The XLDB Group at CLEF 2004

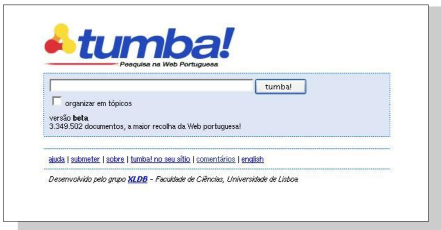
Nuno Cardoso, Mário J. Silva and Miguel Costa - {ncardoso, mjs, mcosta}@xldb.di.fc.ul.pt
Grupo XLDB – Departamento de Informática - Faculdade de Ciências da Universidade de Lisboa

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

tumba! Pesquisa na Web Portuguesa

Screenshots

http://www.tumba.pt


tumba! Pesquisa na Web Portuguesa

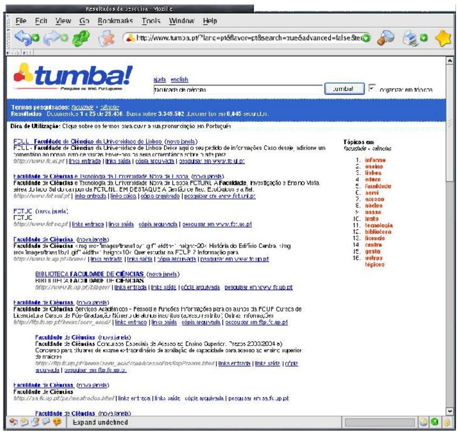
tumba!

□ organizar em tópicos

versão **beta**
3.349.502 documentos, a maior recolha da Web portuguesa!

ajuda | submeter | sobre | tumba! no seu sítio | comentários | english

Desenvolvido pelo grupo ***XLDB*** *– Faculdade de Ciências, Universidade de Lisboa*


Resultados da pesquisa - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.tumba.pt/?lang=pt&flavor=pt&search=true&advanced=false&te

ajuda english

faculdade de ciências

tumba!

☑ organizar em tópicos

Termos pesquisados: faculdade + ciências
Resultados Documentos 1 a 25 de 29.436. Busca sobre 3.349.502 documentos em 0,845 segundos.

Dica de Utilização: Clique sobre os termos para ouvir a sua pronunciação em Português

FCUL - Faculdade de Ciências da Universidade de Lisboa (nova janela)
FCUL - **Faculdade** de **Ciências** da Universidade de Lisboa Deixe aqui o seu pedido de informações Caso deseje, adicione um comentário ao nosso livro de visitas Envie-nos os seus comentários sobre o site para
http://www.fc.ul.pt | links entrada | links saída | cópia arquivada | pesquisar em www.fc.ul.pt

Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (nova janela)
Faculdade de **Ciências** e Tecnologia da Universidade Nova de Lisboa FCTUNL A **Faculdade**, Investigação e Ensino Vista aérea do lado Sul do campus da FCTUNL. EM DESTAQUE A Gestão de Rec. Ecológicos e a Rel.
http://www.fct.unl.pt | links entrada | links saída | cópia arquivada | pesquisar em www.fct.unl.pt

FCTUC (nova janela)
FCTUC
http://www.fct.uc.pt | links entrada | links saída | cópia arquivada | pesquisar em www.fct.uc.pt

Faculdade de Ciências (nova janela)
Faculdade de **Ciências** <img src="images/trans1by1.gif" width=1 height=20> História do Edifício Central <img src="images/trans1by1.gif" width=1 height=10> Quer estudar na FCUP ? Informação para
http://www.fc.up.pt/home/ | links entrada | links saída | cópia arquivada | pesquisar em www.fc.up.pt

BIBLIOTECA FACULDADE DE CIÊNCIAS (nova janela)
BIBLIOTECA FACULDADE DE CIÊNCIAS
http://www.fc.up.pt/bibger/ | links entrada | links saída | cópia arquivada | pesquisar em www.fc.up.pt

