



Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

# Processamento e Recuperação de Informação

## Introduction to Information Retrieval

Departamento de Engenharia Informática  
Instituto Superior Técnico

1º Semestre  
2018/2019



# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem
- 3 IR Tasks and Systems
- 4 Modeling
- 5 Search and The Web



# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem
- 3 IR Tasks and Systems
- 4 Modeling
- 5 Search and The Web



# Concepts to be Covered

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- Basic Concepts of Information Retrieval (IR)
- Classic Information Retrieval Models
- Advanced Information Retrieval Models (BM25, *N*-Gram Language Models, ...)
- Different Types of IR Tasks
- IR Evaluation
- Inverted Indexes and Query Processing
- Web Search and Advanced Search Results Ranking
- Meta-search



# Bibliography

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, **Modern Information Retrieval**, 2nd edition. **Chapter 1**, and beginning of Chapter 3.

Bing Liu, **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data**, 2nd edition. Chapter 6.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval**. Beginning of Chapters 1, 2, 6 and 19.



# Information Retrieval

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

IR deals with the representation, storage, organization of, and access to information items of various types:

- documents,
- Web pages,
- online catalogs,
- structured records,
- multimedia objects,
- ...



# Information Retrieval: Goals

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

Early goals in IR:

indexing text and searching for useful documents in a collection.



# Information Retrieval: Goals

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Early goals in IR:

indexing text and searching for useful documents in a collection.

## Current research goals:

- Modeling,
- Web search,
- text classification,
- systems architecture,
- user interfaces, data visualization,
- filtering and
- human language technologies (e.g., in dialogue systems).





# IR at the Center of the Stage

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- Until recently, IR was an area of interest restricted mainly to librarians and information experts.
- A single fact changed these perceptions:
  - The introduction of the **Web**, the largest repository of knowledge in human history.
- Due to its enormous size, finding useful information on the Web usually requires **running a search**

Searching on the Web is all about IR and its technologies



# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem**
- 3 IR Tasks and Systems
- 4 Modeling
- 5 Search and The Web



# The IR Problem

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- **Ad-hoc retrieval** is perhaps the core problem
- Users of modern IR systems, such as search engine users, have information needs of varying complexity

## An example of complex information need

*Find all documents that address the problem of global warming, including the currently accepted causes and possible mitigation solutions.*



# The IR Problem: Relevance

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- This full description of the user information need is **not necessarily a good query** to be submitted to the IR system
- Instead, the user might want to first translate this information need into a **query**
- This translation process yields a set of keywords, or **index terms**, which summarize the user's information need

Given the user query, the key goal of the IR system is to **retrieve information that is useful or relevant to the user**

The vague notion of **relevance** is of central importance in IR



# Information vs. Data Retrieval

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Data retrieval

Given a specified condition  
(e.g.  $\{\text{global, warming}\} \in \text{document}$ ), find all items that satisfy  
the condition

$\neq$

## Information retrieval

Given a user query, find all items that contain information  
**relevant** to the **user's needs**



# Information vs. Data Retrieval

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Data retrieval

Given a specified condition  
(e.g.  $\{\text{global, warming}\} \in \text{document}$ ), find all items that satisfy  
the condition

$\neq$

## Information retrieval

Given a user query, find all items that contain information  
relevant to the **user's needs**

However...

... how do you characterize the **user's information need**?



# Translating the user information need

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## An example

*Find all Web pages containing information on the ethical treatment of animals for medical experiments. The pages should contain references to recent related scientific articles, together with an enumeration of known existing alternatives for different medical fields.*

try this on Google



# Translating the user information need

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## An example

*Find all Web pages containing information on the ethical treatment of animals for medical experiments. The pages should contain references to recent related scientific articles, together with an enumeration of known existing alternatives for different medical fields.*

try this on Google

Usually this is translated to

ethics + animals + medical + experiments

but is this a convenient translation?  
how do IR systems deal with this?





# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem
- 3 IR Tasks and Systems**
- 4 Modeling
- 5 Search and The Web



# IR Tasks

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Document processing

- Crawling
- Segmenting and annotating
- Indexing
- Query processing
- Distributed IR
- String processing
- ...

## Information processing

- *Ad-hoc* retrieval
- Classification
- Clustering
- Filtering
- Summarization
- Question answering
- ...



