

Sumário

1	Introdução.....	1
1.1	Recuperação de informação	1
1.1.1	Desenvolvimentos iniciais	1
1.1.2	Recuperação de informação em bibliotecas e em bibliotecas digitais.....	3
1.1.3	RI em destaque	3
1.2	O problema de RI	4
1.2.1	A tarefa do usuário	5
1.2.2	Recuperação de informação versus recuperação de dados.....	6
1.3	O sistema de RI.....	6
1.3.1	Arquitetura do software de um sistema de RI	6
1.3.2	Os processos de recuperação e ranqueamento.....	8
1.4	A Web.....	10
1.4.1	Um breve histórico	10
1.4.2	A era da publicação eletrônica	11
1.4.3	Como a Web mudou a busca.....	12
1.4.4	Aspectos práticos da Web	13
1.5	Organização deste livro	14
1.5.1	Foco do livro	14
1.5.2	Conteúdo do livro	14
1.6	Recurso didático para professores	17
1.7	Discussão bibliográfica	17
2	Modelagem.....	21
2.1	Modelos de RI	21
2.1.1	Modelagem e ranqueamento	21
2.1.2	Caracterização de um modelo de RI.....	22
2.1.3	Uma taxonomia de modelos de RI	24
2.2	Recuperação de informação clássica	26
2.2.1	Conceitos básicos	26
2.2.2	O modelo Booleano	29
2.2.3	Ponderação de termos	31
2.2.4	Ponderação TF-IDF	34
2.2.5	Normalização pelo tamanho dos documentos	43
2.2.6	O modelo vetorial	45
2.2.7	O modelo probabilístico.....	48
2.2.8	Breve comparação entre os modelos clássicos	55

2.3	Modelos alternativos da teoria dos conjuntos	56
2.3.1	Modelo baseado em conjuntos	56
2.3.2	Modelo Booleano estendido	62
2.3.3	Modelo de conjunto fuzzy	65
2.4	Modelos algébricos alternativos	69
2.4.1	Modelo vetorial generalizado	69
2.4.2	Modelo de indexação semântica latente	72
2.4.3	Modelo de redes neurais	74
2.5	Modelos probabilísticos alternativos	76
2.5.1	BM25	76
2.5.2	Modelos de linguagem	79
2.5.3	Divergência da aleatoriedade	86
2.5.4	Modelos de redes bayesianas	89
2.6	Outros modelos	97
2.6.1	O modelo de hipertexto	98
2.6.2	Modelos baseados na Web	100
2.6.3	Recuperação de texto estruturado	100
2.6.4	Recuperação multimídia	100
2.6.5	Busca corporativa e vertical	101
2.7	Tendências e aspectos de pesquisa	101
2.8	Discussão bibliográfica	102
3	Avaliação da recuperação	106
3.1	Introdução	106
3.2	O paradigma Cranfield	107
3.2.1	Um breve histórico	107
3.2.2	Coleções de referência	109
3.3	Métricas de recuperação	110
3.3.1	Precisão e revocação	110
3.3.2	Sumários com um único valor: P@N, MAP, MRR, F	116
3.3.3	Medidas orientadas ao usuário	121
3.3.4	DCG: Ganho Acumulado Descontado	122
3.3.5	Bpref: preferências binárias	128
3.3.6	Medidas de correlação de ranking	131
3.4	Coleções de referência	137
3.4.1	As coleções TREC	137
3.4.2	Outras coleções de referência	145
3.4.3	Outras coleções de teste pequenas	146
3.5	Avaliação baseada em usuários	147
3.5.1	Experimentação humana em laboratório	147
3.5.2	Painéis lado a lado	148
3.5.3	Teste A/B	149
3.5.4	Crowdsourcing	149
3.5.5	Avaliação usando dados sobre cliques	151
3.6	Advertências práticas	154
3.7	Tendências e questões de pesquisa	155
3.8	Discussão bibliográfica	155