Faculdade de Ciências (nova janela)
Faculdade de **Ciências** Serviços Académicos - Pessoal e Funções Informações para os alunos da FCUP Cursos de Licenciatura Cursos de Pós-Graduação Número de alunos inscritos (acesso restrito) Outras informações
http://ftp.fc.up.pt/home/serv_acad/ | links entrada | links saída | cópia arquivada | pesquisar em ftp.fc.up.pt

Faculdade de Ciências (nova janela)
Faculdade de **Ciências** Concursos Especiais de Acesso ao Ensino Superior. Prazos 2003/2004 a) Concurso para titulares de exame extraordinário de avaliação de capacidade para acesso ao ensino superior de maiores
http://ftp.fc.up.pt/home/serv_acad/conAcessoEnsSupPrazos.html | links entrada | links saída | cópia arquivada | pesquisar em ftp.fc.up.pt

Faculdade de Ciências (nova janela)
Faculdade de **Ciências**
http://sa.fc.up.pt/pe/mestrados.html | links entrada | links saída | cópia arquivada | pesquisar em sa.fc.up.pt

Faculdade de Ciências (nova janela)

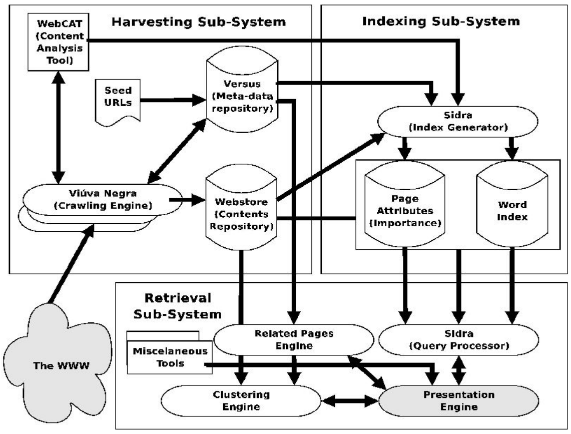
Tópicos em faculdade + ciências
1. informe
2. ensino
3. lisboa
4. educa
5. faculdade
6. servi
7. acesso
8. núcleo
9. nossa
10. insto
11. tecnologia
12. biblioteca
13. licencio
14. centro
15. gesto
16. outros tópicos

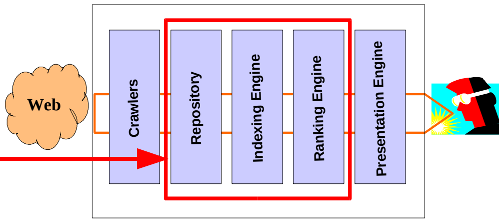
Expand undefined

Features

- The **XLDB Group** is a research unit of **LaSIGE** (Large Scale Information Systems Laboratory) at **FCUL** – Faculdade de Ciências da Universidade de Lisboa.
- The XLDB Group recently joined **Linguateca** (*www.linguateca.pt*), a distributed resource center for Portuguese language processing, which aims and fostering Portuguese-aware systems and applications and increase R&D on Portuguese
- One of our main projects was **tumba!,** a Fully-Functional specialized Search Engine for the Community of Portuguese Web users, offered as a public service since November 2002.
- Indexing over 3.5 million pages from the "Portuguese Web" and serving 20.000 daily queries.
- Similar architecture to global search engines and adopts many algorithms. However, tumba! has a better knowledge of the location and organization of Portuguese Web sites.
- Tumba! profits from annotations extracted from web documents, such as links, anchor texts, titles and headings. They weren't available on the document collection.
- Components of tumba! used in CLEF: Web Repository, Indexing System and Ranking Engine.
- We used the Web search engine tumba! in our first participation in CLEF: Portuguese Monolingual Task

Architecture





CLEF Portuguese Monolingual Task

Overview

- **Unconventional task approach! -** Tumba! is designed for Web Search, it is not optimized for CLEF tasks. Tumba! doesn't use stemmers nor blind feedback / query expansion, and the weighting is tuned for Web documents.

