

# Inteligência Artificial

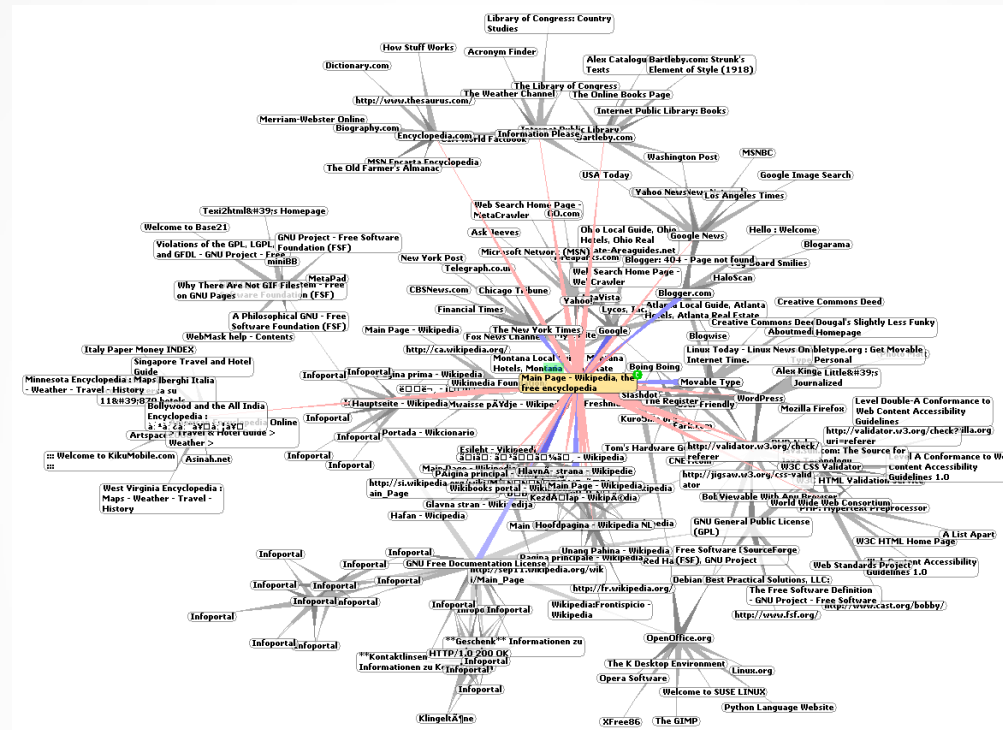
## Recuperação de Informação

Prof. Fabio Augusto Faria

1º semestre 2015



# World Wide Web (WWW)



Existem centenas de milhões de paginas na web tratando de variados assuntos.

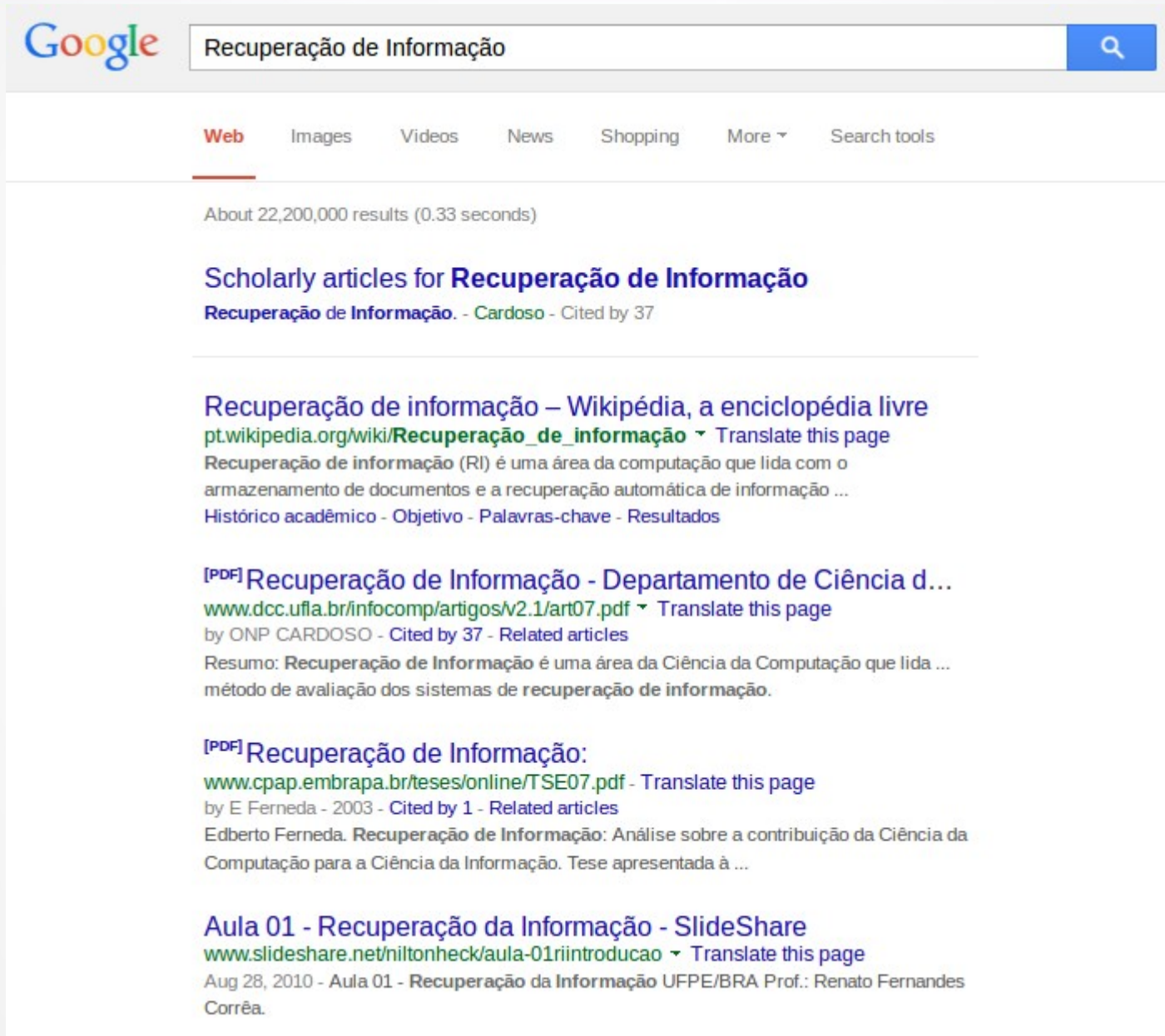
Extraído de [1].

# World Wide Web (WWW)

Dado esse grande número de paginas, como encontrar de forma eficiente a informação desejada?



