Query Processing

PRI 23/24 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Based on Chapters 9 from Introduction to Information Retrieval, Manning et al. (2008)
Based on Chapter 5 from Information Retrieval - Implementing and Evaluating Search Engines, Büttcher et al. (2010)
Based on Chapters 5, 6 and 7 from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Outline

- ➔ Recap
- ➔ Query Processing
  - ➔ Processing strategies, document and term evaluation
  - ➔ Query transformation
  - ➔ Query expansion
  - ➔ Relevance feedback
- ➔ Search User Interface

Recap

Ranking Components

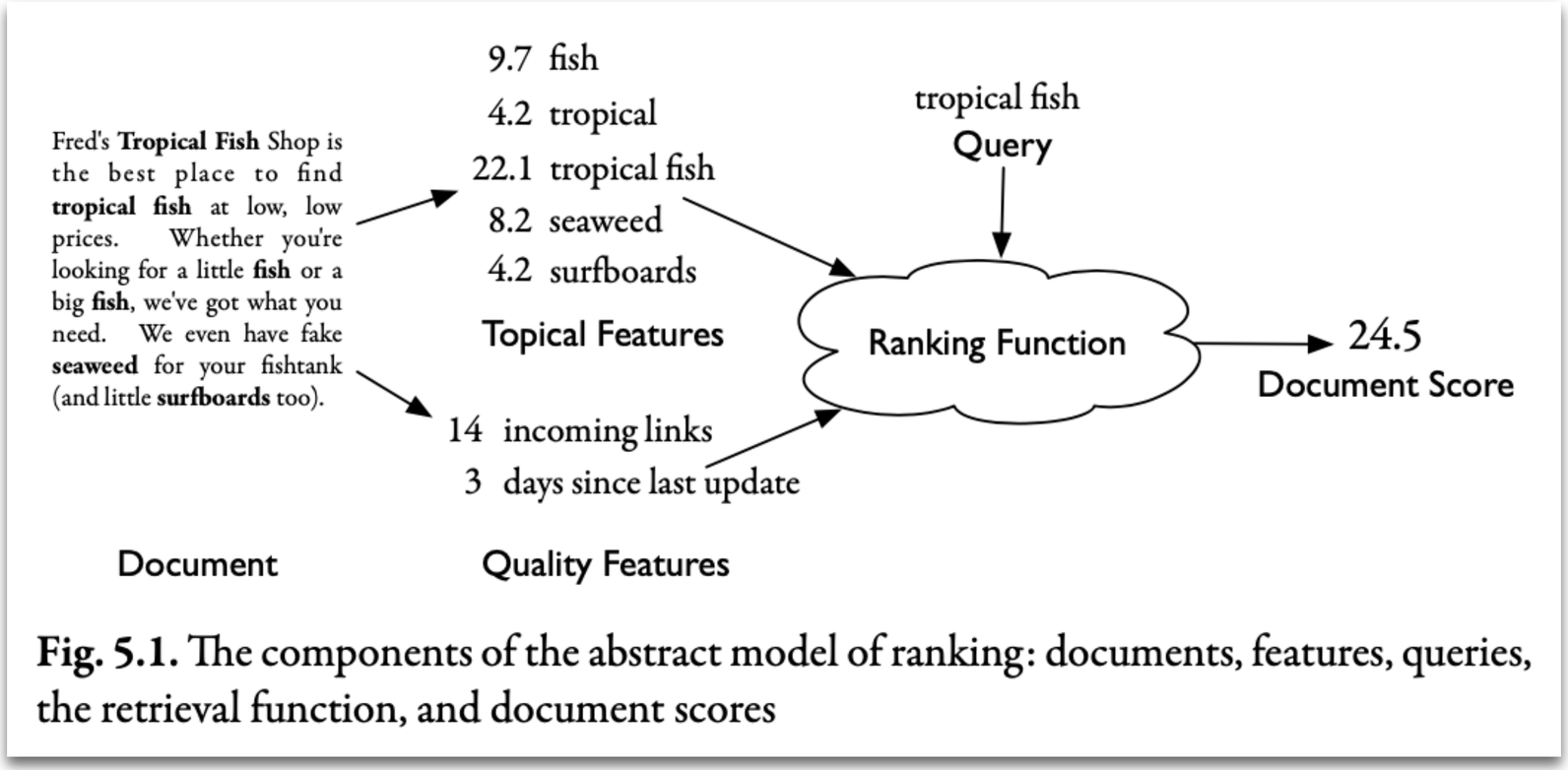


Fig. 5.1. The components of the abstract model of ranking: documents, features, queries, the retrieval function, and document scores

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Inverted Index (with counts)

S_1 Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.

S_2 Fishkeepers often use the term tropical fish to refer only those requiring fresh water, with saltwater tropical fish referred to as marine fish.

S_3 Tropical fish are popular aquarium fish, due to their often bright coloration.

S_4 In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Table 5.1. Four sentences from the Wikipedia entry for *tropical fish*

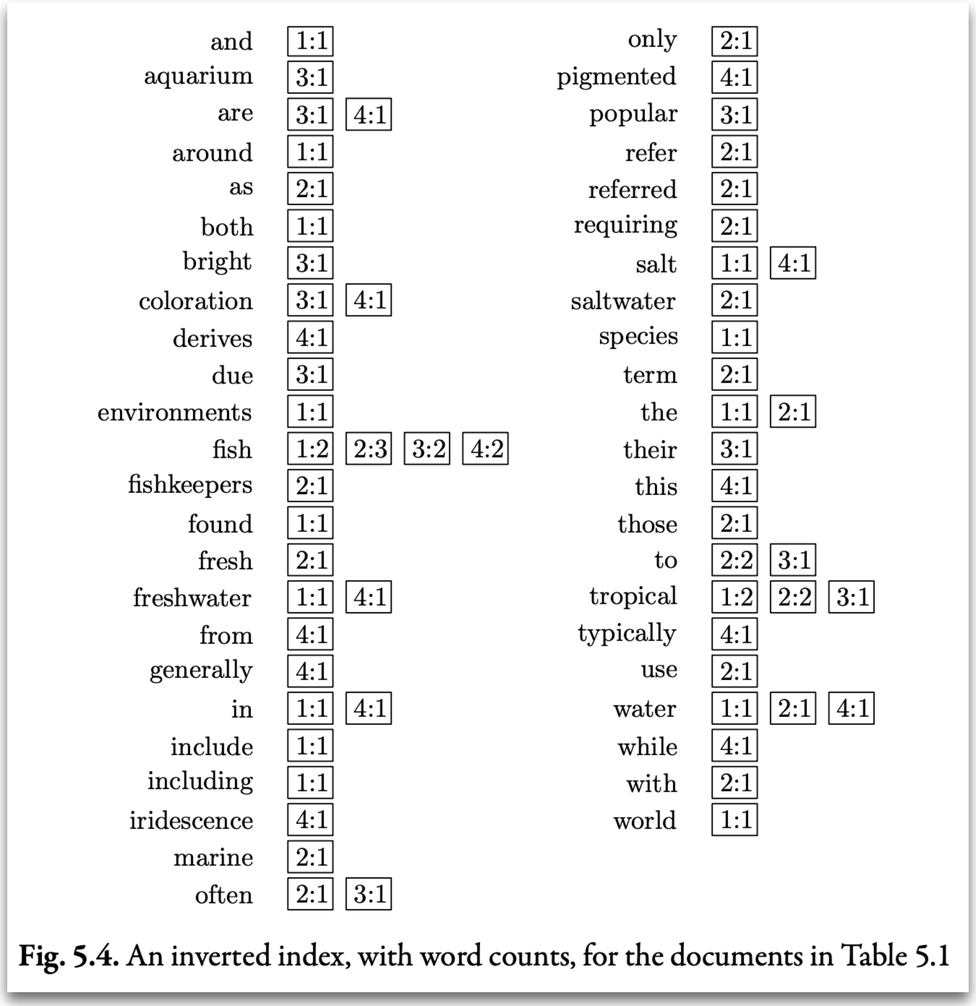


Fig. 5.4. An inverted index, with word counts, for the documents in Table 5.1

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Inverted Index (with positions)

S_1 Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.

S_2 Fishkeepers often use the term tropical fish to refer only those requiring fresh water, with saltwater tropical fish referred to as marine fish.

S_3 Tropical fish are popular aquarium fish, due to their often bright coloration.

S_4 In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Table 5.1. Four sentences from the Wikipedia entry for *tropical fish*

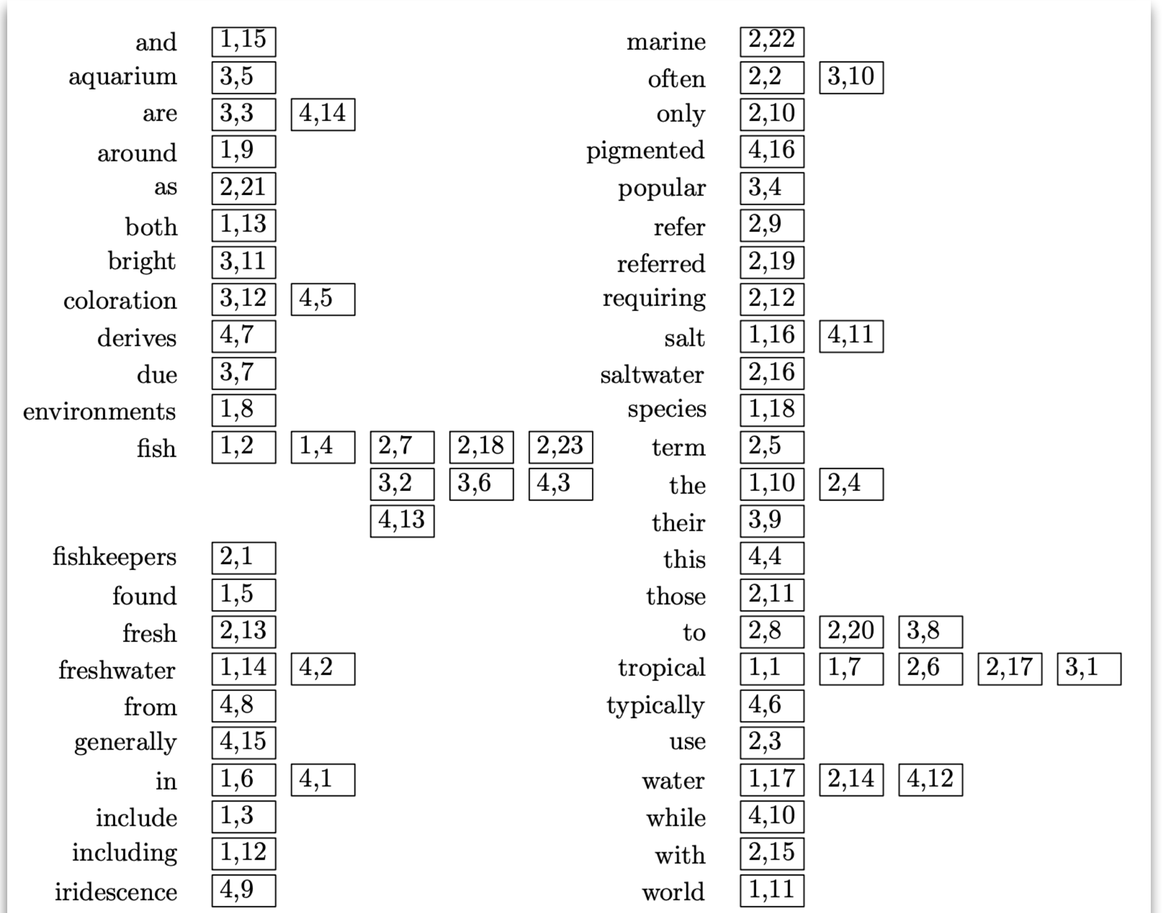


Fig. 5.5. An inverted index, with word positions, for the documents in Table 5.1

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Query Processing

Query Processing

- Once the necessary data structures are in place, we need to efficiently use them to obtain the search results in response to a user query.

- Two main query processing techniques:

  - **Document-at-a-time**, calculates complete scores for documents by processing all term lists, one document at a time. At the end all documents are sorted according to their score.

