



Funcionamento de um Engenho de Busca

Prof. Luciano Barbosa

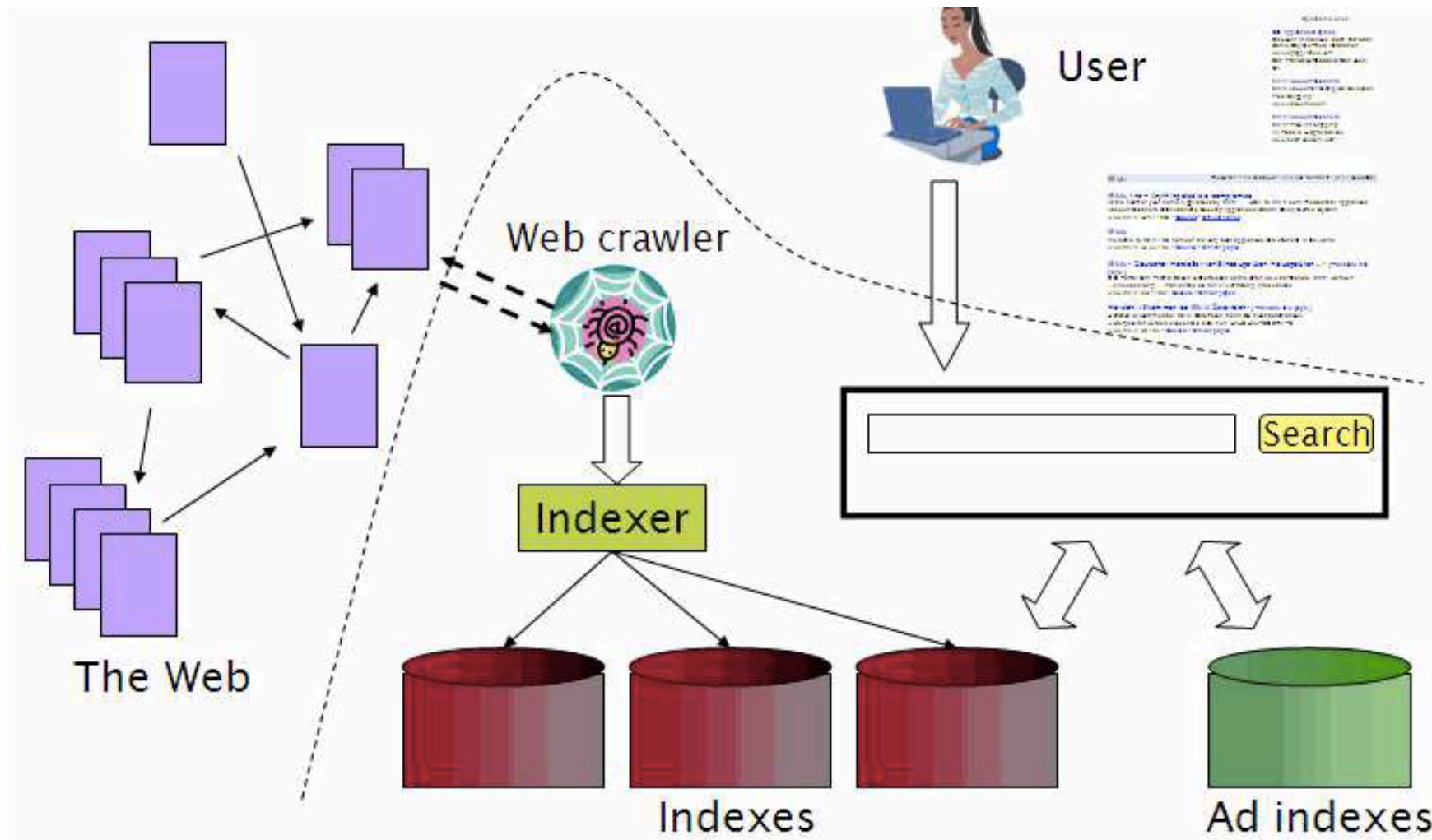
(Parte do material retirado dos slides dos livros adotados)



