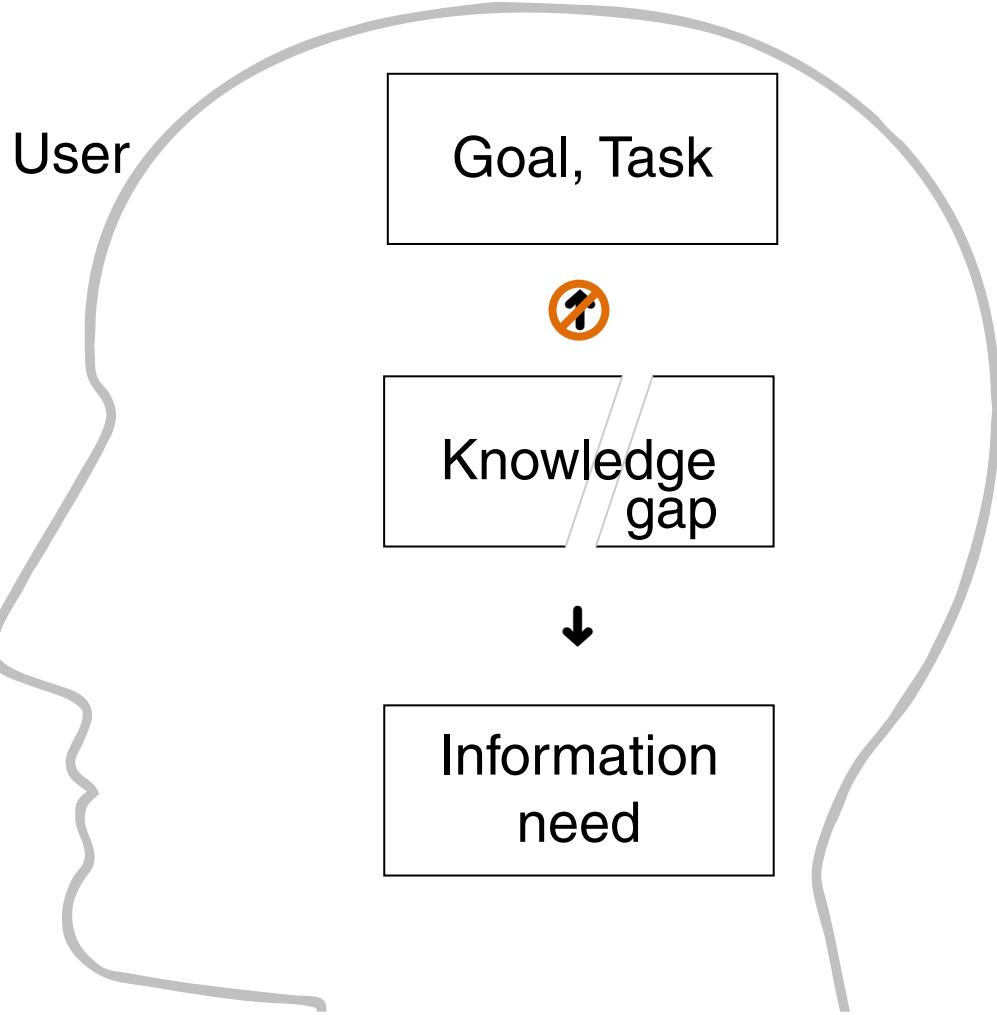


Terminology

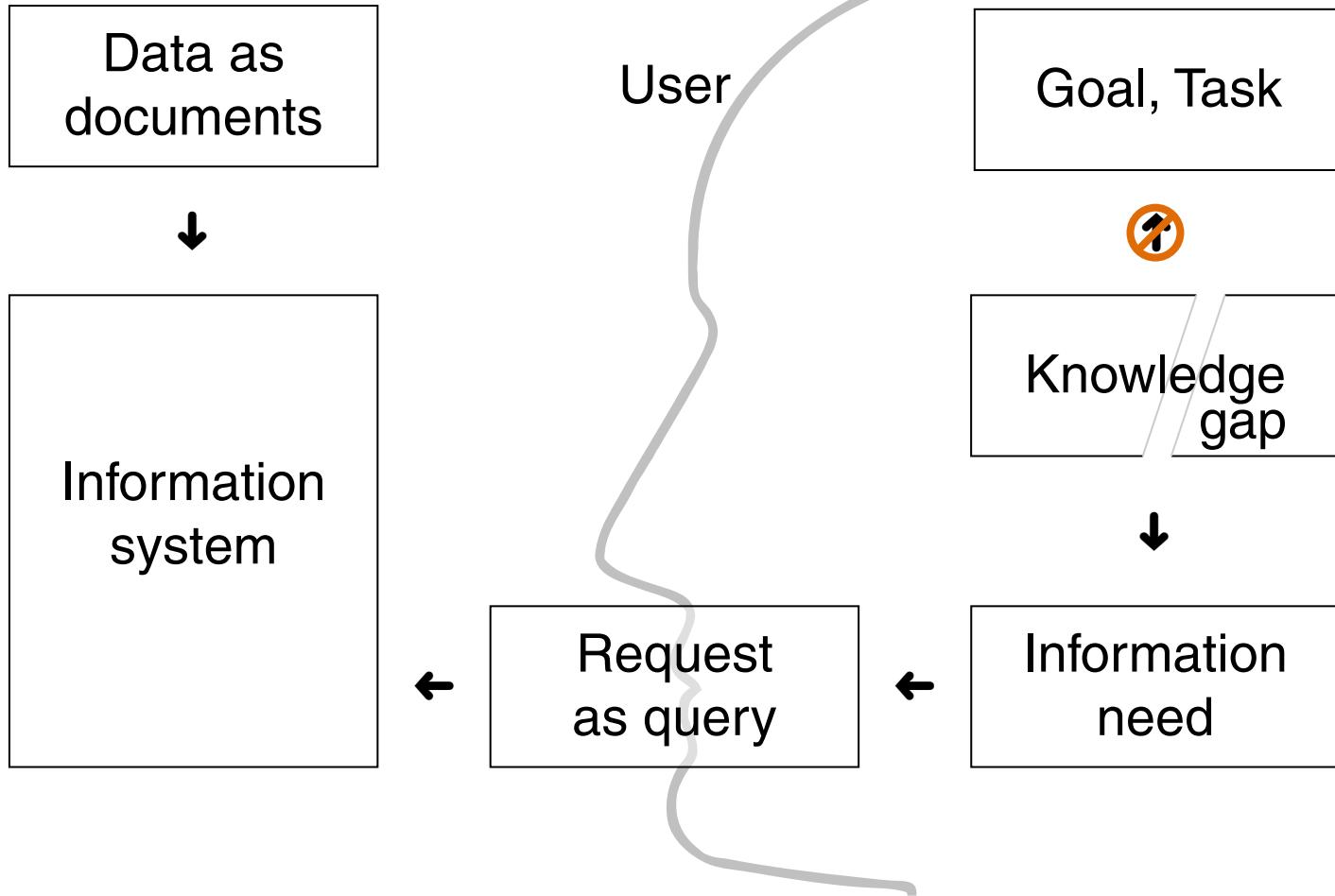
Data as documents



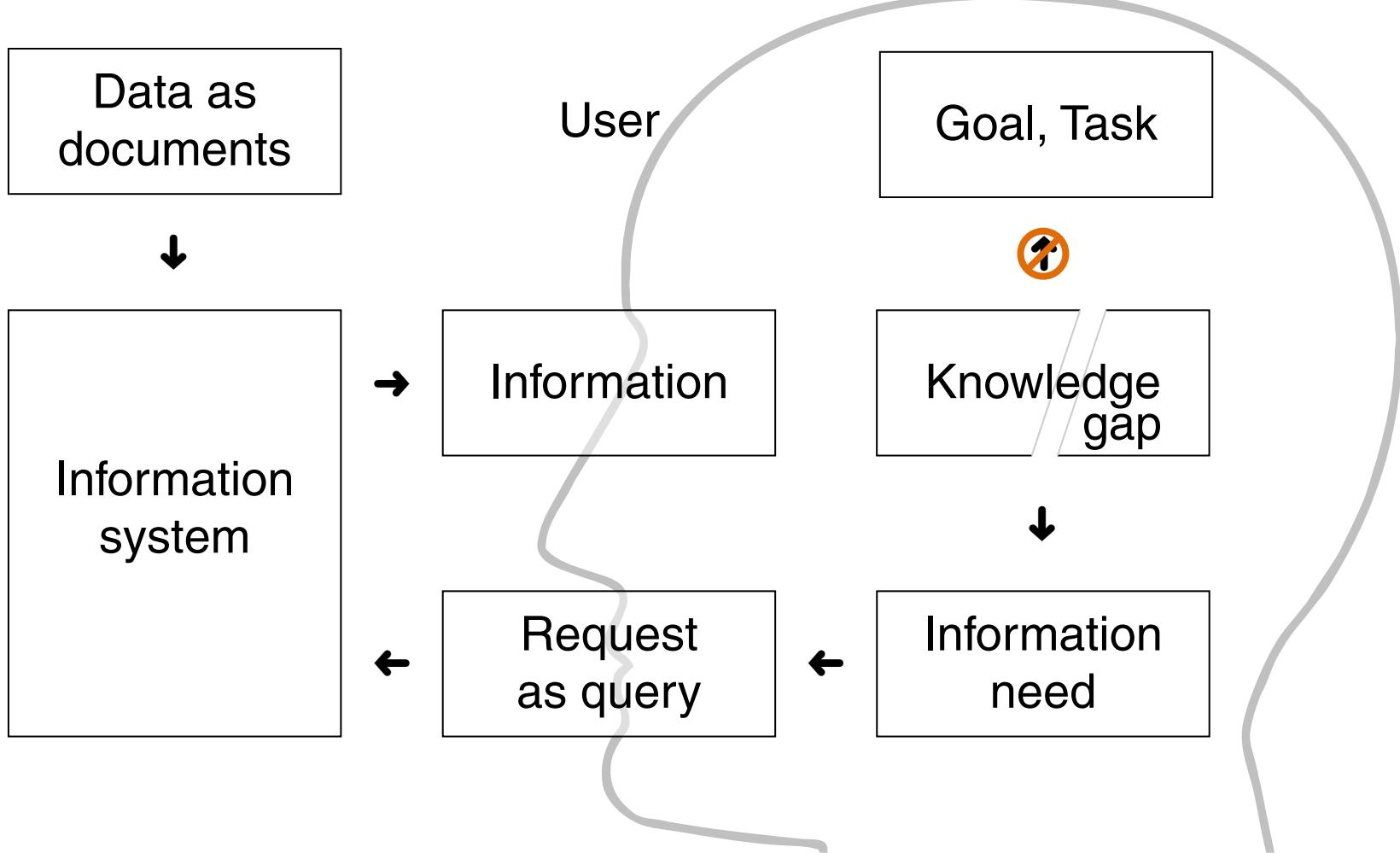
Information system



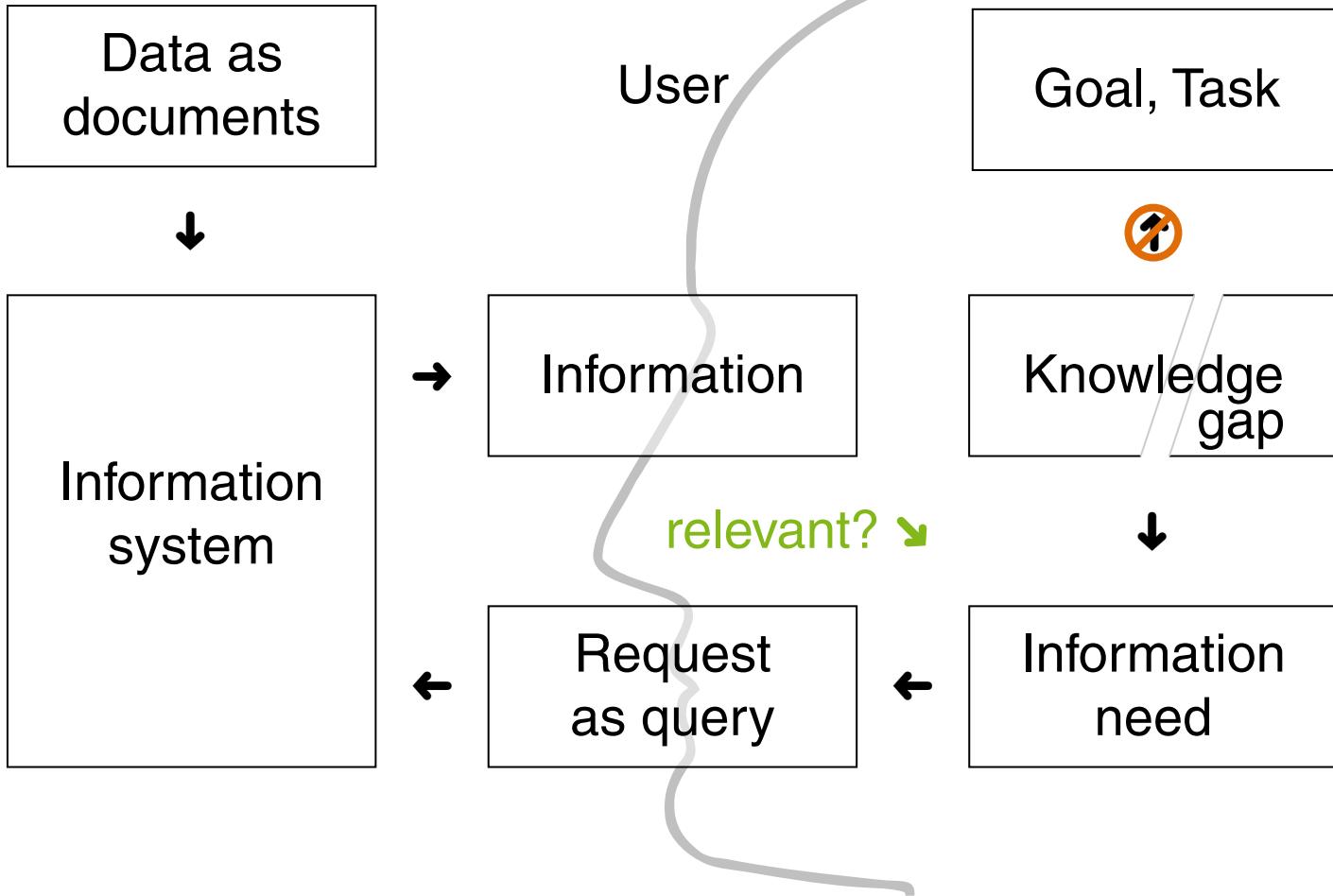
Terminology



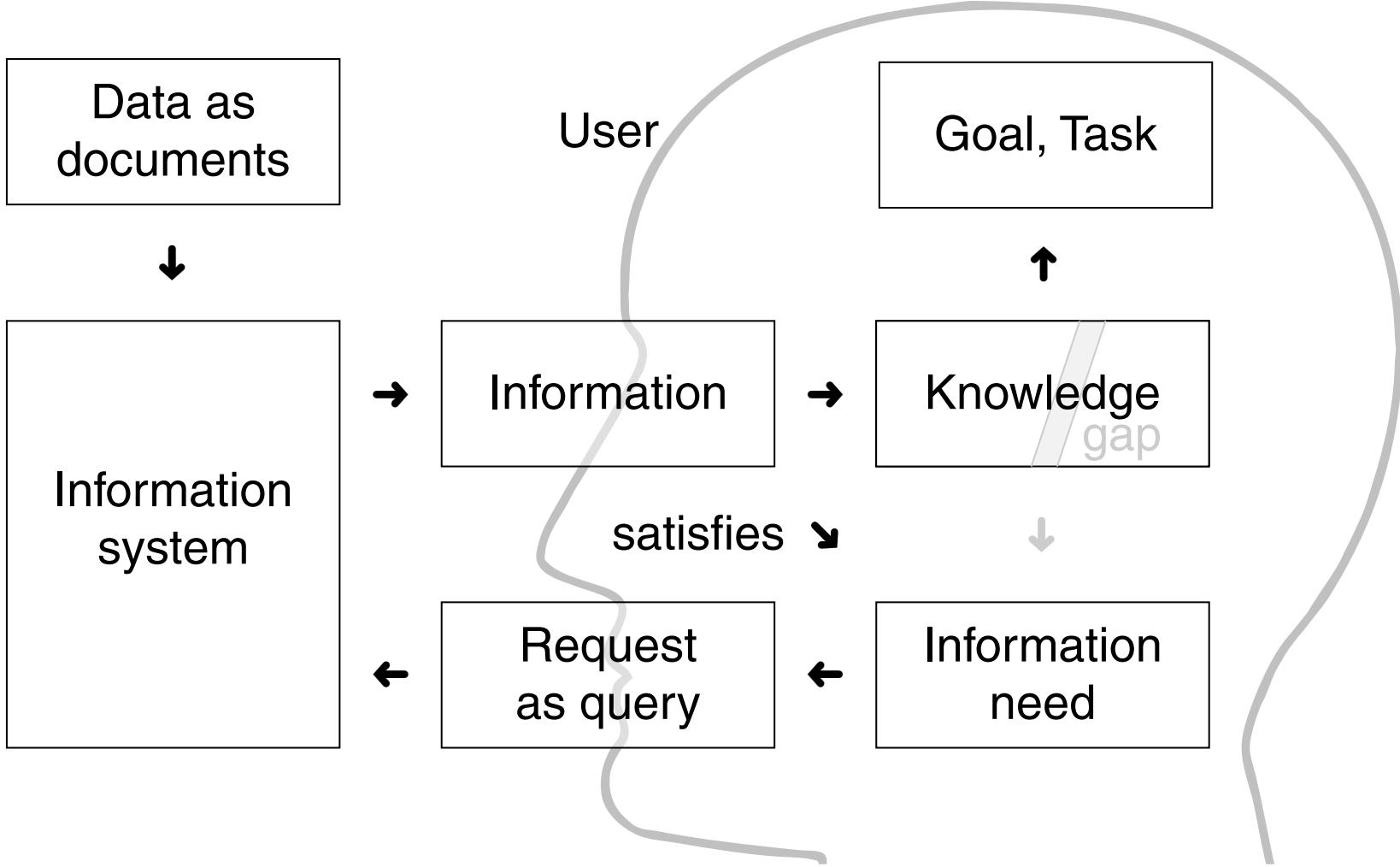
Terminology



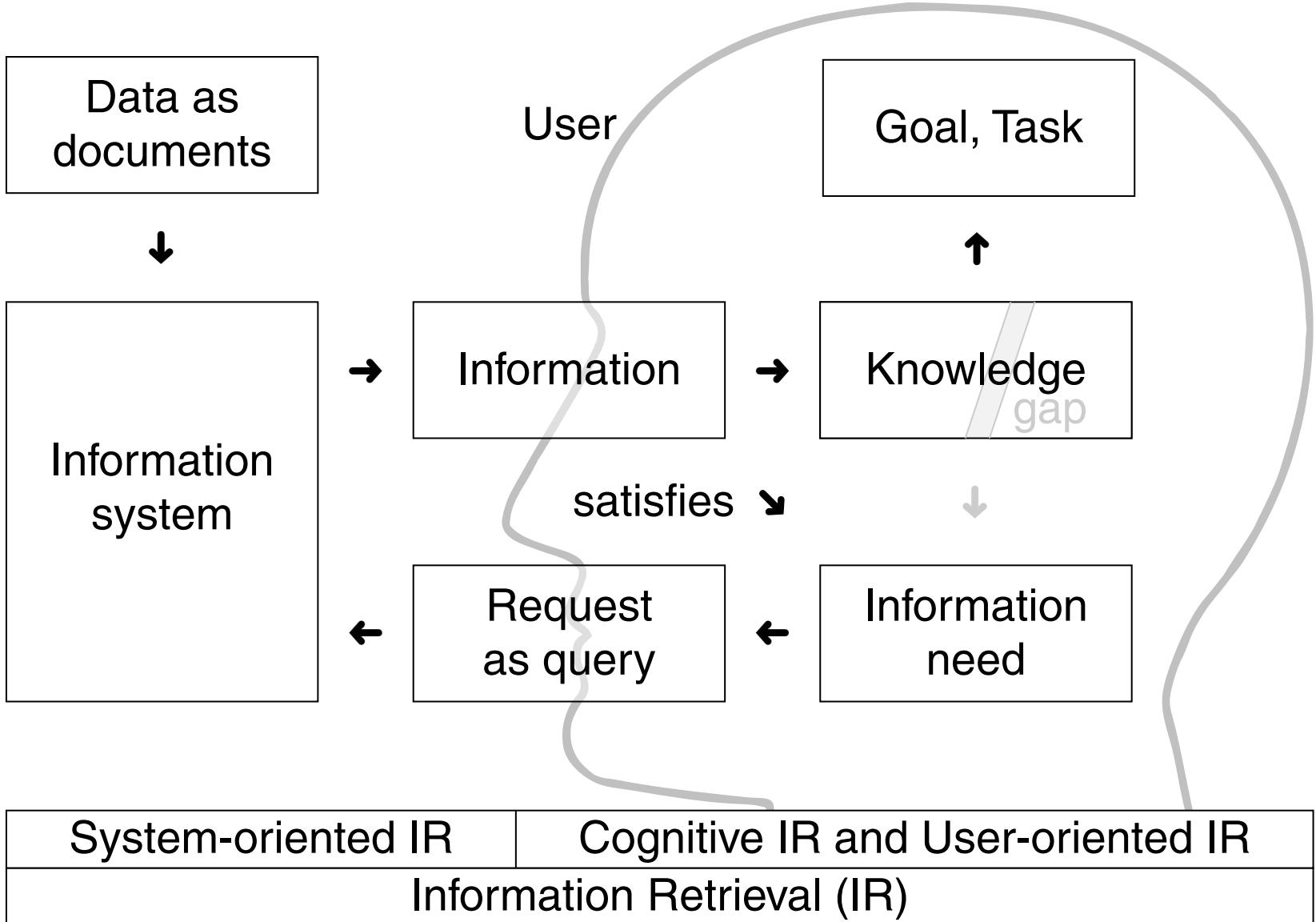
Terminology



Terminology



Terminology



Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

- Role of information systems:

- | | |
|--------------------|---|
| System-oriented IR | information technology |
| Cognitive IR | human interaction with information technology |
| User-oriented IR | information systems as sociotechnical systems |
| □ Design | architecture, algorithms, interfaces |
| □ Implementation | hardware, deployment, maintenance |
| □ Evaluation | effectiveness and efficiency |
| □ Analysis | experiments, user studies, log analysis |

Terminology

Definition 7 (Information Retrieval)

The activity of obtaining information relevant to an information need from data.

