Recuperação de Informação – conceitos iniciais

Paulo Trigo Silva

Uma definição

Recuperação de Informação (RI), ou

Extracção de Informação (EI), ou

Information Retrieval (IR)

Encontrar material (em geral documentos) de natureza não estruturada (em geral texto) que satisfaça determinada necessidade de informação contida em colecções de grande dimensão (em geral armazenadas em computadores).

[An Introduction to Information Retrieval; Manning, Raghavan, Schutze; disponível em `http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html`]

Managing Gigabytes, by I. Witten, A. Moffat, and T. Bell.

Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto

... “informação não estruturada (em geral texto)”

- Mas o texto tem estrutura (e.g. capítulo, secção, parágrafo, etc)
  - ... qualquer frase tem estrutura (sintagma nominal, verbal, etc)!
- ... pois, mas um texto segue uma estrutura geral
  - uma estrutura que não depende de um domínio específico
  - a informação específica do domínio está não (ou semi-) estruturada!
- Mas uma base de dados pressupõe também uma estrutura geral
  - esquemas de relação e restrições de integridade
  - ... será que afinal como qualquer texto?
- ... na base de dados o domínio é “enquadrado” na estrutura
  - a informação específica do domínio está estruturada!
  - é mais apropriada ao tratamento sistemático por um computador
  - ... o texto está mais próximo do homem do que a base de dados!

O papel da informação em “formato texto”

- Forma natural de codificar o conhecimento
  - é natural “escrever o que se pensa” ou “escrever o que se conclui”
  - ... é em texto que se partilha informação e conhecimento
  - ... é em texto que se regista a história e se prevê o futuro!
- Há “excesso de informação” em formato texto!
  - jornais representam 25 terabyte por ano
  - revistas representam 10 terabyte por ano
  - documentos “Office” representam 195 terabyte por ano
- As conversas também são informação em formato texto
  - estima-se 610 biliões emails por ano; cerca de 11000 terabytes\ano
- Há ainda literatura científica, documentos do governo, ...

Alguns desafios

- Como encontrar informação útil?
- Como organizar a informação?
- Como extrair padrões?

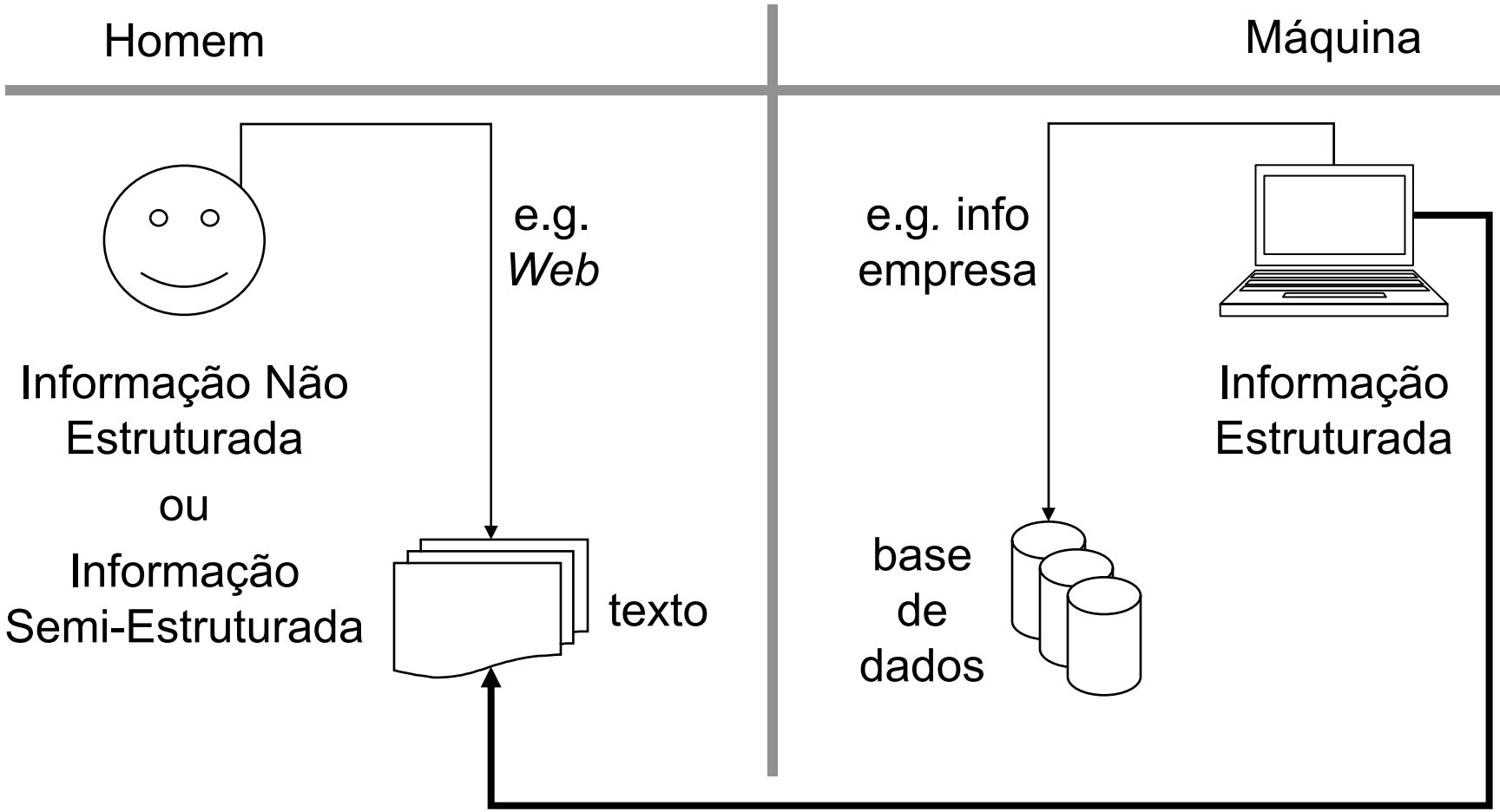
Como gerir a informação textual de forma eficiente?