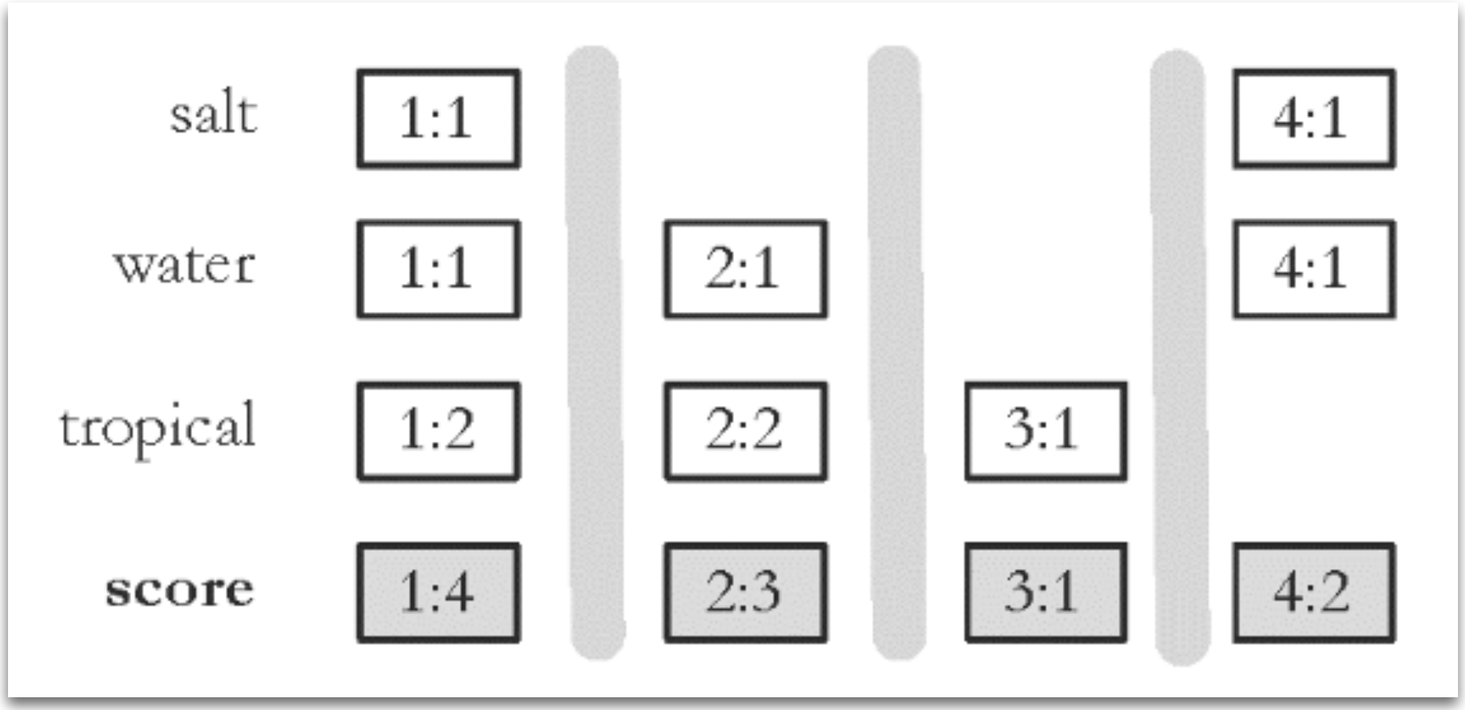
  - **Term-at-a-time**, accumulates scores for documents by processing term lists one at a time. When all terms are processed, the accumulators contain the final scores of all matching documents.

- In both approaches, optimization techniques can significantly reduce the time required.

Document-at-a-Time

➔ query = [salt water tropical]

inverted lists



score is the sum of words in each document

each column represents a different document

1st step: compute all counts for the 1st document

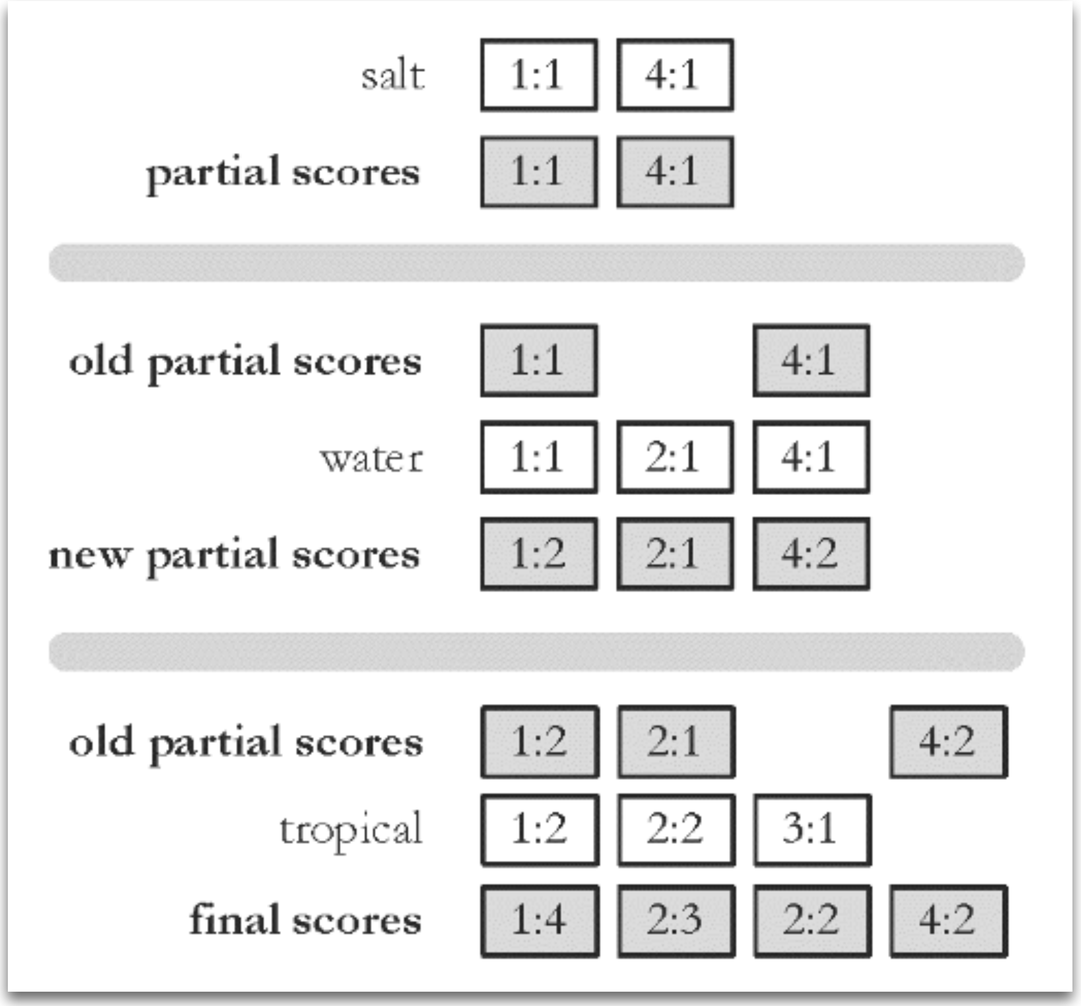
one document in each step

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Term-at-a-Time

➔ query = [salt water tropical]

accumulators keep track of partial scores



1st step: compute all counts for the 1st term [salt]

one term is processed in each step

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Optimization Techniques

- Two classes of optimization techniques for query processing:
  - Read less data from the index.
  - Process fewer documents.
- When using complex feature functions, focusing on scoring fewer documents is the main concert.
- With simple feature functions, performance gains come from ignoring as much of the inverted list data as possible.

Skip Pointers

➔ Skip pointers are use to speed-up inverted index scans.

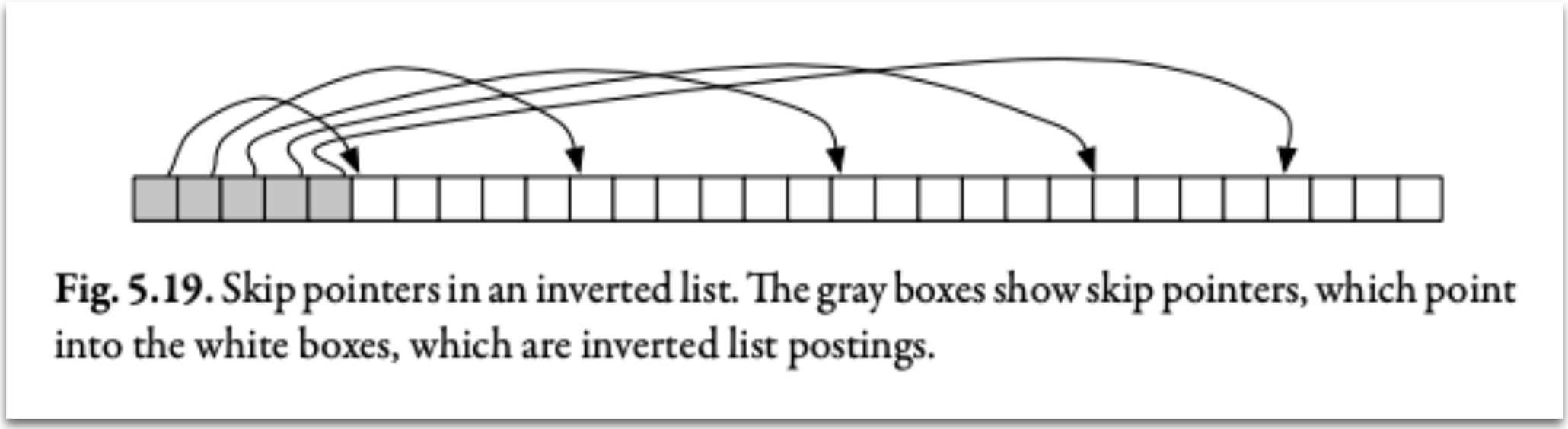


Fig. 5.19. Skip pointers in an inverted list. The gray boxes show skip pointers, which point into the white boxes, which are inverted list postings.

Image from Search Engines - Information Retrieval in Practice, Croft et al. (2015)

Skip Pointers Usage

➔ Skip pointers are use to speed-up inverted index scans,

➔ i.e., reading less data from inverted lists.

➔ In doc-at-a-time, looking for documents in postings benefits with skip pointers, i.e. more quickly get to the document being scored.

➔ In term-at-a-time, skip lists can be used in the term accumulators, and in conjunction with 'conjunctive processing' to skip over common terms.

Conjunctive Processing

- ➔ Base assumption: only return documents that contain all query terms.
- ➔ This is the default in web search engines and the default users' expectations.
- ➔ Conjunctive processing works best when one of the terms is rare.
  - ➔ [wine baga], since 'baga' is rarer, we can skip most of the inverted list for 'wine'.
- ➔ Can be employed for both document-at-a-time or query-at-a-time.
- ➔ In short queries, benefits both efficiency and effectiveness.
- ➔ Long queries, e.g. phrase searches, are not good candidates for conjunctive processing.