As a research field, information retrieval (IR) studies the role of information systems in transferring knowledge via data, as well as the design, implementation, evaluation, and analysis of such systems.

Major challenges of IR:

1. Vague queries

Goal not a priori clear; potential vocabulary mismatch; requires interaction / dialog to refine; dependence on previous results; often combining information from multiple data sources.

2. Incomplete and uncertain knowledge

Results from the limitations of accurately representing semantics; some domains are inherently incomplete / uncertain (e.g., opinion topics like politics, evidence vs. belief topics like religion, interpretation topics like history and news, biased data collections like the web)

3. Accuracy of results

4. Efficiency

Remarks:

- ❑ Definitions of system-oriented IR, cognitive IR, and user-oriented IR are vague.
- ❑ The goal in real-life IR is to find useful information for an information need situation. [...] In practice, this goal is often reduced to finding documents, document components, or document surrogates, which support the user (the actor) in constructing useful information for her / his information need situation. [...]

The goal of systems-oriented IR research is to develop algorithms to identify and rank a number of (topically) relevant documents for presentation, given a (topical) request. On the theoretical side, the goals include the analysis of basic problems of IR (e.g., the vocabulary problem between the recipient and the generator, document and query representation and matching) and the development of models and methods for attacking them. [...]

The user-oriented and cognitive IR research focused [...] on users' problem spaces, information problems, requests, interaction with intermediaries, interface design and query formulation [...].

[Ingwersen 2005]