UNIVERSIDADE
FEDERAL
DE PERNAMBUCO



Funcionamento do Engenho de Busca



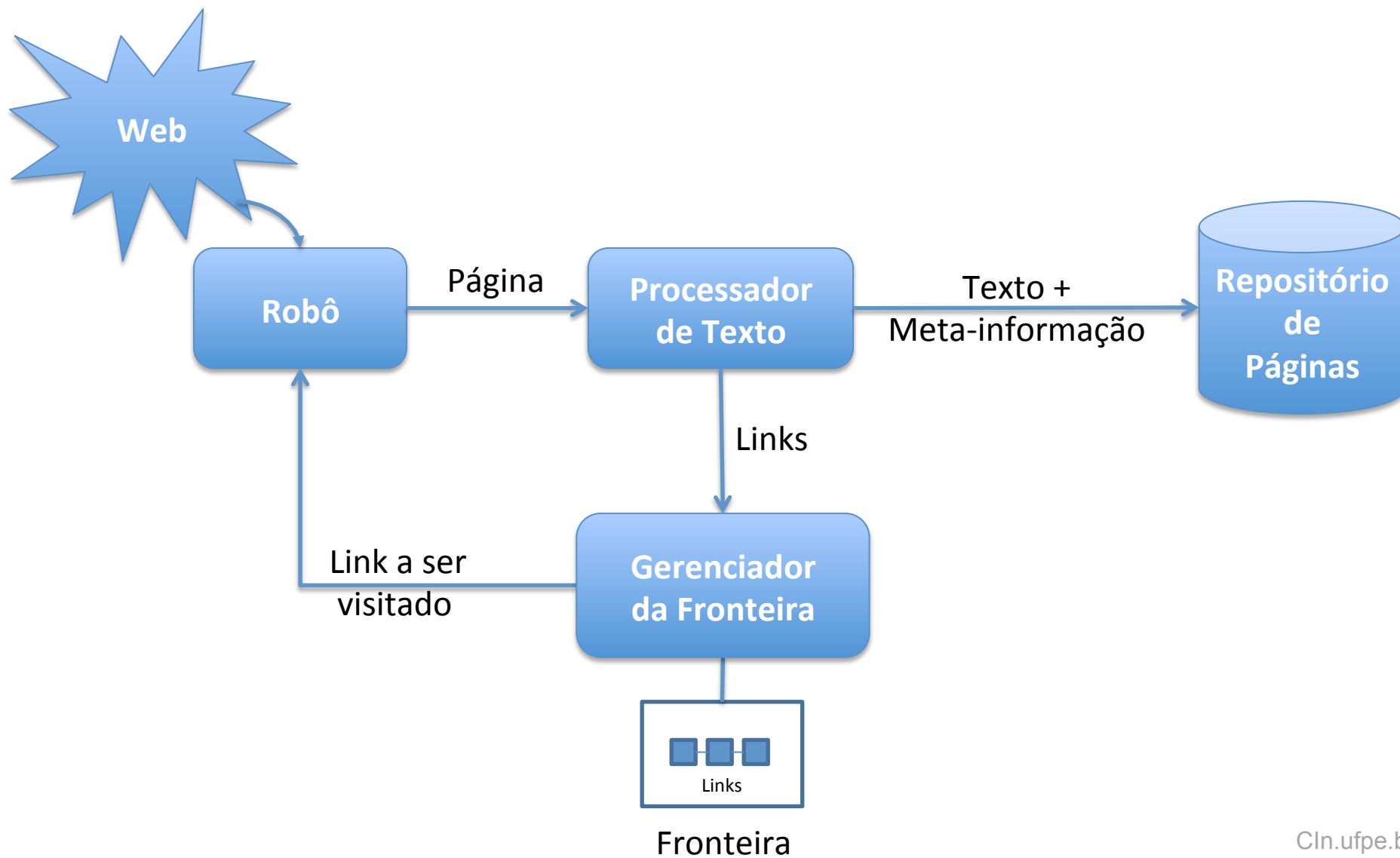


Web Crawler

- Coleta documentos na Web
- Segue links para encontrar documentos
- 2 tarefas:
 - Aumentar cobertura de páginas na base
 - Manter cópias locais sincronizadas com as online



Funcionamento de um Web Crawler





Coletor Focado


- Objetivo: coletar dados em um dado tópico ou domínio
- Exemplo de aplicações: busca vertical na área de saúde

The screenshot shows the Healthline website interface. At the top, there's a search bar with the text 'Healthline' and a dropdown menu labeled 'Topics & Tools'. Below the search bar, the results are titled 'Results for coconut oil' with a subtext 'Showing 1 to 10 of 44 results'. On the left side, there's a sidebar with a purple header 'Everything' and three sub-items: 'Articles', 'Blogs', and 'Interactive Tools'. The main content area displays three article snippets:

- Coconut Oil Diet: Weight Loss Fact or Fiction?**
Talk of **coconut oil** being the latest weight loss sensation is pretty widespread these days. But is this claim grounded in fact or just a bunch of hype? "Some people claim that **coconut oil** increases energy, improves heart health, and helps with weight..."
- Can I Use Coconut Oil for Skin Care?**
Coconut oil has been used to fight dry skin for centuries. It is often recommended to treat chronic dry skin, eczema, psoriasis, and it's also used as an oil massage for infants. It is commonly applied to the skin after a bath or shower to help the skin...
- Can I Use Coconut Oil for Hair Growth?**
Coconut oil is produced when the coconut meat is removed from the outer hard shell and pressed. Lately, **coconut oil** is being touted as a panacea for all sorts of ailments, from indigestion to asthma to autism. Now, some are suggesting a link between coconut...
- The Health Benefits of Coconut Oil**
Coconut oil is derived from the white "meat" of mature coconuts. Coconut oil is extracted by a press that separates the oil from the fruit itself.



Coletor Focado para Páginas Paralelas: Tradução de Texto



A new concept

Company

In NeoAtlas, we know how vitally important it is to care for every detail in your communications strategy. Minimizing errors involves maximizing the number of successes, and that is why we propose a new concept in terms of translation:

CUSTOMIZED TRANSLATION.

Services

We provide translations of every type with the greatest quality: general-purpose and specialized documents, sworn translations, software localization (help file, code and interface), and web sites. We also provide Desktop Publishing (DTP) services. We use the latest technologies to ensure streamlining our translation process and meeting deadlines.


Contact

We place at your disposal a team of highly qualified professionals ready to get the job done professionally and on time. If you wish to hire our services we will be glad to work with you. In this section you will find all the information you need.

Employment

We require the highest quality from our freelance professional collaborators through an initial test and continuous assessment. If you are a freelance translator, and want to join our team of professionals please send us your curriculum vitae to our email address.

<http://www.neoatlas.com/translations/>



Un nuevo concepto

Empresa

En NeoAtlas sabemos lo importante, incluso vital, que para algunas empresas es cuidar su comunicación hasta en el más pequeño detalle. Minimizar los fallos significa maximizar los aciertos, por eso le proponemos un nuevo concepto en cuanto a traducciones se refiere:

la TRADUCCIÓN PERSONALIZADA.

Servicios

Realizamos traducciones de diversa índole y complejidad con la máxima calidad: textos generales y especializados, traducciones juradas, localización de software (ayudas, código...) y páginas web. También ofrecemos servicios de DTP (maquetación). Utilizamos las tecnologías más avanzadas, lo que nos permite agilizar los procesos de traducción y cumplir los plazos previstos.

Contacto

Ponemos a su disposición un equipo de profesionales altamente cualificados capaces de satisfacer sus necesidades y exigencias. Si desea contratar nuestros servicios, estaremos encantados de colaborar con usted. En estas páginas encontrará toda la información que necesita.

