FOUNDATIONS OF INFORMATION RETRIEVAL

Lynda Tamine-Lechani
lynda.lechani@irit.fr
https://www.irit.fr/~Lynda.Tamine-Lechani/

FOUNDATIONS OF INFORMATION RETRIEVAL

- Course description

Study the theory, design, and implementation of **information retrieval systems** from the perspectives of:

✓ **information representation**: focus on ***texts***

✓ theoretical **information retrieval model**: focus on ***language model and learning-based models***

✓ **Performance evaluation**: focus on ***system-centred evaluation***

- Learning objectives

✓ Index and represent textual information;

✓ Recall and discuss well-known information retrieval models;

✓ Design, implement and evaluate the performance of information retrieval systems using retrieval algorithms and models discussed in class.

© L. Tamine-Lechani

FOUNDATIONS OF INFORMATION RETRIEVAL

- Organization
  - 12H course, 6H tutorial: ***Lynda Tamine-Lechani***
  - 10H hands-on work: **Jesus-Lovon Melgajero**, ***José G. Moréno*** and ***Lynda Tamine-Lechani***
- Prerequisites
  - Python programming
  - Basics in probability and statistics
- Course material
  - Copies of the lecture slides are posted on the MOODLE site
  - Book and readings references are provided
- Grading
  - 1st session
    - ✓Hands-on experience with techniques discussed in class: ***assignment of 30% of the final score***
    - ✓Final written exam in class: ***assignment of 70% of the final score***
  - 2nd session
    - ✓Final written exam in class: ***assignment of 100% of the final score***

© L. Tamine-Lechani

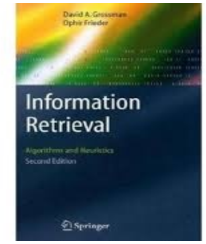
- Schedule

Lecture	Topic
1	Course Introduction; Text indexing, vector semantics
2	Static embeddings, contextual embeddings
3	Infomation retrieval (IR) models: query reformulation, learning to rank
4	Tutorial 1: Text indexing and representation
5	Neural models for IR
6	Page Rank, Performance evaluation
7	Tutorial 2: information retrieval techniques and models
8	Question answering systems and chatbots
9	Tutorial 3: performance evaluation

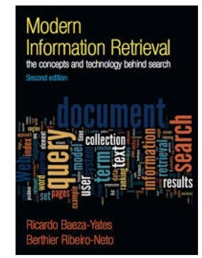
© L. Tamine-Lechani

Books

Information retrieval: Algorithms and Heuristics
David A. Grossamnn, Ophir Frieder, Kluwer
Academic Publishers, 1998

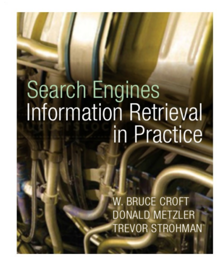


Modern information retrieval
R.B Yates, R. Neto, ACM Press Addisson Wesley, 1999



Recherche d'information, applications, modèles et algorithmes
M.R Amini et E. Gaussier, Eyrolles 2012

Search engines in practice
B. Croft, D. Metzler, T. Trohman, Pearson 2010



© L. Tamine-Lechani

Information Retrieval (IR): definitions

Calvin Mooers 1951 :

Information retrieval (IR) is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. .. Information retrieval is crucial to documentation and organization of knowledge". (Mooers, 1951, p. 25)

Salton, 1980 :