- ❑ User-oriented IR moves the orientation from a “closed system” in which the IR “engine” is tuned to handle a given set of documents and queries, to one that integrates the IR system within a broader information use environment that includes people, and the context in which they are immersed.
- ❑ “Sociotechnical” refers to the interrelatedness of social and technical aspects of an organization. Sociotechnical systems in organizational development is an approach to complex organizational work design that recognizes the interaction between people and technology in workplaces. The term also refers to the interaction between society's complex infrastructures and human behavior.

[Toms 2013]

[Wikipedia]

Delineation

Databases, Data Retrieval [van Rijsbergen 1979]

	Data Retrieval	Information Retrieval
Matching	exact	partial match, best match
Inference	deduction	induction
Model	deterministic	probabilistic
Classification	monothetic	polythetic
Query language	artificial	natural
Query specification	complete	incomplete
Items wanted	matching	relevant
Error response	sensitive	robust

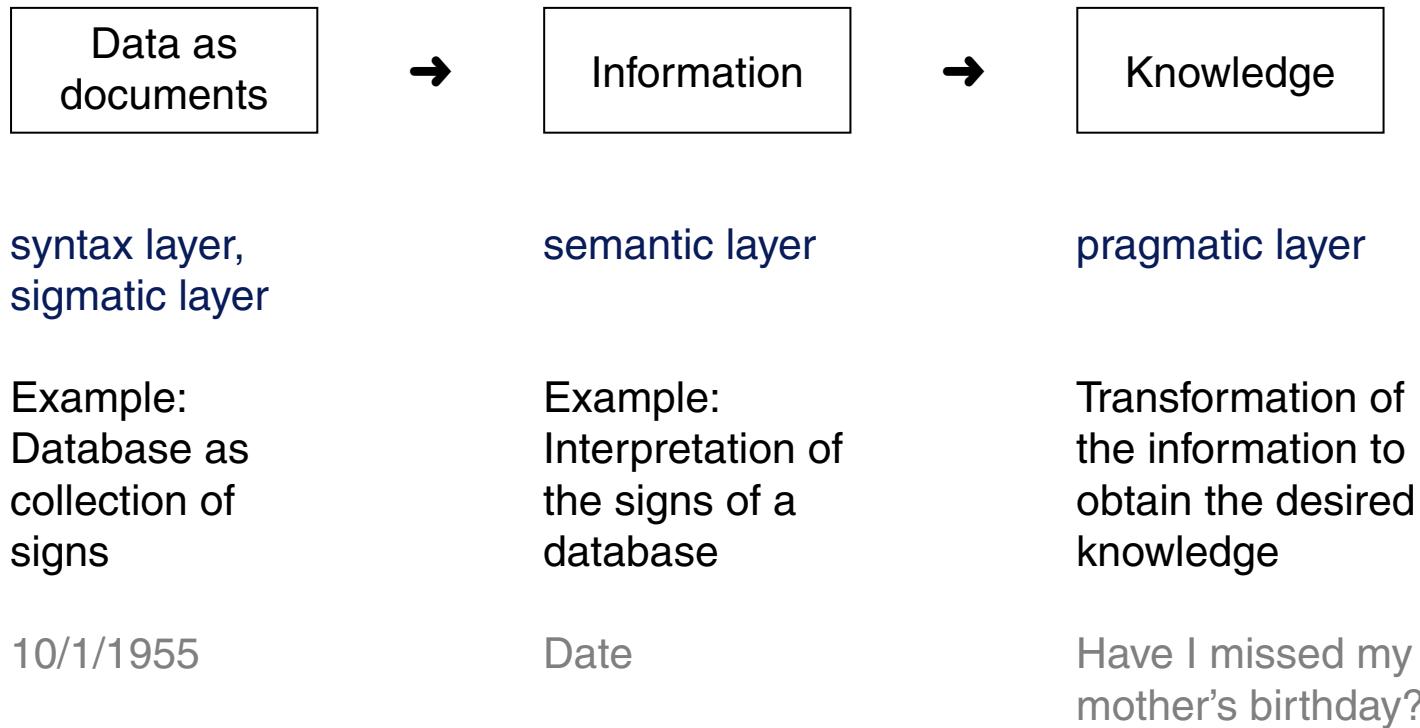
Remarks:

