Informática UFRGS

Avaliação de Desempenho de Sistemas de Recuperação de Informações

CMP254 – Banco de Dados, Web e Recuperação de Informações
Profa. Viviane Moreira Orengo
vmorengo@inf.ufrgs.br

Relembrando

- Uma das maiores diferenças entre IR e Recuperação de Dados é que em IR nem todos os itens recuperados em resposta a uma consulta são relevantes. Além disso, nem todos os itens relevantes são recuperados.

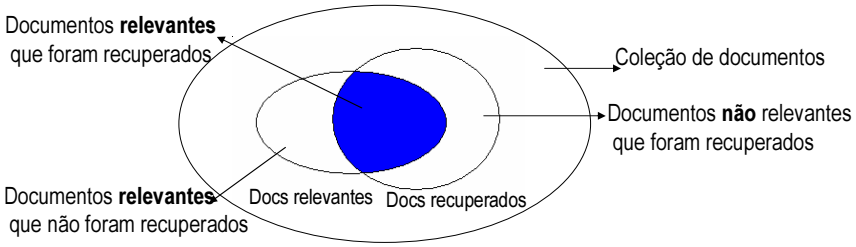
Motivação

- Questões:
  - Como saber se devemos remover stopwords?
  - Como saber se devemos usar stemming?
  - Como saber se devemos usar term weighting?
  - Como comparar dois motores de busca?
  - Como saber se um modelo é melhor do que outro?
- IR é uma disciplina altamente empírica. Todas as técnicas propostas são **avaliadas** por meio de **experimentos** em coleções de tamanho razoável.

Idéia de Relevância

- A base da avaliação de sistemas de IR é a idéia de relevância
- Com base em uma consulta proposta por um usuário, os documentos são classificados como relevantes ou irrelevantes.
- Relevância é tratada como binária – não existem as categorias de muito relevante, razoavelmente relevante etc.
- Críticos afirmam que a relevância é subjetiva
- Esta abordagem, apesar de bastante criticada, ainda é o padrão.

Avaliação de Desempenho



- Precisão (precision): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de recuperados}}$
- Revocação (recall): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de relevantes}}$

Exemplo 1

- 20 documentos relevantes em toda a coleção
- 40 documentos recuperados pela consulta
- 10 relevantes recuperados
- Precisão = 10 ÷ 40 = 0.25 ou 25%
- Revocação = 10 ÷ 20 = 0.5 ou 50%

Exemplo 2

- 20 documentos relevantes em toda a coleção
- 20 documentos recuperados pela consulta
- 10 relevantes recuperados
- Precisão = 10 ÷ 20 = 0.5 ou 50%
- Revocação = 10 ÷ 20 = 0.5 ou 50%

Exemplo 3

- 20 documentos relevantes em toda a coleção
- 1 documento recuperado pela consulta
- 1 relevante recuperado
- Precisão = 1 ÷ 1 = 1 ou 100%
- Revocação = 1 ÷ 20 = 0.05 ou 5%

Conclusão: Altos índices de precisão geralmente são acompanhados por baixos níveis de revocação e vice-versa.

F-measure

- Combina Precisão e Revocação em uma só medida

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

- Se β = 1, a mesma ênfase é dada a P e a R
- Se β = 2, Revocação é enfatizada 2 vezes em relação à Precisão
- Se β = 0.5, enfatiza a Precisão 2 vezes em relação à Revocação

Exercício:

- Calcular a F-measure para os 3 exemplos vistos anteriormente, dando ênfase igual às duas medidas.

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

- Exemplo 1

P = 0.25 & R = 0.5 F = 0.33

- Exemplo 2

P = 0.5 & R = 0.5 F = 0.5

- Exemplo 3

P = 1 & R = 0.05 F = 0.09

Curvas de Precisão - Revocação

- Precisão e revocação são medidas baseadas em conjuntos, ou seja, não levam em conta a ordenação do resultado
  - Problema: um sistema "A" que recuperou 100 documentos, sendo que destes, 5 são relevantes e estão nas posições 1,2,3,4 e 5 do ranking recebe a mesma "nota" que outro sistema "B" que recuperou os mesmos 5 itens relevantes nas posições 96,97,98,99 e 100.
- Solução: Calcular quantos documentos relevantes foram recuperados e o **quão próximos** estão do topo da lista.
- Tradicionalmente calcula-se a precisão para 11 pontos de revocação – 0.0, 0.1, 0.2, ..., 1.0.