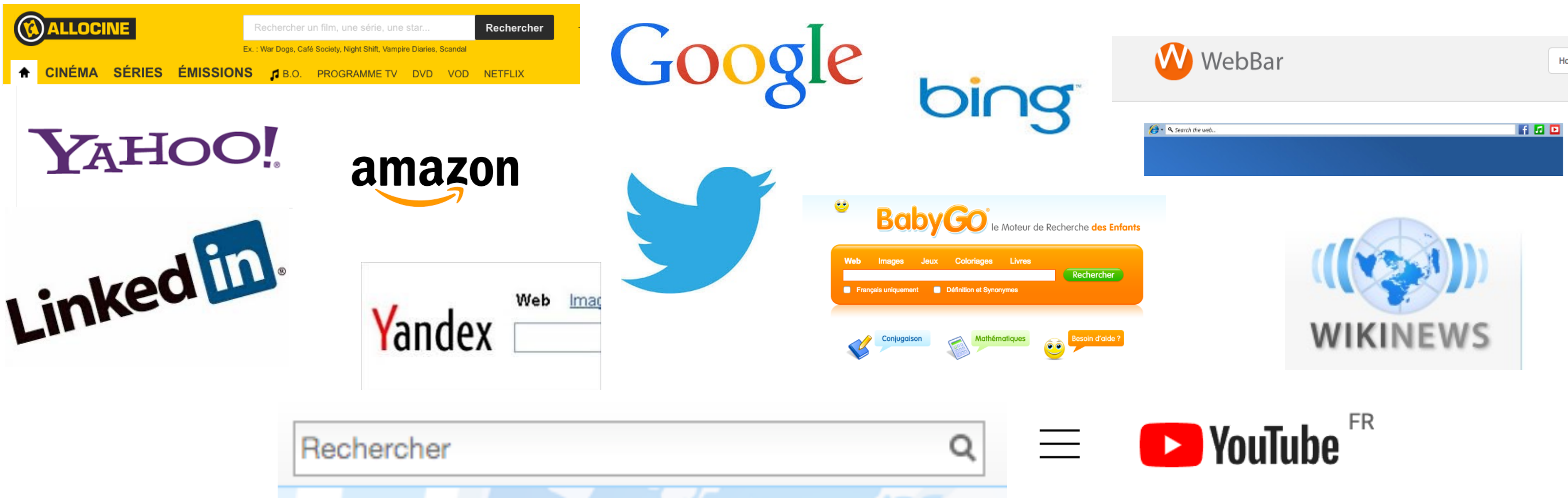
Information retrieval systems are designed to help analyze and describe the items stored in a file, to organize them and search among them, and finally to retrieve them in response to a user's query. Designing and using a retrieval system involves four major activities: information analysis, information organization and search, query formulation, and information retrieval and dissemination.

Information retrieval (IR) in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing.

WIKIPÉDIA
L'encyclopédie libre

© L. Tamine-Lechani

Definitions refer to ....well-known search engines ?



...Yes, but also refer to:

- Search in digital libraries
- Search in campany corpus
- Search in specialized corpus (health, legal, biological –related resources)
- Search for a location
- Search for answers
- Recommend items
- Summarize reviews
- ...

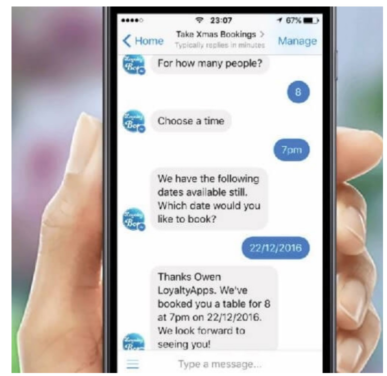
© L. Tamine-Lechani

...and different forms of user-system interactions

- Wide-variety of search systems, interaction environments
  - Web search engines
  - Conversational agents
  - E commmerce: Amazon, AirBnb, ...
  - Media recommendation: Netflix, Spotify, ...

...with voice only!