# The Ad-hoc Retrieval Process

Processamento  
e Recuperação  
de Informação

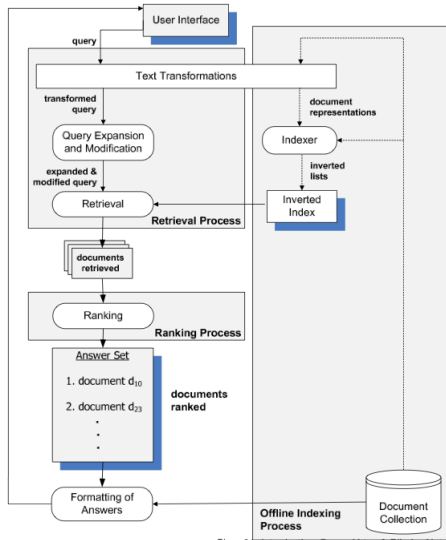
Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web





# A First Proposal: Memex

Processamento  
e Recuperação  
de Informação

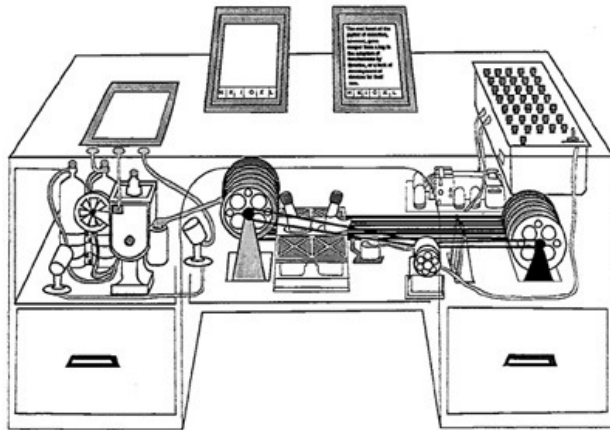
Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web



Bush, V. As We May Think, 1945



# A First Solution: Universal Card Scanner

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web



Luhn, H. P. 1950(-1958)



# The Modern Solution

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

The screenshot shows a Google search for "Hans Peter Luhn". The search bar at the top contains the text "Hans Peter Luhn" and a magnifying glass icon. Below the search bar, there are tabs for "All", "Images", "News", "Maps", "Videos", "More", "Settings", and "Tools". The "All" tab is selected. Below the tabs, it says "About 87,900 results (0.43 seconds)".

The search results are listed below:

- Hans Peter Luhn - Wikipedia**  
[https://en.wikipedia.org/wiki/Hans\\_Peter\\_Luhn](https://en.wikipedia.org/wiki/Hans_Peter_Luhn)  
Hans Peter Luhn (July 1, 1896 – August 19, 1964) was a researcher in the field of computer science, and, Library & Information Science for IBM, and creator of ...
- Hans Peter Luhn | ASIS&T**  
<https://www.asis.org/Pioneers>  
(1896-1964) His SDI and KWICKWOC indexes and many inventions at IBM, while not always commercially adopted, served as catalysts for inventors to ...
- Hans Peter Luhn and the Birth of the Hashing Algorithm - IEEE Spectrum**  
<https://spectrum.ieee.org/tech.../hans-peter-luhn-and-the-birth-of-the-hashing-algorithm>  
Jan 30, 2018 - In November 1958, at a six-day international conference devoted to scientific information, the inventor Hans Peter Luhn demonstrated a series ...
- Mr. Hans Peter Luhn | IT History Society**  
<https://www.ithistory.org/honor-roll/mr-hans-peter-luhn>  
A computer scientist for IBM, he was the creator of the Luhn algorithm and KWIC ( Key Words in Context) indexing. The Luhn algorithm or Luhn formula, also ...
- Hans Peter Luhn and the Birth of the Hashing Algorithm | Hacker News**  
<https://news.ycombinator.com/item?id=16270472>  
Should be: Hans Peter Luhn and the Birth of the Checksum Algorithm. Ctrl+F the article for "hash". What Luhn invented was a checksum, one that is apparently ...
- HANS PETER LUHN, MENTOR, 68, DIES; Data-Processing Specialist ...**  
<https://www.nytimes.com/.../hans-peter-luhn-mentor-68-dies-dataprocessing-specialist-se-19>  
19—Hans Peter Luhn, inventor and a consultant to the International Business Machines Corporation, died in his home here this afternoon. He was 68 years old ...