Paulo Trigo Silva

Exemplos de aplicações para gerir informação textual

- Pesquisa
  - motores de pesquisa na Web (Google, Yahoo,)
  - bibliotecas digitais, ...
- Filtragem & Recomendação
  - notícias
  - recomendações de filmes / música / literatura, ...
- Catalogação
  - encaminhamento automático de e-mail, ...
- Extracção
  - 'Business Intelligence'; Bioinformática, ...
- ...

Informação [Não, ou Semi-] Estruturada

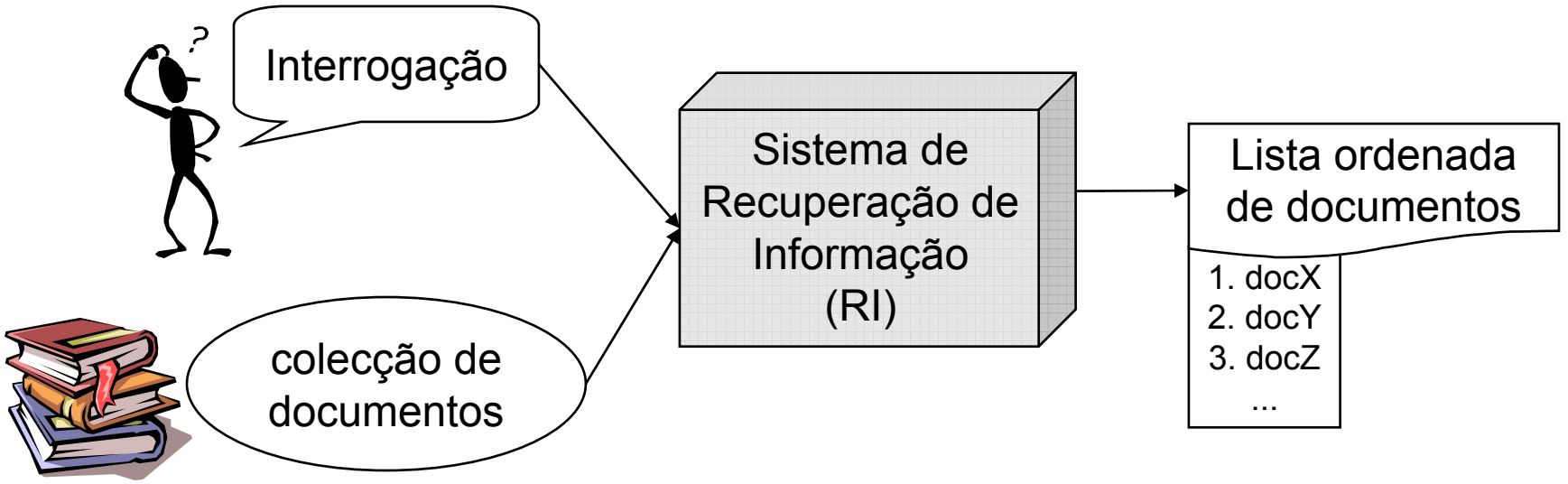


Recuperação de Informação (RI)