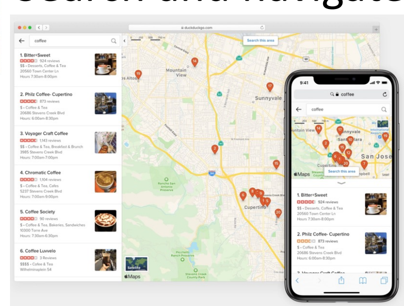
From search to conversation



Heatmaps on SERP

Cross-device search

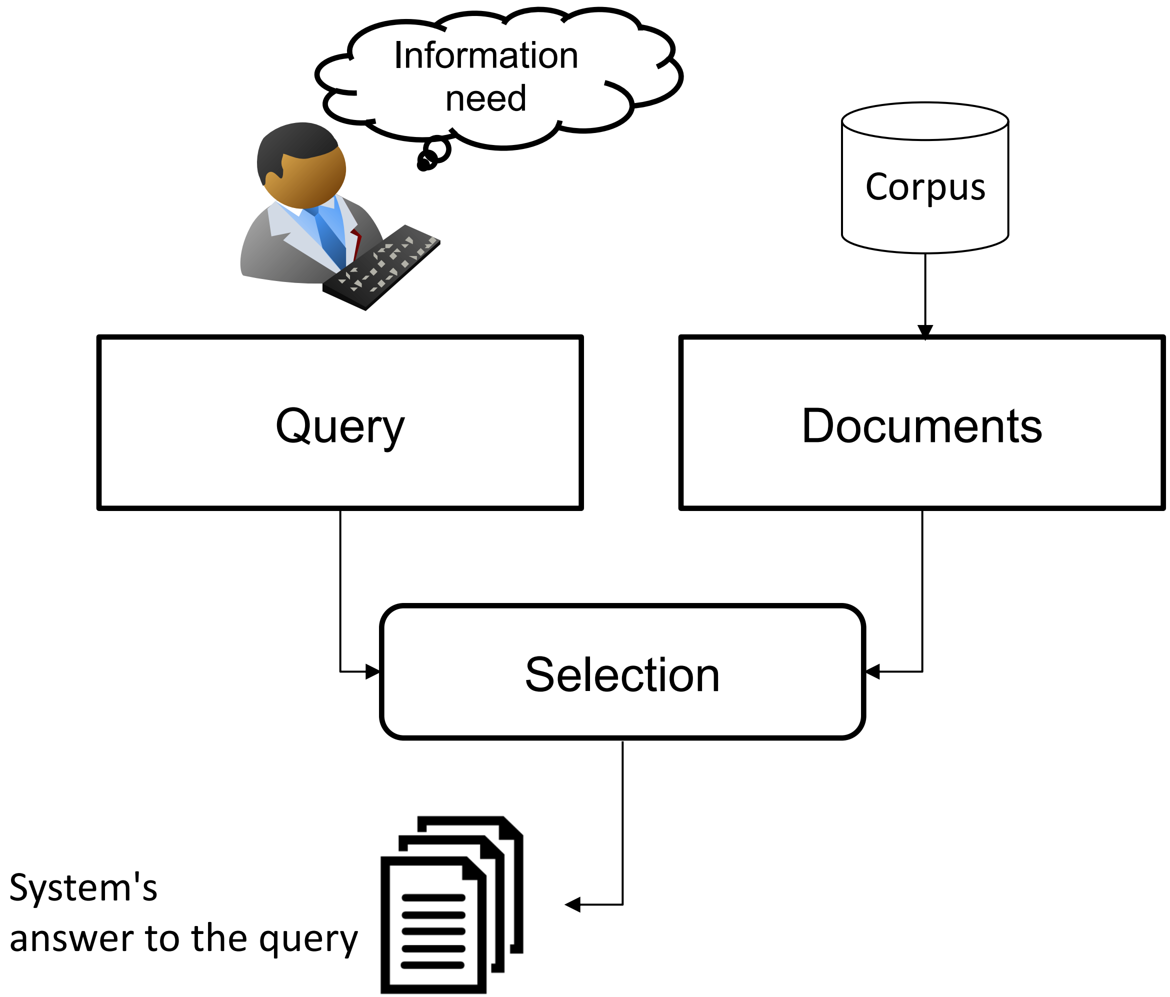
Search and navigate on



© L. Tamine-Lechani

Focus in this lecture

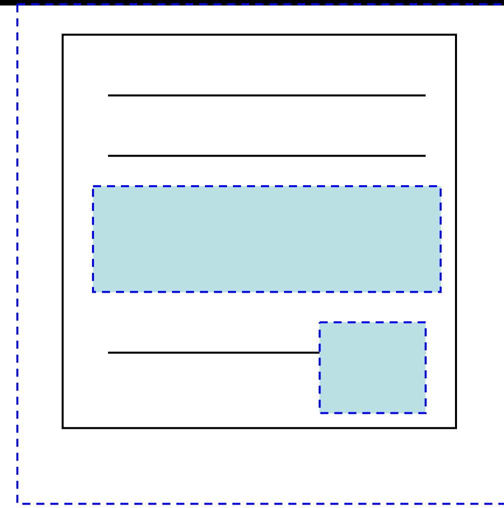
(Web) search systems that select from a corpus of texts **documents** those that are **relevant** to a user information need expersssed by the user using a **query.**



© L. Tamine-Lechani

Basic notions: Document

- **Document:** information unit being searched
  - Document
  - Paragraph
  - Phrase
  - Structure unit (section, chapter,...)



- Different views

Content

This course introduces
the basics of
information retrieval

Structure

1. Introduction
Information
retrieval....

2. Basics
The notion of
query...

Metadata

Date : 15/01/2013
Author : Albert
Langue : Français
...