Other Optimization Techniques

- ➔ Early termination of query processing:
  - ➔ Ignore high-frequency word lists in term-at-a-time,
    common terms have long postings lists, thus high processing costs.
  - ➔ Ignore documents at end of lists in document-at-a-time,
    when documents are ordered by some quality metric.
- ➔ Order postings in inverted indexes
  - ➔ Order inverted lists by quality metric (e.g. number of IN links in web IR).
- ➔ Caching
  - ➔ Cache popular query results

Relevance Feedback and Query Expansion

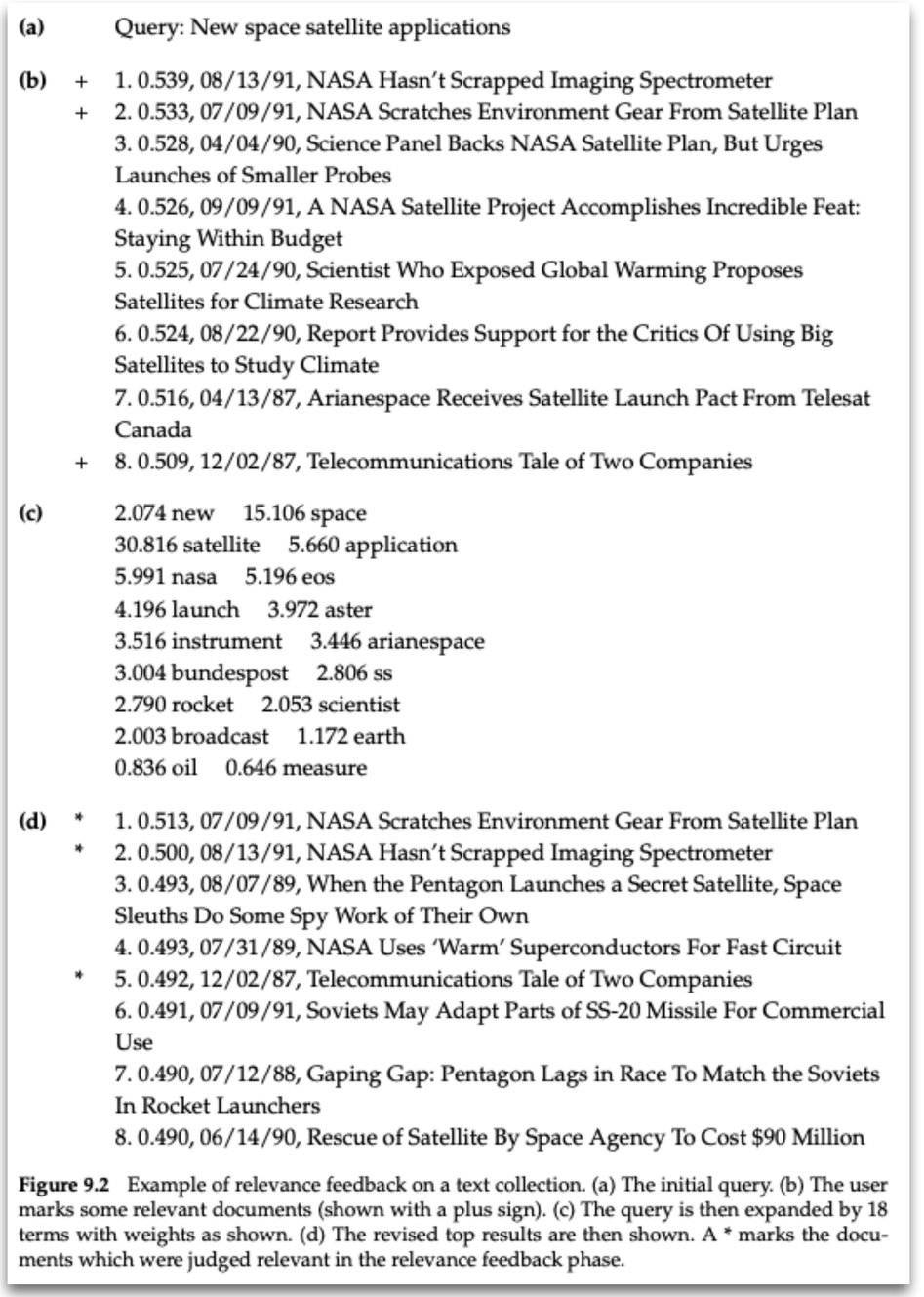
Relevance Feedback

- ➔ Exact matches aren't the only way to obtain relevant results in search systems.
- ➔ The vocabulary mismatch between the user and the collection contribute to this problem.
  Also, the fact that synonyms exist.
- ➔ For example, a search for [aircraft] should also include results for [airplane].
- ➔ This can be addressed by manually refining the query.
- ➔ On the system side, this can be tackled with different techniques, broadly grouped in:
  - ➔ Global methods, expand or reformulate the query terms independently of the query or the results returned from it, e.g. using thesaurus, and using spelling correction.
  - ➔ Local methods, adjust a query relative to the documents that initially appear to match the query, e.g. relevance feedback, and pseudo-relevance feedback.

Local Relevance Feedback Methods

Relevance Feedback

- The idea of relevance feedback is to involve the user in the process to improve the final result set by considering user feedback about the initial set of results.
- Basic procedure (one or more iterations are possible):
  - The user issues a (short, simple) query.
  - The system returns an initial set of retrieval results.
  - The user marks some returned documents as relevant or non relevant.
  - The system computes a better representation of the information need based on the user feedback.
  - The system displays a revised set of retrieved results.



(a) Query: New space satellite applications

(b)
+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

(c)
2.074 new 15.106 space
30.816 satellite 5.660 application
5.991 nasa 5.196 eos
4.196 launch 3.972 aster
3.516 instrument 3.446 arianespace
3.004 bundespost 2.806 ss
2.790 rocket 2.053 scientist
2.003 broadcast 1.172 earth
0.836 oil 0.646 measure

(d)
* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Figure 9.2 Example of relevance feedback on a text collection. (a) The initial query. (b) The user marks some relevant documents (shown with a plus sign). (c) The query is then expanded by 18 terms with weights as shown. (d) The revised top results are then shown. A * marks the documents which were judged relevant in the relevance feedback phase.

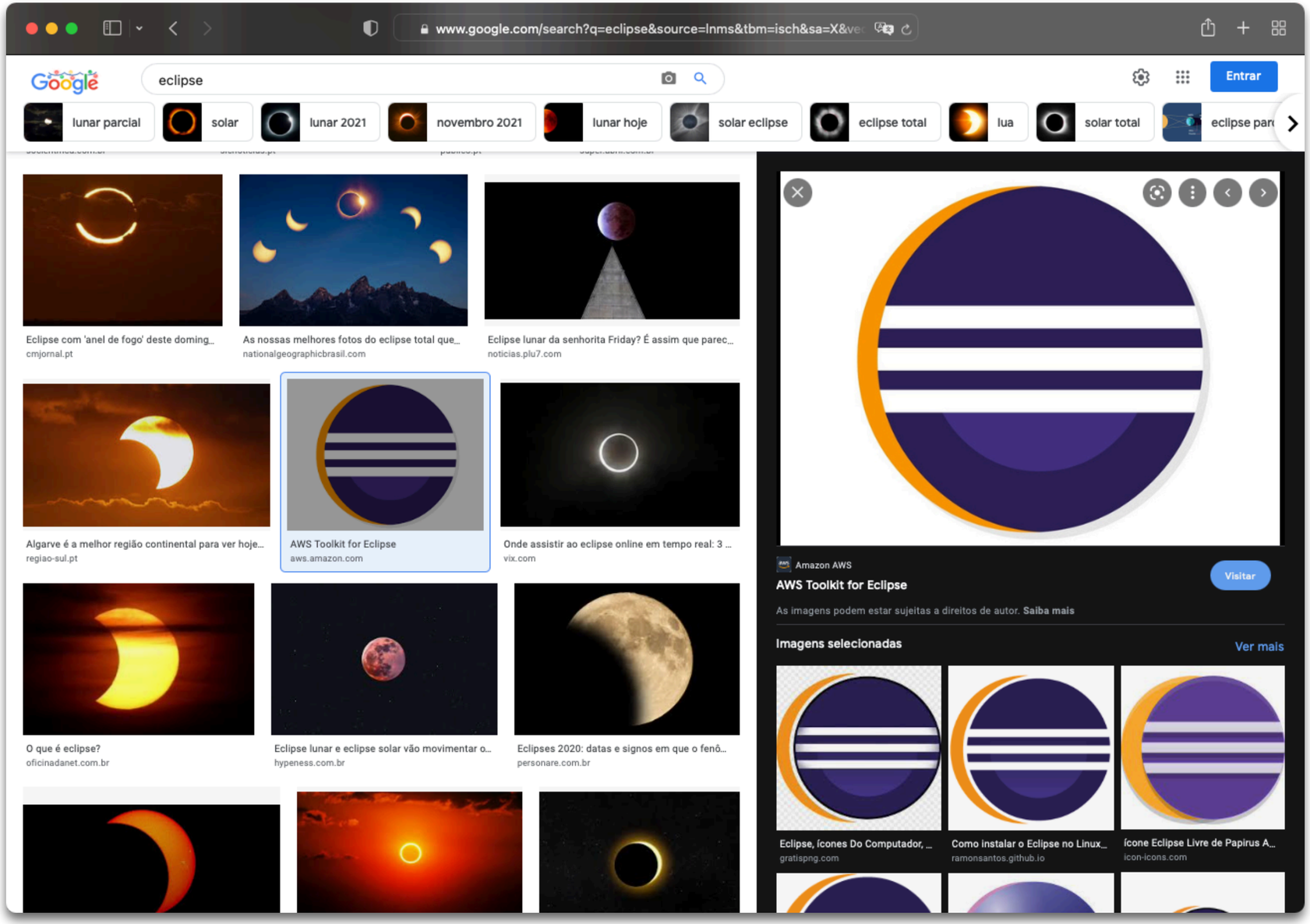
Image from Introduction to Information Retrieval, Manning et al. (2008)

Relevance Feedback

➔ Relevance feedback explores the idea that it may be difficult to formulate a good query when you don't known the collection, but it is easy to judge particular documents.

➔ Relevance feedback can be effective in tracking a user's evolving information need, i.e. seeing some documents may lead users to refine their understanding of the original information need.

➔ Image search is a good example of relevance feedback, a task where it may be difficult to express a query in words, but is easy to determine if an image is relevant.


www.google.com/search?q=eclipse&source=lnms&tbm=isch&sa=X&ve
Google
eclipse
Entrar
lunar parcial
solar
lunar 2021
novembro 2021
lunar hoje
solar eclipse
eclipse total
lua
solar total
Eclipse com 'anel de fogo' deste doming...
cmjornal.pt
As nossas melhores fotos do eclipse total que...
nationalgeographicbrasil.com
Eclipse lunar da senhorita Friday? É assim que parec...
noticias.plu7.com
Algarve é a melhor região continental para ver hoje...
regiao-sul.pt
AWS Toolkit for Eclipse
aws.amazon.com
Onde assistir ao eclipse online em tempo real: 3 ...
vix.com
O que é eclipse?
oficinadanet.com.br
Eclipse lunar e eclipse solar vão movimentar o...
hypeness.com.br
Eclipses 2020: datas e signos em que o fenô...
personare.com.br
Amazon AWS
AWS Toolkit for Eclipse
Visitar
As imagens podem estar sujeitas a direitos de autor. Saiba mais
Imagens selecionadas
Ver mais
Eclipse, ícones Do Computador, ...
gratispng.com
Como instalar o Eclipse no Linux...
ramonsantos.github.io
Ícone Eclipse Livre de Papirus A...
icon-icons.com

Rocchio Algorithm

➔ The Rocchio algorithm is the classic algorithm for implementing relevance feedback.

➔ It models a way of incorporating relevance feedback information into the vector space model.

➔ It is based on the concept of an optimal query vector, which maximizes the difference between the average vector representing the relevant documents, and the average vector representing the non-relevant documents.

➔ Considering that only partial relevance information is available about the collection, the Rocchio algorithm defines the "modified query" as a weighted combination of the initial query and the difference vector between the centroid of the documents marked as relevant and the centroid of the documents marked as non relevant.