- A major difference between information retrieval (IR) systems and other kinds of information systems is the intrinsic uncertainty of IR. Whereas for database systems, an information need can always (at least for standard applications) be mapped precisely onto a query formulation, and there is a precise definition of which elements of the database constitute the answer, the situation is much more difficult in IR; here neither a query formulation can be assumed to represent uniquely an information need, nor is there a clear procedure that decides whether a database object is an answer or not. Boolean IR systems are not an exception from this statement; they only shift all problems associated with uncertainty to the user. [\[Fuhr 1992\]](#)
- In data retrieval we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a class. In IR such a classification is on the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class. [\[van Rijsbergen 1979\]](#)

Example: in a given database, persons are required to possess the attributes name, birth date, gender, etc.; documents about persons may each mention any given subset of these attributes.

Delineation

Semiotics



Information retrieval is an **associative search** that particularly addresses the semantics and pragmatics of documents.

Remarks:

- ❑ Semiotics (“sign theory,” derived from greek) is the study of meaning-making, the study of sign process (semiosis) and meaningful communication. Modern semiotics was defined by C.S. Peirce and C.W. Morris, who divided the field into three basic layers: the relations between signs (syntax), those between signs and the things signified (semantics), and those between signs and their users (pragmatics). [\[Wikipedia\]](#)
- ❑ K. Georg further distinguishes the relations between signs and the object to which they belong (sigmatics). [\[Wikipedia\]](#)
- ❑ The semiotic layers can be aligned with the basic concepts of information science, the latter forming elements of the former:

Layer	Element
Syntax	Sign (e.g., character)
Sigmatics	Data, Document
Semantics	Information
Pragmatics	Message, Knowledge

Delineation

Machine Learning, Data Mining

OLAP, Online Analytical Processing

KDD, Knowledge Discovery in Databases

Data mining, Web mining, Text mining

Scenario: gigabytes, databases, on the
(semantic) Web, in unstructured text

Machine learning

Scenario: in main memory,
specific deduction model

Statistic analysis

Scenario: clean data,
hypothesis evaluation

Explorative data analysis

Delineation

Machine Learning, Data Mining

Analysis	Information visualization	OLAP, Online Analytical Processing
	Data aggregation ...	KDD, Knowledge Discovery in Databases
	Data mining , Web mining, Text mining Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text	
	Machine learning Scenario: in main memory, specific deduction model	
Statistic analysis Scenario: clean data, hypothesis evaluation		
Descriptive data analysis	Explorative data analysis	

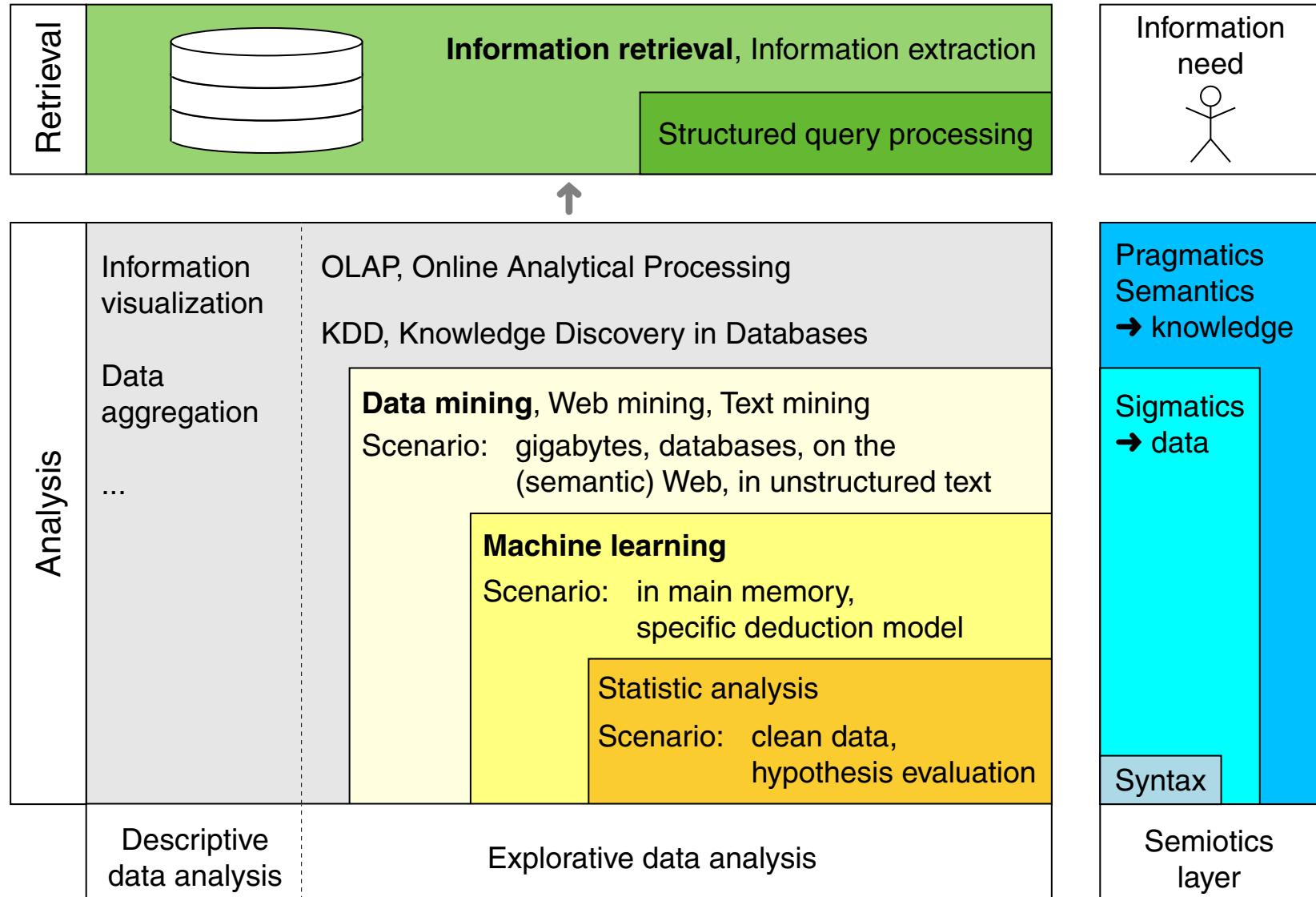
Delineation

Machine Learning, Data Mining

Analysis	Information visualization Data aggregation ...	OLAP, Online Analytical Processing KDD, Knowledge Discovery in Databases	Pragmatics Semantics → knowledge
		Data mining , Web mining, Text mining Scenario: gigabytes, databases, on the (semantic) Web, in unstructured text	Sigmatics → data
		Machine learning Scenario: in main memory, specific deduction model	Syntax
		Statistic analysis Scenario: clean data, hypothesis evaluation	Semiotics layer
	Descriptive data analysis	Explorative data analysis	