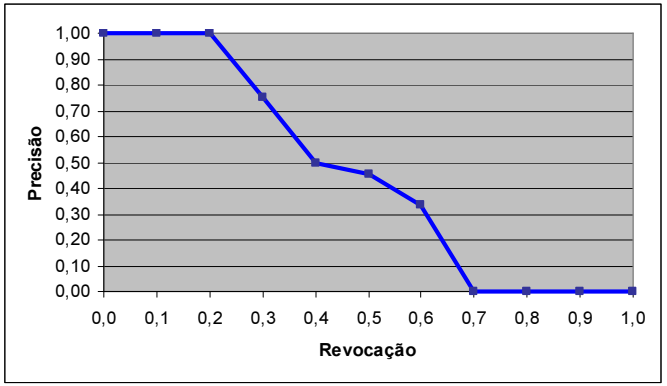
Curvas de Precisão - Revocação

- 20 docs recuperados
- 10 docs relevantes

d1✓	d5	d9	d13	d17

Pergunta: Como seria a curva de Precisão-Revocação para um sistema perfeito?

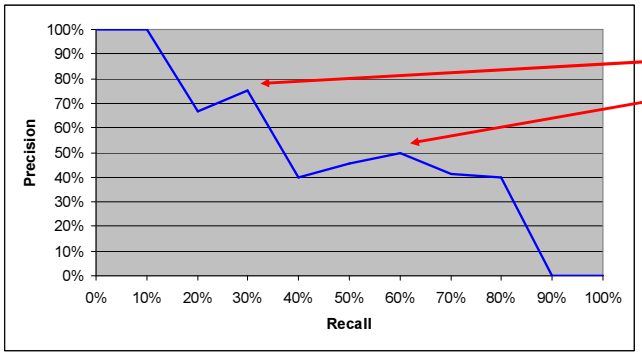
Recall	Precisão.
0%	100%
10%	(1/1) 100%
20%	(2/2) 100%
30%	(3/4) 75%
40%	(4/8) 50%
50%	(5/11) 45%
60%	(6/18) 33%
70%	0%
80%	0%
90%	0%
100%	0%



Exercício

- Desenhar a curva de precisão/revocação para a seguinte situação:
  - 20 docs recuperados
  - 10 docs relevantes
  - 8 relevantes recuperados

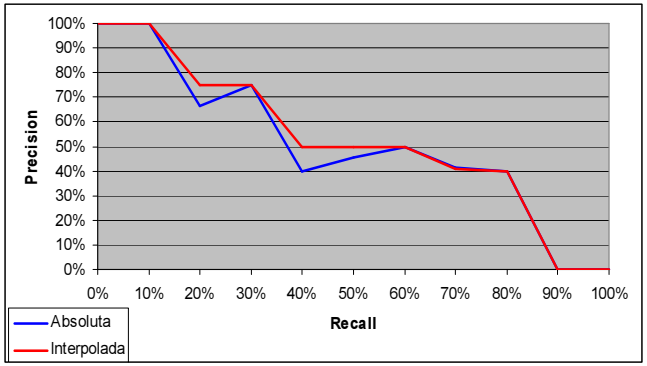
d1 ✓	d5	d9	d13	d17 ✓
d2	d6	d10 ✓	d14	d18
d3 ✓	d7	d11 ✓	d15	d19
d4 ✓	d8	d12 ✓	d16	d20 ✓



Problema: A curva deve ser sempre descendente!

Regra de interpolação

"A precisão interpolada para um nível de recall *j* é o maior valor de precisão para qualquer nível de recall maior ou igual a *j*."