Rocchio Optimal Query

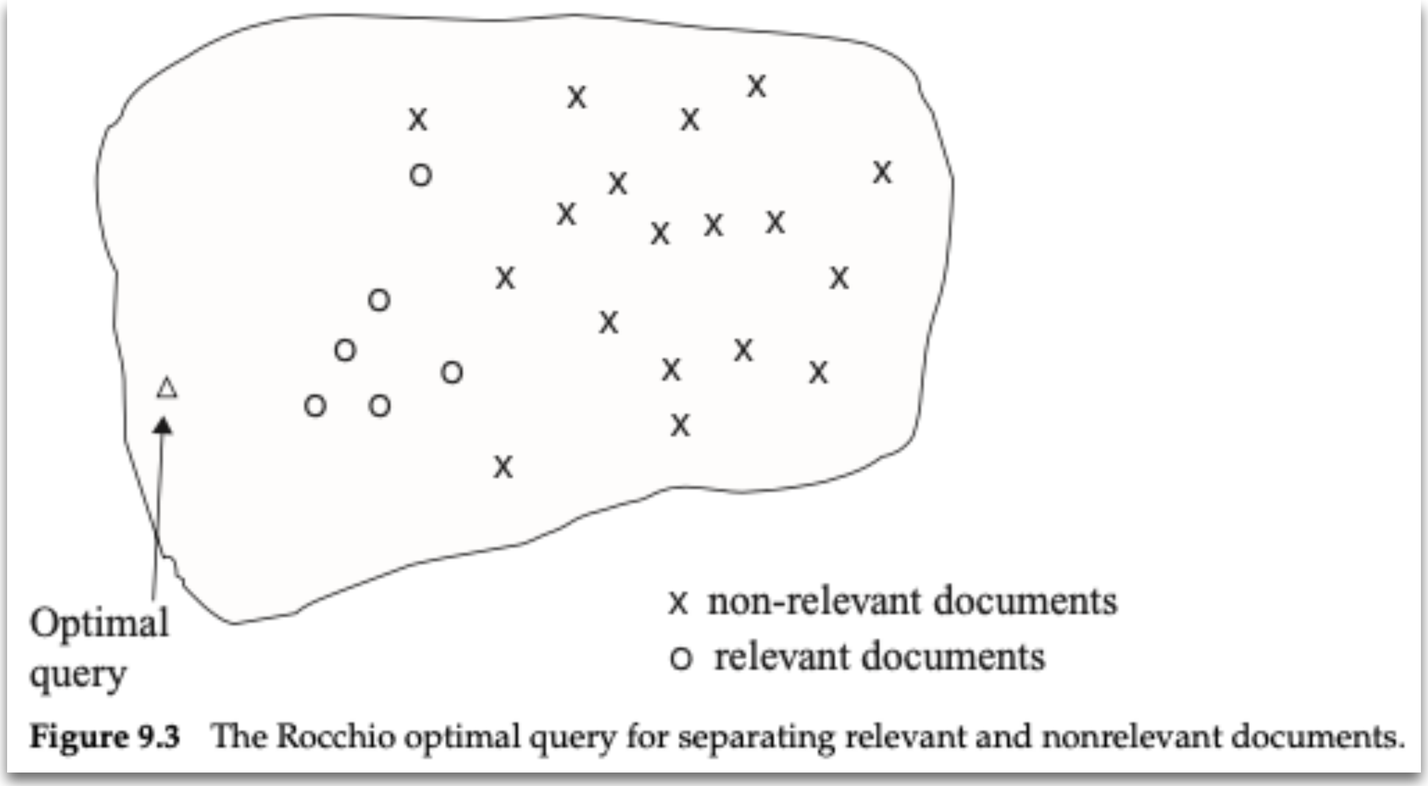


Figure 9.3 The Rocchio optimal query for separating relevant and nonrelevant documents.

Image from Introduction to Information Retrieval, Manning et al. (2008)

Rocchio Adjusted Query

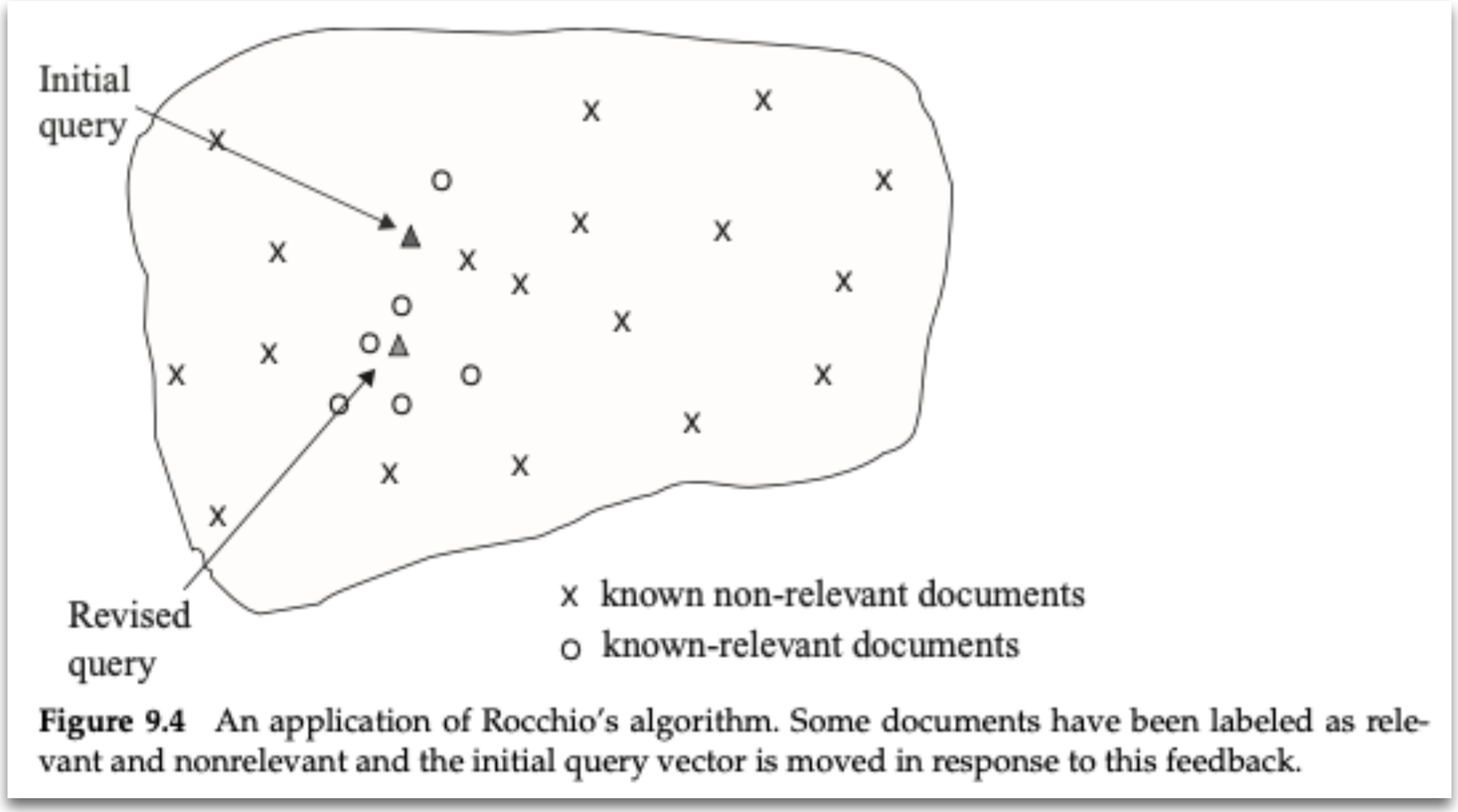


Figure 9.4 An application of Rocchio's algorithm. Some documents have been labeled as relevant and nonrelevant and the initial query vector is moved in response to this feedback.

Image from Introduction to Information Retrieval, Manning et al. (2008)

Rocchio Algorithm Performance

➔ Relevance feedback can improve both recall and precision.

➔ In practice it has been shown to be most useful for increasing recall — partially an effect of the use case: when a user wants high recall (i.e. see all relevant documents), they are expected to take the time to review results and iterate over the search.

➔ Positive feedback is more valuable than negative feedback. Many systems only allow for positive feedback.

Relevance Feedback Limitations

- ➔ Cases where relevance feedback might be insufficient:
  - ➔ **Misspellings**, if the user spells a term in a different way to the way it is spelled in the document collection, relevance feedback is unlikely to be effective.
  - ➔ **Cross-language retrieval**, documents in another language are not near (in a vector space) of other topic related documents, rather documents in the same language are closer.
  - ➔ **Vocabulary mismatch**, in this cases the initial query mostly likely will fail and thus relevant feedback won't be effective.
- ➔ Users are often reluctant to provide explicit feedback. Users expect single interactions in search and the concept of relevance feedback is hard to explain to the average user.
- ➔ Additionally, it is often harder to understand why a particular document was retrieved after relevance feedback was applied.

Pseudo Relevance Feedback

➔ Pseudo relevance feedback provides a method for automatic local analysis.

➔ It automates the manual part so that a relevance feedback algorithm is applied without extended user interaction.

➔ The method is applied to normal retrieval — assume that the top k ranked documents are relevant, and apply a relevance feedback algorithm under this assumption.

Implicit Relevance Feedback

➔ We can use indirect sources of evidence rather than explicit feedback on relevance.

➔ This is called implicit (relevance) feedback.

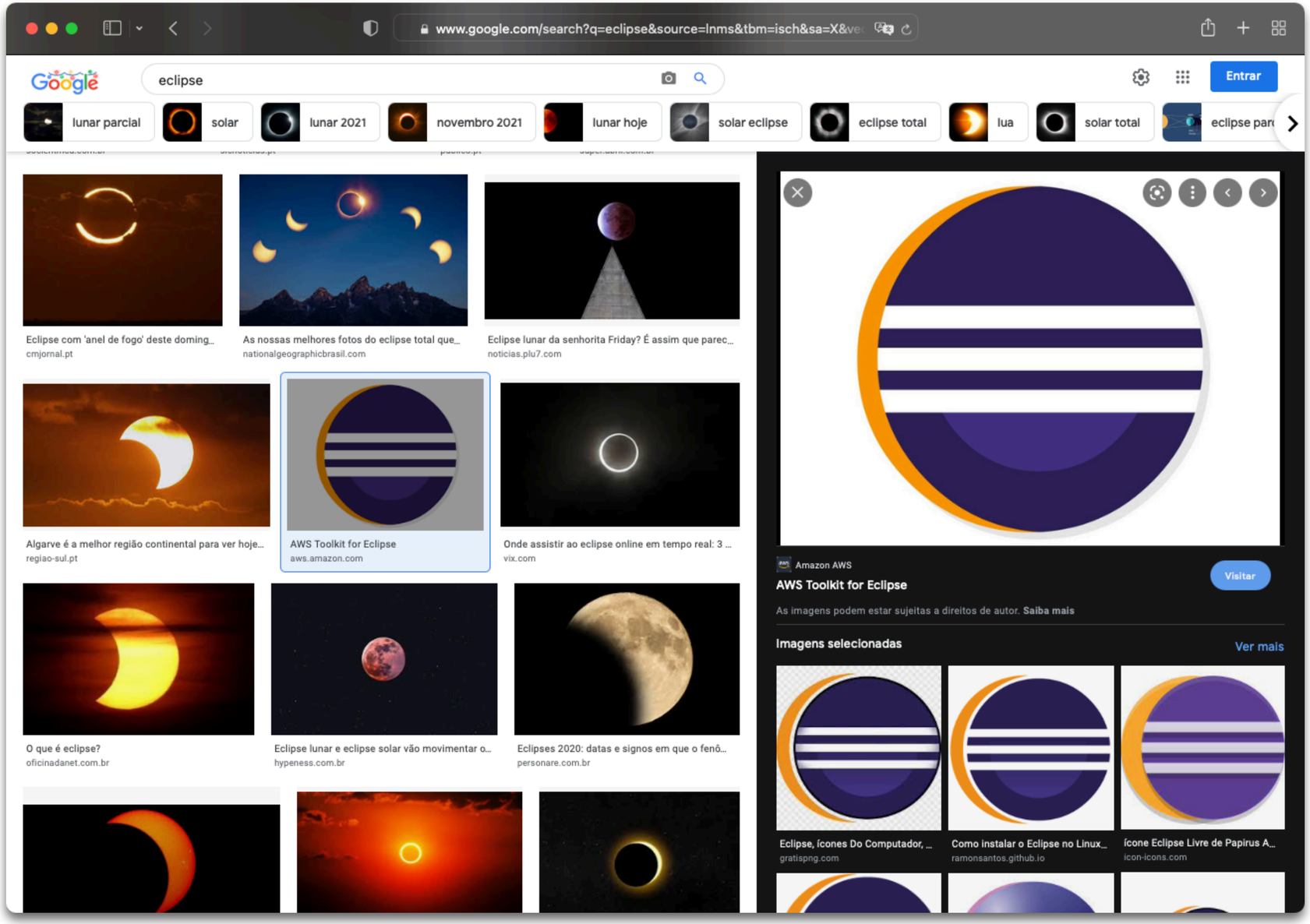
➔ Implicit feedback is less reliable than explicit feedback, but is more useful than pseudo relevance feedback, which contains no evidence of user judgements.