# Sites de Busca



The image shows a Google search results page. At the top, the Google logo is on the left, and the search bar contains the text 'Recuperação de Informação'. To the right of the search bar is a blue button with a magnifying glass icon. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'News', 'Shopping', 'More', and 'Search tools'. The 'Web' tab is selected and underlined. Below the tabs, it says 'About 22,200,000 results (0.33 seconds)'. The first result is 'Scholarly articles for Recuperação de Informação' with a sub-link 'Recuperação de Informação. - Cardoso - Cited by 37'. The second result is 'Recuperação de informação – Wikipédia, a enciclopédia livre' with a sub-link 'pt.wikipedia.org/wiki/Recuperação\_de\_informação' and a 'Translate this page' link. The third result is '[PDF] Recuperação de Informação - Departamento de Ciência d...' with a sub-link 'www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf' and a 'Translate this page' link. The fourth result is '[PDF] Recuperação de Informação:' with a sub-link 'www.cpap.embrapa.br/teses/online/TSE07.pdf' and a 'Translate this page' link. The fifth result is 'Aula 01 - Recuperação da Informação - SlideShare' with a sub-link 'www.slideshare.net/niltonheck/aula-01riintroducao' and a 'Translate this page' link.

Google

Recuperação de Informação

Web Images Videos News Shopping More Search tools

About 22,200,000 results (0.33 seconds)

Scholarly articles for **Recuperação de Informação**  
**Recuperação de Informação.** - Cardoso - Cited by 37

---

**Recuperação de informação – Wikipédia, a enciclopédia livre**  
[pt.wikipedia.org/wiki/Recuperação\\_de\\_informação](http://pt.wikipedia.org/wiki/Recuperação_de_informação) ▾ Translate this page  
Recuperação de informação (RI) é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação ...  
Histórico acadêmico - Objetivo - Palavras-chave - Resultados

**[PDF] Recuperação de Informação - Departamento de Ciência d...**  
[www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf](http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf) ▾ Translate this page  
by ONP CARDOSO - Cited by 37 - Related articles  
Resumo: **Recuperação de Informação** é uma área da Ciência da Computação que lida ...  
método de avaliação dos sistemas de **recuperação de informação**.

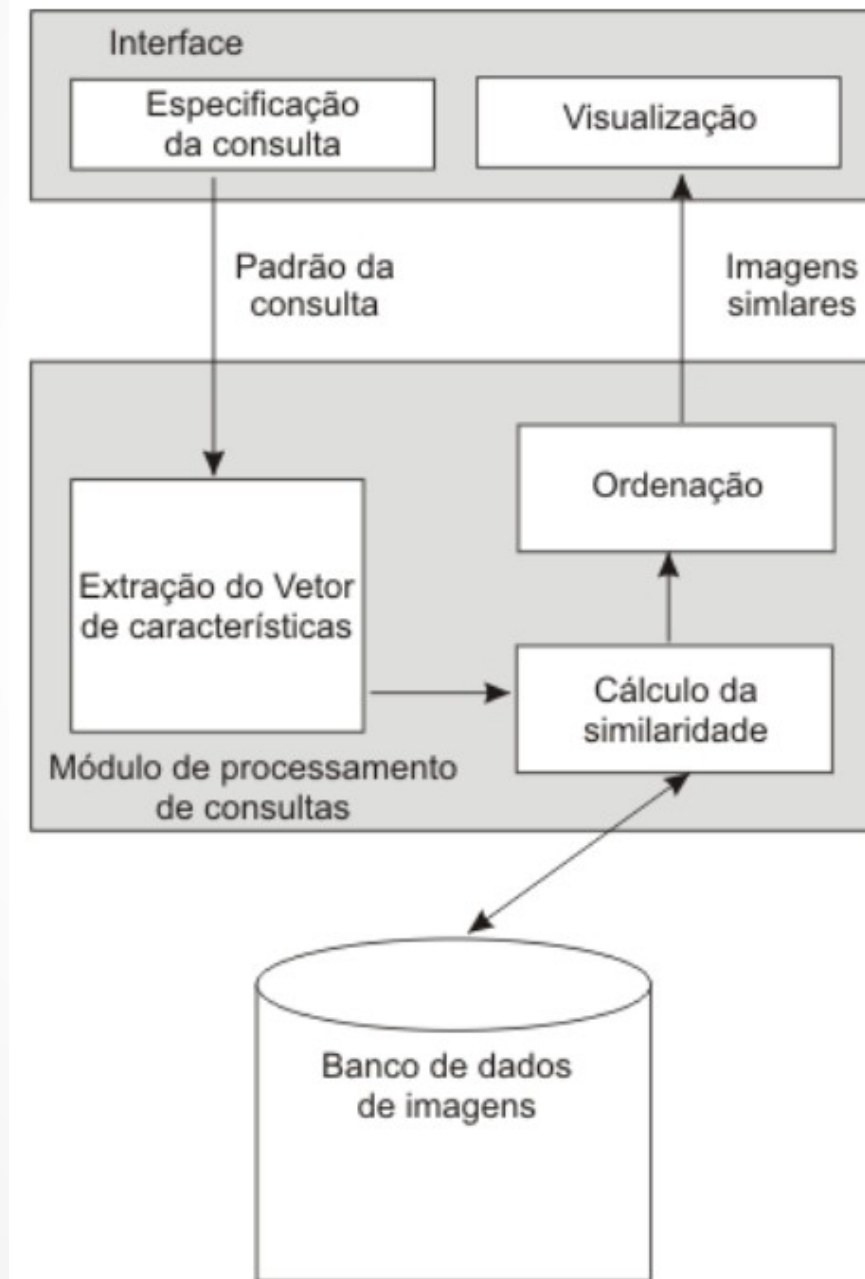
**[PDF] Recuperação de Informação:**  
[www.cpap.embrapa.br/teses/online/TSE07.pdf](http://www.cpap.embrapa.br/teses/online/TSE07.pdf) - Translate this page  
by E Ferneda - 2003 - Cited by 1 - Related articles  
Edberto Ferneda. **Recuperação de Informação**: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. Tese apresentada à ...

**Aula 01 - Recuperação da Informação - SlideShare**  
[www.slideshare.net/niltonheck/aula-01riintroducao](http://www.slideshare.net/niltonheck/aula-01riintroducao) ▾ Translate this page  
Aug 28, 2010 - Aula 01 - **Recuperação da Informação** UFPE/BRA Prof.: Renato Fernandes Corrêa.

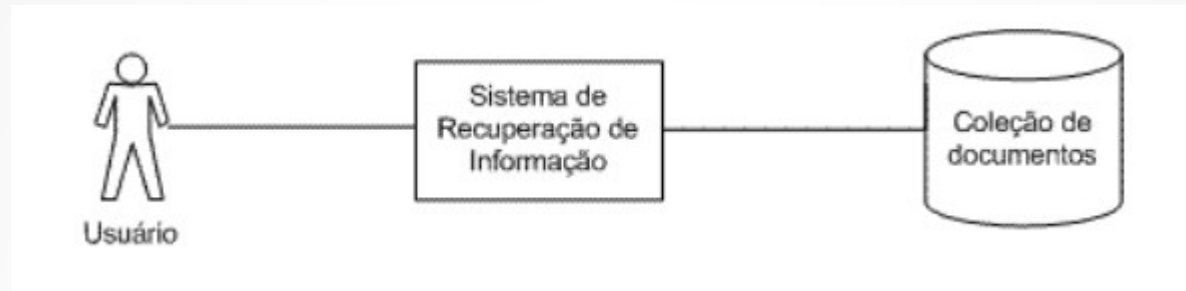
# Recuperação de Informação

- **Corpus de documentos:** é o escopo do sistema (e.g., parágrafo, página ou textos);
- **Consulta:** uma consulta específica sobre o que o usuário quer saber;
- **Resultados:** subconjunto de documentos que o sistema julga ser relevante para a consulta;
- **Apresentação dos resultados:** geralmente uma lista ordenada, mas pode ser utilizadas técnicas de visualização.