Paulo Trigo Silva

Modelo usual de funcionamento

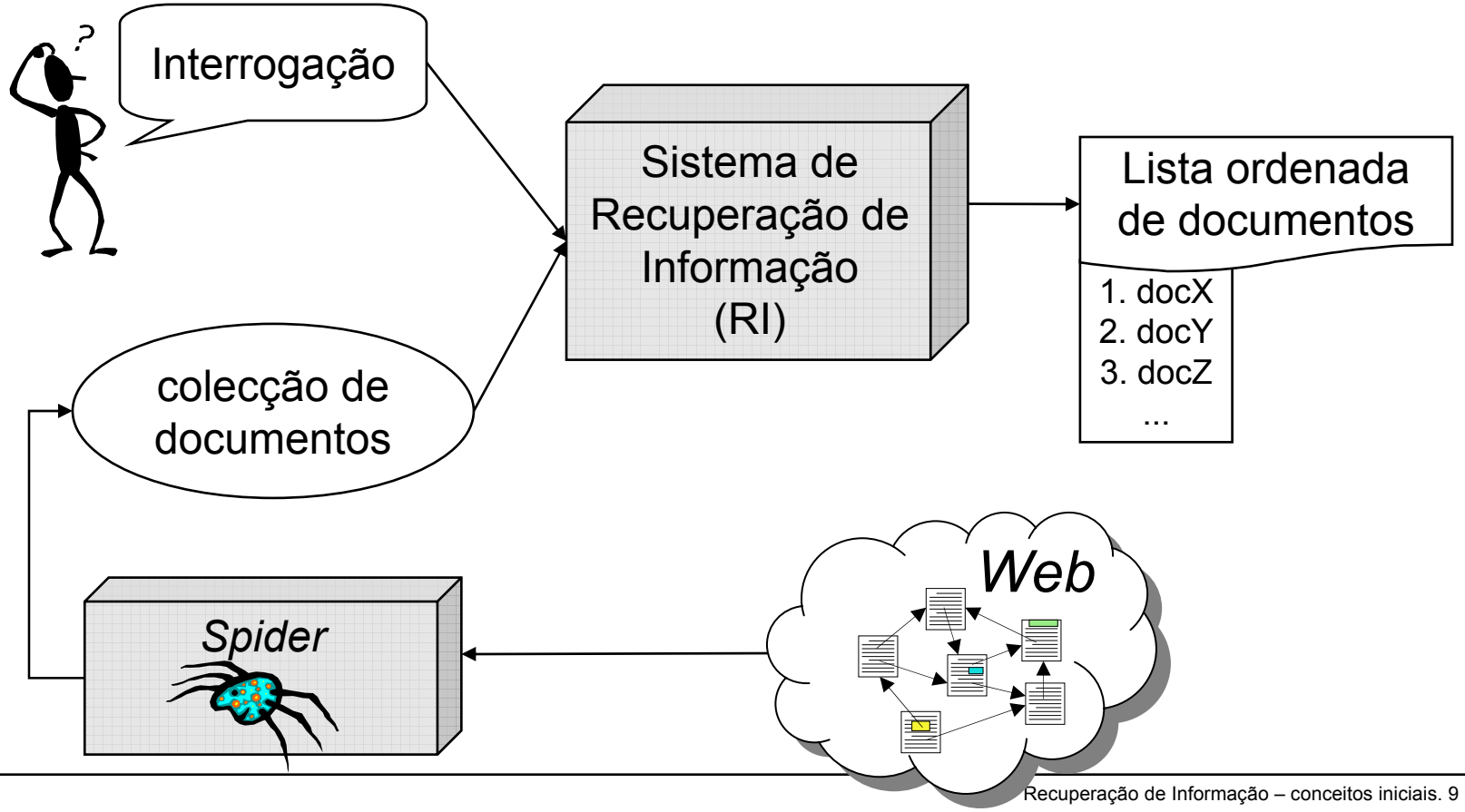
- Entrada:
  - uma colecção de documentos com texto escrito em língua natural
  - uma interrogação do utilizador escrita como sequência de caracteres
- Saída:
  - uma lista ordenada de documentos relevantes para a interrogação



Paulo Trigo Silva

... para pesquisa na Internet

- A colecção de documentos é obtida por “navegação” na Internet
  - recolhendo e armazenando grandes volumes de informação



Características de processamento de um sistema de RI

- Construir respostas rápidas
  - envolvendo grandes quantidades de documentos
- Basear cálculo em operadores flexíveis
  - é impraticável percorrer sempre sequencialmente todos os textos
- Definir relação de ordem entre os documentos identificados
  - o utilizador procura “a melhor resposta” para a interrogação que fez!
- ... oferecer informação relevante
  - mesmo sabendo que o conceito de “relevância” tem diversas leituras!

Aplicação mais recente e mais largamente usada:

procurar páginas na Internet.

Como “oferecer informação relevante”?

- A relevância é uma noção subjectiva e incorpora critérios como
  - o documento incide sobre o tema correcto?
  - o documento está actualizado (refere informação recente)?
  - o documento tem origem numa fonte digna de confiança?
- ... no entanto, o principal critério deve ser
  - o documento satisfaz a necessidade de informação do utilizador?
- O critério mais simples de relevância é o de
  - o documento conter a sequência de caracteres editada pelo utilizador
- Um critério um pouco mais estrito de relevância é o de
  - a sequência de caracteres aparecer frequentemente no documento
  - ... e em qualquer ordem (*bag of words*)

Paulo Trigo Silva

Como “obter informação relevante”?

