Advanced Human Machine Interaction

Personalised information retrieval over the Internet

Alexandre Pauchet

alexandre.pauchet@insa-rouen.fr - BO.B.RC18

Normandie Université

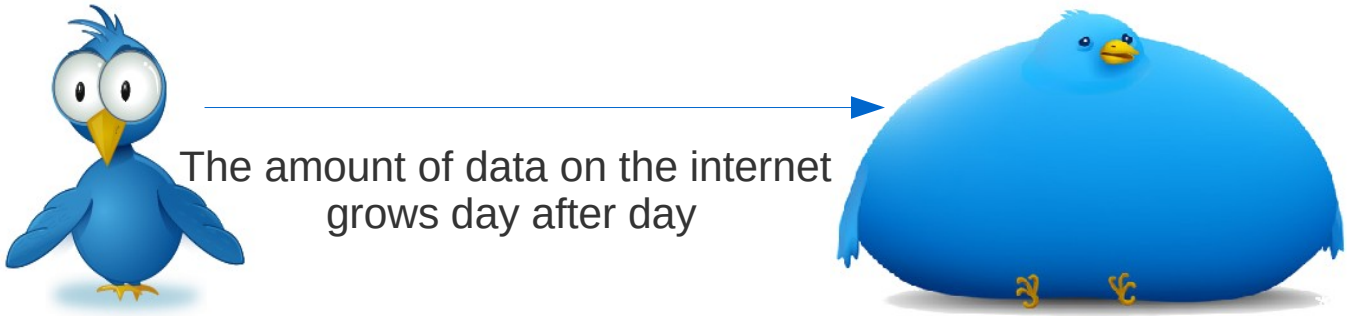
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

litis

Definitions

- **Information Retrieval (IR)**: finding documents relevant to a user query from a huge database.
- **Document**: text, piece of text, webpage, image, video... In IR, a document corresponds to any item that can answer a user query.
- **Query**: formalisation of a user's need of information. Often, it is a conjunction of keywords.
- **Relevance**: metric evaluating the match between a (found) document and a user query.

Internet as an information source



- Content in **constant growth**
- **Heterogeneous** content (texts, videos, pictures...)
- **Dynamic** content (new pages, dynamic pages : blogs/news/forums…)
- **Few structured content** (metadata, hypermedias, tagging semantics…), insufficiently exploited

Internet = Open content

- **Free and simple posting**: unmonitored content (personal webpages, blogs, wiki, forums)
  - Fake information
  - Verified information then modified
  - Information validated by stakeholders
  - Popularity does not mean truth

=> Information relevance and information validity must be evaluated depending on a need

Information retrieval over the Internet

- User satisfaction difficult to measure

  => Exploitation of relevance metrics
- Page “qualities” being very different, relevance should depend on it
- **Quality of research user interface depends on**
  - Speed
  - Index size
  - Error robustness (approximations, bad wording, ambiguities...)
  - Offered "services" (e.g. Google...)

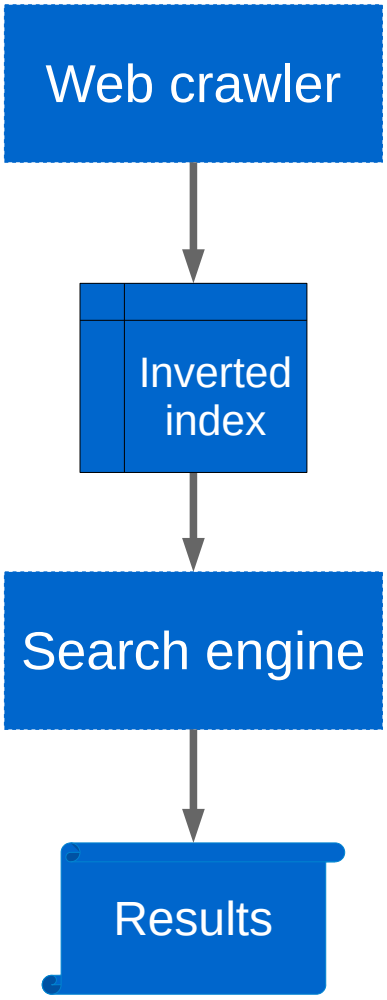
History of IR over the Internet

- **First generation** (~ 1995 : Altavista, Excite, Lycos, etc.) :
  - Indexing from information on the page
- **Second generation** (~ 1998 : Google and others) :
  - Indexing from web structure (entry links, text of links, etc.)
- **Third generation** (~2023) :
  - Answer user's need
  - Semantic and context-specific analysis
  - Help to user: HMI, several languages, autocompletion and spell checking, suggestion of queries...
- **Fourth generation** (ongoing...) :
  - Retrieval-Augmented Generation (RAG)