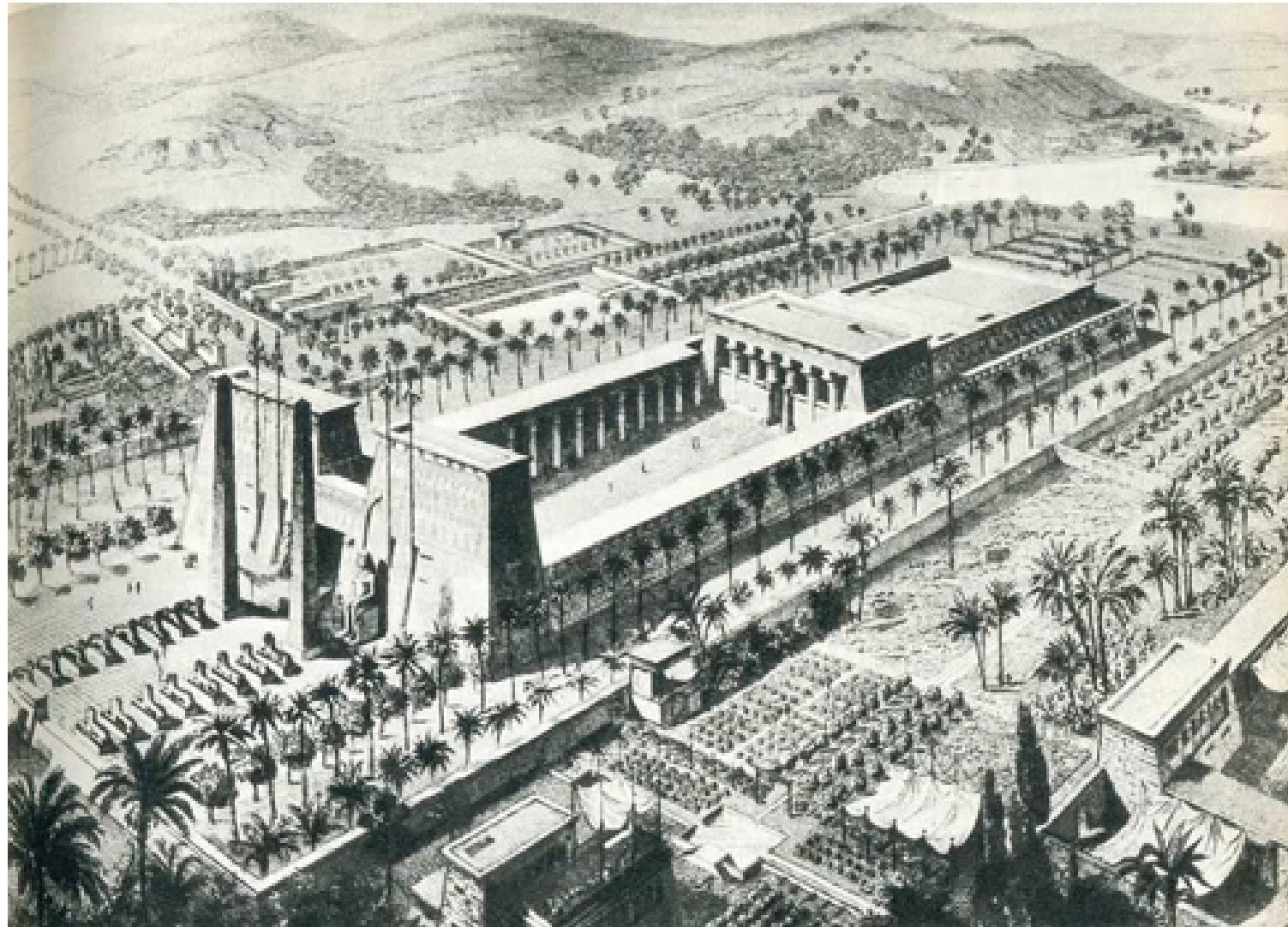
Delineation

Machine Learning, Data Mining



Historical Background

Manual Retrieval



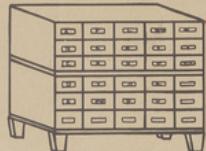
Remarks:

- ❑ The Ancient Library of Alexandria, Egypt, was one of the largest and most significant libraries of the ancient world. It flourished under the patronage of the Ptolemaic dynasty and functioned as a major center of scholarship from its construction in the 3rd century BC until the Roman conquest of Egypt in 30 BC. The library was part of a larger research institution called the Musaeum of Alexandria, where many of the most famous thinkers of the ancient world studied. [\[Wikipedia\]](#)
- ❑ These include Archimedes, father of engineering; Aristarchus of Samos, who first proposed the heliocentric system of the universe; Callimachus, a noted poet, critic and scholar; Eratosthenes, who argued for a spherical earth and calculated its circumference to near-accuracy; Euclid, father of geometry; Herophilus, founder of the scientific method; Hipparchus, founder of trigonometry; Hero, father of mechanics. [\[Wikipedia\]](#)
- ❑ Callimachus' most famous prose work is the *Pinakes* (*Lists*), a bibliographical survey of authors of the works held in the Library of Alexandria. The *Pinakes* was one of the first known documents that lists, identifies, and categorizes a library's holdings. By consulting the *Pinakes*, a library patron could find out if the library contained a work by a particular author, how it was categorized, and where it might be found. Callimachus did not seem to have any models for his *pinakes*, and invented this system on his own. [\[Wikipedia\]](#)
- ❑ The Library held between 400,000 and 700,000 scrolls, grouped together by subject matter. Within the *Pinakes*, Callimachus listed works alphabetically by author and genre. He did what modern librarians would call adding metadata—writing a short biographical note on each author, which prefaced that author's entry. In addition, Callimachus noted the first words of each work, and its total number of lines. [\[Phillips 2010\]](#)

Historical Background

Manual Retrieval

FROM CARD CATALOG TO THE BOOK ON THE SHELF



THE CARD CATALOG
is an alphabetical list of
books found in the Library

THE THREE WAYS OF FINDING A BOOK IN THE CATALOG



UNDER AUTHORS SURNAME

UNDER TITLE OF BOOK

UNDER SUBJECT WITH WHICH BOOK DEALS



**PEABODY
VISUAL AIDS**
PUBLISHED BY
FOLLETT BOOK COMPANY CHICAGO



THE CALL NUMBER

Directs you to the book's location on the shelf
and is found in the upper left hand corner of the
catalog card also on the back of the book
which is on the shelf

ARRANGEMENT OF BOOKS

A numerical system is followed in correct order

CLASSIFICATION

000-099	general works
100-199	Philosophy
200-299	Religion
300-399	Sociology
400-499	Anthropology
500-599	Natural Sciences
600-699	Useful Arts
700-799	Fine Arts
800-899	Literature
900-999	History

Fiction is not classified but is arranged on the shelves alphabetically by author

Prepared under the direction of Miss Ruby Ethel Gurdif for the Peabody Library School course in Teaching the Use of the Library. Planned by Martha Edmondson, lettered by Mr. McCord.