Observação

Algumas palavras são mais comuns que outras.

Estatística

A maior parte das colecções de documentos (textuais) tem características estatísticas (distribuições palavras) semelhantes.

A distribuição estatística influencia a eficiência e eficácia das estruturas de dados usadas para indexar os documentos.

Diversos modelos de recuperação se baseiam nas propriedades estatísticas dos documentos

Paulo Trigo Silva

Frequência dos temos (exemplo)

Considere-se o texto:

Jamie Callan, Characteristics of Text, 1997

Têm 19 milhões de palavras

Próximo slide mostra as 50 palavras mais frequentes no texto, a sua ordem (*r*) e a sua frequência (*f*).

Paulo Trigo Silva

Frequência dos temos (exemplo)

	f		*f*		*f*
the	1,130,021	from	96,900	or	54,958
of	547,311	he	94,585	about	53,713
to	516,635	million	93,515	market	52,110
a	464,736	year	90,104	they	51,359
in	390,819	its	86,774	this	50,933
and	387,703	be	85,588	would	50,828
that	204,351	was	83,398	you	49,281
for	199,340	company	83,070	which	48,273
is	152,483	an	76,974	bank	47,940
said	148,302	has	74,405	stock	47,401
it	134,323	are	74,097	trade	47,310
on	121,173	have	73,132	his	47,116
by	118,863	but	71,887	more	46,244
as	109,135	will	71,494	who	42,142
at	101,779	say	66,807	one	41,635
mr	101,679	new	64,456	their	40,910
with	101,210	share	63,925		

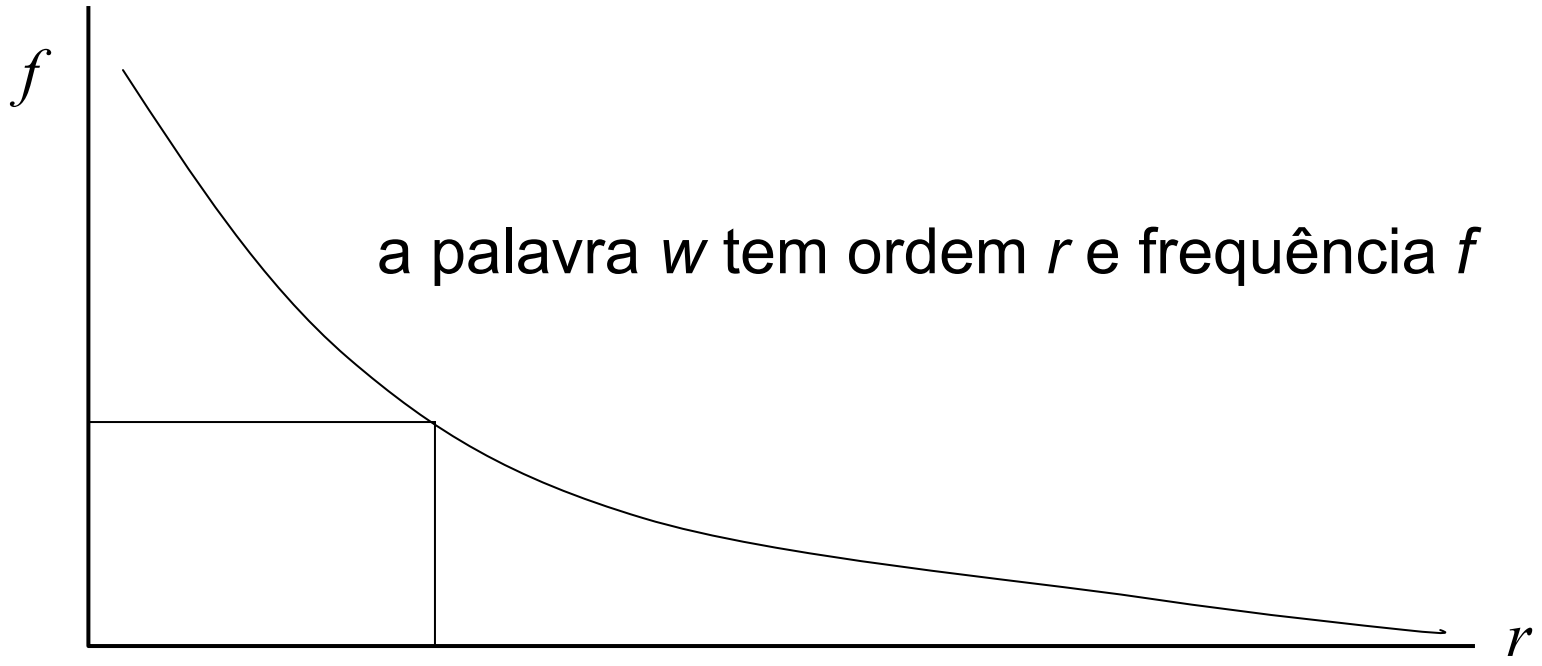
Paulo Trigo Silva

Curva de “frequência – ordem”

Para todas as palavras da colecção de documentos temos:

f : frequência com que a palavra *w* aparece

r : ordem da palavra *w* quanto à sua frequência; a palavra mais comum tem ordem 1, etc.