# Recuperação de Informação (RI)

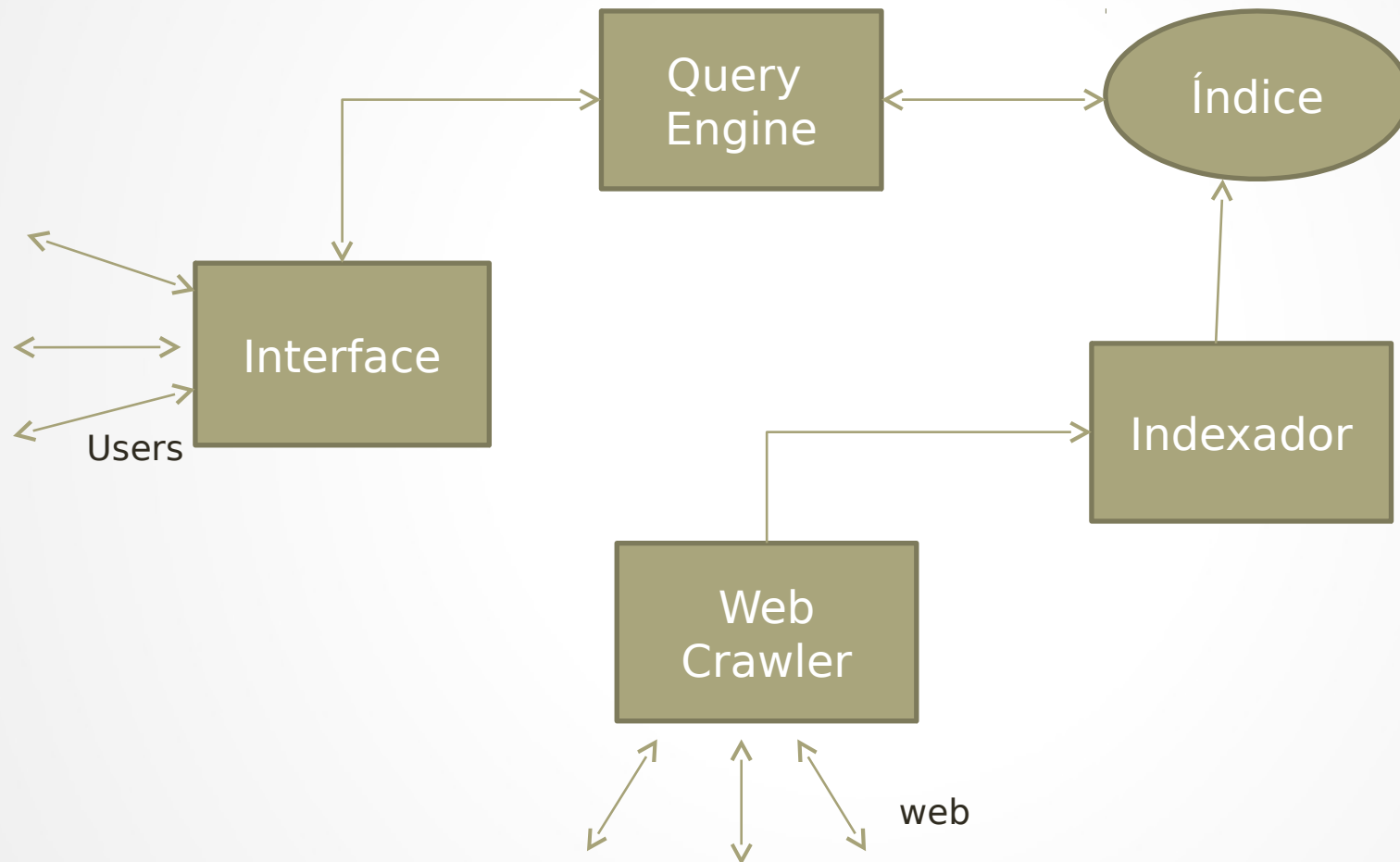


# Recuperação de Informação (RI)



- Dois problemas diferentes e igualmente desafiadores:
  - recuperar informação de forma **eficiente**;
  - estimar a **relevância** dos documentos recuperados para a ordenação do conjunto resposta.

# RI na Web - Engenheiros de Busca





# História

- ALIWEB (outubro de 1993)
  - Desenvolvido por Martijn Koster
  - Permitia que usuários submetessem suas paginas WEB juntamente com a descrição e *keywords*
  - Poucas pessoas submetiam suas paginas o que fazia com que o ALIWEB não fosse muito utilizado

# História

- JumpStation (dezembro de 1993)
  - Primeiro engenho de busca a usar um agente na web (web crawler) para encontrar as paginas
  - Devido a limitações de recursos, o processo de indexação e, conseqüentemente, de busca era limitado ao titulo e cabeçalho das paginas encontradas.

# História

- AltaVista (1995)
  - Possuía recursos de rede ilimitados para época.
  - Primeiro a fornecer recursos de pesquisa em vários idiomas,
  - Primeiro a permitir busca por conteúdo multimídia
  - 300 mil requisições no primeiro dia e após 1 ano recebia 19 milhões de requisições por dia



# História

- Yahoo! (1994)
  - Um dos principais sites utilizados para busca na internet.
  - Funcionava como um diretório web, onde as páginas web eram divididas em categorias e subcategorias.
  - Em 2000 Yahoo comprou a Inktomi e utilizou o seu sistema de busca até 2004, quando lançou seu próprio sistema.
  - Em julho de 2009 o Yahoo! entrou em acordo com a Microsoft para utilizar o Bing como sistema de busca.