HMI and IR

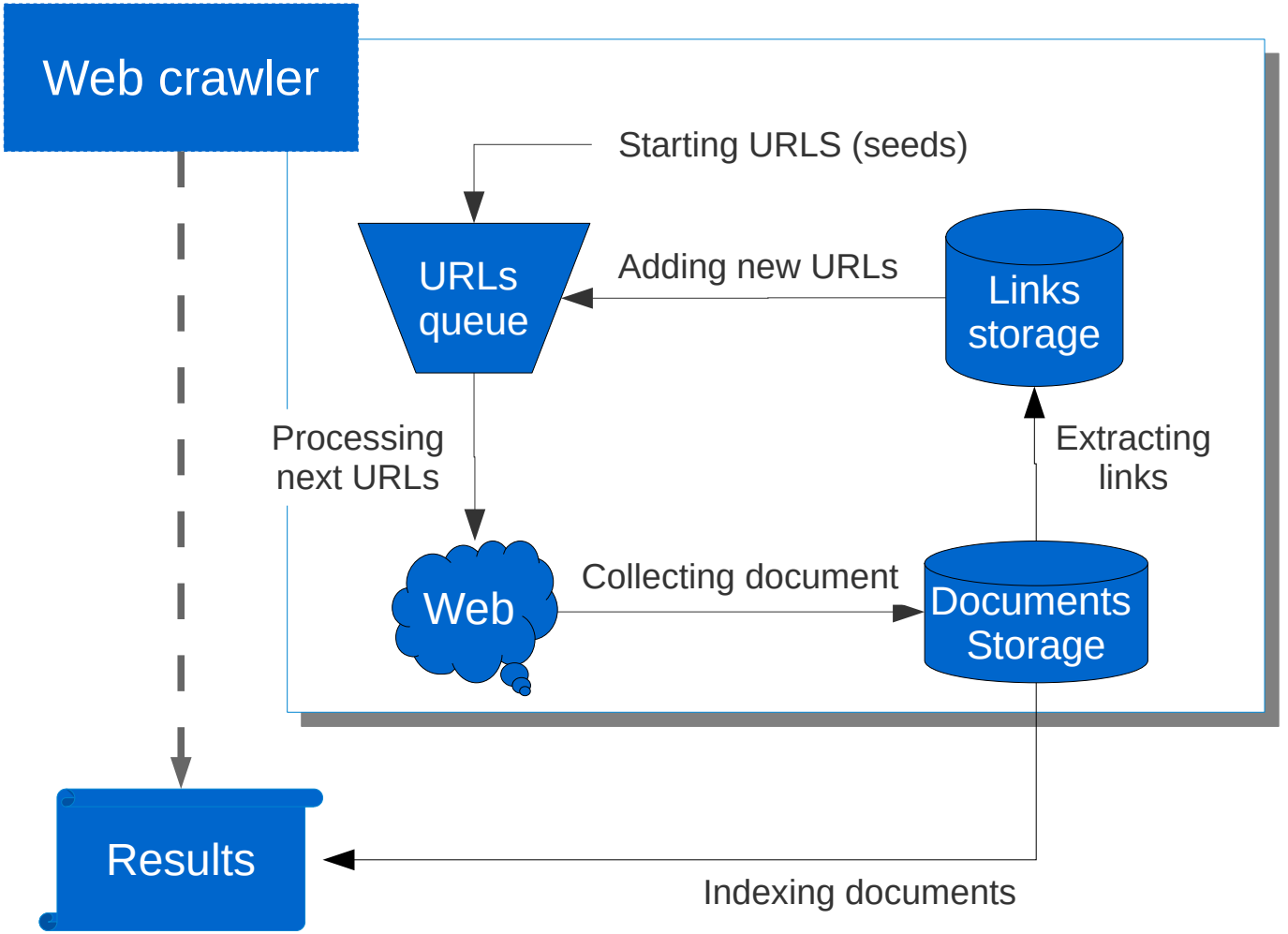
- Representation of documents (index)
- Representation of user’s needs
  - Long term needs (profile, topics of interest)
  - Short term needs (context-specific search)
- User interaction ↔ "Poor" IR systems
  - Little or no access to context
  - Limited query (keywords)
  - Document index not personalized
  - Relevance feedback difficult to catch

IR: basic approach



- Crawlers browse websites and index their content
  - Frequency depends on search engines
  - A page can ask to not be indexed (robots.txt), or to not follow links ("nofollow")
- Index is a summary at time T of webpages
- Results of a search engine are ordered by relevance towards a user request

Web crawler


Web crawler
Starting URLS (seeds)
URLs queue
Adding new URLs
Links storage
Processing next URLs
Extracting links
Web
Collecting document
Documents Storage
Results
Indexing documents

Inverted index

- **Definition**: index data structure used to store a set of documents or elements, mapping them from their content such as words or numbers.