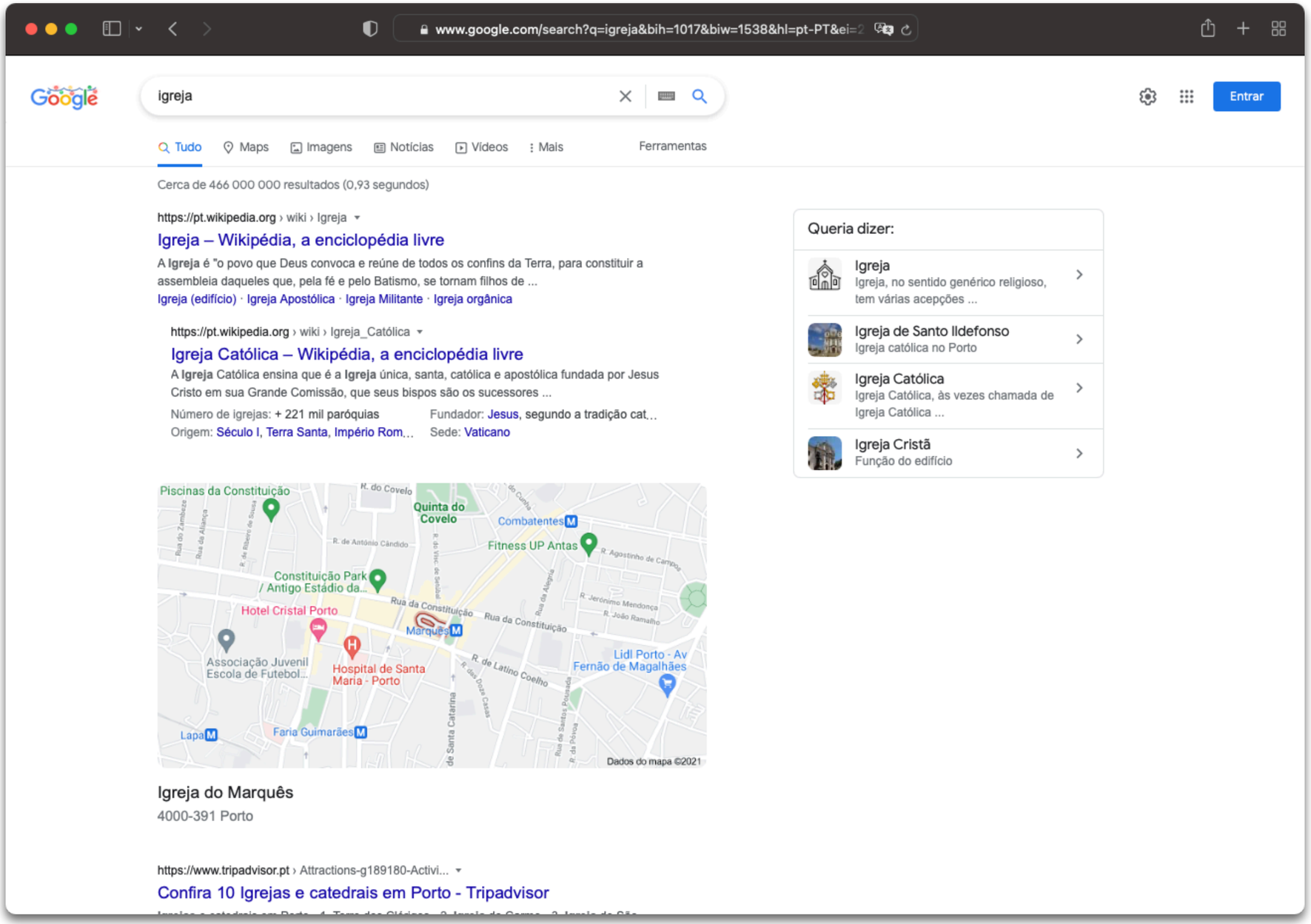
➔ Clickstreams are one of the main examples of indirect relevance information — clicks on links are assumed to indicate that the page was likely relevant for the query.

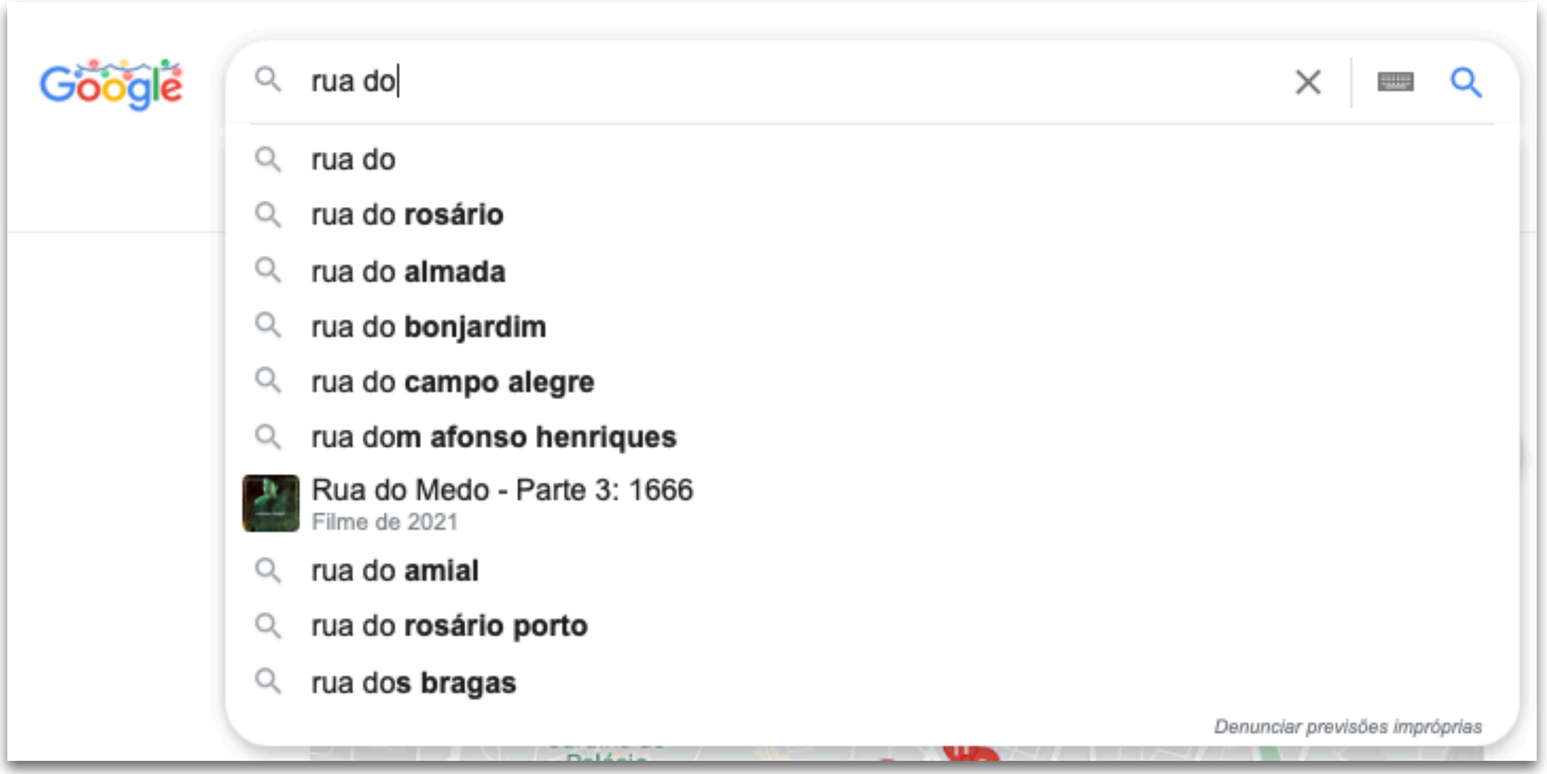
Global Query Reformulation Methods

Query Expansion

- With query expansion, user give additional input on query words or phrases to suggest additional terms. Users opt to use one of alternative query suggestions.
- How to generate alternative question expansions for the user:
  - Use synonyms and related words from a global thesaurus.
    - Use a controlled vocabulary to build a thesaurus;
    - Use a manual thesaurus built by editors;
    - Automatically derive thesaurus, e.g. use text statistics;
    - Use query log mining to find related expansions (global, contextual, user-based);


www.google.com/search?q=eclipse&source=lnms&tbm=isch&sa=X&ve
Google
eclipse
Entrar
lunar parcial
solar
lunar 2021
novembro 2021
lunar hoje
solar eclipse
eclipse total
lua
solar total
Eclipse com 'anel de fogo' deste doming...
cmjornal.pt
As nossas melhores fotos do eclipse total que...
nationalgeographicbrasil.com
Eclipse lunar da senhorita Friday? É assim que parec...
noticias.plu7.com
Algarve é a melhor região continental para ver hoje...
regiao-sul.pt
AWS Toolkit for Eclipse
aws.amazon.com
Onde assistir ao eclipse online em tempo real: 3 ...
vix.com
O que é eclipse?
oficinadanet.com.br
Eclipse lunar e eclipse solar vão movimentar o...
hypeness.com.br
Eclipses 2020: datas e signos em que o fenô...
personare.com.br
Amazon AWS
AWS Toolkit for Eclipse
Visitar
As imagens podem estar sujeitas a direitos de autor. Saiba mais
Imagens selecionadas
Ver mais
Eclipse, ícones Do Computador, ...
gratispng.com
Como instalar o Eclipse no Linux...
ramonsantos.github.io
Ícone Eclipse Livre de Papirus A...
icon-icons.com


www.google.com/search?q=igreja&bih=1017&biw=1538&hl=pt-PT&ei=2
Google
igreja
Entrar
Tudo
Maps
Imagens
Notícias
Vídeos
Mais
Ferramentas
Cerca de 466 000 000 resultados (0,93 segundos)
https://pt.wikipedia.org › wiki › Igreja
Igreja – Wikipédia, a enciclopédia livre
A Igreja é "o povo que Deus convoca e reúne de todos os confins da Terra, para constituir a assembleia daqueles que, pela fé e pelo Batismo, se tornam filhos de ...
Igreja (edifício) · Igreja Apostólica · Igreja Militante · Igreja orgânica
https://pt.wikipedia.org › wiki › Igreja_Católica
Igreja Católica – Wikipédia, a enciclopédia livre
A Igreja Católica ensina que é a Igreja única, santa, católica e apostólica fundada por Jesus Cristo em sua Grande Comissão, que seus bispos são os sucessores ...
Número de igrejas: + 221 mil paróquias
Fundador: Jesus, segundo a tradição cat...
Origem: Século I, Terra Santa, Império Rom...
Sede: Vaticano
Queria dizer:
Igreja
Igreja, no sentido genérico religioso, tem várias acepções ...
Igreja de Santo Ildefonso
Igreja católica no Porto
Igreja Católica
Igreja Católica, às vezes chamada de Igreja Católica ...
Igreja Cristã
Função do edifício
Piscinas da Constituição
Quinta do Covelo
Combatentes
Fitness UP Antas
Constituição Park / Antigo Estádio da...
Hotel Cristal Porto
Marquês
Rua da Constituição
Associação Juvenil Escola de Futebol...
Hospital de Santa Maria - Porto
Lidl Porto - Av Fernão de Magalhães
Lapa
Faria Guimarães
Dados do mapa ©2021
Igreja do Marquês
4000-391 Porto
https://www.tripadvisor.pt › Attractions-g189180-Activi...
Confira 10 Igrejas e catedrais em Porto - Tripadvisor