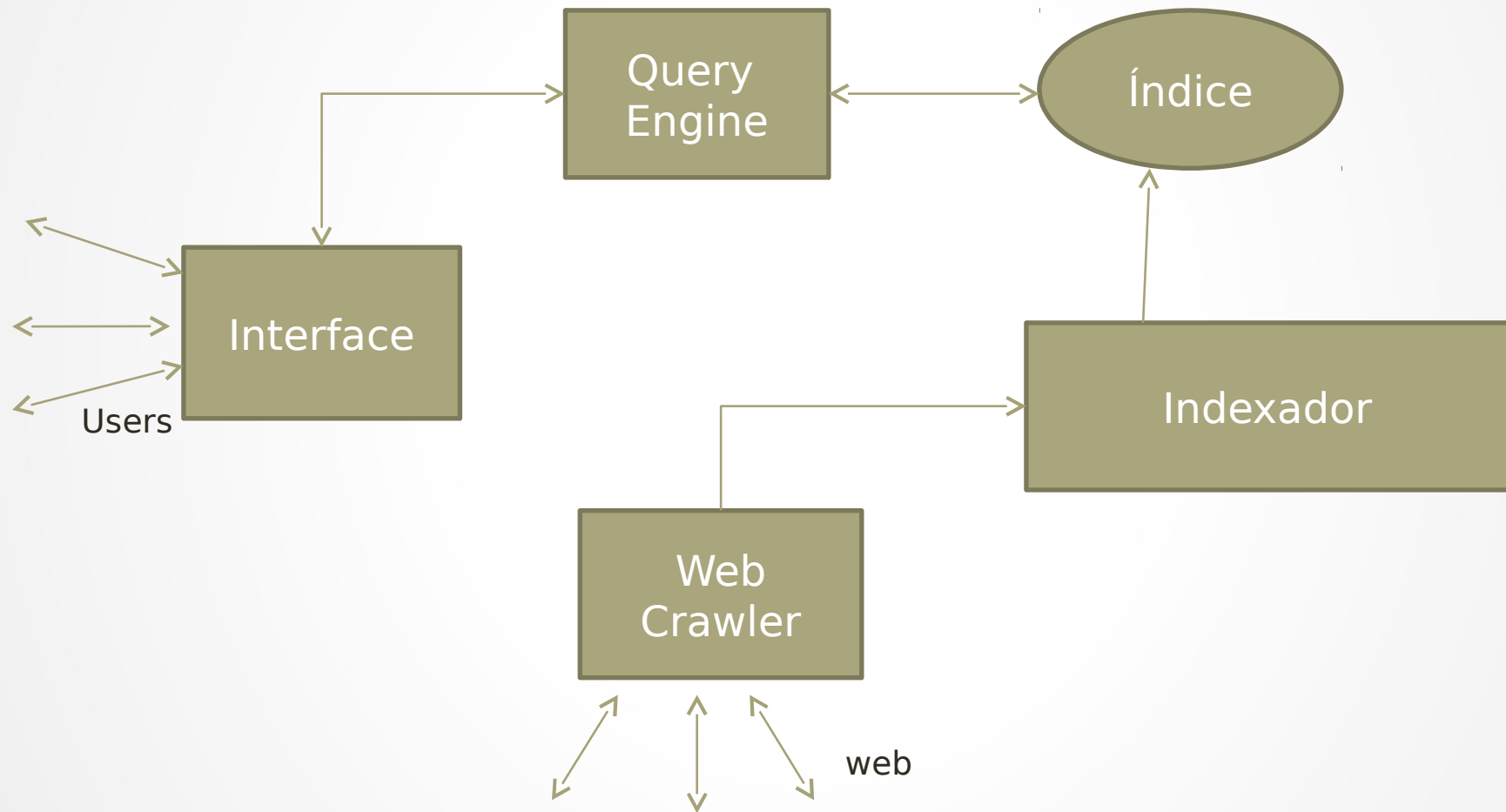


# História

- Google (2000)
  - Melhores resultados que os concorrentes ao utilizar o algoritmo **PageRank**
  - Diferentemente de seus concorrentes, fornecia uma interface de busca simples ao invés de um portal web
  - Se tornou o maior engenho de busca do mundo.



# RI na Web - Engenheiros de Busca

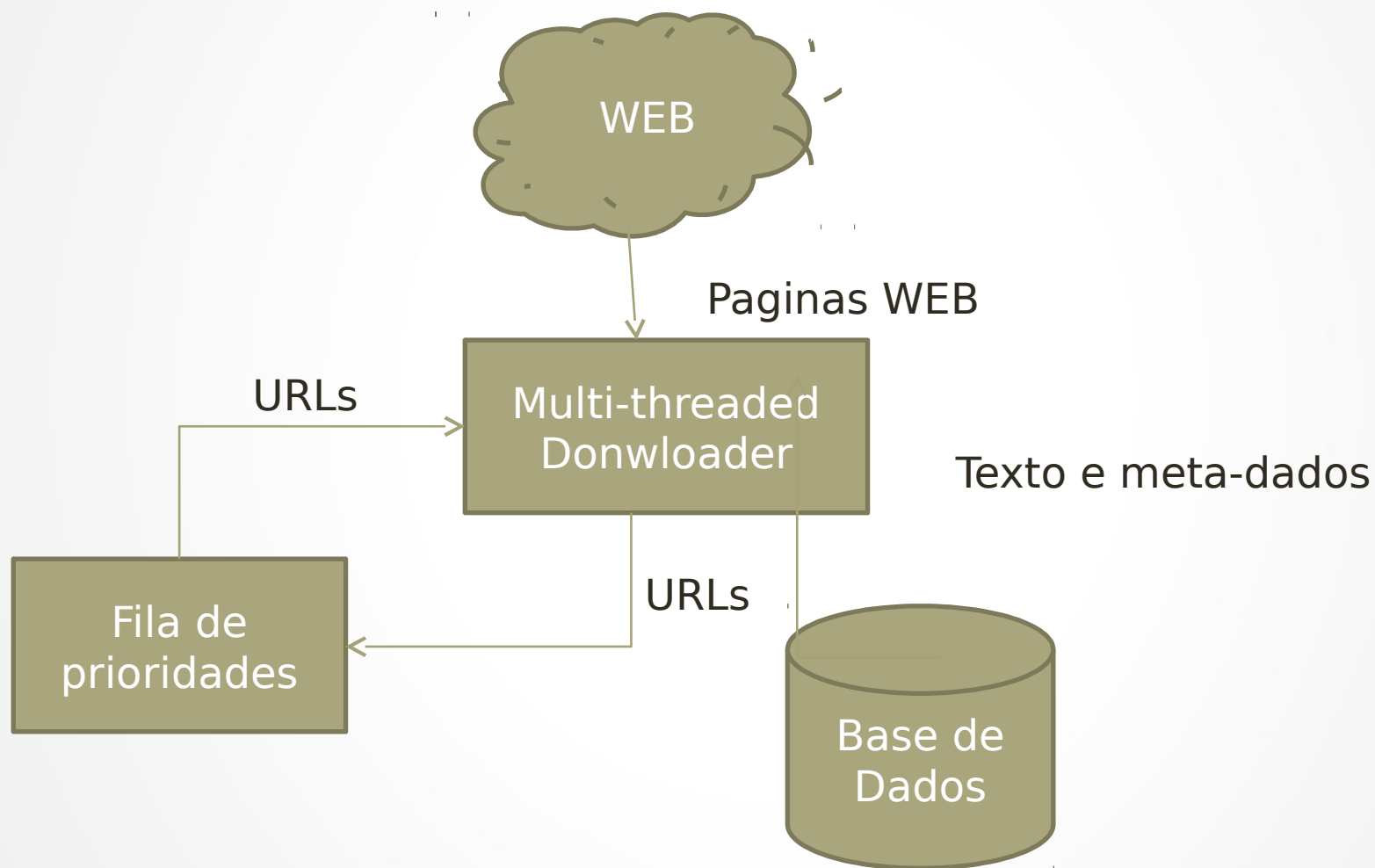


# Web Crawler

- Agente que navega pela web de maneira automática e sistematicamente;
- Captura informações que serão utilizadas na etapa de indexação.
- Utiliza um conjunto de URL iniciais e segue para outra páginas através de hiperlinks