Historical Background

Manual Retrieval



Remarks:

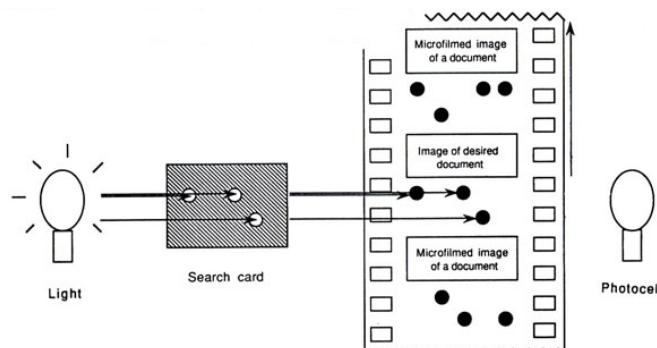
- ❑ Today, [WorldCat](#) is a union catalog that itemizes the collections of 72,000 libraries in 170 countries and territories. It contains more than 420 million records, representing over 2.6 billion physical and digital assets in 491 languages, as of August 2018. [\[OCLC\]](#)
- ❑ What are problems when sorting by author?
- ❑ What is necessary to organize library cards by subject?
- ❑ Librarians can find books by author, by title, and by subject. What is still missing?

Historical Background

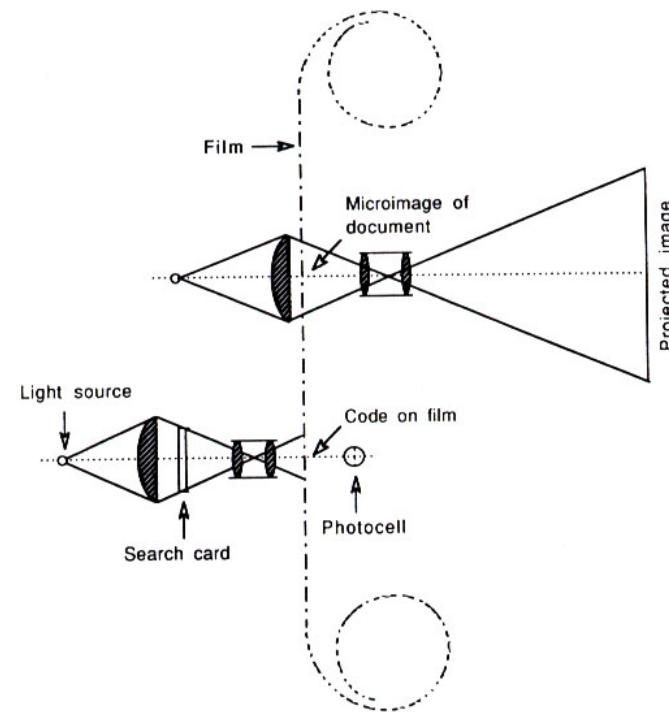
Mechanical Retrieval

Emanuel Goldberg's Statistical Machine [Buckland 1995]:

- Documents on microfilm with associated patterns of holes
- Punch cards as search patterns
- US patent No. 1,838,389, applied 1927, issued 1931



Searching

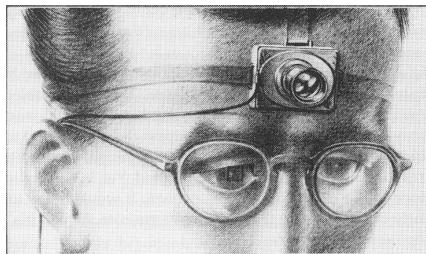


Result presentation

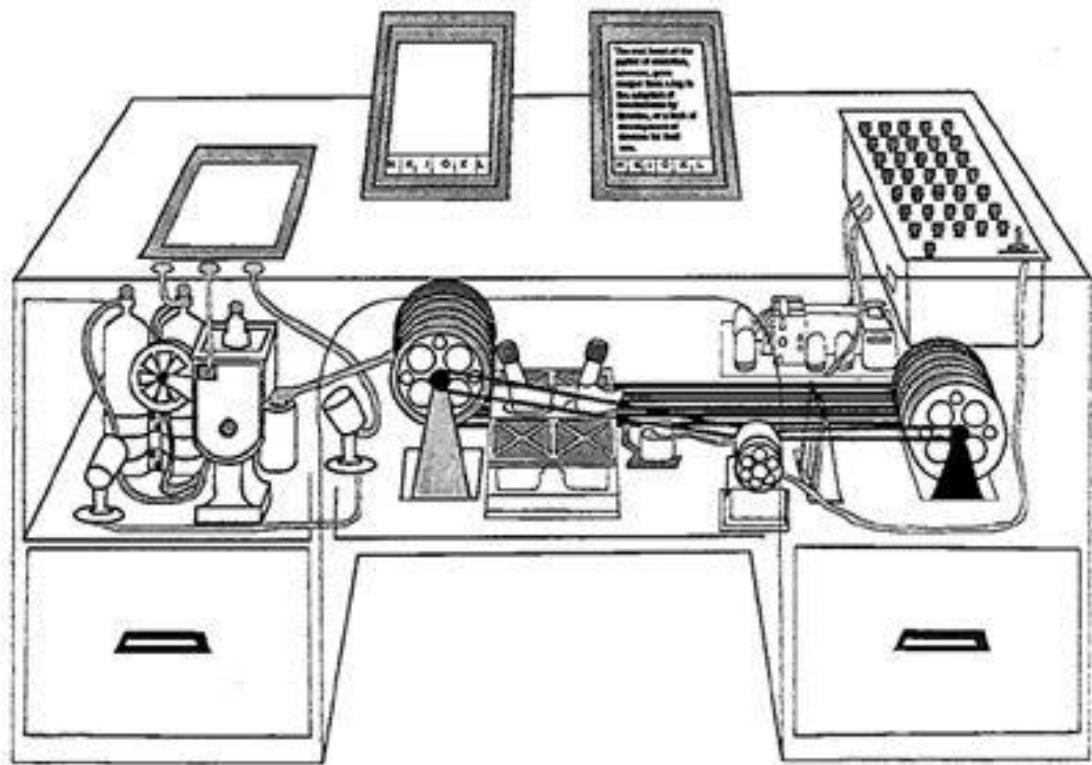
Historical Background

Mechanical Retrieval

Vannevar Bush's Memex [Bush 1945]:



Recording via camera
(early life logging)



Retrieval, Commenting, Browsing, Cross-referencing

Historical Background

Computerized Retrieval

First reference to computer-based search [Holmstrom 1948]:

Then there is also in America a **machine called the Univac** which has a typewriter keyboard connected to a device whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape.