© L. Tamine-Lechani

Basic notions: Document

- Different media

Text (monomedia)

Présentation

Wikipédia est un projet d'encyclopédie collective établie sur Internet, universelle, multilingue et fonctionnant sur le principe du wiki. Wikipédia a pour objectif d'offrir un contenu librement réutilisable, objectif et vérifiable, que chacun peut modifier et améliorer.

Le cadre du projet est défini par des principes fondateurs. Son contenu est sous licence Creative Commons by-sa et peut être copié et réutilisé sous la même licence — même à des fins commerciales — sous réserve d'en respecter les conditions.

Multimedia

Rugby : le XV de France s'incline en Irlande (18-11)

Malgré un essai inscrit en fin de match, les Bleus ont concédé à Dublin leur première défaite dans ce Tournoi des six nations, une semaine après leur victoire poussive contre l'Ecosse.

Video

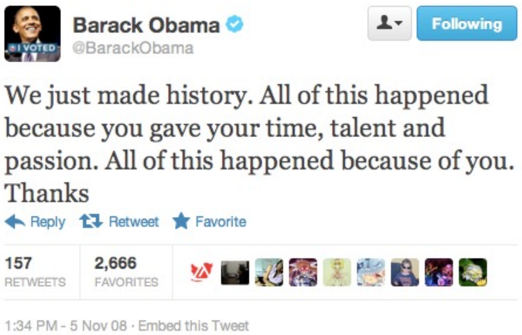
Image

© L. Tamine-Lechani

Basic notions: Document

- Different forms

  -Document
  -Blog
  -Tweet
  -News
  -Presentation
  -E-mail
  --..



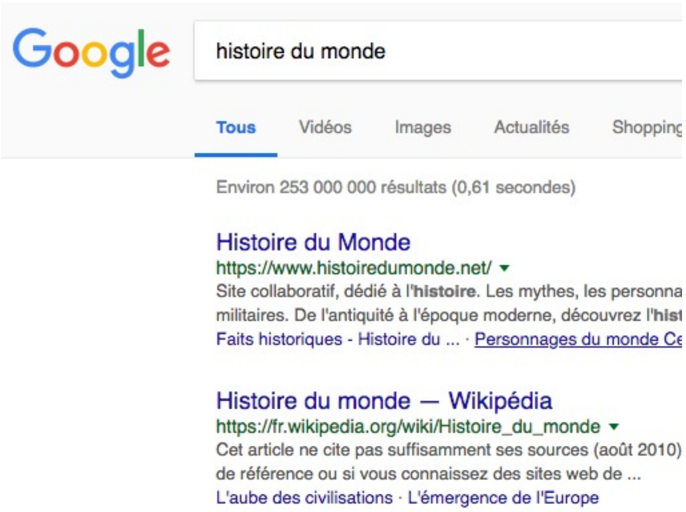
© L. Tamine-Lechani

Basic notions: information need, query

- What the user seeks for: **an information need**

- How the user expresses his information need : **a query**

In this course: **a query is a list of keywords**



© L. Tamine-Lechani

Basic notions: Relevance

- A key concept in information retrieval

  A document is **relevant** if it **matches the information need**. Numerous types of relevance:

  - Topical (aboutness) relevance: the document covers the query topic
  - Situational relevance: the document matches the user's situation (e.g., task, location, ...)
  - Cognitive relevance : the documents fits with the user's knowledge state
  - ...

  and numerous criteria of relevance:

  - Novelty
  - Fresheness
  - Language
  - Specificity
  - Trust
  - ...

The main focus in this course is topical relevance: useful and "easy" to define and to measure, but it does not cover everything related to relevance

© L. Tamine-Lechani

What makes information retrieval challenging ?

© NIST (TREC)

What makes information retrieval challenging ?

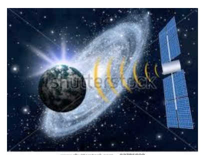
- Deluge of information
  - Large-scale information
  - Often little ratio of information is relevant and/or useful for a query
  - Information is noisy
  - Information is not always trusty
  - Hetrogeneous information forms and sources
  - ...

Information is every where

Increasing volumes of information
available on increasing information sources: social applications
mobile devices, sensors, ...