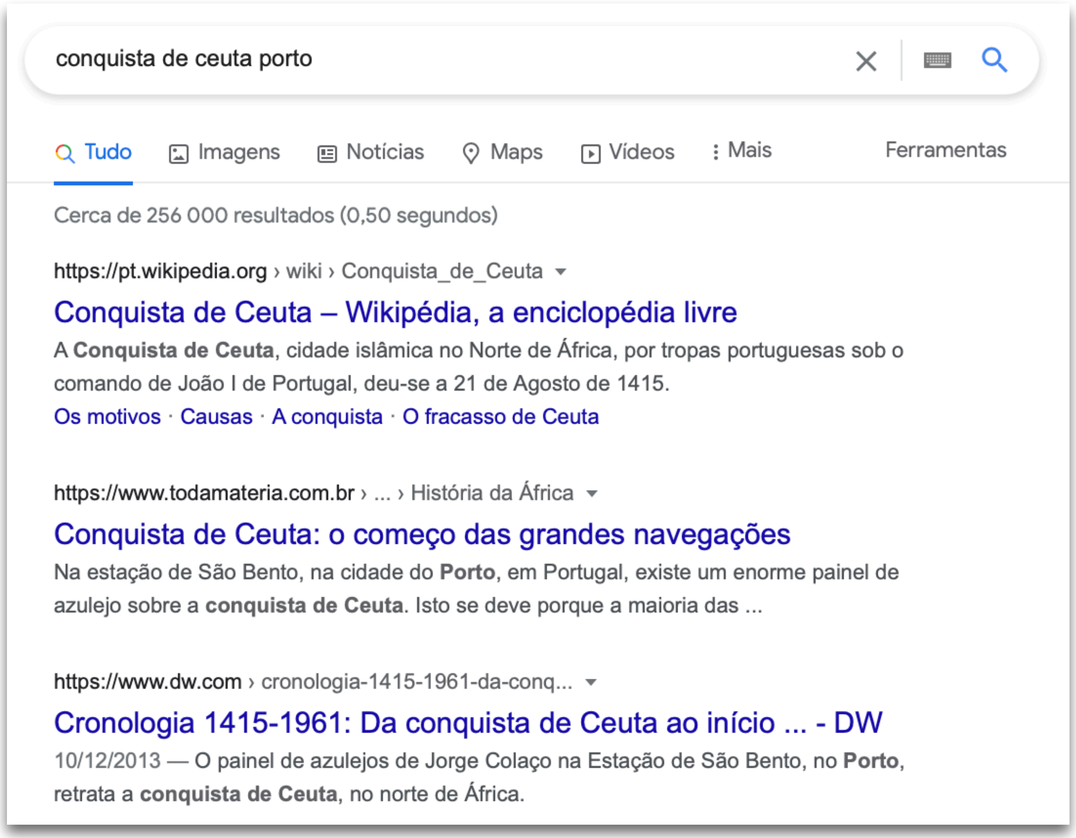

Google
rua do
rua do
rua do rosário
rua do almada
rua do bonjardim
rua do campo alegre
rua dom afonso henriques
Rua do Medo - Parte 3: 1666
Filme de 2021
rua do amial
rua do rosário porto
rua dos bragas
Denunciar previsões impróprias

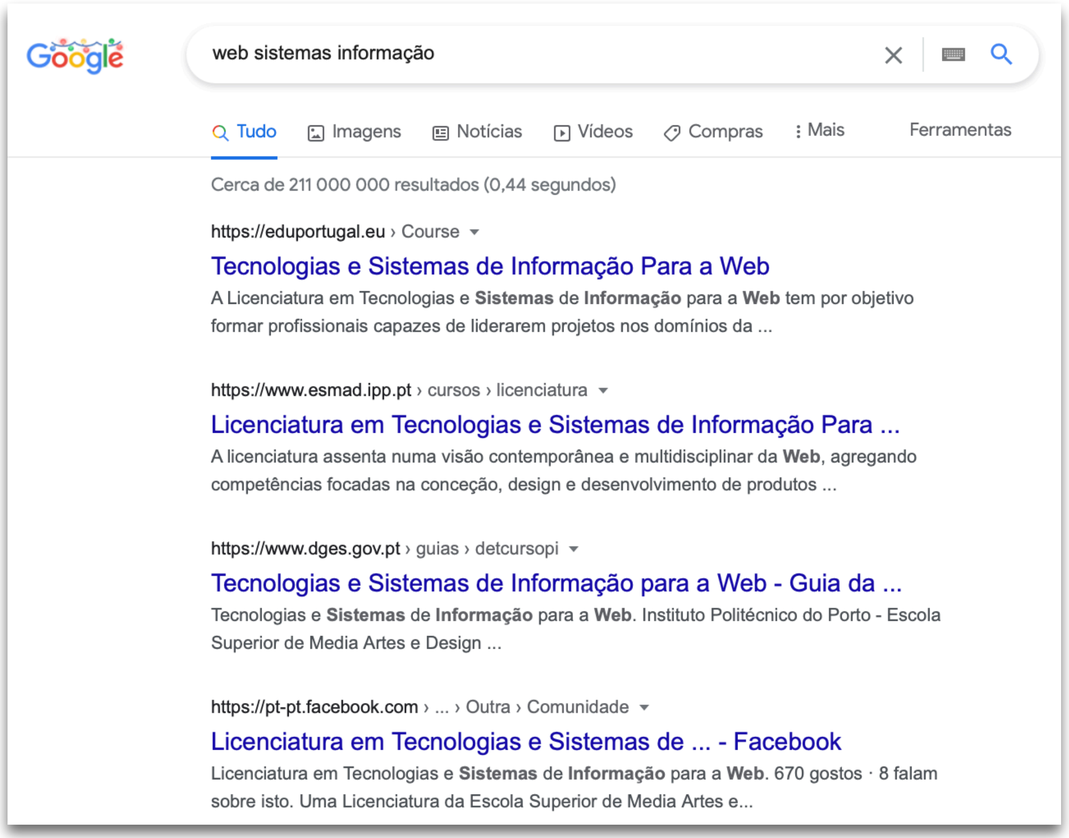
Search User Experience

Search User Experience

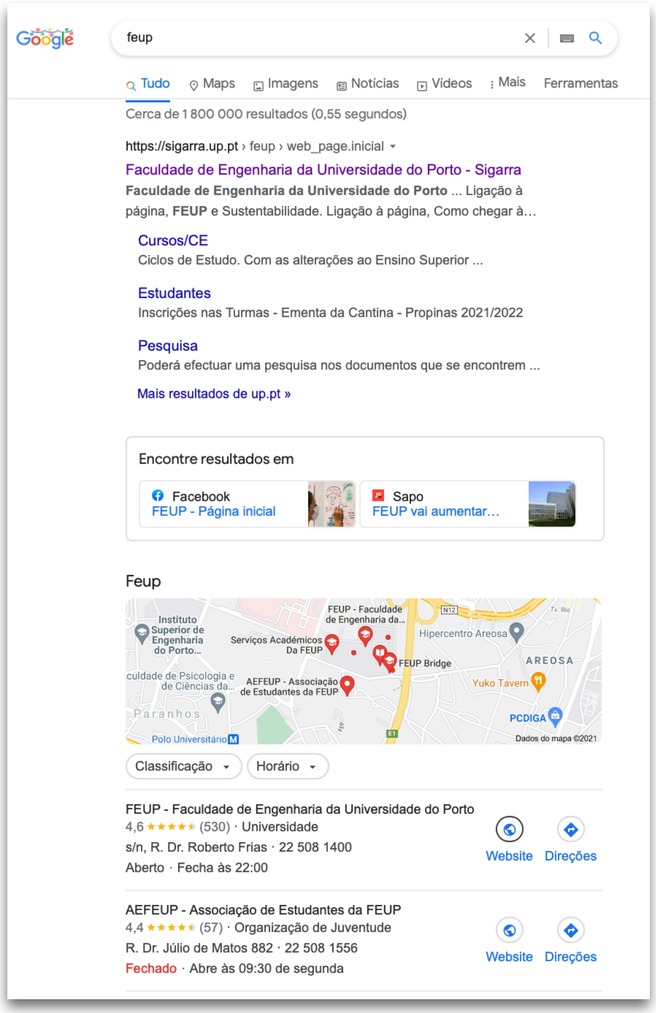
- ➔ Ranking is only a part of the information retrieval process, user interaction is central to the overall user experience and success of the user task.
- ➔ Designing the search user experience involves designing how the user interacts with the system during query formulation and reformulation, and while browsing the results.
- ➔ Example of search user experience techniques:
  - ➔ Support natural language queries.
  - ➔ Query auto-complete and suggestions;
  - ➔ Results snippets, e.g. query-dependent result snippets.
  - ➔ Clustering results.
  - ➔ Support site search.