- **Examples**:

- "crepes" | {flour, egg, milk}
- "genoise" | {egg, sugar, flour}
- "caramel" | {sugar, butter}
- "custard" | {egg, milk, sugar}

- "flour" | {crepes, genoise}
- "egg" | {crepes, genoise, custard}
- "milk" | {crepes, custard}
- "sugar" | {genoise, caramel, custard}
- "butter" | {caramel}

Status of IR over the Web

Representation of document:
Terms or groups of terms
[Salton & Yang, 1973]

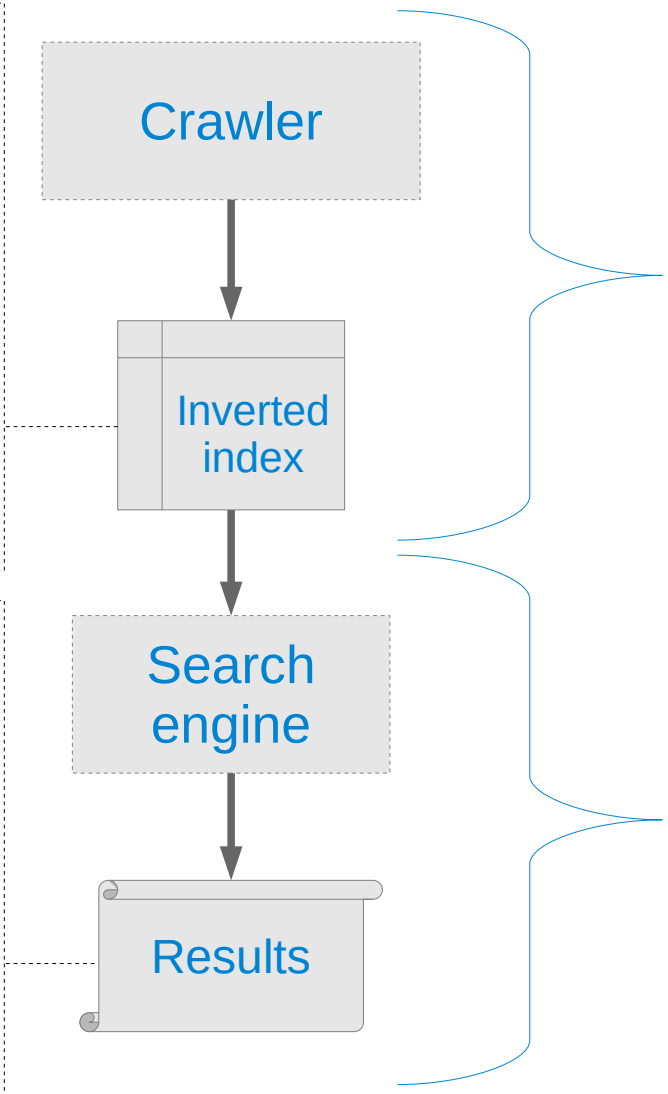
TF*IDF
[Salton & McGill, 1983]

Automated or semi-automated
[Desmontils & Jacquin, 2002]

Google, Bing & Yahoo !

Wording, representation, matching
[Maisonnasse, 2008]

Boolean, vectorial, probabilistic models
[Salton, 1969], [Salton, 1971], [Nottelmann & Fuhr, 2003]



A crawler browses and collects webpages which are then indexed

A search engine provides indexed webpages answering a given query

Boolean model

- First IR model, based upon set theory.
- A document is represented as a set of terms
  - Example: d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)
- Query = set of words and boolean operators ($\wedge$, $\vee$, $\neg$)
  - Example: q = t1 $\wedge$ (t2 $\vee$ $\neg$t3)
- **Exact** document/query matching relies on the presence or absence of query terms in the document
  - Example: Match(q,d1)=1 ; Match(q,d2)=0
- Drawbacks:
  - Selection of a document is a binary decision
  - Selected documents are not sorted
  - Query formulation is difficult for numerous users
  - Size problem: a lot of documents are returned

Vectorial model

- Proposed by Salton in the SMART system (Salton G., 1970).
- Working principle: documents and queries are represented as vectors in the space of document collection terms.
  - Document *j* : $d_j = (w_{1j}, w_{2j}, ..., w_{Mj})$
  - Query: $q = (w_{1q}, w_{2q}, ..., w_{Mq})$
  - With w_{ij}: weight of term t_i in document d_j (e.g.: tf*idf)
- A collection of n different documents and M different terms can be represented as a n*M matrix, and the query as a vector
- Relevance is a similarity measure between vectors (e.g.: cosine similarity)
- Pros:
  - Weighting enhances search results
  - The similarity measure allows to sort the documents by relevance
- Cons:
  - Do not take into account the word order (bag of words)
  - Vectorial representation necessitates term independance

Model which leads to RAG !