1972	1990	1994	1995	1998	1999	2001	2003
ARPANET	WWW	E-commerce	Annuaire	Recherche	Blogs	Wiki	Réseaux sociaux

Source : Infographic

© L. Tamine-Lechani

Focus on Web 3.0: The digital world today

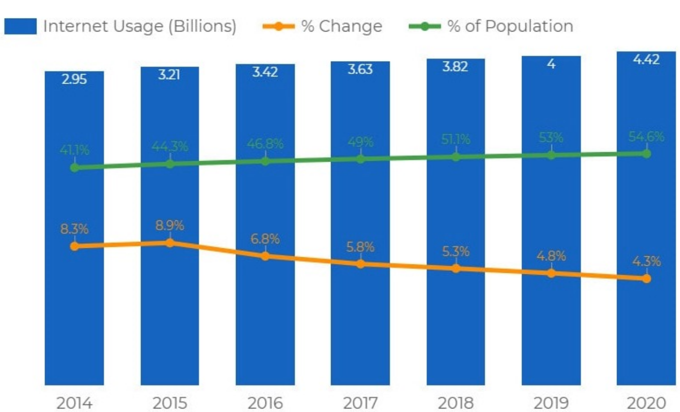
- 1st place: platforms for **publication/sharing of texts** (mostly), newsletters, podcasts, videos, photos,
  - Wikipedia, Blogger, Google Poadcast, youtube, Flickr, TripAdvisor, ...
- 2nd place: platforms for **messaging**
  - Facebook, Messenger, telegram,...
- 3rd place: platforms for **conversations**
  - Quora, StackExchange, Reddit, Facebook groups, Google Groups, ...
- 4th place: platforms for **collaboration**
  - Facebook workplace, TeamWork, Chatter, ...

Social media landscape 2021

image credit https://fredcavazza.net/2021/05/06/panorama-des-medias-sociaux-2021/

Some statistics 2020-2021: information and users

- Google processes in 2020 more than **7 milliards of queries every day** among which 15% have never been submitted before (new queries)
- The number of users in the world is estimated as **2.77 milliards on social media**, 2.46 milliards in 2017
- **51%, or more than 240 milliards of dollars, de tout l'argent publicitaire** dépensé dans le monde en 2019 seront basés sur les médias numériques.
- Les ventes en ligne devraient atteindre **3.45 billions de dollars** de ventes en 2020
- 47.3% de la population mondiale devrait



Statitistics on usage of information access systems 2014-2020

https://datastudio.google.com/embed/reporting/1sImC_rjeWqNXdgQt5MtmrQMbH44qFjtA/page/1fzh

- Users and information shared in live 2021

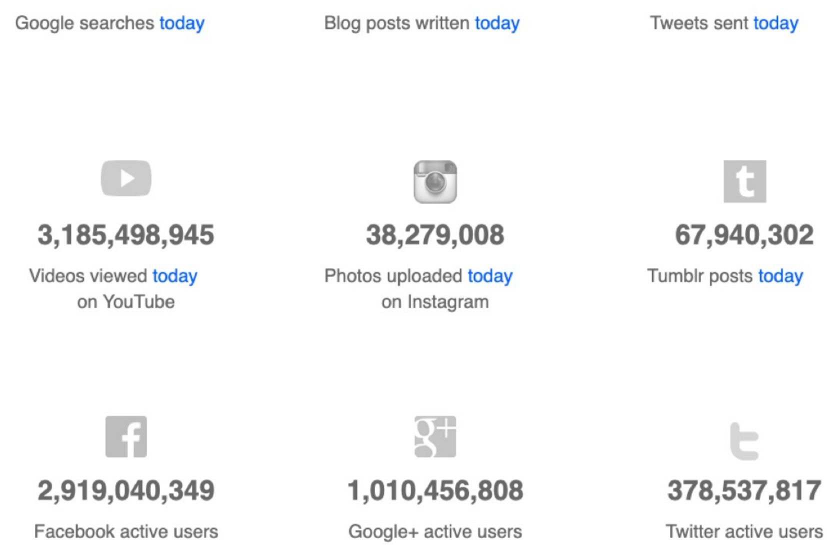


image credit https://www.internetlivestats.com/

What makes information retrieval challenging?

- Information needs are ambiguous
  - Queries are generally short, ambiguous
  - The matching between queries and intents is M-N

Roi lion

?

1 Query→ N intents

- Master UPS Intelligence artificielle
- Université paul Sabatier IA
- Formation IA Toulouse
- Matsre IAFA

..