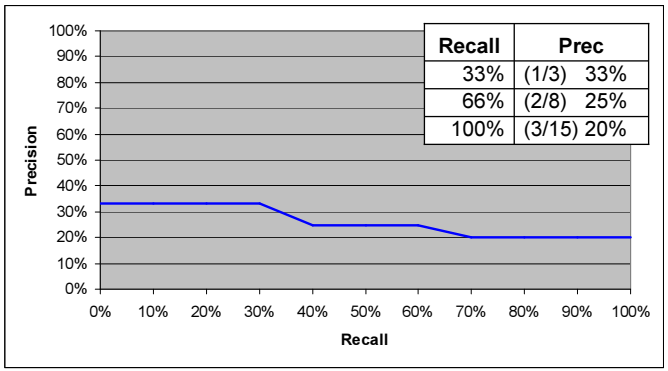


Recall	Abs	Int
0%	100%	100%
10%	100%	100%
20%	67%	75%
30%	75%	75%
40%	40%	50%
50%	45%	50%
60%	50%	50%
70%	41%	41%
80%	40%	40%
90%	0%	0%
100%	0%	0%

Exercício

- 20 docs recuperados
- 3 docs relevantes
- 3 relevantes recuperados

d1	d5	d9	d13	d17
d2	d6	d10	d14	d18
d3 ✓	d7	d11	d15 ✓	d19
d4	d8 ✓	d12	d16	d20



Recall	Prec
33%	(1/3) 33%
66%	(2/8) 25%
100%	(3/15) 20%

Recall	Pr Interp.
0%	33%
10%	33%
20%	33%
30%	33%
40%	25%
50%	25%
60%	25%
70%	20%
80%	20%
90%	20%
100%	20%

Observações

- Quando várias consultas forem realizadas, o valor da precisão será o valor médio para cada nível de recall, calculado por:

$$\overline{P}(r) = \sum_{i=1}^{Nc} \frac{P_i(r)}{Nc}$$

onde:

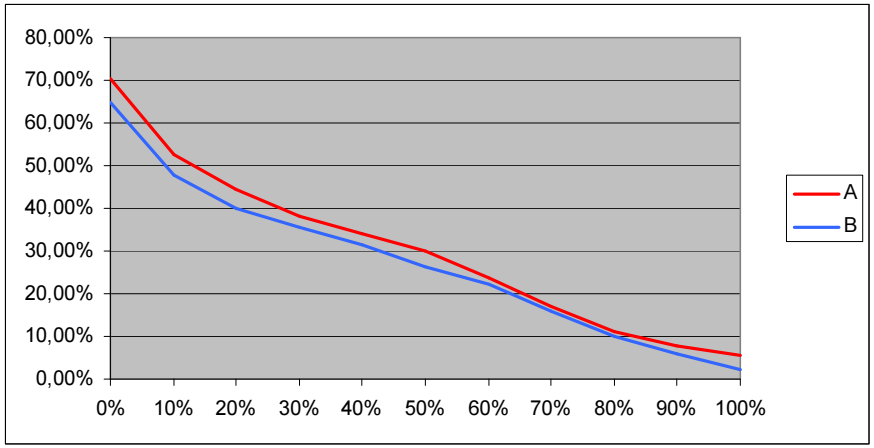
$\overline{P}(r)$ é a precisão média para o nível de recall r

Nc é o número de consultas

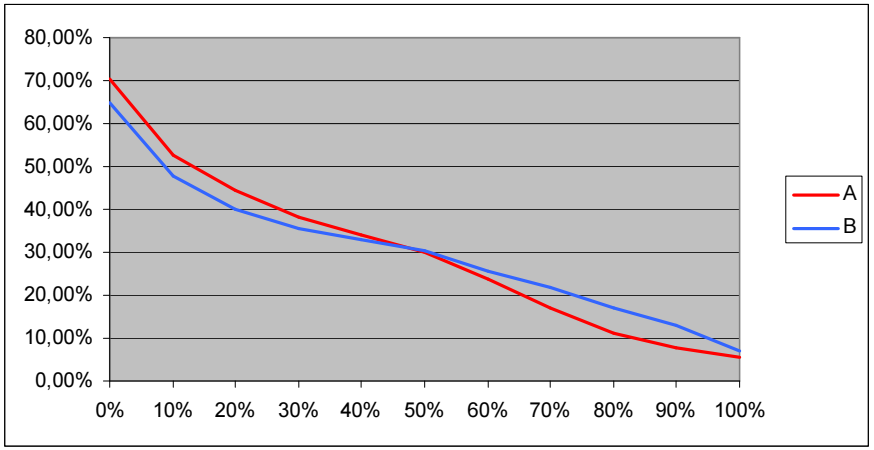
$P_i(r)$ é a precisão no nível de recall r para a i-ésima consulta

- Considerando-se que os níveis de recall para cada consulta podem ser diferentes dos 11 níveis padrão, a utilização da regra de interpolação é necessária.