Limitations of IR on the web

Jaguar

Queries ambiguity
Search context

Cat

Crawler

Inverted index

Search engine

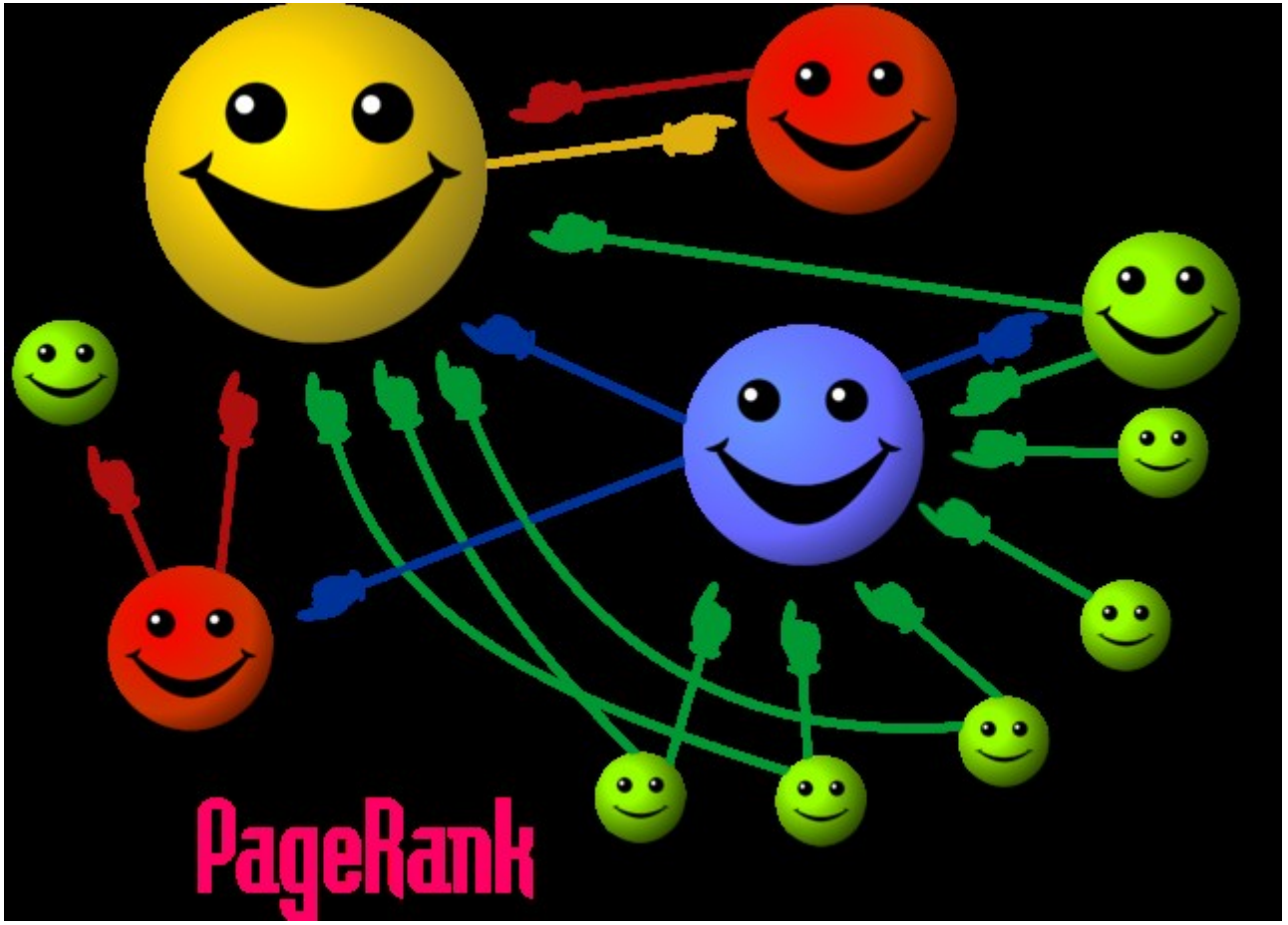
Results

Document corpus whose size comes close to the web one

Indexing

- Analysis (parsing) of pages/documents
  - Depends on the type of document
  - Conversion into ***Hits*** = weighting according to:
    - Number of appearances of each word
    - Position in the document
    - Relative font size and word letter case
- Only the surface web is indexed
  - Deep web: not-linked-to websites, restricted access content, content avoiding indexing, dynamic webpages
  - Deep web = 500 times bigger than surface web according to BrightPlanet, 2001

Google: PageRank (Page et al., 1998)



http://en.wikipedia.org/wiki/File:PageRank-hi-res.png

A webpage is important if linked to other important webpages

PageRank ~ probability to access a webpage by pure chance

PageRank formula

3 factors determine PageRank of a webpage P:

- Number of links pointing to P
- Number of links contained into webpages that have a link to P
- PageRank of pages containing a link to P

- PageRank is the summation of PageRanks of pages containing a link to Pi, weighted by the total number of output links:

$$r(P_i) = \sum_{P_j \in \mathscr{B}_{P_i}} \frac{r(P_j)}{|P_j|}$$

- Calculation is iterative

IR at Google's

Indexing documents

- Inverted index (word(s) → document(s))
- Relevance of words weighted by their type (title, anchor, URL, font…) and their position into the webpage
- Equivalent to TF*IDF