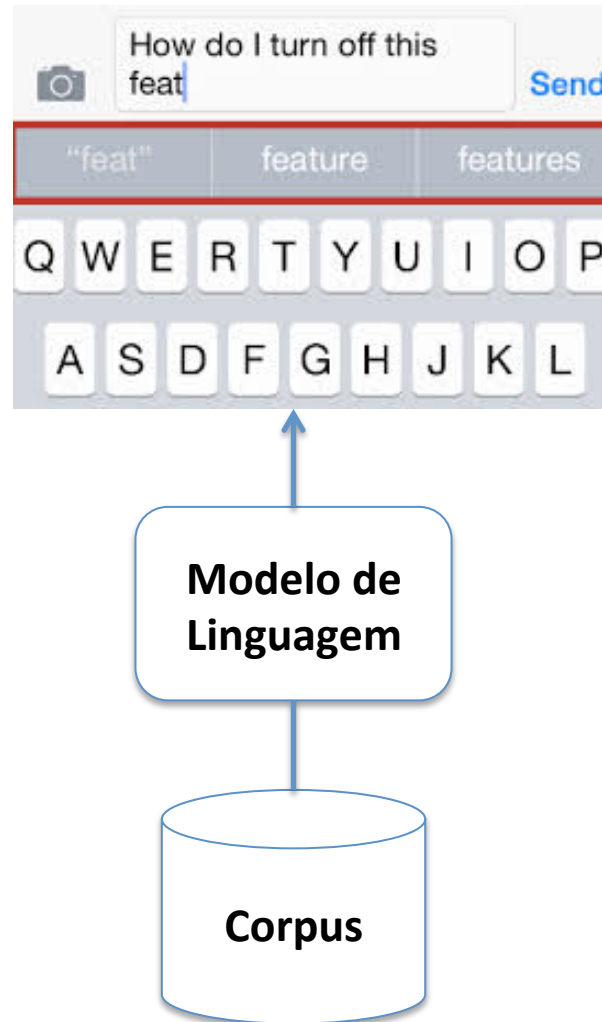
Empleo

En NeoAtlas exigimos la mayor calidad de los profesionales autónomos (mediante un examen inicial y una evaluación continua). Si eres traductor autónomo y quieres formar parte de nuestra base de colaboradores, envíanos tu CV a la dirección de contacto.

<http://www.neoatlas.com/traduccion/>



Coletor Focado para Criação de Corpus de Modelos de Linguagem





Outros Tipos: Feeds

- Streams de documentos: radio, vídeos, blogs, notícias
- RSS (Rich Site Summary) é padrão
- Ex.:
 - <http://www.reuters.com/tools/rss>
 - <http://g1.globo.com/dynamo/rss2.xml>



Processamento de Texto

- Parsing
- Stopwords
- Stemming
- Classificação
- Extração



Parsing

- Extrair dos documentos palavras e estrutura (título, links, headings, âncoras)
 - Tokenização
 - Normalização: considerar capitalização, hífen, separadores, links relativos etc
- Ex.: NYT
 - Título: The New York Times - Breaking News, World News & Multimedia
 - Palavras frequentes: the, to, of, a, in, on, and, for, by, new, times, york, news, opinion, at, as



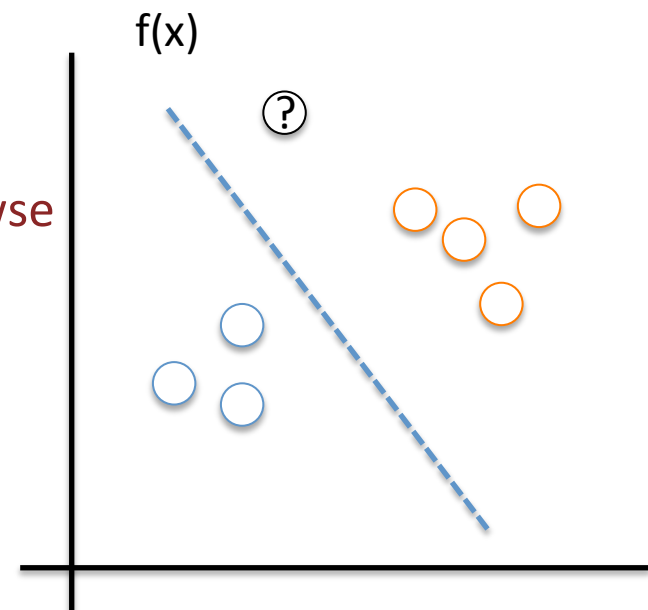
Processamento de Texto

- Stopwords
 - Ex: de, com, em
 - Impacta qualidade (palavra pouco discriminativa) e eficiência (aumenta o tamanho da base)
 - Remoção pode ser problemática para algumas consultas
- Stemming
 - Agrupa palavras com o mesmo radical
 - Ex.: computer, computers, computing, compute -> comput
 - Diminui vocabulário
 - Pode agrupar palavras com significados diferentes
- Demo: <http://text-processing.com/demo/>



Processamento de Texto

- Classificação
 - Classificar documentos em classes
 - Ex.: categorias, sentimento etc
 - Demo: <https://www.uclassify.com/browse>





Extração de Informação

- Identifica classes de termos importantes em texto semi-estruturado e não-estruturado
- Pode ser usado em:
 - Artigos de notícias, páginas Web, artigos científicos e classificados



HTML de Página de Livro da Amazon

```
....  
</td></tr>  
</table>  
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>  
<font face=verdana,arial,helvetica size=-1>  
by <a href="/exec/obidos/search-handle-url/index=books&field-author=  
Kurzweil%2C%20Ray/002-6235079-4593641">  
Ray Kurzweil</a><br>  
</font>  
<br>  
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">  
</a>  
<font face=verdana,arial,helvetica size=-1>  
<span class="small">  
<span class="small">  
<b>List Price:</b> <span class=listprice>$14.95</span><br>  
<b>Our Price: <font color=#990000>$11.96</font></b><br>  
<b>You Save:</b> <font color=#990000><b>$2.99 </b>  
(20%)</font><br>  
</span>  
<p><br>
```



Resultado da Extração

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence

Author: Ray Kurzweil

List-Price: \$14.95

Price: \$11.96

:
:

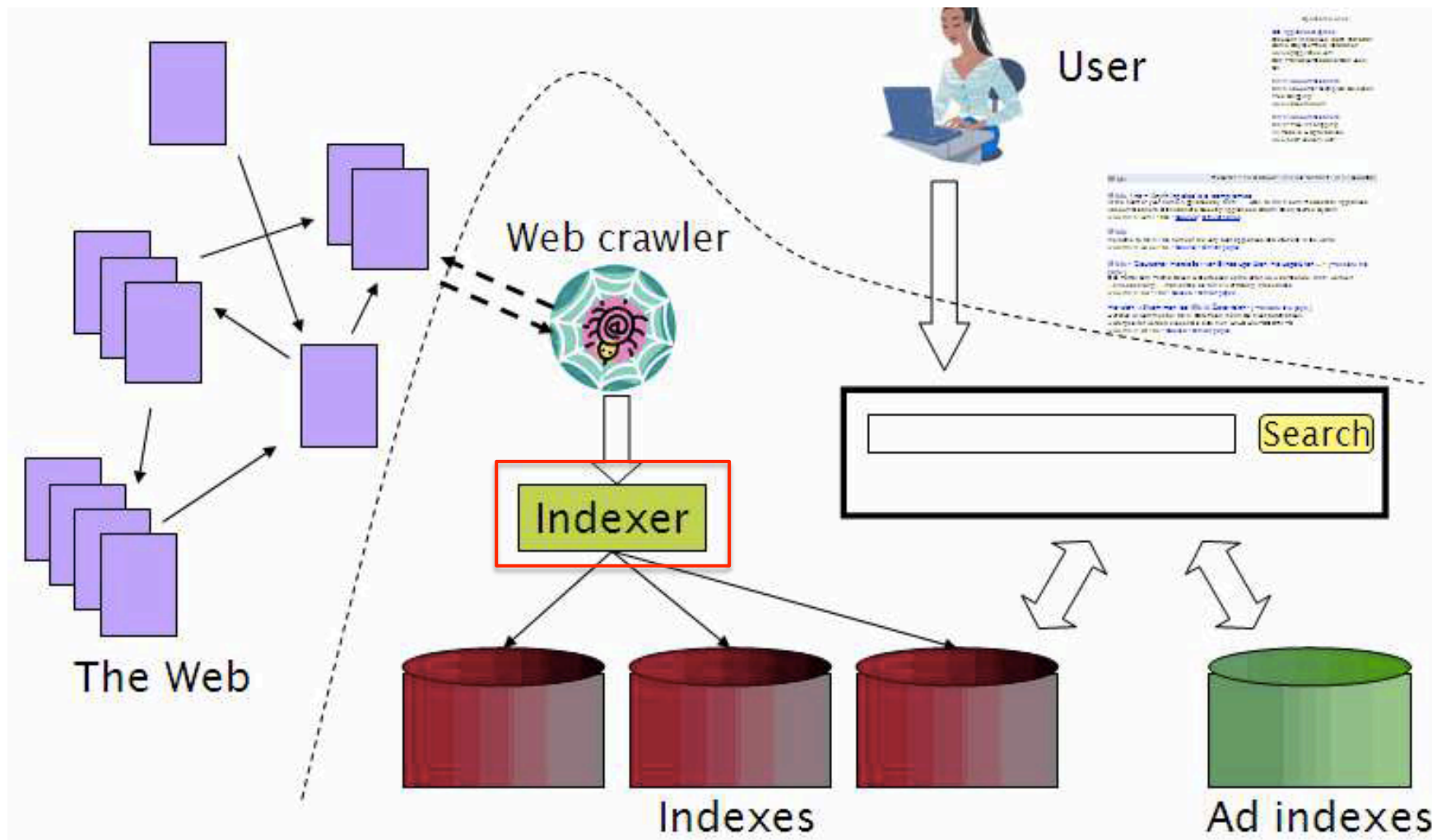


Exemplos de Padrões de Extração

- Expressão regular para extração
 - Preço: “\\$\d+(\.\d{2})?”
- Pré-filtro para identificar contexto
 - Extração do preço do livro da Amazon:
 - Padrão pré-filtro: “List Price: ”
 - Padrão do filtro: “\\$\d+(\.\d{2})?”
- Pós-filtro para identificar fim do campo
 - Extração do preço do livro da Amazon:
 - Padrão pré-filtro: “List Price: ”
 - Padrão do filtro: “\\$\d+(\.\d{2})?”
 - Padrão pós-filtro: “”
- Demos:
 - <http://openie.allenai.org/>
 - <http://services.gate.ac.uk/annie/>



Funcionamento do Engenho de Busca





Indexador

1. Computa estatísticas do documento
 - Frequência e posição das palavras
 - Usadas no ranqueamento e consultas de proximidade
2. Inversão
 - Converte de documento-termo para termo-documento
 - Aumenta velocidade consulta
 - Compressão usada para eficiência
 - Deve lidar com atualização



Índice Invertido

Vocabulário

Postings

Vocabulary	n_i
to	2
do	3
is	1
be	4
or	1
not	1
I	2
am	2
what	1
think	1
therefore	1
da	1
let	1
it	1

Occurrences as inverted lists

[1,4],[2,2]

[1,2],[3,3],[4,3]

[1,2]

[1,2],[2,2],[3,2],[4,2]

[2,1]

[2,1]

[2,2],[3,2]

[2,2],[3,1]

[2,1]

[3,1]

[3,1]

[4,3]

[4,2]

[4,2]