Paulo Trigo Silva

Relação entre “frequência & ordem”

Proximo slide mostra as palavras do texto Jamie Callan, *Characteristics of Text*, 1997, *normalizadas:*

f é a frequência da palavra w

r é a ordem da palavra w quanto à sua frequência

n é o numero total da ocorrência das palavras

Paulo Trigo Silva

Relação entre “frequência & ordem” (cont.)

*rf**1000/*n*		*rf**1000/*n*		*rf**1000/*n*	
the	59	from	92	or	101
of	58	he	95	about	102
to	82	million	98	market	101
a	98	year	100	they	103
in	103	its	100	this	105
and	122	be	104	would	107
that	75	was	105	you	106
for	84	company	109	which	107
is	72	an	105	bank	109
said	78	has	106	stock	110
it	78	are	109	trade	112
on	77	have	112	his	114
by	81	but	114	more	114
as	80	will	117	who	106
at	80	say	113	one	107
mr	86	new	112	their	108
with	91	share	114		

Lei de Zipf

Se as palavras (*w)*, numa colecção são ordenadas (r), pela sua frequência (*f)*, temos a seguinte relação (aproximada):

$$r * f \approx c$$

c depende da colecção.

Paulo Trigo Silva

Métodos baseados na "Lei de Zipf"

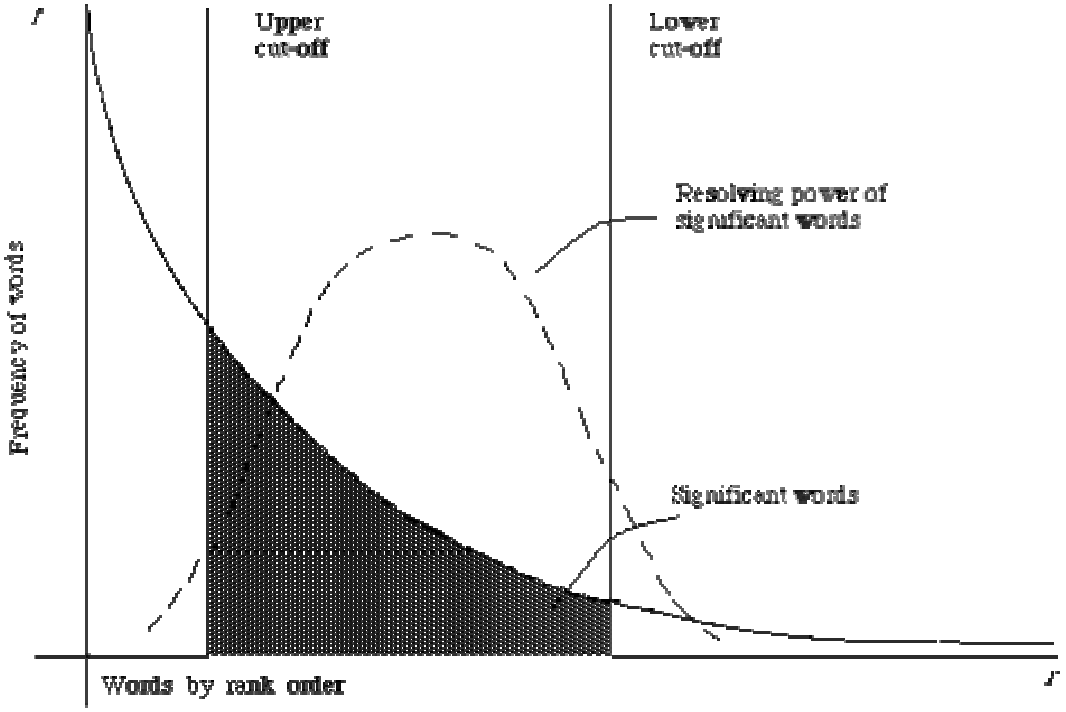


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adopted from Schultz[44] page 120)

Stop lists:

Ignoram-se a palavras mais frequentes (upper cut-off), i.e., usadas na maior parte dos documentos.

Significant words:

Ignoram-se as palavras com menor frequência (upper and lower cut-off), i.e., raramente usadas nos documentos.

Peso Termos: Damos diferentes pesos aos termos baseado na frequência das palavras; as palavras com maior frequência a "valerem menos".