- Query/document similarity measure (dot product of the 2 vectors)
- Combination of IR score with PageRank in order to sort the results

Examples of relevance measure

Measure relative to terms (cosine similarity)

0 or 1

Herbal — 0 — Herbalism

Herbalism — 1 — Herbalism

Conceptual measure

Between 0 and 1

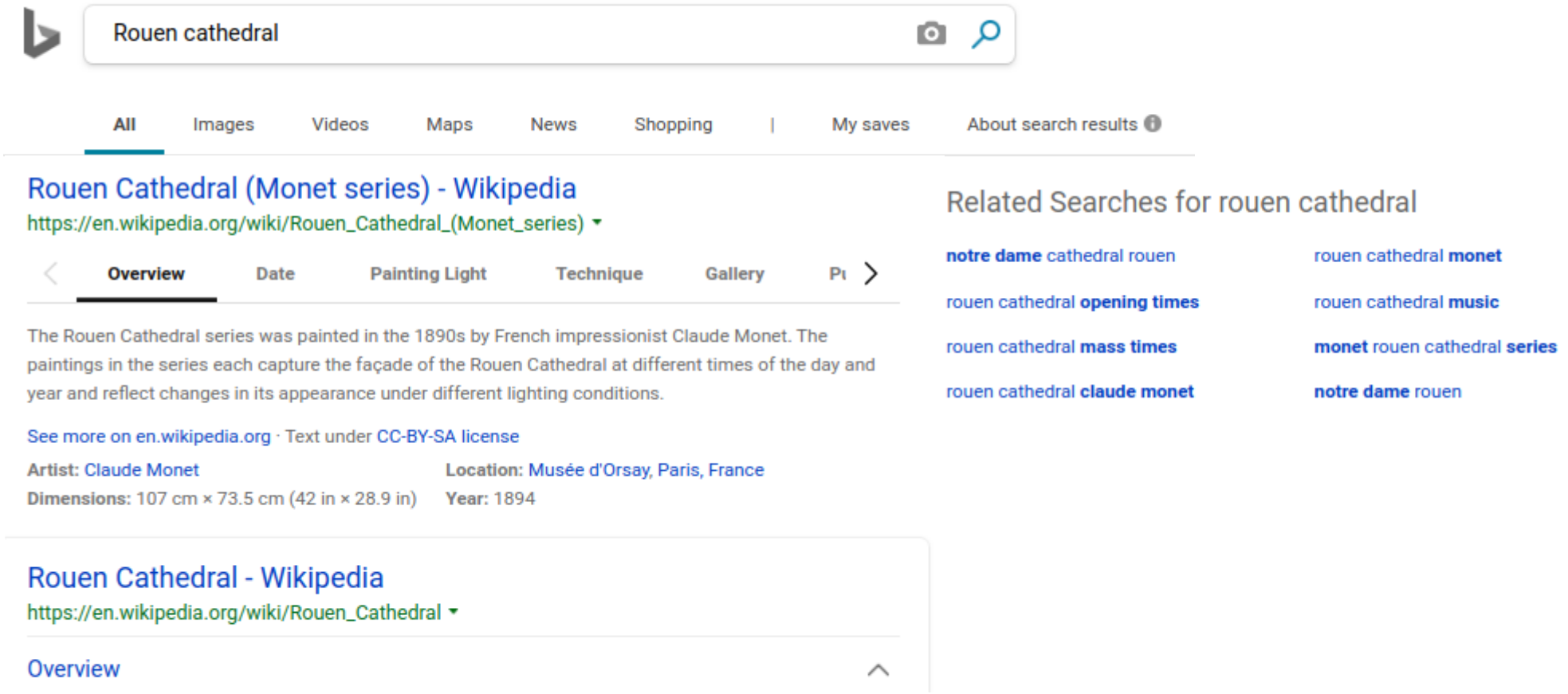
Herbal_tonic — 0,33 — Herbalism

Category:Herbalism

Category:Dietary_supplements
Category:Herbalism
Category:Biologically_based_therapies

Basic HMI

- Minimalistic HMI: 1 text field, 1 result page (snippet) including extracts
- Advanced search is usually proposed