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

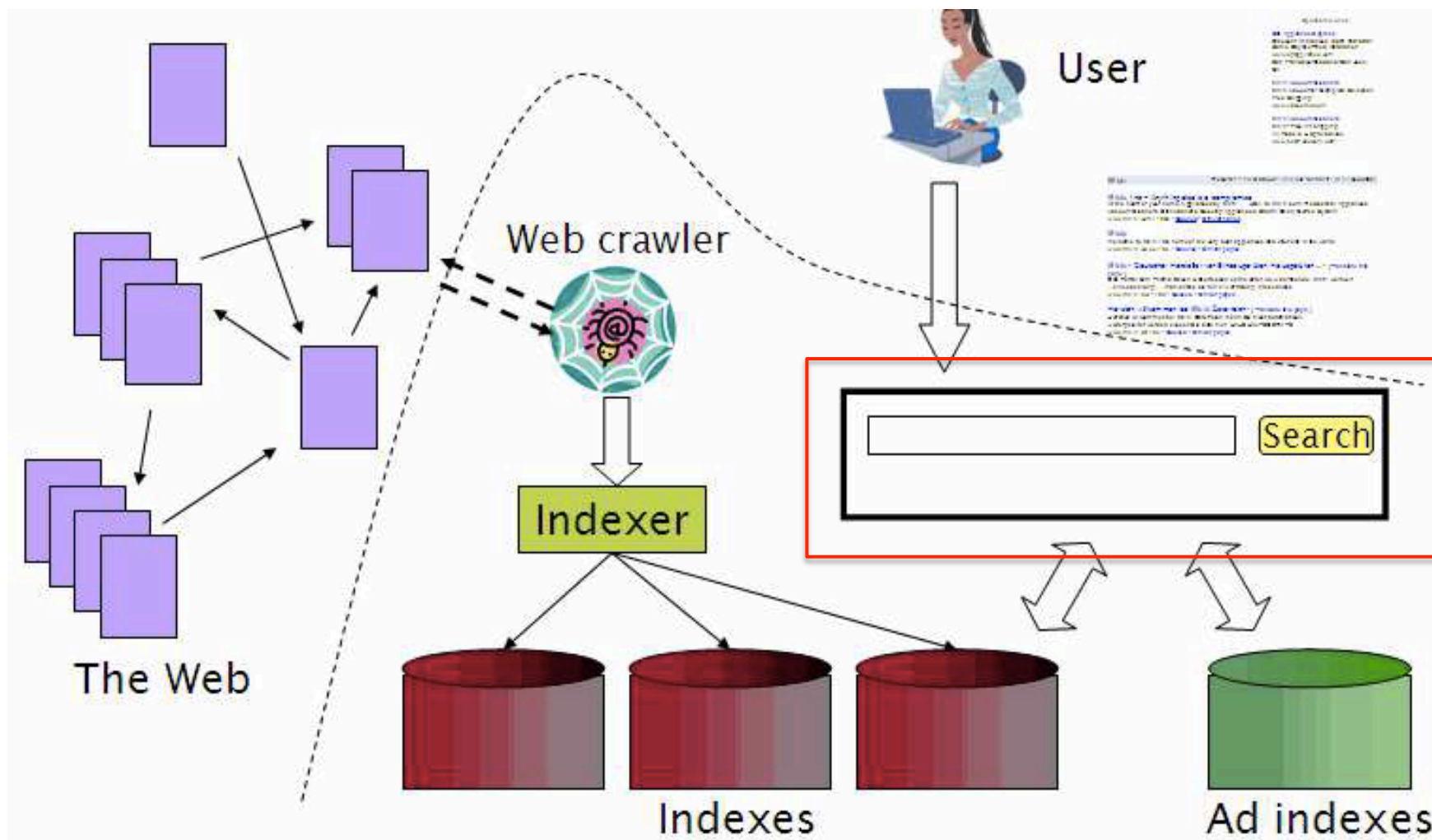


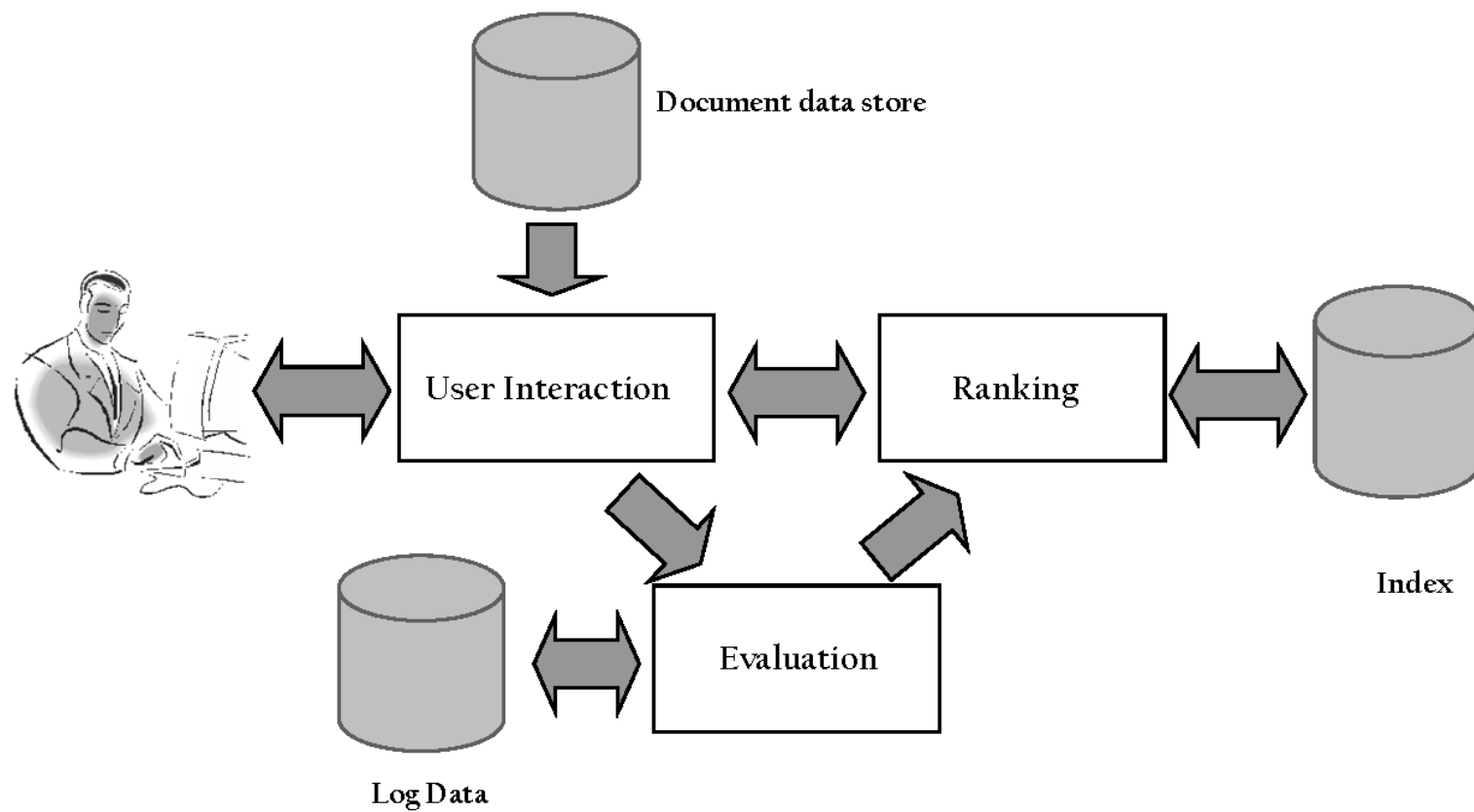
Indexador: Distribuição do Índice

- Distribuído em vários computadores ou lugares
- Aumenta velocidade de consulta
- Ex.: Google Data Center
 - Distribuído por todo mundo
 - 1 milhão de servidores
 - 3 milhões processadores/cores



Funcionamento do Engenho de Busca







- Interação com usuário
 - Criação e refinamento de consultas
 - Mostrar resultados
- Ranqueamento
 - Ranquear documentos na base a partir da consulta
- Avaliação
 - Monitorar e medir qualidade e velocidade (offline)



Interação com Usuário

- Entrada
 - Provê interface e linguagem de consulta
 - Muita variação no tipo de consulta
 - https://www.google.com/advanced_search
- Transformação da consulta
 - Processamento de texto
 - Checagem ortográfica e sugestão de consulta
 - Expansão de consulta



Interação com Usuário: Saída

- Contrói a página de resultados ranqueados
- Gera snippets
- Enfatiza palavras importantes
- Recupera anúncios relacionados
- Pode apresentar outro tipo de visualização