Manual Run: XLDBTumba01

- We created several different queries related to each topic and we used them to retrieve documents matching the query terms.
- The returned results were manually examined and classified as irrelevant and relevant according to topic criteria.
- This run showed us how difficult it is to formulate queries that correctly match an information need.

This was our manual baseline run.

Flat Ranking Run: XLDBTumba02

- For each topic, we chose a single query from the different queries used for the XLDBTumba01 run.
- Note that we didn't use more than one query per topic, neither we did any kind of query expansion.
- The Indexing and ranking Engine were configured to perform an exact match (flat-ranking algorithm), returning only the documents that match all the query terms.
- This run was our automatic baseline run.

Distances + Titles Run: XLDBTumba04

- Created using *distMinTerms* and the following algorithm, *termsInTiles*:

$$termsInTitle(d,q) = \frac{|T \cap Q|}{max(|T|,|Q|)}$$

- this is a similarity function between the terms in the title of each document *d*, denoted *T*, and the query terms in a query *d*, denoted *Q*.
- This run evaluated the importance of the title in the document ranking, but resulted in the worst performance.
- This was probably due to the naïve heuristic approach to extract titles from documents, which might mislead the ranking engine.

Distances Run: XLDBTumba05

- Created using the *distMinTerms* algorithm:

$$distMinTerms(d,q) = \begin{cases} 1 & minDist = 1 \\ 1 - \frac{minDist-1}{9} & 1 < minDist < 10 \\ 0 & minDist \geq 10 \end{cases}$$

uses the minimum distances between any pair of query terms *q* in documents *d*, *minDist*, to increase the ranking of documents whose query terms are closer on the document.

- This function indeed improved the results, as the queries with more than one term we used for the topic tend to be adjacent.

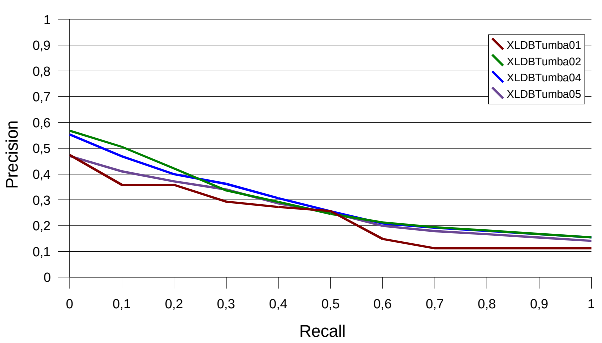
http://xldb.di.fc.ul.pt

Results

Portuguese Monolingual Task results

Run	Manual Run (XLDBTumba01)	Flat Ranking (XLDBTumba02)	Distances (XLDBTumba05)	Distances+Titles (XLDBTumba04)
Nr. Docs retrieved	209	2350	2350	2350
Nr. relevant Docs	678	678	678	678
Relevant Docs retrieved	79	168	168	168
Overall Precision	37,8%	7,1%	7,1%	7,1%
Overall Recall	11,6%	24,8%	24,8%	24,8%
Average Precision	21,8%	28,1%	25,1%	27,8%
R-Precision	22,4%	26,3%	26,7%	27,3%

XLDB Tumba Recall-Precision Values



Conclusion

- Our main objective: test, compare and improve the quality of tumba!'s results, and gather ideas on how to do it.
- The environment that we work on, the Web, is different from the flat and small collections of document texts that we used on the CLEF task.
- Tumba! does not perform stemming or query expansion and relies heavily on detecting the presence of query terms in document titles and URLs; as these weren't available for this evaluation, our results had to reflect that. Tumba! is effective on named-page finding tasks, in particular when these have properly chosen titles and multiple links.
- We intend to extend our Web Search system to provide better results in situations where the documents are not rich in HTML features, such as hyper links and meta-tags.