M Queries→ 1 intent

Parcours mention Informatique

Intelligence Artificielle : Fondements et Applications (IAFA)

L'Intelligence Artificielle révolutionne de nombreux domaines de l'informatique, comme l'imagerie numérique (traitement, analyse et reconnaissance d'images, vision par ordinateur), l'informatique graphique (analyse d'acquisitions 3D, synthèse d'images 2D et 3D), le traitement automatique des langues (traduction, *chatbots*…), la recherche d'informations (moteurs de recherche, réseaux sociaux), les jeux vidéo, la robotique, le commerce électronique (recommandation, configuration de produits), le traitement de données massives… Les points communs entre ces différentes problématiques constituent les deux piliers de l'IA moderne : l'apprentissage automatique (notamment les réseaux de neurones) d'une part, et la représentation de connaissances et le raisonnement d'autre part.

What makes information retrieval challenging ?

- Relevance is subjective
  - Relevance is subjective
    - ✓ User-dependent
    - ✓ Situation-dependant
    - ✓ Topicality is often the **threshold relevance**

- Relevance faces vocabulary mismatch between queries and documents
  - Matching as word overlap: is it really semantic overlap?

    Q: "most jurisdictions exercise a high degree of regulation over **banks**" **[financial institution]**
    D_1: "I have been stolen when I withdrew the money from the **bank**" **[Building]**
    D_2: "fish lined the **bank** of the stream**" [The land alongside or sloping down to a river or lake]**

  - Matching is not exact, rough matching between queries and documents

    Q: "Presidential Elections in France"

    D_1 : "Election campaign is running"
    [relevant, but missing 'presidential' and 'France']

    D_2 : "Macron, the President of France is attending COP21"

    [irrelevant, and matching 'France' and 'President']

What makes information retrieval challenging ?

- Queries and documents vary in length
  - Models must handle variable length input
  - Relevant documents have irrelevant content

Q: "variant Omicron symptomes"

D: "Le variant Omicron a déjà atteint plusieurs patients en France après avoir fait son apparition en Afrique du Sud. S'il semble plus transmissible, il ne serait pas plus virulent. Mais quels sont ses symptômes ?
Le 26 novembre dernier, l'Organisation mondiale de la Santé (OMS) qualifiait le variant Omicron, nouvellement apparu en Afrique du Sud, de « préoccupant » sur la base de sa rapidité de propagation. De nombreux cas commencent depuis à émerger à travers le monde, dont quelques-uns en France.
Mais concernant sa dangerosité ou ses symptômes, le grand flou règne. Alors, que savons-nous ?
En se basant sur les situations en Afrique du Sud et au Royaume-Unis, l'OMS a indiqué dans une mise au point technique que le variant Omicron semble se propager plus vite que Delta.
Néanmoins, contrairement à ce dernier, les symptômes seraient moins sévères.
Pas de perte de goût ou d'odorat
Interrogée par la BBC, le Dr Angelique Coetzee, présidente de l'Association médicale sud-africaine, qui fut l'une des premières à être confrontée à Omicron, a indiqué que les symptômes qu'elle a pu observer semblent moins spécifiques que ceux de la maladie originelle. « Cela a débuté avec un patient de sexe masculin âgé d'environ 33 ans », a-t-elle expliqué lors de cet entretien.
« Il a déclaré qu'il était **extrêmement fatigué ces derniers jours et se plaignait de courbatures et de légers maux de tête.** » Mais **l'homme n'a pas perdu son sens du goût ni celui de l'odorat ; il avait la « gorge qui le grattait », et non pas un mal de gorge et une toux comme avec les variants précédents**.
Elle a également déclaré que les autres patients auscultés le même jour « présentaient les mêmes symptômes bénins ".