Interação com Usuário: Saída

Google [Search](#) [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online **discount brokers** emphasizing rates, charges, and customer comments and complaints.
www.broker-reviews.us/ - 94k - [Cached](#) - [Similar pages](#)

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ **Brokerage/** Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/brokers/index.cfm?story=2004-discount-table - 121k - [Cached](#) - [Similar pages](#)

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 **Brokers** headlines. 10. Don't Pay Your **Broker** for Free Funds May 15 at 3:39 PM. 5. Don't **Discount** the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/brokers/index.aspx - 44k - [Cached](#) - [Similar pages](#)

Discount Broker
Discount Broker - Definition of **Discount Broker** on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/discountbroker.asp - 31k - [Cached](#) - [Similar pages](#)

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock **broker** SogoTrade offers the best in **discount brokerage** investing. Get stock market quotes from this internet stock trading company.
www.sogotrade.com/ - 39k - [Cached](#) - [Similar pages](#)

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a **discount broker** can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/Startinvesting/P66171.asp - 34k - [Cached](#) - [Similar pages](#)

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee
Transfer to Firsttrade for Free!
www.firsttrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

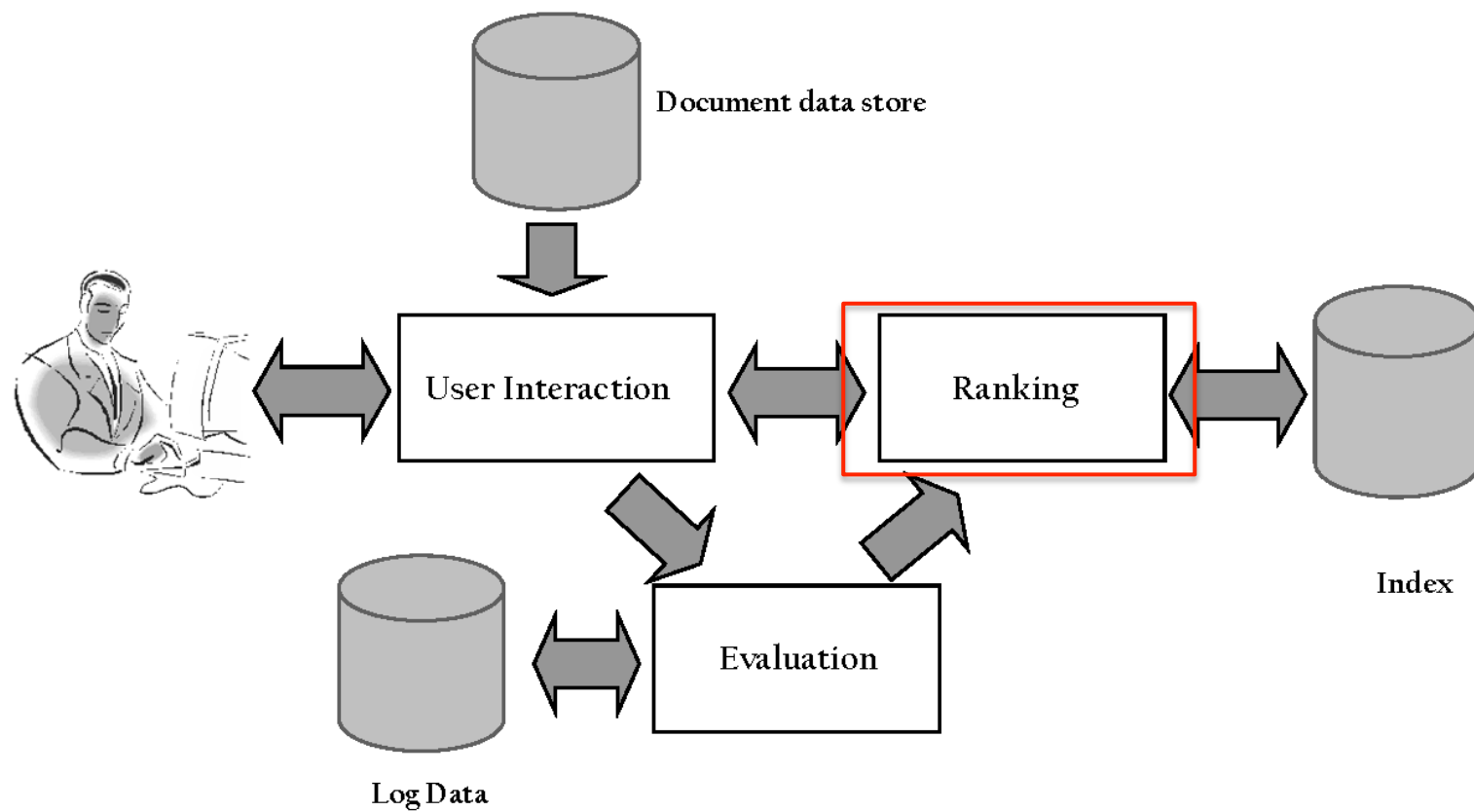
TradeKing - Online Broker
\$4.95 per Trade, Market or Limit
SmartMoney Top **Discount Broker** 2007
www.TradeKing.com

Scottrade Brokerage
\$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades \$1.50 - \$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
www.sogotrade.com

\$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

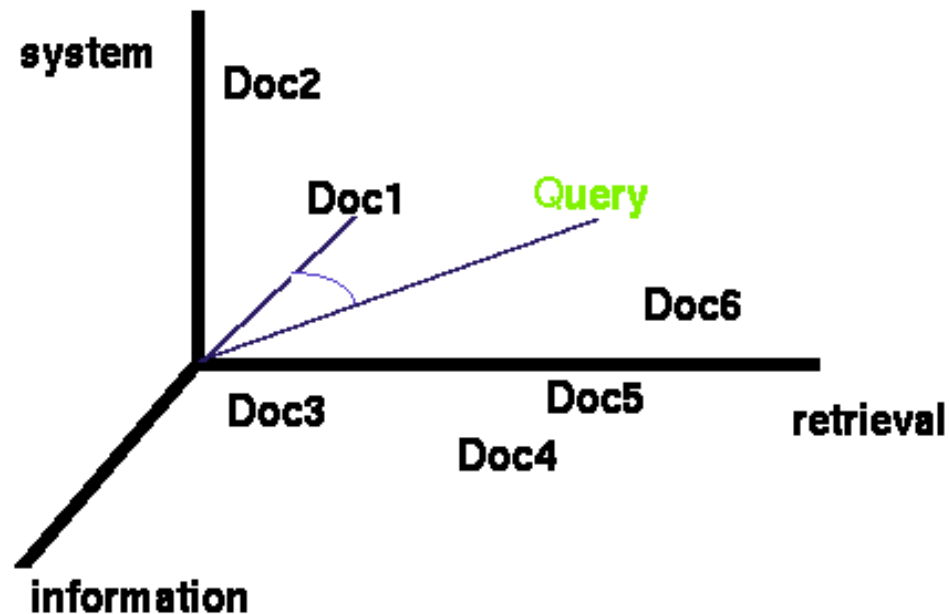
INGDIRECT | ShareBuilder
[Business Office](#) [Contact Us](#)