Usado na maioria dos métodos de comparação.

Como “construir respostas rápidas”?

- Evitando pesquisar linearmente todos os textos
  - em resposta a cada interrogação
- Identificando os termos contidos nos documentos
  - ... qual o conjunto de palavras deste documento?
  - eliminar palavras repetidas e pontuação (e.g. vírgulas, pontos)
- Definido estruturas que integrem termos e documentos
  - ... que termos estão contidos em que documentos?
  - estas estruturas designam-se genericamente por índices
- Formulando operadores sobre índices
  - e.g. que aceitam uma lista de termos e devolvem documentos

Paulo Trigo Silva

Métricas de apreciação de um sistema de RI

- A relevância é uma noção subjectiva mas pode ser aferida
  - admitindo que o utilizador avalia dos resultados de cada interrogação
- ... podem extrair-se duas medidas estatísticas
  - precisão, e
  - cobertura
- A precisão ("precision") mede, para cada resposta do sistema,
  - a fracção de documentos relevantes contidos nessa resposta
- A cobertura ("recall") mede, para cada resposta do sistema,
  - a fracção dos documentos relevante na colecção obtida nessa resposta
- ... precisão: "destes documentos quantos são relevantes?"
- ... cobertura: "de todos os documentos relevantes quantos obtive?"

Paulo Trigo Silva

... métricas de avaliação do resultado da pesquisa

avaliação do utilizador (humano)

resposta da pesquisa (máquina)	Relevante	Não Relevante
Recuperado	positivo (p)	falso positivo (fp)
Não Recuperado	falso negativo (fn)	negativo (n)

Métricas:

Precisão (“Precision”) ≡ fracção dos documentos recuperados que é relevante

P(Relevante | Recuperado) =

= #(documentos relevantes e recuperados) / #(documentos recuperados)

Cobertura (“Recall”) ≡ fracção os documentos relevantes que são recuperados

P(Recuperado | Relevante) =

= #(documentos relevantes e recuperados) / #(documentos relevantes)

Paulo Trigo Silva

... métricas de avaliação – cálculo

avaliação do utilizador (humano)

resposta da pesquisa (máquina)	Relevante	Não Relevante
Recuperado	positivo (p)	falso positivo (fp)
Não Recuperado	falso negativo (fn)	negativo (n)

Precisão (“Precision”) ≡ fracção dos documentos recuperados que é relevante

= #(documentos relevantes e recuperados) / #(documentos recuperados)

= p / (p + fp)

Cobertura (“Recall”) ≡ fracção os documentos relevantes que são recuperados

P(Recuperado | Relevante) =

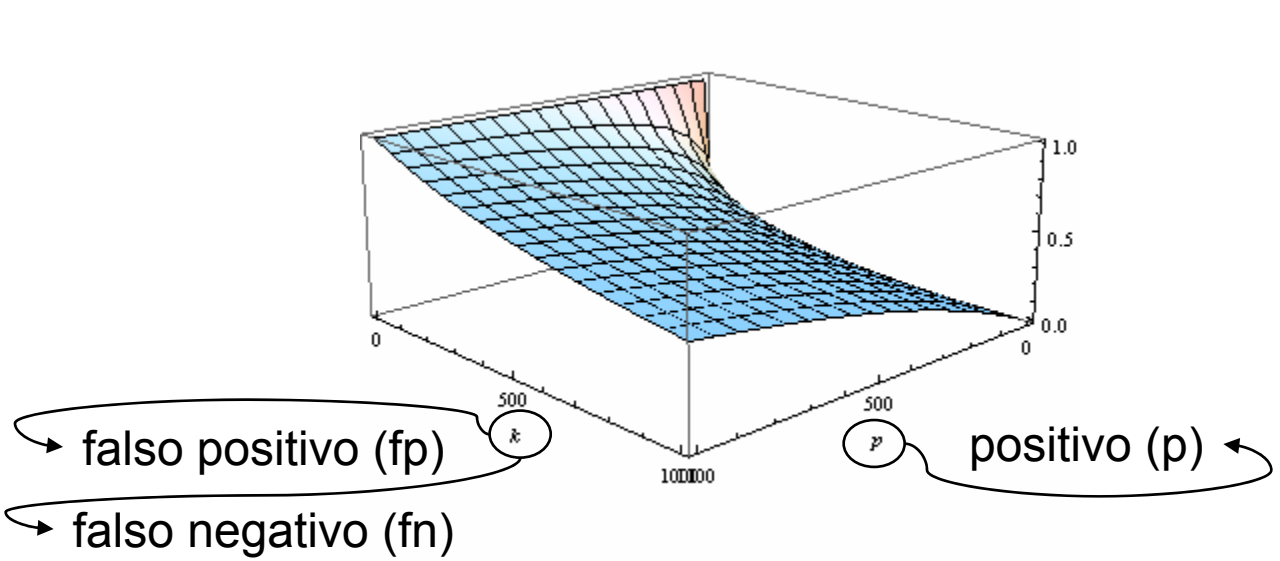
= #(documentos relevantes e recuperados) / #(documentos relevantes)

= p / (p + fn)

Perspectiva de variação – precisão e cobertura

Precisão (“Precision”) ≡ p / (p + fp)

Cobertura (“Recall”) ≡ p / (p + fn)



A **precisão** aumenta quando diminuem os falsos positivos.