4 Realimentação de relevância e expansão de consultas	157
4.1 Introdução	157
4.2 Um arcabouço para métodos de realimentação	158
4.3 Realimentação de relevância explícita	161
4.3.1 Realimentação de relevância para o modelo vetorial: método de Rocchio	161
4.3.2 Realimentação de relevância para o modelo probabilístico	163
4.3.3 Avaliação da realimentação de relevância	165
4.4 Realimentação explícita por meio de cliques	165
4.4.1 Rastreamento ocular e julgamentos de relevância	166
4.4.2 Comportamento do usuário	167
4.4.3 Cliques como métrica de preferências do usuário	168
4.5 Realimentação implícita por meio de análise local	171
4.5.1 Realimentação implícita por meio de clustering local	171
4.5.2 Realimentação implícita por meio da análise do contexto local	176
4.6 Realimentação implícita por meio de análise global	178
4.6.1 Expansão de consulta baseada em um tesouro de similaridade	178
4.6.2 Expansão de consultas com base em um tesouro estatístico	181
4.7 Tendências e questões de pesquisa	183
4.8 Discussão bibliográfica	184
5 Documentos: linguagens e propriedades	187
<i>com Gonzalo Navarro e Nivio Ziviani</i>	
5.1 Introdução	187
5.2 Metadados	189
5.3 Formatos de documentos	190
5.3.1 Texto	190
5.3.2 Multimídia	191
5.3.3 Gráficos e realidade virtual	193
5.4 Linguagens de marcação	193
5.4.1 SGML	194
5.4.2 HTML	196
5.4.3 XML	199
5.4.4 RDF	202
5.4.5 HyTime	203
5.5 Propriedades do texto	204
5.5.1 Teoria da informação	205
5.5.2 Modelando a linguagem natural	206
5.5.3 Similaridade de textos	209
5.6 Pré-processamento de documentos	210
5.6.1 Análise léxica do texto	211
5.6.2 Eliminação de stopwords	213
5.6.3 Stemming	214
5.6.4 Seleção de palavras-chave	215
5.6.5 Tesouros	216
5.7 Organização de documentos	219
5.7.1 Taxonomias	219
5.7.2 Folksonomias	220

5.8	Compressão de texto	221
5.8.1	Conceitos básicos	222
5.8.2	Métodos estatísticos	223
5.8.3	Métodos estatísticos: modelagem	224
5.8.4	Métodos estatísticos: codificação	228
5.8.5	Métodos de dicionário	236
5.8.6	Pré-processamento para a compressão	238
5.8.7	Comparação das técnicas de compressão de textos	239
5.8.8	Compressão de textos estruturados	241
5.9	Tendências e questões de pesquisa	242
5.10	Discussão bibliográfica	245
6	Consultas: linguagens e propriedades	248
	<i>com Gonzalo Navarro</i>	
6.1	Linguagens de consulta	248
6.1.1	Consulta baseada em palavras-chave	249
6.1.2	Além de palavras-chave	253
6.1.3	Consultas estruturais	257
6.1.4	Protocolos de consulta	260
6.2	Propriedades de consulta	261
6.2.1	Caracterização das consultas na Web	262
6.2.2	Comportamento de busca do usuário	265
6.2.3	Intenção da consulta	266
6.2.4	Tópico da consulta	267
6.2.5	Sessões de consulta e missões	268
6.2.6	Dificuldade de consulta	269
6.3	Tendências e questões de pesquisa	274
6.4	Discussão bibliográfica	276
7	Classificação de textos	277
	<i>com Marcos Gonçalves</i>	
7.1	Introdução	277
7.2	Uma caracterização da classificação de textos	278
7.2.1	Aprendizado de máquina	278
7.2.2	O problema da classificação de textos	279
7.2.3	Algoritmos de classificação de textos	280
7.3	Algoritmos não supervisionados	282
7.3.1	Clustering	283
7.3.2	Classificação de textos ingênua	287
7.4	Algoritmos supervisionados	288
7.4.1	Árvores de decisão	291
7.4.2	O classificador k -NN	297
7.4.3	O classificador de Rocchio	299
7.4.4	Classificação de documentos usando Bayes ingênuo probabilístico	301
7.4.5	O classificador SVM	306
7.4.6	Classificadores do tipo ensemble	316
7.4.7	Considerações finais sobre algoritmos supervisionados	319
7.5	Seleção de características ou redução de dimensionalidade	320
7.5.1	Tabela de incidência termo-classe	321
7.5.2	Frequência de documentos de um termo	322