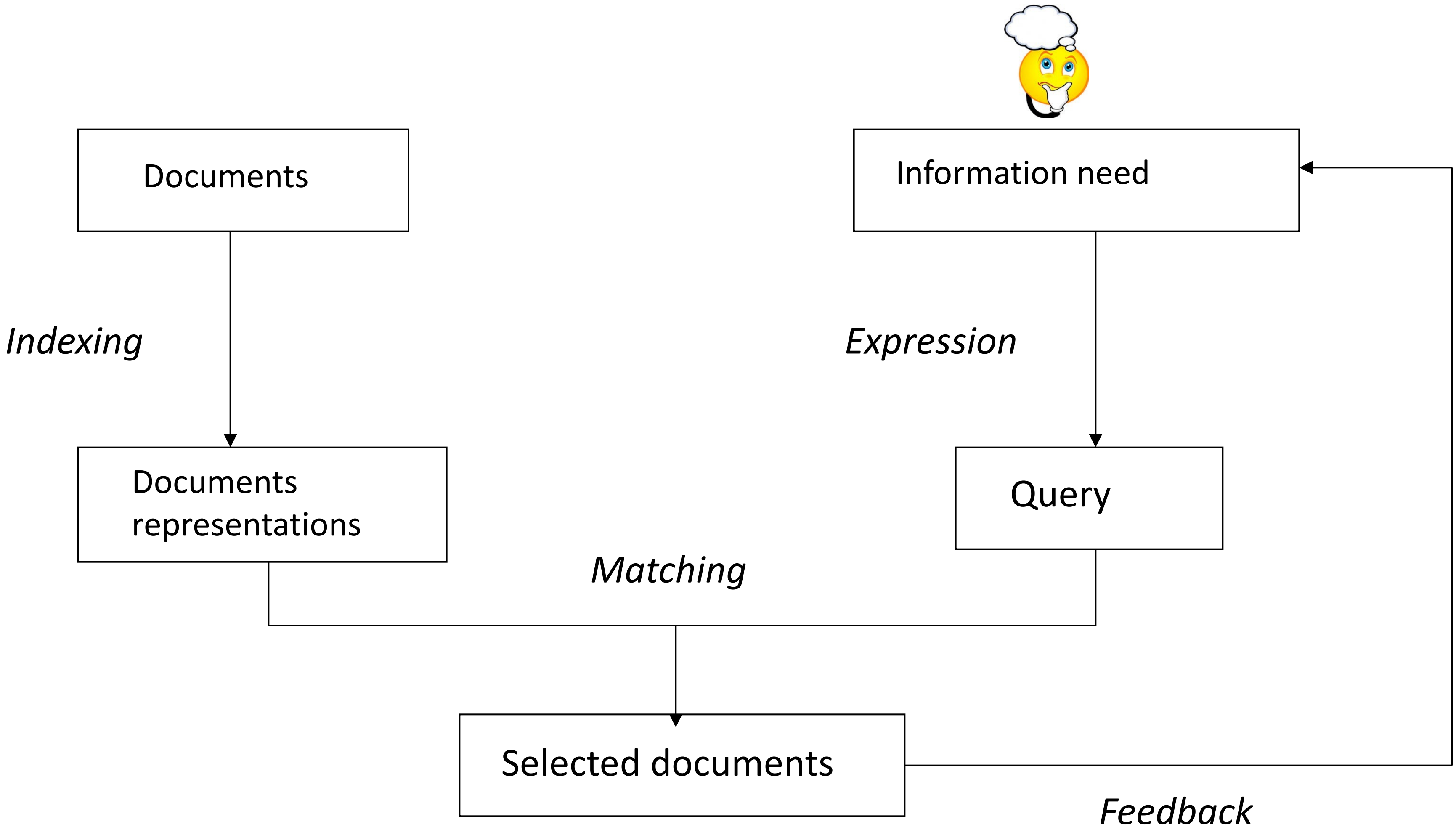
Source: https://www.leprogres.fr/magazine-sante/2021/12/13/variant-omicron-quels-sont-les-premiers-symptomes-detectes

What makes information retrieval similar vs. different from data retrieval (Databases)?

	Information retrieval	**Data retrieval**
Information unit	Information	Data (attribute-value)
Query	Vague expression of an information need	Vague expressio
Language of the query	Natural language	Formel language
Matching query-information	Approximatif	Exact
Selected information	Information relevant to the query	All the data that satifies the query

© L. Tamine-Lechani

The basic process of information retrieval



- Lecture structure
  - Introduction
  - Chapter 1: Text indexing and representation

    "*How to transform raw texts into machinable representations?*

    Keywords: indexation, words, documents, representation learning of texts
  - Chapter 2: Information retrieval (IR) models

    "How to score the relevance of a document as an answer to a user's query?"

    Keywords: relevance status value, retrieval model
  - Chapter 3: Performance evaluation of an IR system

    "How to measure the performance of an information retrieval system?"

    Keywords: evaluation metrics, test collections
  - Chapter 4: From question-answering systems to chatbots

    "How to interact with systems while searching for information?"

    Keywords: conversation, turn, clarification