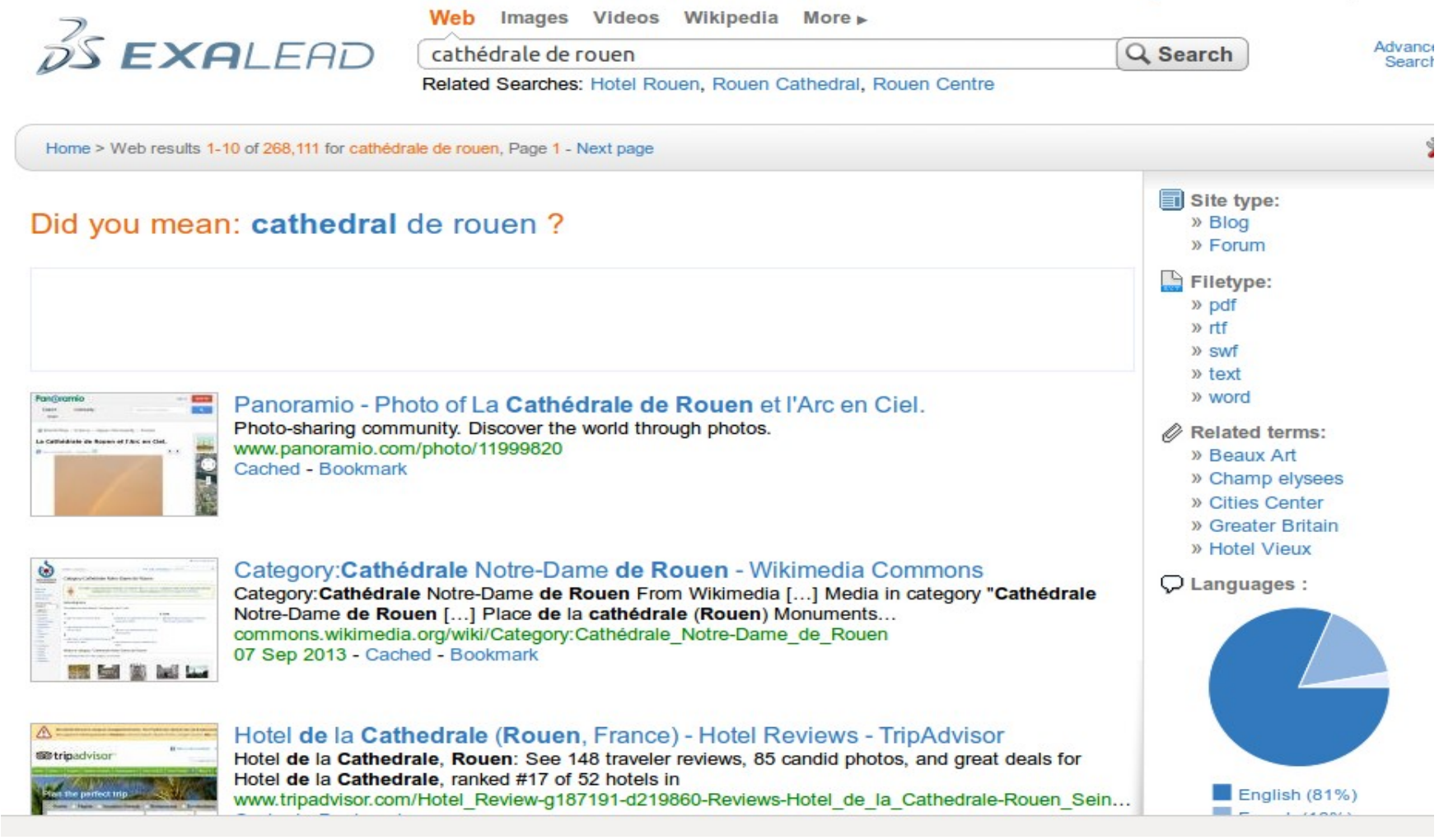


HMI: input help

- Spell check
- Automatic completion proposing the most searched queries

HMI: result analysis help

- Preview of results
- Sponsored links
- Terms and directories
- Language choice

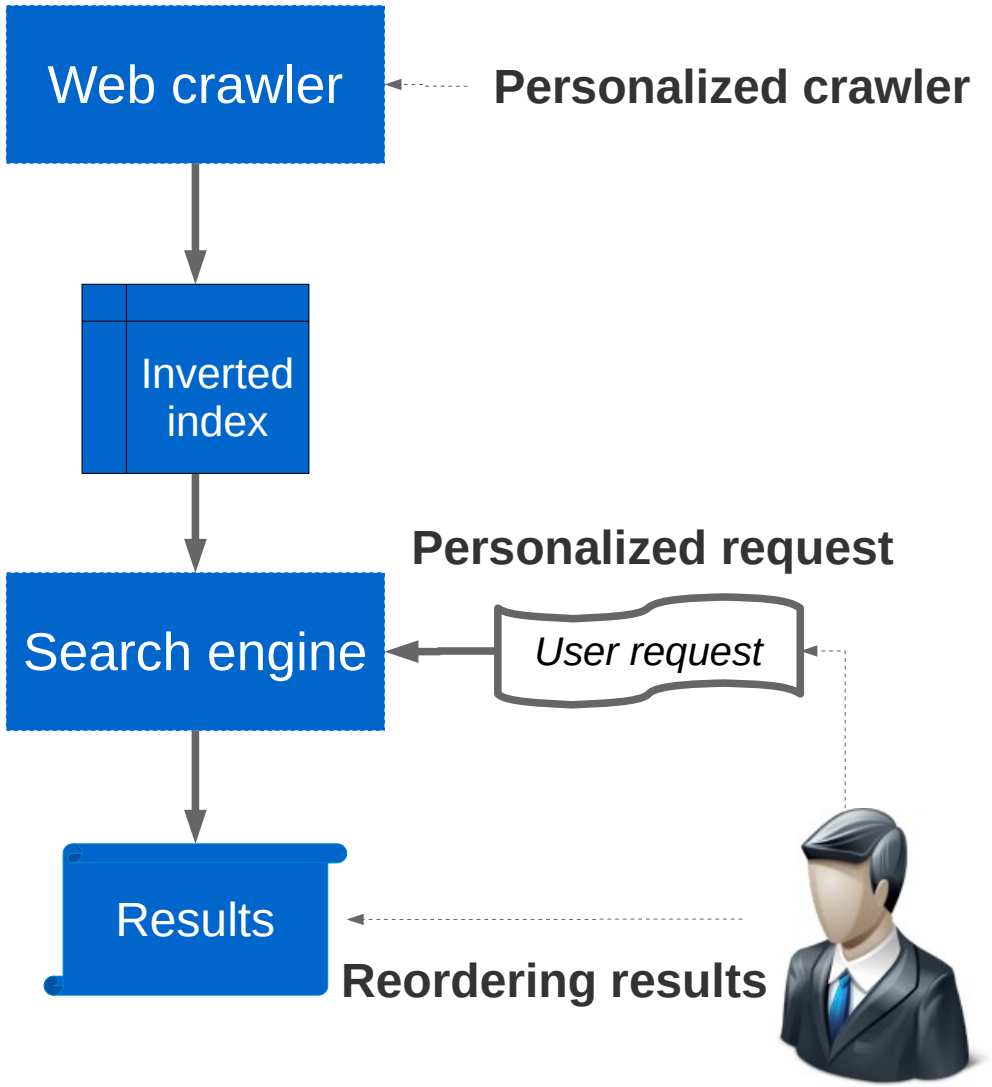


Drawbacks of classic approaches

- PageRank problems
  - Google: 82% of queries in France
  - A webpage is well referenced if popular ; a webpage is popular if well referenced
  - Difficult appearance of new pages
  - SEO is very lucrative
- Content quality is not taken into account
- Ambiguities? Synonymies?
- Search context?
- Generalization (concepts) ?

⇒ **Personalization!**

Personalization: founding approaches



Statistical model

C. Aggarwal, F. Al-Garawi, and P. Yu. Intelligent crawling on the world wide web with arbitrary predicates. 2001.

