ÍNDICE

PREFACIO **19**

LISTA DE AUTORES **25**

CAPÍTULO 1. INTRODUCCIÓN A LA RECUPERACIÓN DE INFORMACIÓN **33**
Benjamin Piwowarski y Roi Blanco González (Traducido por Nahir Seijo Saavedra)

1.1 DEFINICIÓN DE RECUPERACIÓN DE INFORMACIÓN 33
1.1.1 Introducción 34
1.1.2 Las tres dimensiones de la Recuperación de Información 36
1.1.3 Componentes de un sistema de RI 40
1.2 BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN 42
1.2.1 Las bibliotecas se digitalizan 42
1.2.2 La web 46
1.3 CONCEPTOS IMPORTANTES 48
1.3.1 Relevancia 48
1.3.2 Representación 50
1.3.3 Modelo 52
1.3.4 Evaluación 54
1.3.5 Índices 57

CAPÍTULO 2. INDEXACIÓN DE DOCUMENTOS Y PROCESADO DE CONSULTAS **61**
Roi Blanco González

2.1 INTRODUCCIÓN 61

2.1.1 Fuentes de datos y adquisición de documentos 63
2.1.2 Procesamiento de textos 64
2.1.3 Procesado de términos 65
2.2 INDEXACIÓN MEDIANTE FICHEROS INVERTIDOS 68
2.2.1 Diccionario 69
2.2.2 Fichero de ocurrencias 74
2.3 PROCESADO DE CONSULTAS 77
2.3.1 Modificación de los términos de la consulta 77
2.3.2 Consultas booleanas y emparejamiento coordinado 78
2.3.3 Optimizaciones al procesado de consultas 81
2.3.4 Implementación del emparejamiento en un sistema real 83

CAPÍTULO 3. MODELOS DE RECUPERACIÓN DE INFORMACIÓN CLÁSICOS 87
Fidel Cacheda Seijo y Juan Antonio Martínez Comeche

3.1 INTRODUCCIÓN 87
3.2 MODELO BOOLEANO 90
3.3 MODELO VECTORIAL 95
3.3.1 Esquemas de ponderación 103
3.4 MODELO PROBABILÍSTICO 105

CAPÍTULO 4. EVALUACIÓN DE LA EFICACIA DE LA RECUPERACIÓN 125
Juan Antonio Martínez Comeche

4.1 INTRODUCCIÓN 126
4.2 CONCEPTOS BÁSICOS 128
4.2.1 Eficacia, eficiencia 128
4.2.2 Relevancia 129
4.2.3 Exhaustividad, precisión 130
4.3 MEDIDAS DE LA EFICACIA A PARTIR DE LA CURVA PRECISIÓN-EXHAUSTIVIDAD 136
4.3.1 Medidas basadas en puntos de la curva 136
4.3.2 Medidas que emplean interpolación y valores medios 139
4.4 MEDIDAS ORIENTADAS AL USUARIO 147
4.5 COLECCIONES DE PRUEBA 150

CAPÍTULO 5. RECUPERACIÓN DE INFORMACIÓN WEB 157
Juan Antonio Martínez Comeche y Fidel Cacheda Seijo

5.1 INTRODUCCIÓN 158

5.2 LA WORLD WIDE WEB 159
5.2.1 Componentes 160
5.2.2 Conceptos básicos 161
5.2.3 Retos 165
5.3 RECOPILACIÓN DE PÁGINAS WEB 166
5.4 PROCESAMIENTO DE PÁGINAS WEB 170
5.5 ORDENACIÓN DE RESULTADOS BASADA EN EL ANÁLISIS DE ENLACES 175
5.5.1 HITS 175
5.5.2 PageRank 179
5.6 MEDIDAS ESPECÍFICAS DE EVALUACIÓN DE LA EFICACIA 182

CAPÍTULO 6. SISTEMAS DE BÚSQUEDA Y OBTENCIÓN DE INFORMACIÓN 191
Lluís Codina Bonilla

6.1 INTRODUCCIÓN 191
6.2 LA BÚSQUEDA COMO SECTOR ECONÓMICO Y SOCIAL 193
6.3 CARACTERÍSTICAS GENERALES Y ESTRATEGIAS EN LA BÚSQUEDA DE INFORMACIÓN COGNITIVA 195
6.3.1 Qué es la información cognitiva 195
6.3.2 La búsqueda de información en el ciclo de vida de un proyecto 197
6.3.3 Componentes universales de los sistemas de búsqueda 199
6.4 NECESIDADES DE INFORMACIÓN Y LENGUAJES DE BÚSQUEDA 203
6.4.1 Tipos de búsquedas 206
6.5 LA BÚSQUEDA DE INFORMACIÓN EN LA WEB 212
6.5.1 Los motores de búsqueda 212
6.5.2 Búsqueda avanzada 214
6.5.3 Motores de búsqueda especializados: buscadores académicos 217
6.6 BASES DE DATOS 220
6.6.1 El concepto de registro 220
6.6.2 Bases de datos profesionales 222
6.6.3 Bases de datos académicas 222
6.6.4 Búsqueda avanzada 223
6.7 LA BÚSQUEDA MULTIMEDIA 224
6.7.1 Bancos de imágenes y vídeo 228
6.7.2 Repositorios *Creative Commons* 228
6.7.3 Búsqueda avanzada 229
6.8 CONCLUSIONES 230