A **cobertura** aumenta diminuem os falsos negativos.

A **precisão** e a **cobertura** diminuem quando diminuem os positivos.

Paulo Trigo Silva

... precisão e cobertura – "tensão" entre ambas

Para obter óptima cobertura basta recuperar todos os documentos da colecção!

No entanto teria baixa precisão (muitos não seriam relevantes).

A cobertura não decresce com o número de documentos recuperados.

A precisão pode decrescer com o número de documentos recuperados.

A precisão decresce, em sistemas reais (e.g. comerciais) usualmente com:

- o aumento de documentos recuperados, ou
- o aumento da cobertura.

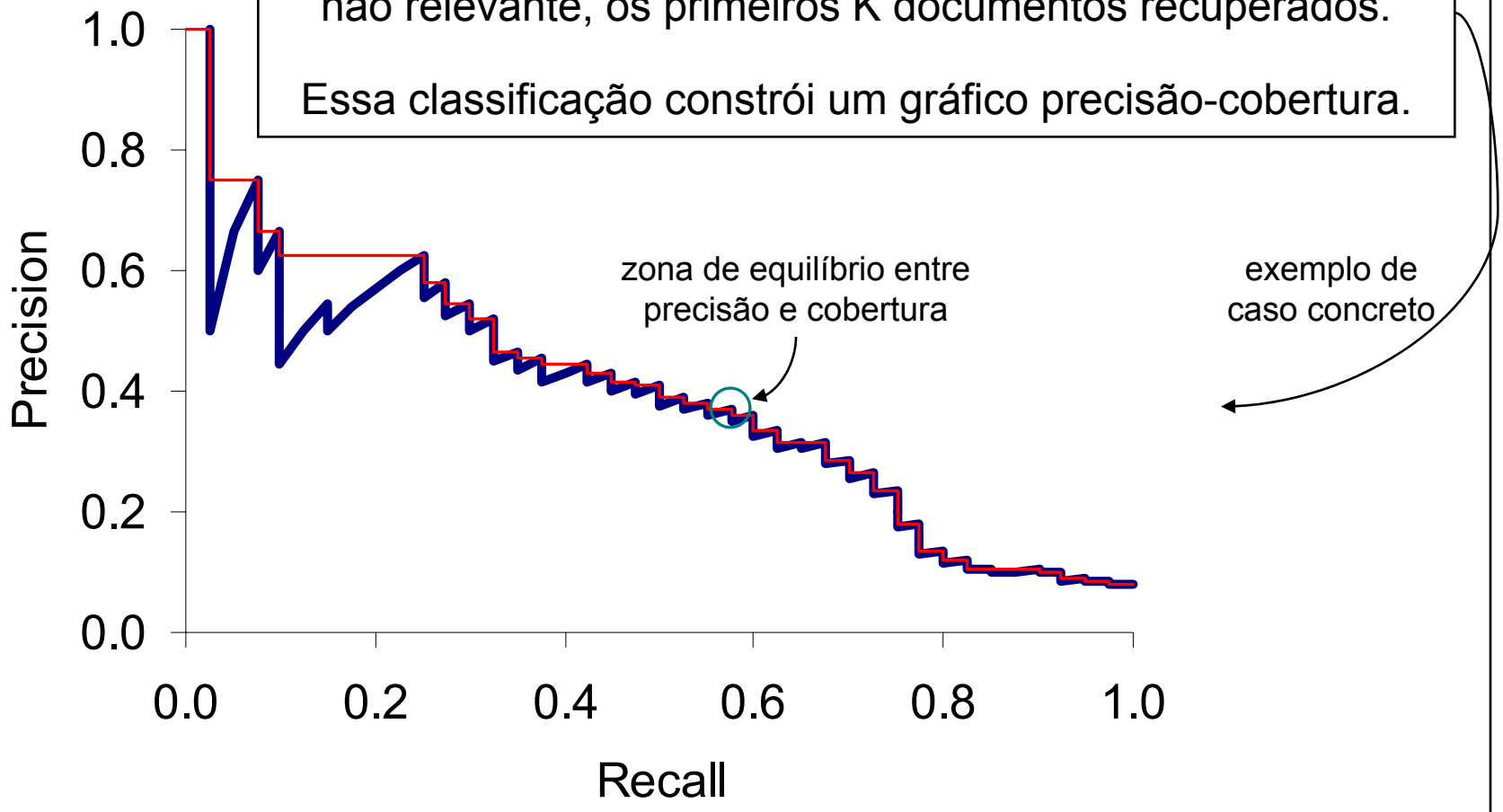
... empiricamente pode comprovar-se num qualquer um motor de busca.