Reinforcement learning

S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. 2002.

Multi-agent : genetic model

F. Menczer and R. Belew. Adaptive retrieval agents : Internalizing local context and scaling up to the web. 2000.

Multi-agent : biological model

F. Gasparetti and A. Micarelli. Swarm intelligence : Agents for adaptive web search. In ECAI, 2004.

System – Human interaction in IR

- Example: R="animal", D1="horse"
  - D1 will not be returned as it does not contain "animal"
  - How to return relevant documents even if they do not contain any of the query terms?

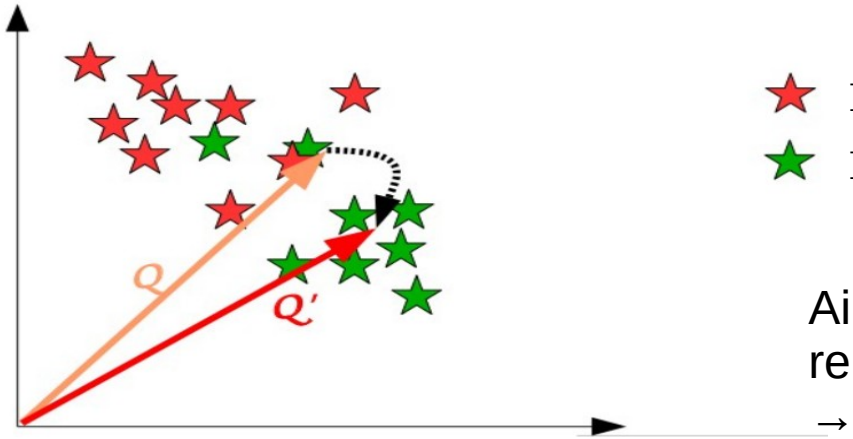
Methods improving recall:

- Query expansion
- Relevance feedback
- Query disambiguation: concepts, detection of synonyms, query validation through H-M "dialogue"...