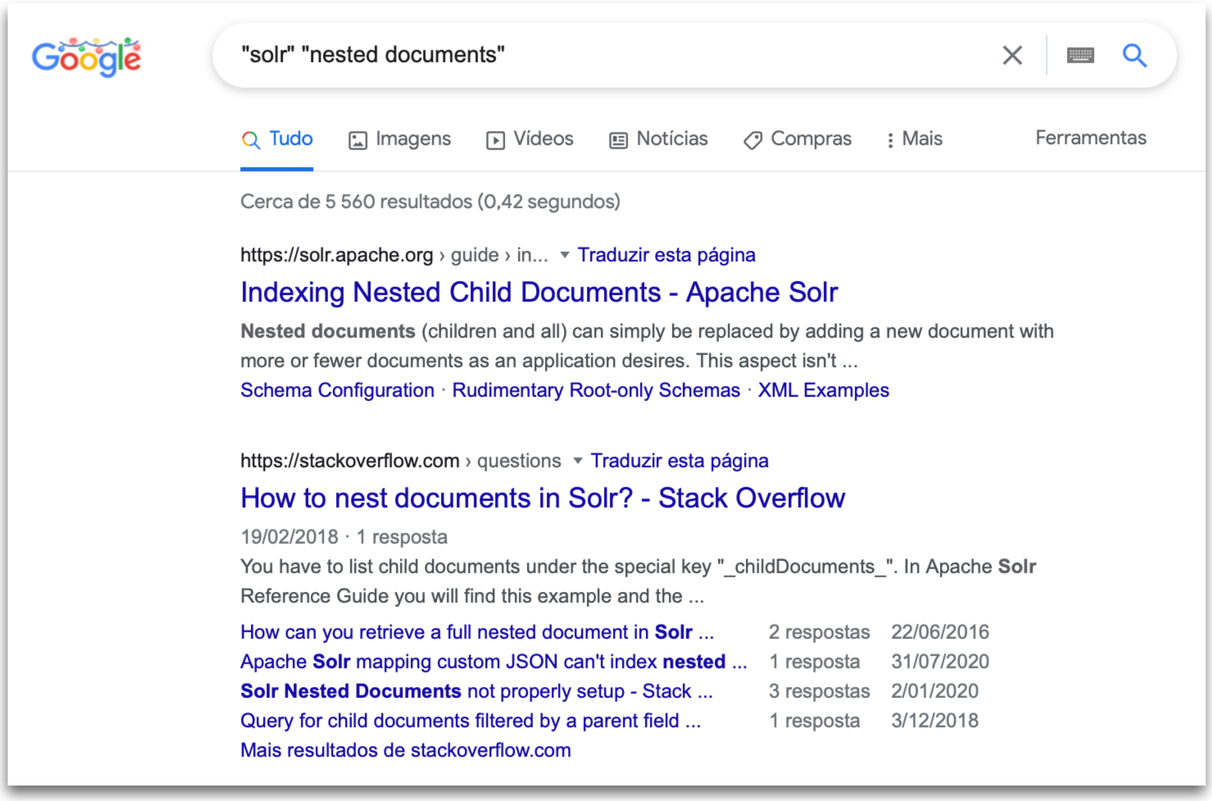
Contextual Snippets



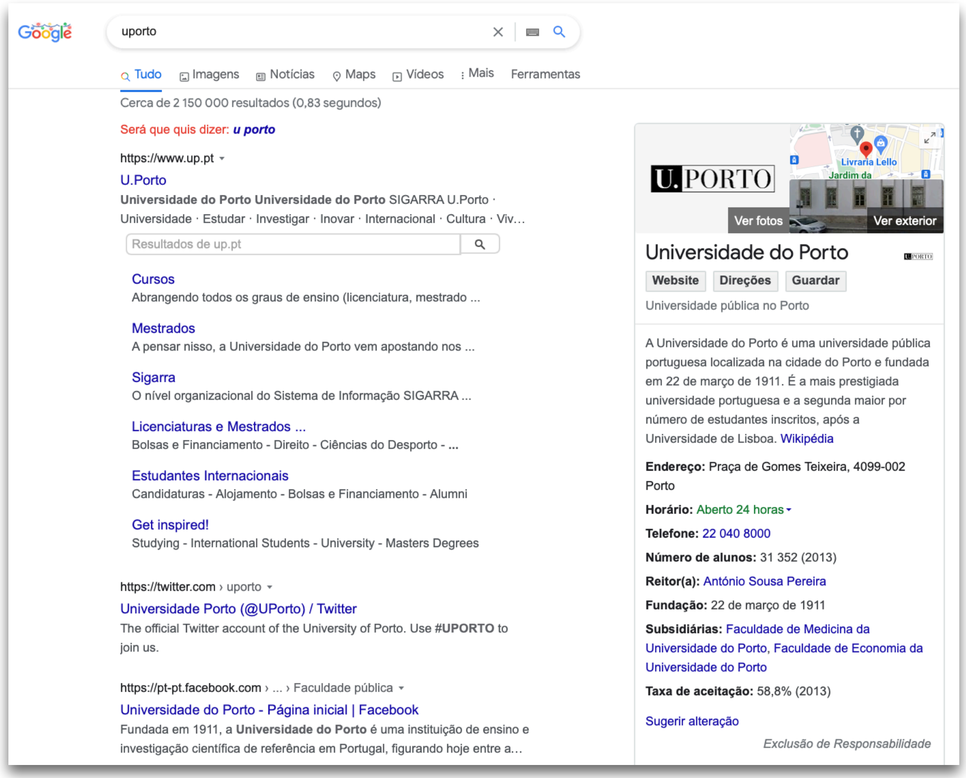


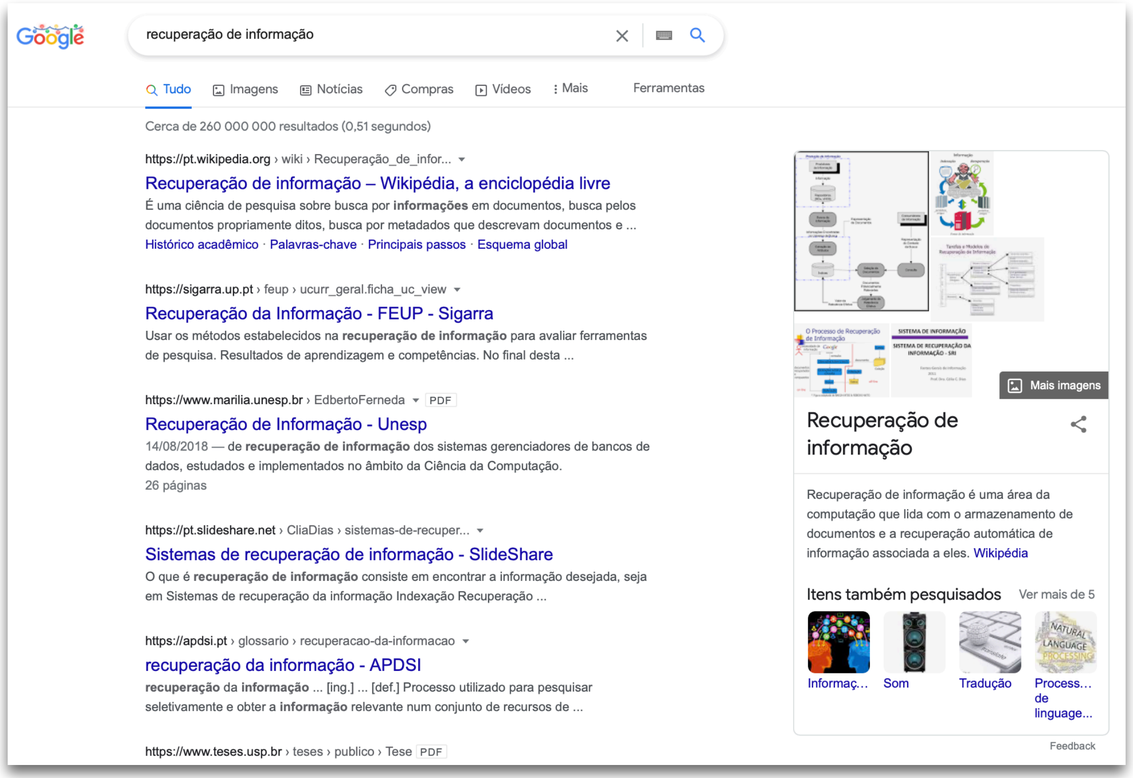
Clustering Results



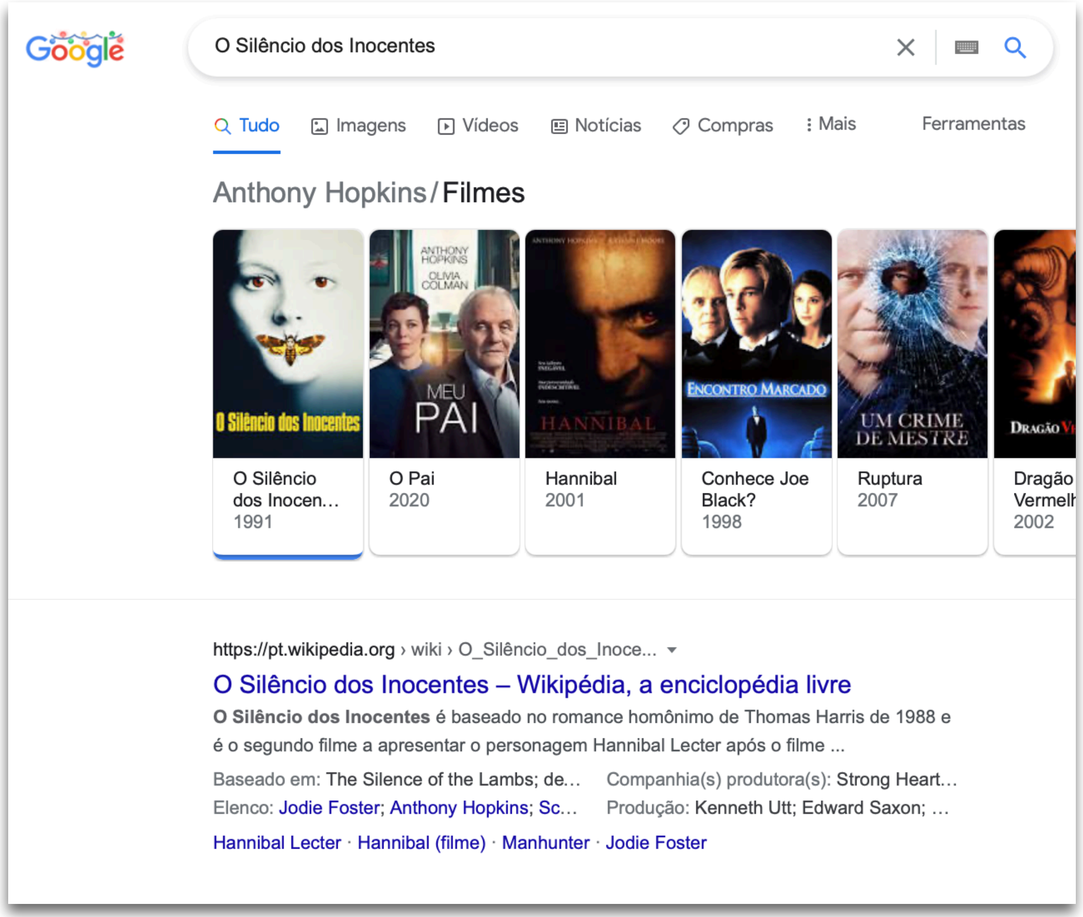


Infoboxes

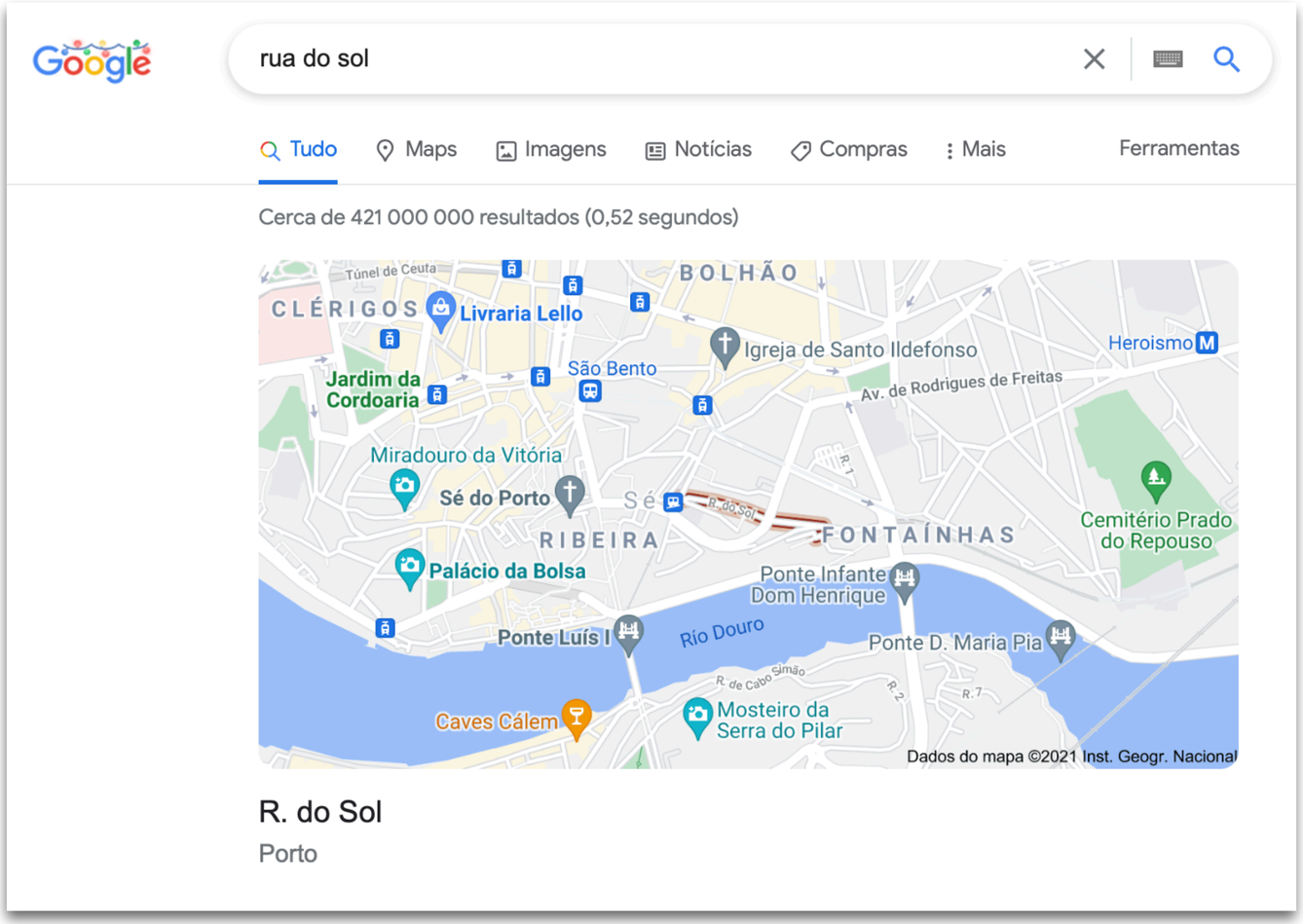

uporto
Tudo
Imagens
Notícias
Maps
Vídeos
Mais
Ferramentas
Cerca de 2 150 000 resultados (0,83 segundos)
Será que quis dizer: u porto
https://www.up.pt
U.Porto
Universidade do Porto Universidade do Porto SIGARRA U.Porto · Universidade · Estudar · Investigar · Inovar · Internacional · Cultura · Viv…
Resultados de up.pt
Cursos
Abrangendo todos os graus de ensino (licenciatura, mestrado ...
Mestrados
A pensar nisso, a Universidade do Porto vem apostando nos ...
Sigarra
O nível organizacional do Sistema de Informação SIGARRA ...
Licenciaturas e Mestrados ...
Bolsas e Financiamento - Direito - Ciências do Desporto - ...
Estudantes Internacionais
Candidaturas - Alojamento - Bolsas e Financiamento - Alumni
Get inspired!
Studying - International Students - University - Masters Degrees
https://twitter.com › uporto
Universidade Porto (@UPorto) / Twitter
The official Twitter account of the University of Porto. Use #UPORTO to join us.
https://pt-pt.facebook.com › ... › Faculdade pública
Universidade do Porto - Página inicial | Facebook
Fundada em 1911, a Universidade do Porto é uma instituição de ensino e investigação científica de referência em Portugal, figurando hoje entre a…
U.PORTO
Livraria Lello
Jardim da
Ver fotos
Ver exterior
Universidade do Porto
Website
Direções
Guardar
Universidade pública no Porto
A Universidade do Porto é uma universidade pública portuguesa localizada na cidade do Porto e fundada em 22 de março de 1911. É a mais prestigiada universidade portuguesa e a segunda maior por número de estudantes inscritos, após a Universidade de Lisboa. Wikipédia
Endereço: Praça de Gomes Teixeira, 4099-002 Porto
Horário: Aberto 24 horas
Telefone: 22 040 8000
Número de alunos: 31 352 (2013)
Reitor(a): António Sousa Pereira
Fundação: 22 de março de 1911
Subsidiárias: Faculdade de Medicina da Universidade do Porto, Faculdade de Economia da Universidade do Porto
Taxa de aceitação: 58,8% (2013)
Sugerir alteração
Exclusão de Responsabilidade