Qual o melhor sistema de IR?



E agora?



Mean Average Precision

- A **Precisão Média** (*average precision*) de uma consulta é a média das precisões após cada documento relevante recuperado (sem interpolação).
- A **Mean Average Precision** (MAP), para uma série de consultas, é a média das precisões médias de cada consulta.
- Esta é a medida padrão em vários experimentos (TREC e CLEF)

Recall	Precisão
33%	33%
66%	25%
100%	20%
AvP	26%

Recall	Precisão
0%	100%
10%	100%
20%	100%
30%	75%
40%	50%
50%	45%
60%	33%
70%	0%
80%	0%
90%	0%
100%	0%
AvP	46%

Recall	Precisão
0%	100%
10%	100%
20%	67%
30%	75%
40%	40%
50%	45%
60%	50%
70%	41%
80%	40%
90%	0%
100%	0%
AvP	51%

$$MAP = \frac{26+46+51}{3}$$

$$MAP = 41\%$$

Como saber com certeza se um sistema é melhor do que outro?

- Análise estatística - Teste de Hípotese
  - T-teste Pareado
  - ANOVA
- Executa-se um número significativo de consultas nos dois sistemas (as mesmas consultas nos dois sistemas)
- Compara-se a precisão média de cada consulta
- Executa-se o teste estatístico
- Se a variável *P* calculada for menor do que o threshold de significância (geralmente 0.05), então existe uma diferença significativa entre os dois sistemas.

Exemplo 1

Teste-t: duas amostras em par para médias

	sw	*s*
Média	0,2063	0,2310
Variância	0,0350	0,0364
Observações	30	30
Correlação de Pearson	0,9815	
Hipótese da diferença de média	0	
gl	29	
Stat t	-3,7013	
P(T<=t) uni-caudal	0,0004	
t crítico uni-caudal	1,6991	
P(T<=t) bi-caudal	0,0008	
t crítico bi-caudal	2,0452	

O sistema S é significativamente melhor do que o sistema SW, pois 0,0008 < 0,05

Query	sw	s
1	0,1434	0,1617
2	0,0184	0,0838
3	0,0134	0,0187
4	0,2451	0,2478
5	0,3394	0,4201
6	0,2683	0,2988
7	0,0025	0,0103
8	0,5979	0,6797
9	0,1836	0,2310
10	0,4862	0,4496
11	0,3244	0,3347
12	0,0991	0,1158
13	0,0035	0,1020
14	0,0667	0,0474
15	0,0290	0,0160
16	0,2864	0,3127
17	0,0467	0,0716
18	0,1259	0,1154
19	0,1244	0,1196
20	0,0218	0,0354
21	0,6499	0,6872
22	0,0546	0,0850
23	0,2042	0,2815
24	0,2103	0,2829
25	0,3107	0,3370
26	0,2387	0,2345
27	0,1434	0,1388
28	0,6451	0,6131
29	0,0710	0,0794
30	0,2354	0,3189
AVGPrec	**0,2063**	**0,2310**

Exemplo 2

Teste-t: duas amostras em par para médias

	s	*st*
Média	0,2310	0,2283
Variância	0,0364	0,0363
Observações	30	30
Correlação de Pearson	0,9779	
Hipótese da diferença de média	0	
gl	29	
Stat t	0,3697	
P(T<=t) uni-caudal	0,3571	
t crítico uni-caudal	1,6991	
P(T<=t) bi-caudal	0,7142	
t crítico bi-caudal	2,0452	

Não há diferença significativa entre os sistemas S e ST, pois 0,7142 > 0,05

Query	s	st
1	0,1617	0,2171
2	0,0838	0,0523
3	0,0187	0,0183
4	0,2478	0,2436
5	0,4201	0,4299
6	0,2988	0,2928
7	0,0103	0,0237
8	0,6797	0,6319
9	0,2310	0,2256
10	0,4496	0,3774
11	0,3347	0,3637
12	0,1158	0,1246
13	0,1020	0,0483
14	0,0474	0,0481
15	0,0160	0,0239
16	0,3127	0,4088
17	0,0716	0,0424
18	0,1154	0,1280
19	0,1196	0,0860
20	0,0354	0,0321
21	0,6872	0,6692
22	0,0850	0,0976
23	0,2815	0,3385
24	0,2829	0,1982
25	0,3370	0,3333
26	0,2345	0,2344
27	0,1388	0,0701
28	0,6131	0,6258
29	0,0794	0,1405
30	0,3189	0,3232
AVGPrec	**0,2310**	**0,2283**

Problemas com Revocação e Precisão

- É necessário saber de antemão todos os documentos relevantes para cada consulta.
- Para grandes coleções, é impossível conhecer todos os documentos relevantes, por isso estas medidas são imprecisas.
- Estas duas medidas não servem para consultas interativas (por exemplo em um motor de busca na web)
- Apesar dos problemas, precisão e revocação ainda são a norma.