Paulo Trigo Silva

Gráfico precisão-cobertura

Um avaliador (humano) pode classificar como relevante, ou não relevante, os primeiros K documentos recuperados.

Essa classificação constrói um gráfico precisão-cobertura.



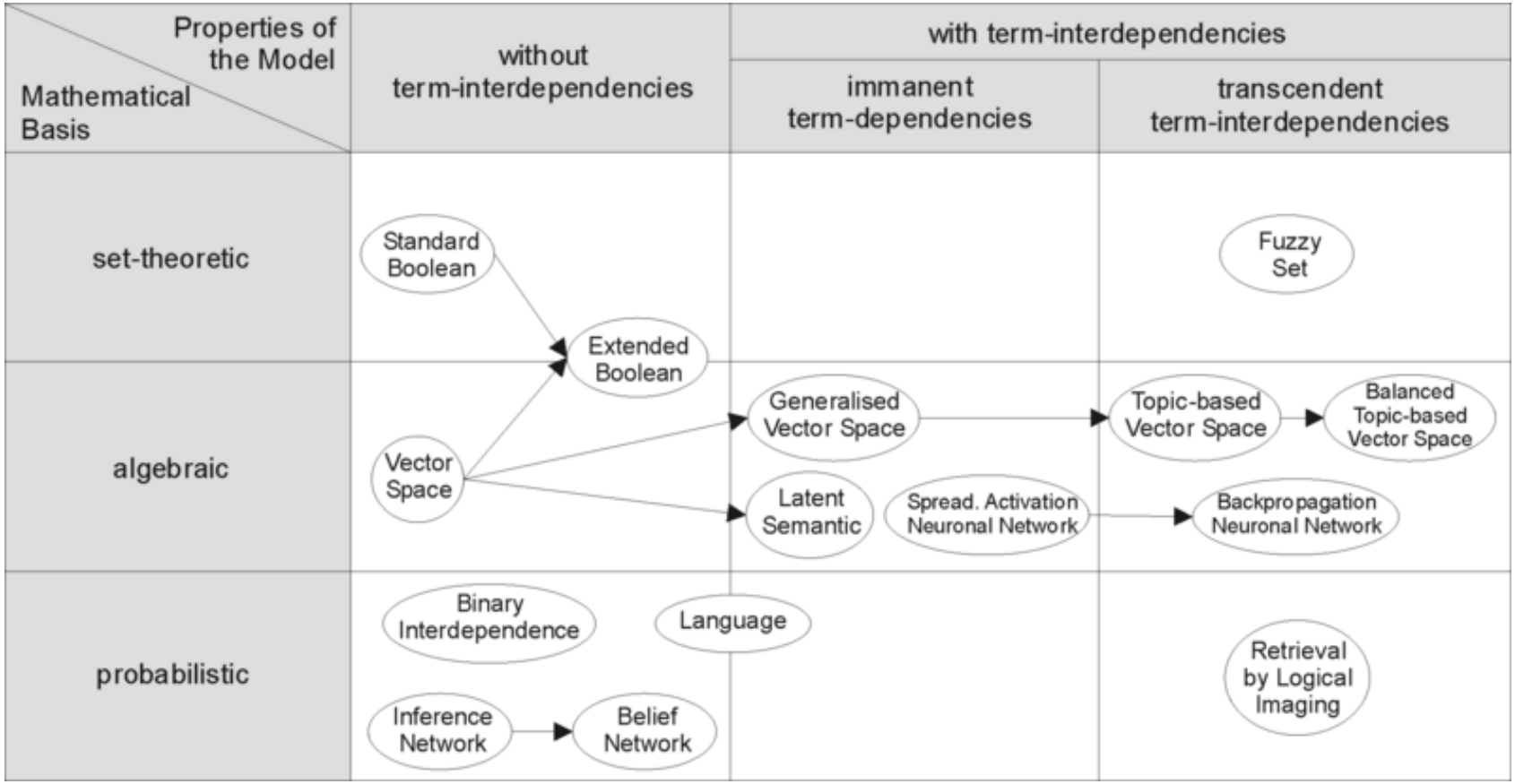
Paulo Trigo Silva

Recuperação de informação em documentos de texto

- Processos
  - indexar documentos
  - extrair os documentos com determinada informação
- Preocupações
  - 1. identificar os documentos relevantes
  - 2. extrair eficazmente grandes quantidades de documentos
- ... existem diferentes modelos para implementar os processos
  - contemplando as preocupações!

Paulo Trigo Silva

Classificação dos modelos de recuperação de informação



Paulo Trigo Silva