On the right side of the search results, there is a knowledge panel for Hans Peter Luhn. It includes a large portrait photo of him, a smaller collage of photos, and a "More images" link. Below the photos, his name "Hans Peter Luhn" is displayed, followed by the subtitle "Computer science researcher". A brief biography follows: "Hans Peter Luhn was a researcher in the field of computer science, and, Library & Information Science for IBM, and creator of the Luhn algorithm, KWIC indexing, and Selective dissemination of information. [Wikipedia](#)". His birth and death dates are listed: "Born: July 1, 1896, Barmen, Germany" and "Died: August 19, 1964, Ammonk, New York, United States". His known work is noted as "Known for: Key Word in Context" and his institution as "Institution: IBM".

Below the knowledge panel, there is a section titled "People also search for" with a "View 10+ more" link. It features four small portrait photos of other individuals: Gerard Salton, Karen Spärck Jones, Calvin Moores, and Eugene Garfield, each with their name written below.

Google Search, 2018



# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem
- 3 IR Tasks and Systems
- 4 Modeling**
- 5 Search and The Web



# An Example

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Example document

*A spider conducts operations that resemble those of a weaver, and a bee puts to shame many an architect in the construction of her cells. But what distinguishes the worst architect from the best of bees is this, that the architect raises his structure in imagination before he erects it in reality.*





# An Example

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Index terms

a	but	and	architect	bee	before
best	cell	conduct	construction	distinguish	erect
from	he	her	his	imagination	in
is	it	many	of	operation	put
raise		reality	resemble	shame	spider
structure	that	the	this	those	to
weaver	what	worst			



# An Example

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Index terms

a 4	but 1	and 2	architect 3	bee 2	before 1
best 1	cell 1	conduct 1	construction 1	distinguish 1	erect 1
from 1	he 1	her 1	his 1	imagination 1	in 3
is 1	it 1	many 1	of 3	operation 1	put 1
raise 1		reality 1	resemble 1	shame 1	spider 1
structure 1	that 2	the 4	this 1	those 1	to 1
weaver 1	what 1	worst 1			



# An Example

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Index terms

			architect 3	bee 2	before 1
best 1	cell 1	conduct 1	construction 1	distinguish 1	erect 1
	he 1	her 1	his 1	imagination 1	
is 1	it 1	many 1		operation 1	put 1
raise 1		reality 1	resemble 1	shame 1	spider 1
structure 1					
weaver 1		worst 1			



# An Example

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Logical view of the documents

	architect	bee	before	...	weaver	worst	...
$d_1$	3	2	1	...	1	1	...
$d_2$	0	1	0	...	2	2	...
$d_3$	0	2	0	...	1	0	...
$d_4$	2	0	2	...	2	1	...

...



# Bag-of-Words Models

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web



Art installation of a bag of words @ CMU



# Term Vectors

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

In general, documents can be represented as vectors, or sets, of term weights

$$\vec{d}_1 = (0, 1, 2, 3, 0, 1, \dots)$$

$$\vec{d}_2 = (3, 0, 1, 1, 0, 2, \dots)$$

...

Operations on the documents are operations on the vectors

Other, more complex, models are also possible:

- e.g., probabilistic language models, approximating the probability of word sequences as:

$$P(t_1 t_2 t_3) = P(t_1)P(t_2 | t_1)P(t_3 | t_1 t_2)$$



# Building the Logical View of the Documents

Processamento  
e Recuperação  
de Informação

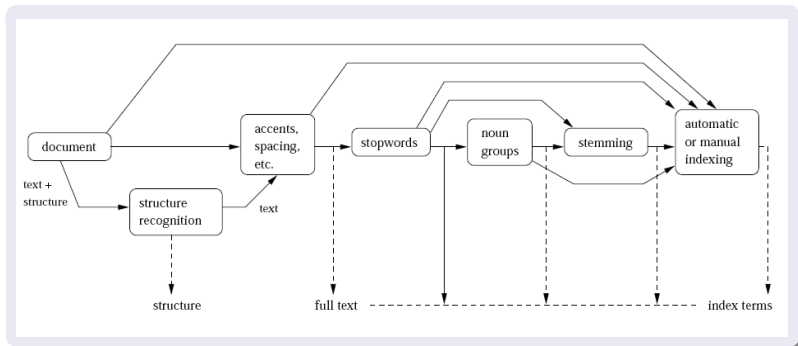
Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web