# Arquitetura Geral Web Crawler





# Indexação

- Processo de atribuir termos/códigos a um documento com objetivo de recuperá-lo mais rapidamente quando necessário;
- “Índices são, portanto, estruturas de dados auxiliares cujo único propósito é tornar mais rápido o acesso a registros baseado em certos campos, chamados campos de indexação”;
- O índice possui uma descrição sobre a pagina como: data de criação, tamanho, o titulo e as primeiras linhas;
- “Atualmente os **engenhos de busca** oferecem buscas por frases, porém detalhes dessa funcionalidade não é publicamente conhecida”, Berthier Ribero-Neto.

# Modelo de Recuperação de Informação

## ▮ Definição Formal de modelo em IR:

– É definido pela quádrupla [ **D**, **Q**, **f**,  $r(q_i, d_j)$  ]

**D** - é um conjunto composto por representações para os documentos em uma coleção

**Q** - é um conjunto formado por representações (consultas) para uma necessidade de informação do usuário

**f** - é um arcabouço para modelagem de representações de documentos, consultas e seus relacionamentos

$r(q_i, d_j)$  - é uma função de ordenação que associa um número real a uma consulta  $q_i \in Q$  uma representação de documento  $d_j \in D$  para ordenar os documentos de acordo com a consulta.

# Modelo de Recuperação de Informação

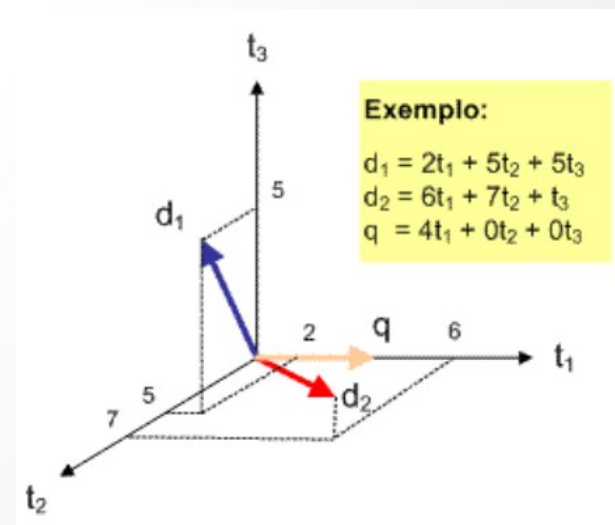
- Booleano
- TF-IDF
- BM25 (Projeto Okapi)
- PageRank (Google)
- HITS (Hiperlink Induced Topic Search)

# Modelo Booleano (palavra chave)

- Mais simples de entender e implementar;
- Cada palavra do documento é tratada como uma característica booleana (V ou F);
- Não reflete o grau de relevância de um documento com relação aos outros (apresentação dos resultados);
- Consulta com mais de um termo de-se usar **AND** e **OR**.

# Modelo Vetorial TF-IDF

- **TF (term frequency)**: é a frequência de um termo em um documento ou o número de vezes que um termo **ki** ocorre em um documento **dj**;
- **IDF (inverse document frequency)**: é o inverso da frequência do documento ou o número de documentos nos quais um termo **ki** é encontrado, considerando **toda** uma coleção de documentos.



# Modelo Vetorial TF-IDF

$$w_{i,j} = f(tf_{i,j}) \times idf_i = (1 + \log tf_{i,j}) \times \log \left( 1 + \frac{N}{df_i} \right) \quad (2.2)$$

onde  $tf_{i,j}$  é a frequência de um termo  $k_i$  em um documento  $d_j$ ,  $N$  é o número de documentos da coleção e  $df_i$  é o número de documentos onde um termo  $k_i$  ocorre;

$$w_{i,q} = f(tf_{i,q}) \times idf_i = (1 + \log tf_{i,q}) \times \log \left( 1 + \frac{N}{df_i} \right) \quad (2.3)$$

onde  $N$  é o número de documentos da coleção,  $df_i$  é o número de documentos onde um termo  $k_i$  ocorre e  $tf_{i,q}$  é o número de ocorrências de um termo  $k_i$  em uma consulta  $q$ .

# Modelo Probabilístico BM25

“Modelo probabilístico tenta estimar a relevância de um documento baseada na idéia de que os termos das consultas têm diferentes distribuições nos documentos relevantes e não-relevantes” [2]

# Modelo Probabilístico BM25

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{\text{TF}(q_i, D) \cdot (k_1 + 1)}{\text{TF}(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

$$\text{IDF}(q_i) = \log \frac{N - \text{DF}(q_i) + 0.5}{\text{DF}(q_i) + 0.5},$$

**k** e **b** são parâmetros que podem ser encontrados por validação cruzada ( $k=2$  e  $b=0.75$ );  
**DF** é número de documentos que o termo  $q_i$  ocorre;  
**avgdl** é o comprimento médio do documento no corpus.



# PageRank

- Criado por um dos fundadores da Google, Larry Page;
- Estima a importância de um site pela contagem e qualidade dos links da página;
- Um site de alta qualidade está ligado a outros sites de alta qualidades;

# PageRank

- Dado que o usuário está em uma determinada pagina é possível:
  - Pular pra uma pagina aleatória com probabilidade  $d$
  - Seguir um dos hiperlinks da pagina com probabilidade  $1 - d$
- A pagina  $p_i$  é apontada pelas paginas  $p_j$
- $L(p_j)$  = N° de hyperlinks de saída em  $p_j$
- $N$  é o número total de páginas consideradas
- $M(p_i)$  é o número de páginas que apontam  $p_i$

$$PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

# HITS - Hypertext Induced Topic Search

- Cada pagina possui dois scores
  - **Autoridades** = Paginas que são apontadas por muitas outras
  - **Hubs** = Paginas que apontam para muitas outras.



$$H(p) = \sum_{p \rightarrow u} A(u)$$

$$A(p) = \sum_{v \rightarrow p} H(v)$$

# Sistema de Extração de Informações

## Texto Livre

4 de abril em Dallas – cedo na noite passada, um tornado varreu todo o noroeste da área de Dallas, causando extensos danos. Testemunhas confirmam que o ciclone passou sem advertência, aproximadamente às 7:15 da noite, e destruiu dois *motor-homes*. O posto Texaco, na Rua Principal, 102, Farmers Branch, TX, também foi severamente danificado, mas nenhuma morte foi informada. O valor total calculado dos danos é de US\$200.000.