CAPÍTULO 7. MOTORES DE BÚSQUEDA DE CÓDIGO ABIERTO..... 233
Sergio Cleger Tamayo, Carlos G. Figuerola y Julio César Rodríguez Cano

7.1 INTRODUCCIÓN 234
7.2 ¿POR QUÉ CÓDIGO ABIERTO? 235
7.2.1 Licencias de distribución 236
7.3 MOTORES DE BÚSQUEDA 238
7.3.1 Apache Lucene 239
7.3.2 Minion 239
7.3.3 Terrier 240
7.3.4 Indri 240
7.3.5 DataParkSearch 240
7.3.6 Swish-e 240
7.3.7 MG4J 241
7.3.8 mnGoSearch 241
7.3.9 Solr 241
7.4 HERRAMIENTAS COMPLEMENTARIAS 244
7.5 DESARROLLO DE SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN 247
7.6 CONCLUSIONES 257

CAPÍTULO 8. CONSTRUCCIÓN Y COMPRESIÓN DE ÍNDICES 261
Roi Blanco González

8.1 INTRODUCCIÓN 261
8.1.1 Restricciones hardware 263
8.2 CONSTRUCCIÓN DE ÍNDICES 263
8.2.1 Métodos básicos de construcción de índices 265
8.2.2 Inversión en memoria 267
8.2.3 Indexación incremental 269
8.3 COMPRESIÓN DE ÍNDICES 271
8.3.1 Compresión de datos 271
8.3.2 Modelos y códigos 275
8.3.3 Compresión de diccionarios 279
8.3.4 Compresión del fichero de ocurrencias 283

CAPÍTULO 9. MODELOS DE RECUPERACIÓN DE INFORMACIÓN II 295
David E. Losada

9.1 INTRODUCCIÓN 296

9.2 EXTENSIONES DE LOS MODELOS CLÁSICOS 297
9.2.1 Modelo vectorial con normalización por longitud basada en pivote 298
9.2.2 El modelo BM25 302
9.2.3 Los modelos basados en indexación por semántica latente 306
9.3 LOS MODELOS DE LENGUAJE ESTADÍSTICOS 310

CAPÍTULO 10. TÉCNICAS DE MODIFICACIÓN DE LA CONSULTA 321
Ángel Zazo Rodríguez, Carlos García-Figuerola Paniagua y José Luis Alonso Berrocal

10.1 INTRODUCCIÓN 321
10.2 REALIMENTACIÓN DE CONSULTAS 324
10.2.1 Realimentación de consultas para el modelo vectorial 326
10.2.2 Inconvenientes de la realimentación de consultas 327
10.2.3 Evaluación de la realimentación de consultas 328
10.3 EXPANSIÓN DE CONSULTAS 331
10.3.1 Pseudo-realimentación de consultas 333
10.3.2 Utilización de diccionarios y tesauros manuales 335
10.3.3 *Clustering* en expansión de consultas 336
10.4 TESAUROS AUTOMÁTICOS 337
10.4.1 Expansión de la consulta original 339
10.4.2 Tesauros construidos utilizando medidas de coocurrencia 342
10.4.3 Tesauros de similitud 347
10.4.4 Asociación de términos y frases: Phrase-finder 354
10.4.5 Tesauros de términos infrecuentes 355

CAPÍTULO 11. CLASIFICACIÓN DOCUMENTAL 359
Luis M. de Campos Ibáñez y Alfonso E. Romero López

11.1 INTRODUCCIÓN A LA CLASIFICACIÓN DOCUMENTAL 359
11.1.1 El proceso de clasificación documental 361
11.1.2 Representaciones de documentos 361
11.1.3 El problema de la clasificación documental 362
11.1.4 Dificultades del problema 363
11.1.5 Notación 364
11.2 EVALUACIÓN 366
11.2.1 Medidas desde el punto de vista de las categorías 367
11.2.2 Medidas desde el punto de vista de los documentos 371
11.3 MÉTODOS PARA LA CLASIFICACIÓN DOCUMENTAL 372
11.3.1 El método k-NN 372