By this means the **text of a document**, preceded by its subject code symbol, **can be recorded** on the tape by any typist.

For **searching**, the tape is run through the machine which thereupon automatically selects and types out those references which have been coded in any desired way **at a rate of 120 words a minute**--complete with small and capital letters, spacing, paragraphing, indentations and so on.

(If the tape is run through the other way, it obediently types out the text backwards at the same rate!)

Historical Background

Computerized Retrieval

First use of the term “information retrieval” [[Mooers 1950](#)]:

The problem under discussion here is machine searching and retrieval of information from storage according to specification by subject. An example is the library problem of selection of technical abstracts from a listing of such abstracts. It should not be necessary to dwell upon the importance of **information retrieval** before a scientific group such as this, for all of us have known frustration from the operation of our libraries – all libraries, without exception.

Remarks:

- ❑ Serious research into information retrieval started after World War II ended, when scientists of the allied forces turned their attention away from warfare, realizing that the vast quantities of scientific results and other information accumulated throughout the war was too much to make sense of for any individual scientist.
- ❑ [Bagley 1951] observed in his Master thesis that “recently published statistics relating to chemical publication show that a search of Chemical Abstracts would have been complete in 1920 after considering twelve volumes containing some **184,000 abstracts**. But in 1935 there would have been fifteen more volumes to search, and these new volumes alone contain about **382,000 abstracts**. By the end of 1950 the forty-four volumes of Chemical Abstracts to be searched contained well **over a million abstracts**. If the present trend in publication continues, the total abstracts published in this one field by 1960 will be almost **1,800,000**.”
- ❑ [Sanderson and Croft 2012] compiled a brief history of information retrieval research.

Historical Background

Information Retrieval (1950s)

Indexing and ranked retrieval:

- ❑ **Coordinate Indexing**

Mortimer Taube proposes “Coordinate Indexing” of documents based on a selection of independent “uniterms,” called (index) terms or keywords today, departing from traditional subject categorization schemes. Assigning uniterms to documents is called indexing. Adding a reference to a document to the specific catalog cards for its uniterms is called posting. Retrieval works by looking up a set of uniterms of interest, collecting documents to which at least a subset of them has been assigned.

- ❑ **Cranfield paradigm**

Cyril Cleverdon starts the Cranfield projects, introducing lab evaluation of indexing and retrieval based on (1) a document collection, (2) a set of queries, and (3) relevance judgments for pairs of queries and documents, later known as the Cranfield paradigm of IR evaluation.

- ❑ **Term frequency-based ranking**

Hans Peter Luhn proposes to score and rank documents based on their relevance to a query. He suggests term frequency of terms in a document as an approximate measure of term importance during scoring.

Historical Background

Information Retrieval (1960s)

Gerard Salton:

- Eminent IR researcher: “father of Information Retrieval”
- Many seminal works

Invention of / key contributions to automatic indexing, full-text indexing (i.e., using all words of a document as index terms), term weighting, relevance feedback, document clustering, dictionary construction, term dependency, phrase indexing, semantic indexing via thesauri, passage retrieval, summarization, ...



- Cosine similarity

The Vector Space Model, proposed by Paul Switzer, represents documents and queries in high-dimensional space. Salton suggests to measure the similarity between query and document vectors via the cosine of the angle between them, the cosine similarity.

- Integration of the state of the art into the SMART retrieval system.
- First laureate of the Gerard Salton Award in 1983, named in his honor.

Remarks:

- ❑ A funny side note; as per [\[Salton 1968\]](#), “information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”
What is wrong with this definition?
- ❑ Interestingly, commercial applications that emerged around this time largely ignored the insights gained from IR research. Not even ranked retrieval was adopted, but basic Boolean retrieval models were employed. This situation did not change until the mid-1990s and even today Boolean search is still very important in some domains like patent retrieval or prior art search, systematic reviews, etc.

Historical Background

Information Retrieval (1970s)

tf · idf-weighted Vector Space Model:

- ❑ Inverse document frequency

Karen Spärck Jones proposes the Inverse Document Frequency to measure term importance within document collections, complementing Luhn's term frequency to form the well-known *tf · idf* term weighting scheme.

- ❑ Vector space model

Supposed formalization of “A Vector Space Model for Information Retrieval” by Salton, Wong, and Yang; this attribution has been debunked [\[Dubin 2004\]](#).

Probabilistic retrieval:

- ❑ Probability ranking principle

Stephen Robertson formalizes the probability ranking principle, stating that “documents should be ranked in such a way that the probability of the user being satisfied by any given rank position is a maximum.”

- ❑ C.J. “Keith” van Rijsbergen proposes to incorporate term dependency into probabilistic retrieval models.

Historical Background

Information Retrieval (1980s - mid-1990s)

- ❑ BM25

Stephen Robertson et al. introduce BM25 (Best Match 25) as an alternative to $tf \cdot idf$.

- ❑ Latent semantic indexing

Scott Deerwester et al. propose to embed document and query representations in low-dimensional space using singular value decomposition of the term-document matrix.

- ❑ Stemming

Introduction of Porter's stemming algorithm into the indexing pipeline to conflate words sharing the same stem.

- ❑ TREC-style evaluation: shared tasks

Ellen Vorhees and Donna Harman organize the first Text REtrieval Conference (TREC), focusing on large-scale IR systems evaluation under the Cranfield paradigm, repeating it annually to this day.



- ❑ Learning to rank

Norbert Fuhr describes the foundations of learning to rank, the application of machine learning to ranked retrieval, where relevance is learned from training samples of pairs of queries and (ir)relevant documents.

Historical Background

Information Retrieval (mid-1990s - 2000s)

Web search:

- Web crawlers are developed for the rapidly growing web.

- PageRank and HITS

Spam pages increasingly pollute search results. Sergey Brin and Larry Page propose PageRank to identify authoritative web pages based on link structure, laying the foundation of Google. In parallel, John M. Kleinberg proposes HITS.

- Query log analysis

Thorsten Joachims renders learning to rank feasible, exploiting clickthrough data for training. Others develop query suggestion, spell correction, query expansion, etc. based on logs.

- Anchor text indexing

Oliver A. McBryan proposes the use of anchor text indexing to gain additional information about a web page, and to undo spam.

- Maximum marginal relevance for diversity

Jaime Carbonell and Jade Goldstein propose maximum marginal relevance (MMR) to allow for search result diversity.

- Language modeling for IR

Jay M. Ponte and W. Bruce Croft first apply language modeling to IR.

It's been a long way

Web Search



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Web Search



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Web Search



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