7.5.3	Pesos TF-IDF	322
7.5.4	Informação mútua	323
7.5.5	Ganho de informação	324
7.5.6	Chi-quadrado	325
7.5.7	Impacto da seleção de características	325
7.6	Métricas de avaliação	326
7.6.1	Tabela de contingência	326
7.6.2	Acurácia e erro	327
7.6.3	Precisão e revocação	328
7.6.4	Medida-F e F_1	328
7.6.5	Validação cruzada	330
7.6.6	Coleções-padrão	330
7.7	Organização das classes – construindo taxonomias	331
7.8	Tendências e questões de pesquisa	334
7.9	Discussão bibliográfica	335
8	Indexação e busca	339
	<i>com Gonzalo Navarro</i>	
8.1	Introdução	339
8.2	Índices invertidos	342
8.2.1	Conceitos básicos	342
8.2.2	Índices invertidos completos	344
8.2.3	Busca	347
8.2.4	Ranqueamento	352
8.2.5	Construção	354
8.2.6	Índices invertidos compactados	359
8.2.7	Buscas estruturais	361
8.3	Arquivos de assinatura	362
8.4	Árvores de sufixos e arranjos de sufixos	365
8.4.1	Estrutura: tries e árvores de sufixos	366
8.4.2	Busca por sequências simples	368
8.4.3	Busca por padrões complexos	369
8.4.4	Construção	371
8.4.5	Arranjos de sufixos compactados	374
8.5	Busca sequencial	379
8.5.1	Strings simples: Horspool	380
8.5.2	Padrões complexos: autômatos e paralelismo de bits	383
8.5.3	Algoritmos mais rápidos de paralelismo de bits	387
8.5.4	Expressões regulares	389
8.5.5	Padrões múltiplos	392
8.5.6	Busca aproximada	392
8.5.7	Busca em texto comprimido	396
8.6	Indexação multidimensional	399
8.7	Tendências e questões de pesquisa	401
8.8	Discussão bibliográfica	403
9	Recuperação na Web	408
	<i>com Yoelle Maarek</i>	
9.1	Introdução	408
9.2	Um problema desafiador	410

9.3	A Web	412
9.3.1	Características	412
9.3.2	Estrutura do grafo da Web	414
9.3.3	Modelando a Web	416
9.3.4	Análise de links	419
9.4	Arquiteturas de máquinas de busca	421
9.4.1	Arquitetura básica	421
9.4.2	Arquitetura baseada em clusters	423
9.4.3	Mecanismo de cache	426
9.4.4	Múltiplos índices	428
9.4.5	Arquiteturas distribuídas	430
9.5	Ranqueamento em máquinas de busca	433
9.5.1	Sinais de ranqueamento	434
9.5.2	Ranqueamento baseado em links	435
9.5.3	Funções de ranqueamento simples	438
9.5.4	Aprendendo a ranquear	439
9.5.5	Aprendendo a função de ranqueamento	440
9.5.6	Avaliação de qualidade	441
9.5.7	Spam na Web	442
9.6	Gerenciando dados da Web	444
9.6.1	Atribuindo identificadores a documentos	444
9.6.2	Metadados	444
9.6.3	Comprimindo o grafo Web	445
9.6.4	Manipulando dados duplicados	445
9.7	Interação de usuários em máquinas de busca	447
9.7.1	O paradigma do retângulo de busca	447
9.7.2	A página de resultados da máquina de busca	456
9.7.3	Educando o usuário	467
9.8	Navegação	468
9.8.1	Navegação plana	469
9.8.2	Navegação guiada pela estrutura e diretórios Web	469
9.9	Além da navegação	472
9.9.1	O hipertexto e a Web	472
9.9.2	Combinando busca com navegação	472
9.9.3	Linguagens de consulta na Web	474
9.9.4	Busca dinâmica	474
9.10	Problemas relacionados	475
9.10.1	Publicidade computacional	475
9.10.2	Mineração na Web	478
9.10.3	Metabusca	481
9.11	Tendências e questões de pesquisa	482
9.11.1	Além de dados textuais estáticos	482
9.11.2	Desafios atuais	483
9.12	Discussão bibliográfica	486

10 Coleta na Web 489

com Carlos Castillo

10.1	Introdução	489
10.2	Aplicações de um coletor Web	491
10.2.1	Busca geral na Web	491
10.2.2	Coleta por tópicos	492

10.2.3	Caracterização da Web	492
10.2.4	Espelhamento	493
10.2.5	Análise de sites	493
10.3	Uma taxonomia de coletores	494
10.3.1	Tipos de páginas Web	494
10.4	Arquitetura e implementação	496
10.4.1	Arquitetura de coleta	496
10.4.2	Questões práticas	498
10.4.3	Coleta paralela	501
10.5	Algoritmos de escalonamento	502
10.5.1	Política de seleção	503
10.5.2	Política de revisitação	506
10.5.3	Política de boa educação	511
10.5.4	Combinando políticas	515
10.6	Avaliação	516
10.6.1	Avaliando o uso de largura de banda	516
10.6.2	Avaliando o escalonamento de longo prazo	517
10.7	Tendências e questões de pesquisa	518
10.7.1	Coletando a Web “oculta”	518
10.7.2	Coletando com o auxílio de sites	519
10.7.3	Coleta distribuída	520
10.8	Discussão bibliográfica	520
	Referências	523
	Créditos	577
	Índice	579