## Jornal



## Sistema de Extração de Informações

## Template

Evento: tornado  
Data: 4/4/2000  
Hora: 19:15  
Local: Farmers Branch : "noroeste de Dallas" : TX : USA  
Danos: "*motor-homes*" (2) : "Posto Texaco" (1)  
Perdas Estimadas: US\$200.000  
Mortes: nenhuma

FIGURA 3.1 – *Template* de um sistema de extração de informações.

# Sistema de Extração de Informações

## **Baseados em PLN**

Extrair informações de textos em linguagem natural (livre)

Padrões lingüísticos

# Processamento de Linguagem Natural

## Processo de extração

Extração de fatos (unidades de informação)

- Através da análise local do texto

Integração e combinação de fatos

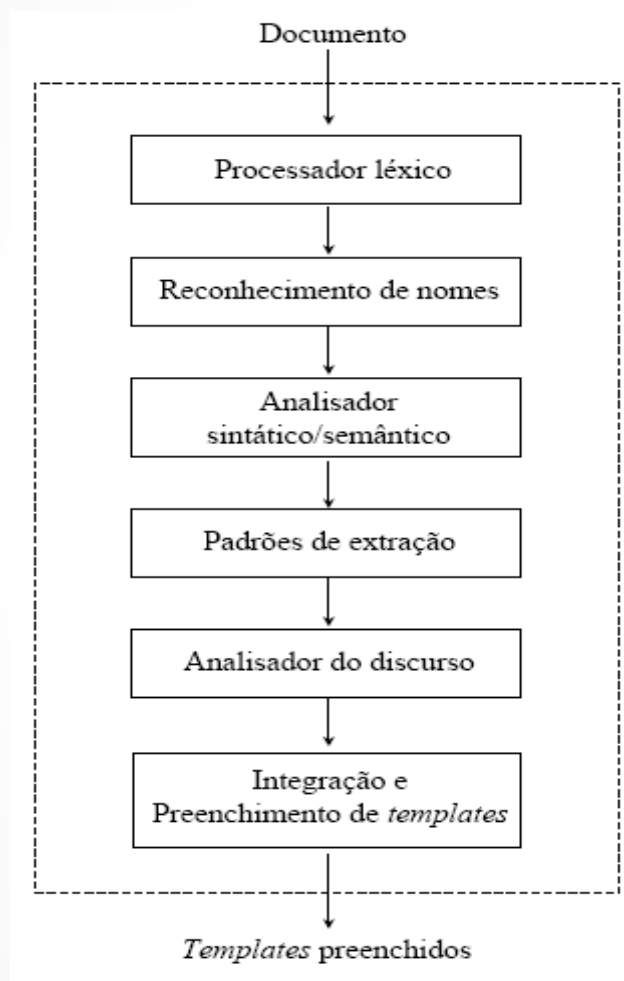
- Produção de fatos maiores ou novos fatos

Estruturação de fatos relevantes

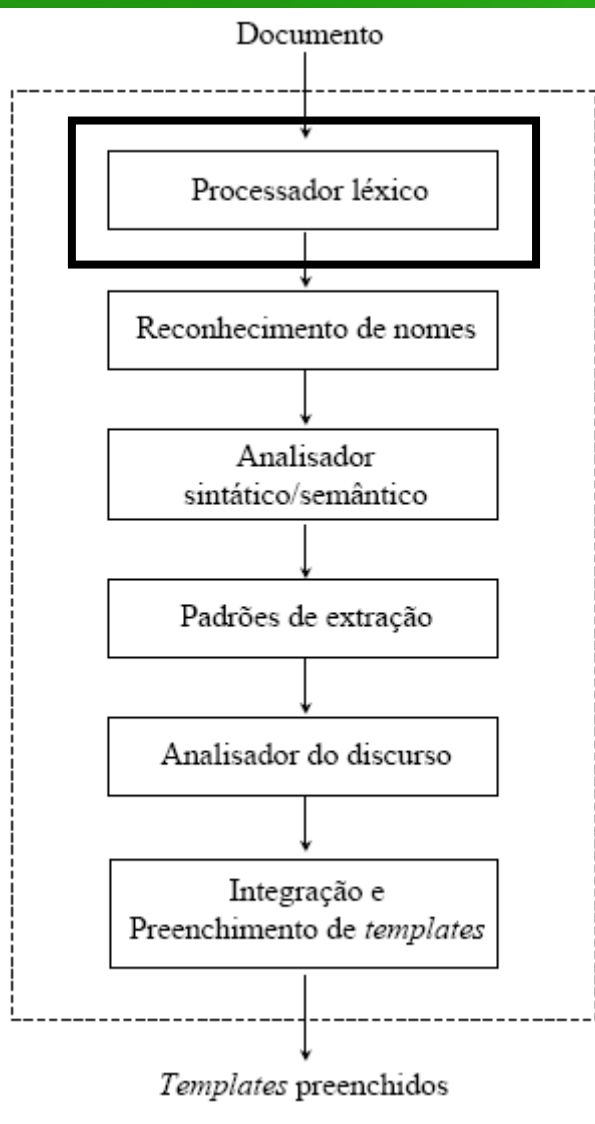
- Padrão de saída

# Processamento de Linguagem Natural

## Arquitetura



# Processador Léxico



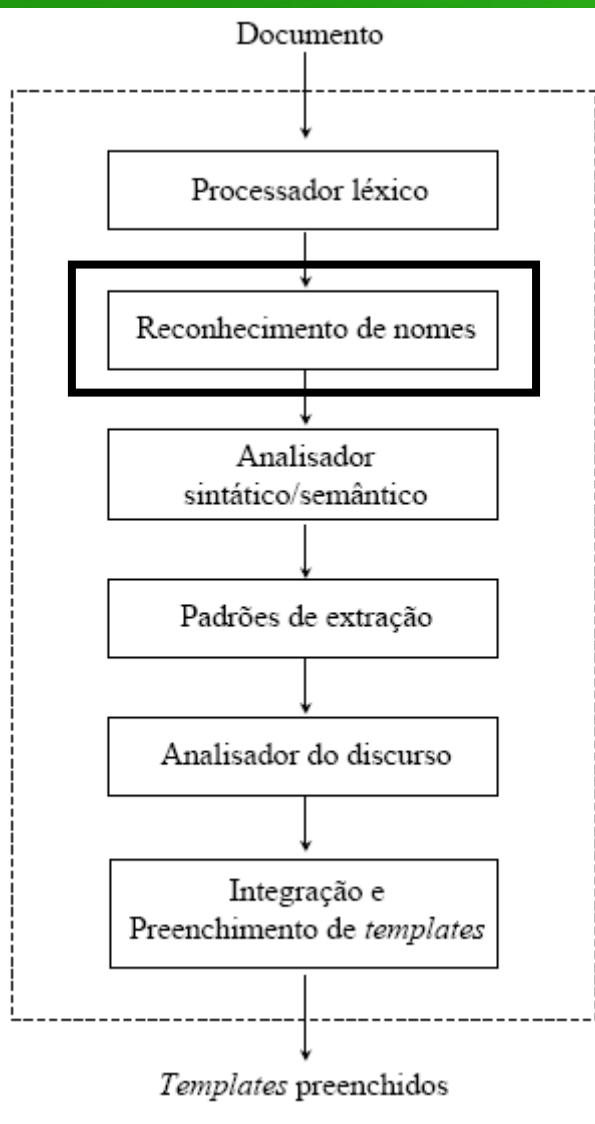
Separação dos termos (*tokenization*) pelo reconhecimento de espaços em branco e sinais de pontuação que delimitam o texto;