Outras Medidas

- **R-Precision** - Precisão após R documentos recuperados. R é o número de documentos relevantes para a consulta.

• 20 docs recuperados
• 15 docs relevantes
• 8 relevantes recuperados

d1 ✓	d5	d9	d13	d17 ✓
d2	d6	d10 ✓	d14	d18
d3 ✓	d7	d11 ✓	d15 ✓	d19
d4 ✓	d8	d12 ✓	d16	d20

$$\text{R-Precision} = \frac{7}{15} = 47\%$$

Número de relevantes recuperados até a posição R do ranking

Número total de relevantes (R)

Outras Medidas

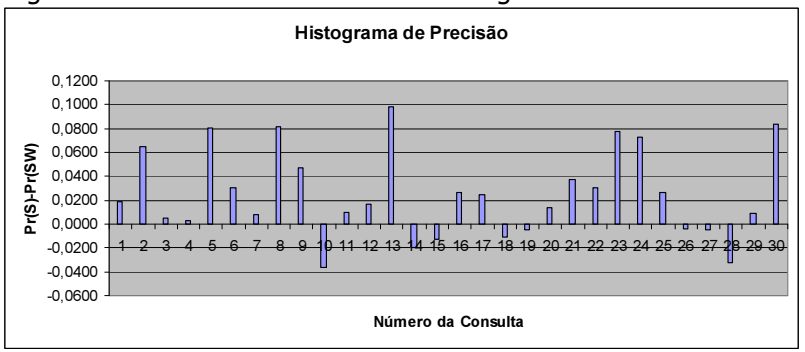
- **Precision at 10** – precisão após 10 documentos terem sido recuperados
- Útil para consultas na Web, onde se sabe que os usuários raramente olham além do décimo documento recuperado
- Vantagem – não é necessário saber o número total de documentos relevantes

$$\text{Pr@10} = \frac{4}{10} = 40\%$$

d1 ✓	d5	d9	d13	d17 ✓
d2	d6	d10 ✓	d14	d18
d3 ✓	d7	d11 ✓	d15 ✓	d19
d4 ✓	d8	d12 ✓	d16	d20

Histogramas

- Útil para comparar a performance de dois algoritmos
- Calcula-se Pr(A)-Pr(B)
- Nas consultas com resultado positivo, então A foi melhor do que B
- Também pode ser usado para comparar o desempenho de um algoritmo com a média de vários outros algoritmos



Coleções de Teste

- Composta por:
  - Coleção de Documentos
  - Consultas (chamadas de tópicos)
  - Julgamentos de relevância – quais documentos são relevantes para cada consulta.
- Exemplos:
  - Cranfield
  - Time Magazine
  - CACM
  - LA Times
  - Folha de São Paulo
  - Wall Street Journal

Exemplos de Tópicos (consultas)

<top>
<num> C260 </num>
<PT-title> Legislação anti-tabagista </PT-title>
<PT-desc> Encontrar documentos que descrevam leis ou qualquer legislação anti-tabaco que proíba o fumo em locais públicos. </PT-desc>
<PT-narr> Documentos relevantes devem fornecer informações sobre leis, regras ou políticas anti-tabaco atualmente em vigor num país. A pena para as transgressões também pode ser especificada. Porém documentos sobre propostas de lei anti-tabagistas não são relevantes. </PT-narr>
</top>
<top>
<num> C300 </num>
<PT-title> Prêmios na loteria </PT-title>
<PT-desc> Encontrar documentos sobre os vencedores de prêmios na loteria. </PT-desc>
<PT-narr> Apenas documentos relatando uma atribuição de prêmios na loteria e fornecendo informação sobre os vencedores, tal como nomes, passado, ou origens são relevantes. Documentos que apenas listam prêmios a ser atribuídos não são relevantes.
</PT-narr>
</top>