Query expansion

- Query modification using external resources
  - Dictionaries, thesauri, ontologies…
  - Increase recall
  - Precision will lower consequently
- Examples:
  - Hospital → medical (thematic)
  - Interest rate → interesting (grammatical)
  - Shiny → bright (synonymy)
  - Horse → animal (generalization)
  - Animal → horse, dog, ... (specialization)

Relevance feedback



Irrelevant documents

Relevant documents

Aim: (implicitly) move query vector to bring it closer to relevant documents
→ Rocchio's formula

$$\vec{Q}' = \alpha \vec{Q} + \beta \vec{P} + \gamma \vec{N} \vec{P}$$

Mean of irrelevant documents vectors

Negative value (e.g.: -0.25)

Mean of relevant documents vectors

Positive value (e.g.: 0.5)

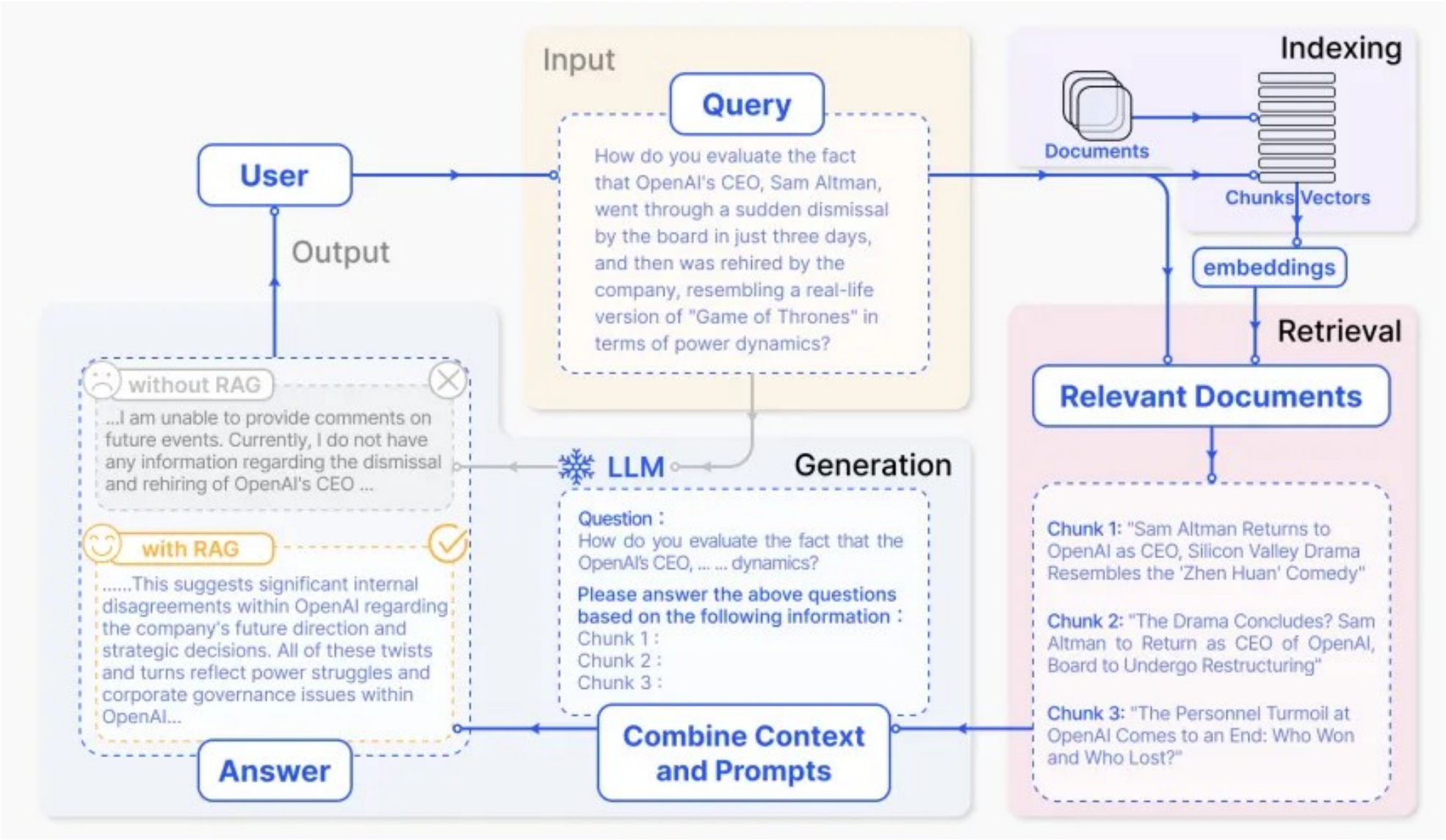
Initial query vector

Positive value greater than the others

New query vector

Translated from http://www2.ifi.auf.org/personnel/Alain.Boucher/cours/recherche_information/RI-03.pdf

Personalised RAG?



Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv

Hugely inspired by

Lectures

- Xavier Tannier’s lectures
  http://perso.limsi.fr/amax/enseignement/iri/M2PRO_IRI_6_RIWeb.pdf
- http://math.univ-lyon1.fr/homes-www/malbos/lib/exe/fetch.php?media=ens:algapp12:algapiichapiiisec9.pdf
- http://www.iro.umontreal.ca/~nie/IFT6255/Introduction.html
- https://www.irit.fr/~Mohand.Boughanem/Enseignements_RI.php