# Retrieval Models

Processamento  
e Recuperação  
de Informação

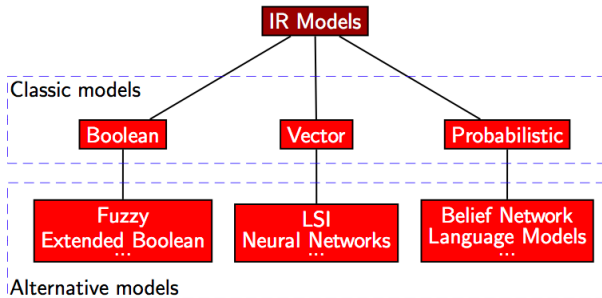
Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web







# Outline

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- 1 Scope of IR topics
- 2 The IR Problem
- 3 IR Tasks and Systems
- 4 Modeling
- 5 Search and The Web**



# Motivation for IR Research & Development

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

- Growing importance of access to information
- Growing volume of electronically stored information



- Growing need of efficient and effective means to **organize**, **store** and **provide access** to information
- None of these are trivial tasks



# How the Web Changed Search

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Search is everywhere

- Web search is today the most prominent application of IR and its techniques;
- The ranking and indexing components of any search engine are fundamentally IR pieces of technology.
- A **search box** is now common in every web site, app, ...

Five major impacts...



# First Major Impact

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

The first major impact of the Web on search is related to the **characteristics of the document collection** itself

- The traditional Web is composed of pages distributed over millions of sites and connected through hyperlinks
- This requires collecting all documents and storing copies of them in a central repository, prior to indexing
- This new phase in the IR process, introduced by the Web, is called **crawling**
  - Many challenges related to the fact that Web contents are very dynamic, and are often hidden behind forms



## Second Major Impact

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

The second major impact of the Web on search is related to:

- The size of the collection
- The volume of user queries submitted on a daily basis

As a consequence, **performance and scalability** have become critical characteristics of IR systems

Modern Web-based services (e.g., social media platforms) are further extending the scope of these scalability problems



# Third Major Impact

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

In a very large and diverse collection (e.g., documents in multiple languages, of different types, often quite short and ambiguous), **predicting relevance is much harder** than before

Fortunately, the Web also includes new sources of evidence

- hyperlinks
- user clicks in documents in the answer set
- complex user profiles
- ...



# Fourth Major Impact

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

The Web is also a medium to do business

Search problem has been extended beyond the seeking of text information to also encompass **entities and other user needs**:

- the price of a book,
- the phone number of a hotel,
- the location of a restaurant,
- the link for downloading a software,
- ...

Mobile and voice-based access to the Web has further extended the scope of the problem

- Increasing importance of **automated assistants** leveraging natural language technologies



# Fourth Major Impact and Relation to IE

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

Advanced search services (e.g., focusing on entities and their properties) often involve **Information Extraction** (IE)

Search is made over structured knowledge bases, build through large-scale extraction (e.g., from the Web, e-mails, ...)

- Knowledge graphs associated to major search engines,
- Price comparison services,
- Services like Google Trips,
- Dialogue services: Siri, Cortana, Alexa, Google Assistant,
- ...

Many different Information Extraction challenges:

- recognizing and disambiguating entities in text,
- extracting properties and relations
- ...





# Fifth Major Impact

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

The fifth major impact of the Web on search is **Web spam**

## Web spam:

abusive availability of commercial information disguised in the form of informational content

This difficulty is so large that today we even talk of **Adversarial Web Retrieval**.



# Practical Issues in the Web

Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

## Security

Commercial transactions over the Internet are not yet a completely safe procedure

## Privacy

Frequently, people are willing to exchange information as long as it does not become public

## Copyright and patent rights

It is far from clear how the widespread data on the Web affects copyright and patent laws in the various countries

## Interfaces

Modern trends relate to multi-modality (e.g., voice interaction through mobile devices) and going beyond *search boxes* (e.g., tools that assist in solving complex tasks)



Processamento  
e Recuperação  
de Informação

Scope of IR  
topics

The IR  
Problem

IR Tasks and  
Systems

Modeling

Search and  
The Web

# Questions?