Tópicos da Campanha CLEF2005

Exemplo de Documento

<DOC>
<DOCNO>FSP951229-038</DOCNO>
<DOCID>FSP951229-038</DOCID>
<DATE>951229</DATE>
<CATEGORY>DINHEIRO</CATEGORY>
<TEXT>
Os carros da GM ficam entre 1,5% e 3% mais caros a partir do dia 1º. O Corsa Wind sobe 3% e passa a custar R$ 9.987,00. Também aumentam 3% o Corsa Wind Super, Kadett e Vectra. O Monza tem reajuste de 2% e o Omega, entre 1,5% e 2,5%. A Volks aumenta em 3% a sua tabela a partir do dia 1º.
Cai 36% rentabilidade dos fundos de pensão
A rentabilidade dos fundos de pensão brasileiros caiu 36,4% neste ano. A estimativa da Abrapp (Associação Brasileira das Entidades Fechadas de Previdência Privada) é que os fundos terminem o ano com rentabilidade de 0,21%. Em 94, esse índice foi de 57,6%.
Ufir valerá R$ 0,8287 entre janeiro e junho
A Secretaria da Receita Federal fixou em R$ 0,8287 o valor da Ufir (Unidade Fiscal de Referência) para o período de 1º de janeiro a 30 de junho de 1996. A Ufir não indexa mais a tabela do IR, mas ainda será usada na atualização de impostos em atraso com fato gerador (origem da dívida) anterior a 95 e no cálculo do ganho de capital na venda de bens e direitos.
</TEXT>
</DOC>

Folha de São Paulo (CLEF)

Julgamentos de Relevância

Número da consulta	Número do documento	Relevante?
251	FSP940128-148	0
251	FSP940130-099	1
251	FSP940130-100	0
251	FSP940131-030	0
251	FSP940131-079	0
251	FSP940204-141	0
251	FSP940205-099	0
251	FSP940206-094	1
251	FSP940221-132	0
251	FSP940227-193	0
251	FSP940228-080	0
251	FSP940304-100	0
251	FSP940317-084	0
251	FSP940318-123	0
251	FSP940319-099	0
251	FSP940319-100	0
251	FSP940321-033	0
251	FSP940329-082	0
251	FSP940330-099	1

Cranfield Collection

- Primeira coleção de teste desenvolvida no Cranfield College of Aeronautics (Inglaterra) no final dos anos 60
- 1400 abstracts de documentos sobre aeronáutica
- 225 consultas
- Cada documento julgado por mais de um avaliador – mais de meio milhão de julgamentos de relevância
- $225 \times 1400 \times 2 = 630\,000$ julgamentos de relevância!!!

Construindo uma coleção de teste

- Problema: Julgar todos os documentos para todas as consultas é impraticável.
  - Ex: LA Times = 50 consultas X 113 mil docs X 2 avaliadores = 11.300.000 julgamentos
- Solução: **Método Pooling**
  - Repositórios de documentos são criados para cada consulta a partir dos resultados de experimentos enviados por diferentes grupos.
  - Apenas os top *n* documentos de cada grupo são adicionados ao repositório
  - Os avaliadores julgam somente documentos que estão no repositório.
  - Os documentos não julgados são considerados irrelevantes.
  - Como o repositório é construído com documentos retornados por vários sistemas, existe uma probabilidade de que todos os documentos relevantes sejam encontrados.

Campanhas de Avaliação de Sistemas de IR

Sumário

- A avaliação em IR baseia-se em Precisão e Revocação
- Medida mais utilizada em campanhas de avaliação
  - Mean Average Precision
- Curvas de Precisão-Revocação bastante utilizadas para mostrar graficamente o comportamento de 2 sistemas
- Coleções de teste criadas pelo método pooling

Na próxima aula

- Modelos Tradicionais de IR

Referências

- Modern Information Retrieval
  Ricardo Baeza-Yates & Berthier Ribeiro Neto
  Adisson Wesley
  Seção 3.2
- Managing Gigabytes
  Ian Witten, Alistair Moffat, Tim Bell
  Morgan Kaufmann
  Seção 4.5
- An Introduction to Information Retrieval
  Christofer Manning, Prabhakar Raghavan, Hinrich Schütze
  Cambridge University Press (draft)
  Capítulo 8