Ranqueamento

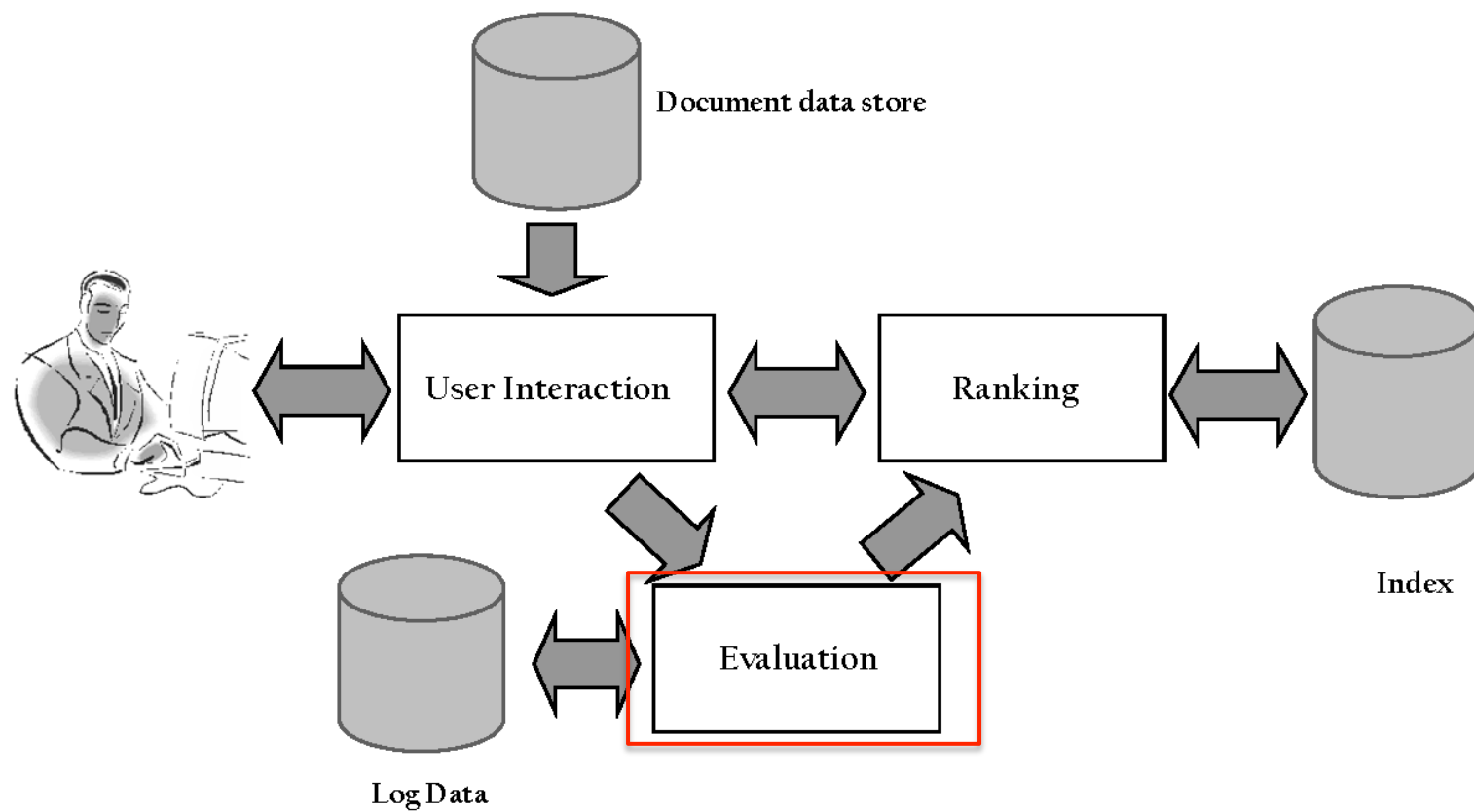
- Calcula o score dos documentos para uma consulta usando um algoritmo de ranqueamento
- Medida de similaridade
- Ex.: Modelo de Espaço de Vetores





Ranqueamento

- Tempo de resposta
- Distribuição
 - Consultas processadas num ambiente distribuído
 - Broker distribui consultas e une os resultados
 - Cache de resultados





Avaliação: Logs

- Logar consultas e interações

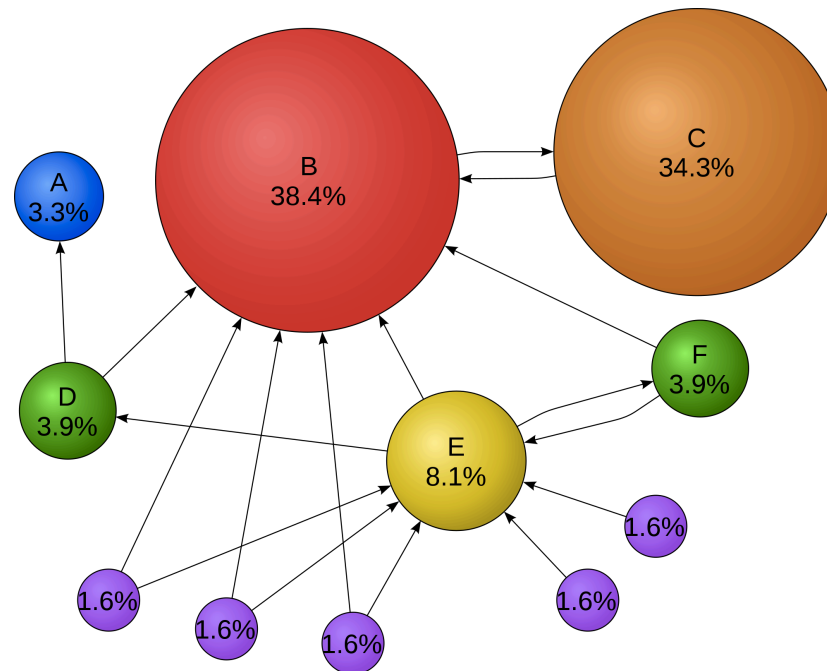
AnonID	Query	QueryTime	ItemRank	ClickURL
1268	ozark horse blankets	2006-03-01 17:39:28	8	http://www.blanketsnmore.com
1268	www.ghostrockranch.com	2006-03-04 13:58:23		
1268	openrangeht.zachsairforce.com	2006-03-09 22:38:45		
1268	sstack.com	2006-03-11 00:17:09		
1268	www.mecab.org	2006-03-12 18:59:26		
1268	www.raindanceexpress.com	2006-03-18 20:13:01		
1268	www.victoriacostumiere.com	2006-03-19 00:26:51		
1268	osteen-schaztberg.com	2006-03-21 17:55:25		
1268	osteen-schatzberg.com	2006-03-21 17:55:42	1	http://www.osteen-schatzberg.com
1268	osteen-schatzberg.com	2006-03-21 17:55:42	2	http://www.osteen-schatzberg.com
1268	www.buckmountianestates.com	2006-03-24 18:53:10		
1268	idx.techsolsc.com	2006-05-07 00:58:21		
1268	www.bridleandhit.com	2006-05-09 21:34:23		