Análise léxica e morfológica dos termos para determinar suas possíveis classes (substantivo, verbo, etc.) e outras características (masculino, feminino);

É comum o uso de autômatos finitos para o reconhecimento das informações



# Reconhecimento de Nomes

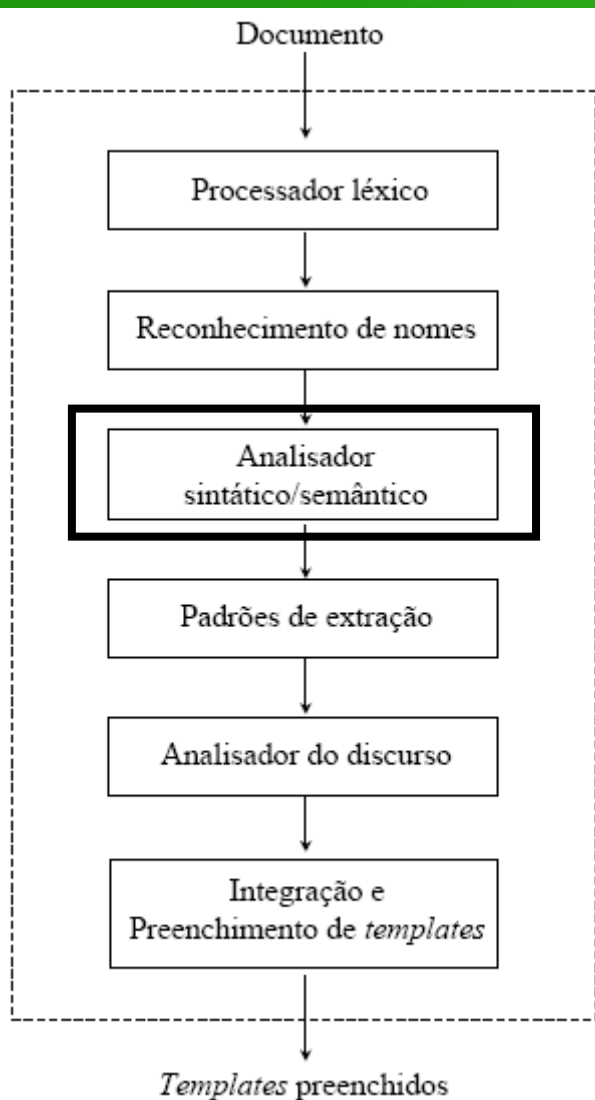


Identifica nomes próprios;

Itens que têm estrutura interna como da data e hora;

Nomes são identificados por expressões regulares expressos em função das classes morfosintáticas (part-of-speech) e características sintáticas e ortográficas (letras maiúsculas) presentes nos termos.

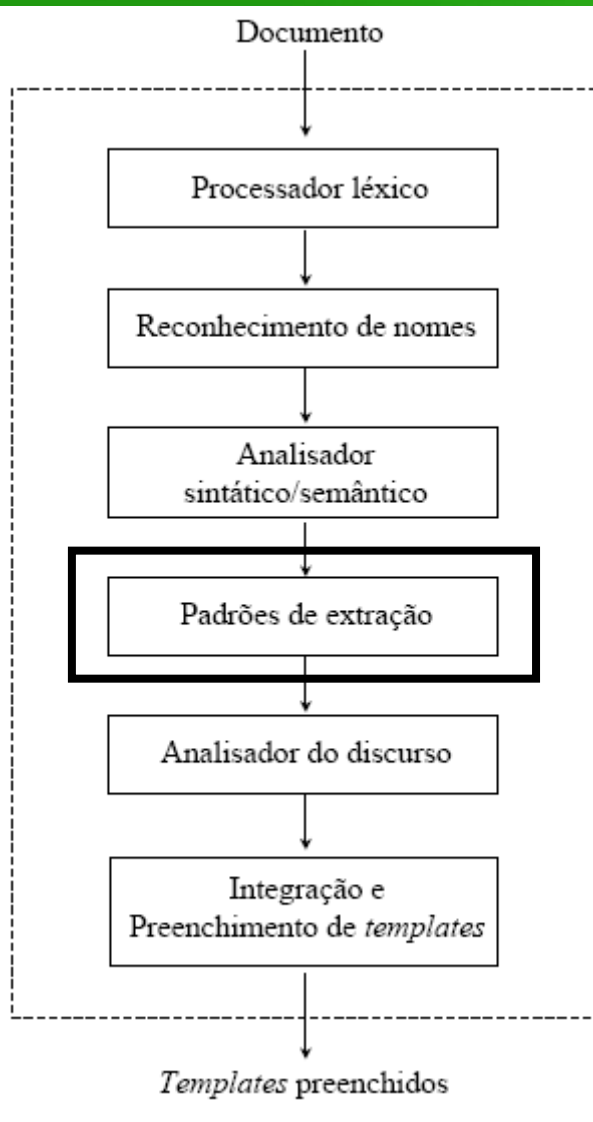
# Analizador sintático/semântico



Recebe uma seqüência de itens léxicos e tenta construir uma estrutura sintática junto com alguma semântica;

Identifica os segmentos de texto e para cada um associa alguma característica que podem ser combinadas na fase seguinte.