recuperação de informação
Tudo
Imagens
Notícias
Compras
Vídeos
Mais
Ferramentas
Cerca de 260 000 000 resultados (0,51 segundos)
https://pt.wikipedia.org › wiki › Recuperação_de_infor...
Recuperação de informação – Wikipédia, a enciclopédia livre
É uma ciência de pesquisa sobre busca por informações em documentos, busca pelos documentos propriamente ditos, busca por metadados que descrevam documentos e ...
Histórico acadêmico · Palavras-chave · Principais passos · Esquema global
https://sigarra.up.pt › feup › ucurr_geral.ficha_uc_view
Recuperação da Informação - FEUP - Sigarra
Usar os métodos estabelecidos na recuperação de informação para avaliar ferramentas de pesquisa. Resultados de aprendizagem e competências. No final desta ...
https://www.marilia.unesp.br › EdbertoFerneda PDF
Recuperação de Informação - Unesp
14/08/2018 — de recuperação de informação dos sistemas gerenciadores de bancos de dados, estudados e implementados no âmbito da Ciência da Computação.
26 páginas
https://pt.slideshare.net › CliaDias › sistemas-de-recuper...
Sistemas de recuperação de informação - SlideShare
O que é recuperação de informação consiste em encontrar a informação desejada, seja em Sistemas de recuperação da informação Indexação Recuperação ...
https://apdsi.pt › glossario › recuperacao-da-informacao
recuperação da informação - APDSI
recuperação da informação ... [ing.] ... [def.] Processo utilizado para pesquisar seletivamente e obter a informação relevante num conjunto de recursos de ...
https://www.teses.usp.br › teses › publico › Tese PDF
Mais imagens
Recuperação de informação
Recuperação de informação é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. Wikipédia
Itens também pesquisados
Ver mais de 5
Informaç…
Som
Tradução
Process… de linguage…
Feedback

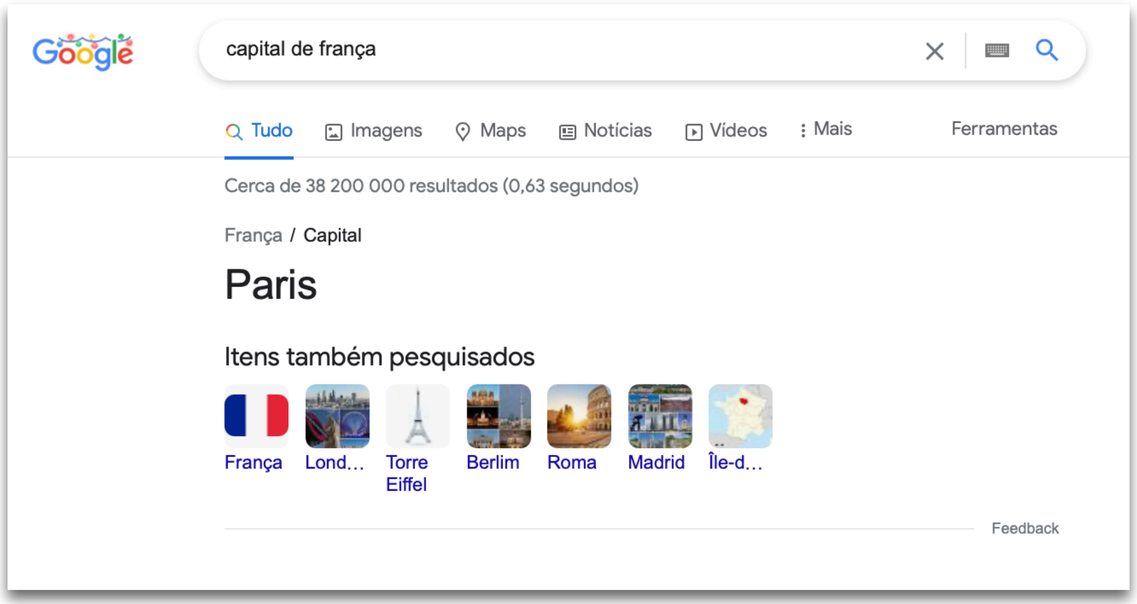
Results Browsing

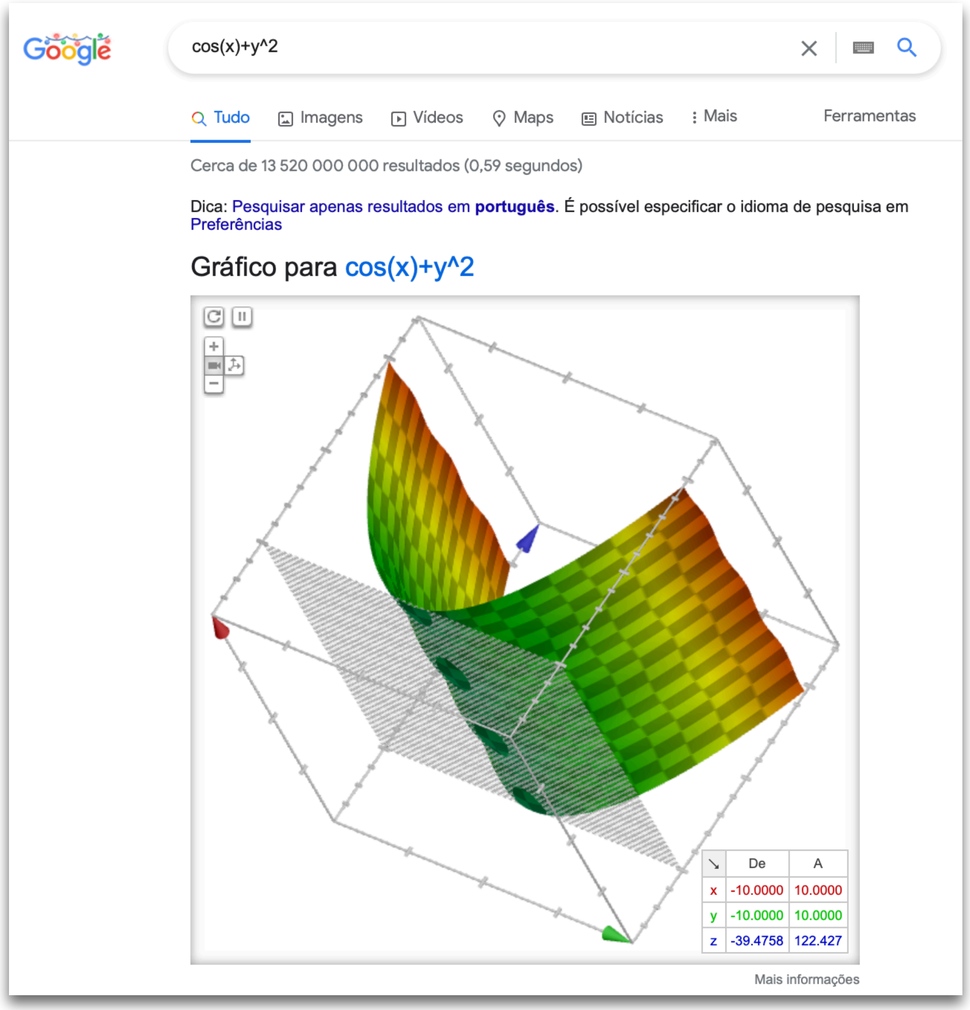


Geographical Results

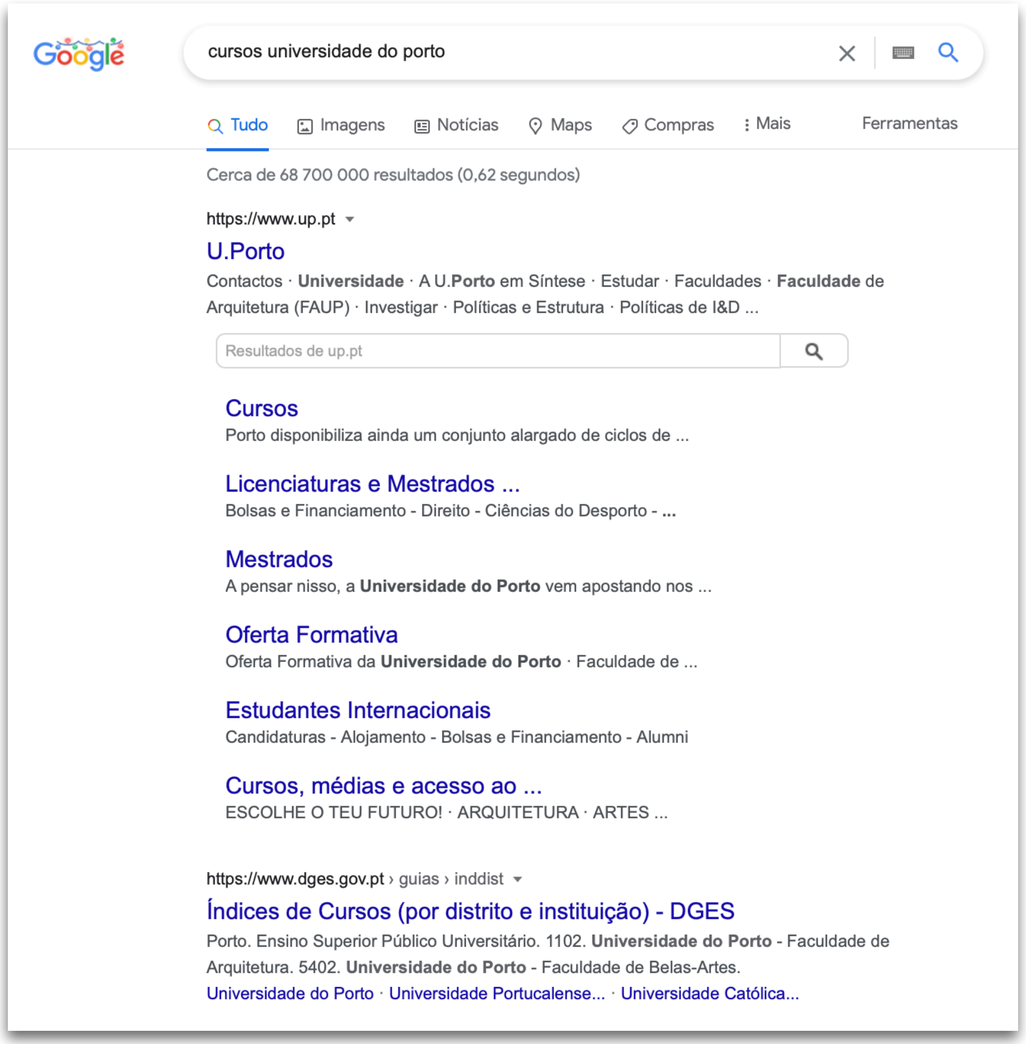


Immediate Answers





Domain Specific Search



References

- ➔ Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval
  - ➔ Chapter 9: Relevance feedback and query expansion
- ➔ W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines - Information Retrieval in Practice
  - ➔ Chapter 5: Ranking with Indexes + Slides
  - ➔ Chapter 6: Queries and Interfaces + Slides
  - ➔ Chapter 7: Retrieval Models + Slides
- ➔ Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search
  - ➔ Chapter 2: User Interfaces for Search + Slides