- Melhoraria da qualidade dos resultados
- Usado para sugestão de consultas, cache de consultas, ranqueamento



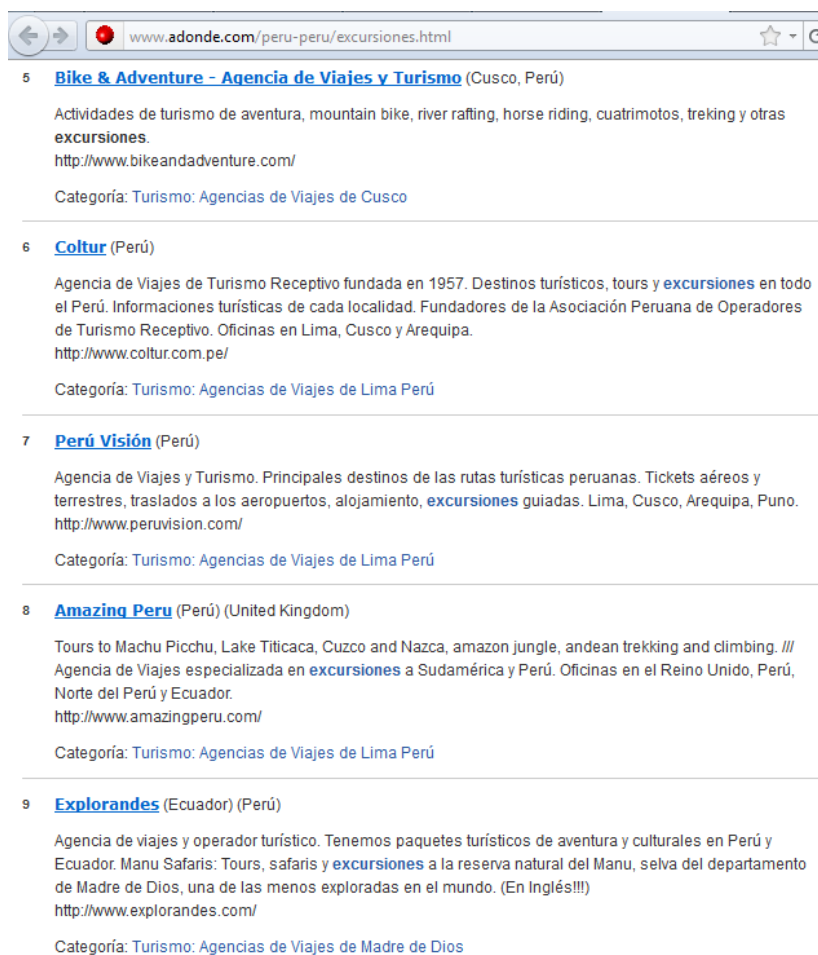
Análise de Links

- Links ajudam a identificar popularidade: ex. PageRank



- Ex.:
<http://doheth.co.uk/info/list-of-web-sites-with-high-page-rank.php>

- Muito úteis para consultas navegacionais



5 [Bike & Adventure - Agencia de Viajes y Turismo](#) (Cusco, Perú)
Actividades de turismo de aventura, mountain bike, river rafting, horse riding, cuatrimotos, trekking y otras **excursiones**.
<http://www.bikeandadventure.com/>
Categoría: [Turismo: Agencias de Viajes de Cusco](#)

6 [Coltur](#) (Perú)
Agencia de Viajes de Turismo Receptivo fundada en 1957. Destinos turísticos, tours y **excursiones** en todo el Perú. Informaciones turísticas de cada localidad. Fundadores de la Asociación Peruana de Operadores de Turismo Receptivo. Oficinas en Lima, Cusco y Arequipa.
<http://www.coltur.com.pe/>
Categoría: [Turismo: Agencias de Viajes de Lima Perú](#)

7 [Perú Visión](#) (Perú)
Agencia de Viajes y Turismo. Principales destinos de las rutas turísticas peruanas. Tickets aéreos y terrestres, traslados a los aeropuertos, alojamiento, **excursiones** guiadas. Lima, Cusco, Arequipa, Puno.
<http://www.peruvision.com/>
Categoría: [Turismo: Agencias de Viajes de Lima Perú](#)

8 [Amazing Peru](#) (Perú) (United Kingdom)
Tours to Machu Picchu, Lake Titicaca, Cuzco and Nazca, amazon jungle, andean trekking and climbing. /// Agencia de Viajes especializada en **excursiones** a Sudamérica y Perú. Oficinas en el Reino Unido, Perú, Norte del Perú y Ecuador.
<http://www.amazingperu.com/>
Categoría: [Turismo: Agencias de Viajes de Lima Perú](#)

9 [Explorandes](#) (Ecuador) (Perú)
Agencia de viajes y operador turístico. Tenemos paquetes turísticos de aventura y culturales en Perú y Ecuador. Manu Safaris: Tours, safaris y **excursiones** a la reserva natural del Manu, selva del departamento de Madre de Dios, una de las menos exploradas en el mundo. (En Inglés!!!)
<http://www.explorandes.com/>
Categoría: [Turismo: Agencias de Viajes de Madre de Dios](#)