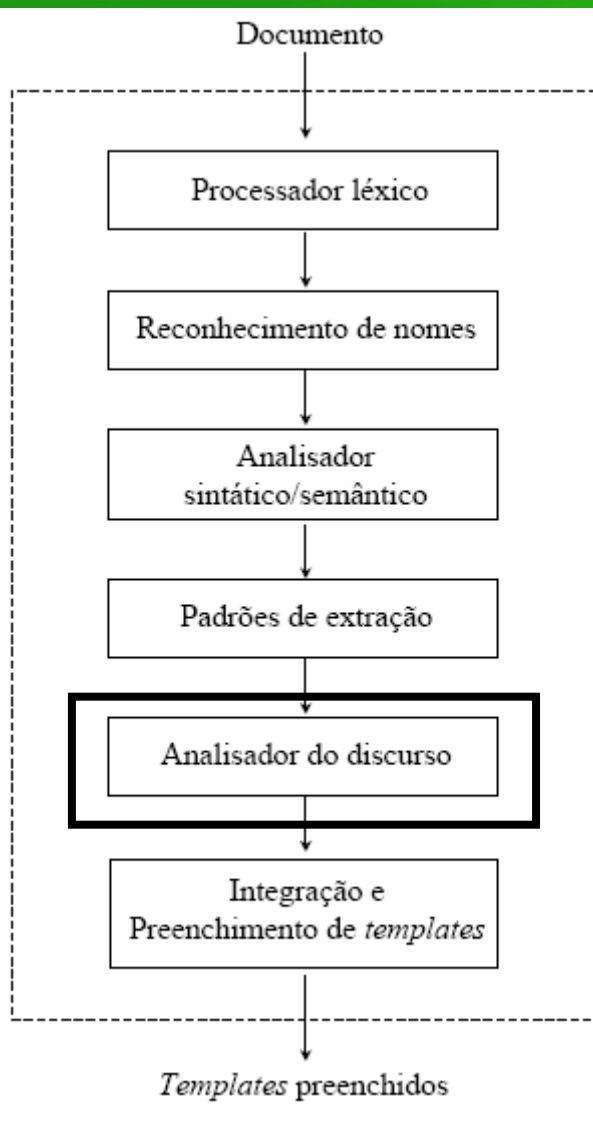
# Padrões de extração



Consiste na indução de um conjunto de regras de extração para o domínio tratado;

Esses padrões baseiam-se em restrições sintáticas e semânticas aplicadas as sentenças.

# Analizador do Discurso

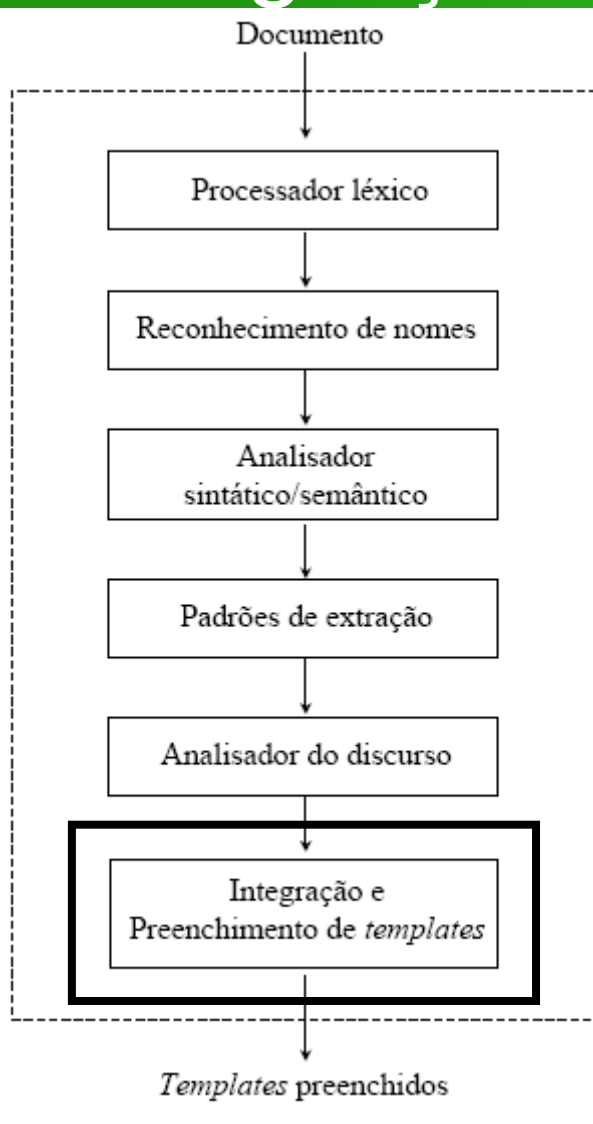


Relaciona diferentes elementos do texto;  
Análise de frases nominais, reconhece apostos e outros grupos nominais complexos;

Resolução de conferência, identifica quando uma frase nominal se refere a outra já citada;

Descoberta de relacionamento entre as partes do texto, para estruturar palavras do texto em uma rede associativa.

# Integração e Preenchimento de templates



As informações são combinadas

Os templates são preenchidos com as informações relevantes ao domínio

# Como Avaliar um SRI?

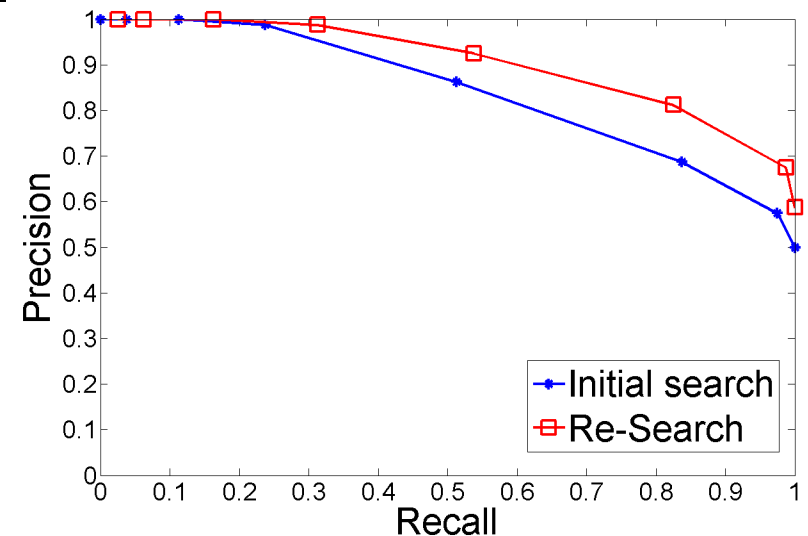
- Precisão (Precision)

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

- Revocação (Recall)

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- Curva Precision x Recall



# Referência

- 1- Aulas Profa. Flávia de Almeida Barros em <http://www.cin.ufpe.br/~in1152/aulas/> .**
- 2- Dissertação de H. M. de Almeida (UFMG). Uma Abordagem de Componentes Combinados para a Geração de Funções de Ordenação usando Programação Genética, 2007.**
- 3- Dissertação de C. Zambenedetti (UFRGS). Extração de Informação sobre Bases de Dados Textuais, 2002.**
- 4- R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval: The concepts and technology behind search.**

# RI

- “Recuperação de informação lida com os aspectos de representação, organização e acesso à informação e sua especificação para busca, e tanto para humanos quanto para máquinas que são empregadas para realizar essas tarefas” (Baeza-Yates e Ribeiro-Neto, 1999)
- apud Saracevic, 1999)
- Para Baeza-Yates e Ribeiro-Neto [3], recuperação de informação é o processo de lidar com a representação, armazenamento, organização e acesso à informação, de modo que os usuários possam acessar de forma fácil a informação na qual estão interessados.



# RI

- Um tópico essencial em RI é o conceito de **relevância**;
- **Relevância** está associada à necessidade de informação de um u
- a relevância descreve o grau de aceitação ou rejeição de um documento retornado por um sistema em relação a uma consulta fornecida por um usuário.