11.3.2 El método Rocchio 378
11.3.3 El método Naive Bayes multinomial 382
11.3.4 Otros métodos 387
11.4 COLECCIONES DOCUMENTALES 388
11.4.1 Reuters-21578 388
11.4.2 Ohsumed 388
11.4.3 20 Newsgroups 389
11.4.4 RCV1 389

CAPÍTULO 12. AGRUPAMIENTO DOCUMENTAL 393
M. Eduardo Ares Brea, Javier Parapar López y Álvaro Barreiro García

12.1 INTRODUCCIÓN 394
12.1.1 Una breve definición 394
12.1.2 Aplicaciones de técnicas de agrupamiento documental 394
12.2 REPRESENTACIÓN DE DOCUMENTOS Y MEDIDAS DE SIMILITUD 396
12.2.1 Representación de documentos textuales 397
12.2.2 Medidas de distancia 401
12.3 ALGORITMOS DE AGRUPAMIENTO 402
12.3.1 Batch k-Means 403
12.3.2 Algoritmos jerárquicos aglomerativos (*-link) 406
12.4 EVALUACIÓN DE LOS ALGORITMOS 409
12.4.1 Importancia de la evaluación 409
12.4.2 Metodología 409
12.4.3 Colecciones 411
12.4.4 Métricas 413
12.5 RECUPERACIÓN DE INFORMACIÓN BASADA EN *CLUSTERS* 415
12.6 OTROS ALGORITMOS DE AGRUPAMIENTO DE DOCUMENTOS 416

CAPÍTULO 13. RECUPERACIÓN XML 419
Juan Manuel Fernández Luna y Juan Francisco Huete Guadix

13.1 INTRODUCCIÓN 420
13.2 EXTENSIBLE MARKUP LANGUAGE (XML) 423
13.2.1 ¿Qué es XML? 423
13.2.2 Elementos XML 426
13.2.3 Declaración de tipo de documento 428
13.2.4 Conceptos básicos de XPath 429
13.3 INDEXACIÓN 430

13.4 CONSULTA 435
13.4.1 NEXI 436
13.5 RECUPERACIÓN 440
13.5.1 Propagación 441
13.5.2 Mezcla 443
13.5.3 Resolución de consultas estructuradas 444
13.6 PRESENTACIÓN DE RESULTADOS 445
13.6.1 Eliminación de solapamientos 446
13.6.2 Elementos en contexto 448
13.6.3 Puntos de entrada 450
13.7 RESUMEN 451

CAPÍTULO 14. COMPORTAMIENTO INFORMACIONAL Y ESTUDIOS DE USUARIO 457
Jesús Tramullas Saz y Piedad Garrido Picazo

14.1 LOS ESTUDIOS DE USUARIO EN LA RECUPERACIÓN DE INFORMACIÓN 458
14.2 EL COMPORTAMIENTO INFORMACIONAL 459
14.3 MODELOS DE COMPORTAMIENTO INFORMACIONAL 461
14.4 LA BÚSQUEDA DE INFORMACIÓN EN EL ENTORNO DIGITAL 465
14.5 PLANTEAMIENTO DE LA INVESTIGACIÓN EN COMPORTAMIENTO INFORMACIONAL 467
14.6 TÉCNICAS BÁSICAS PARA TOMA DE DATOS 470
14.7 CONCLUSIONES 472

CAPÍTULO 15. SISTEMAS DE RECOMENDACIÓN 475
Pablo Castells Azpilicueta y Juan Francisco Huete Guadix

15.1 INTRODUCCIÓN 475
15.2 PLANTEAMIENTO Y DEFINICIONES 477
15.2.1 La tarea de recomendación 477
15.2.2 Tipos de sistemas de recomendación 479
15.3 FILTRADO COLABORATIVO 480
15.3.1 Vecinos más próximos 480
15.3.2 Métodos basados en modelo 485
15.4 RECOMENDACIÓN BASADA EN CONTENIDO 493
15.4.1 Recomendación basada en modelos probabilísticos 495
15.5 MÉTODOS HÍBRIDOS 498
15.6 LIMITACIONES DE LOS MÉTODOS DE RECOMENDACIÓN 499

15.7 EVALUACIÓN 502
15.7.1 Metodologías 503
15.7.2 Métricas 504
15.7.3 Colecciones 507

CAPÍTULO 16. BIBLIOTECAS DIGITALES 511
Pablo de la Fuente Redondo y Jesús M. Vegas Hernández

16.1 INTRODUCCIÓN 512
16.2 DEFINICIÓN Y PROPÓSITO DE LAS BIBLIOTECAS DIGITALES 513
16.2.1 Bibliotecas digitales nativas, evolucionadas e híbridas 515
16.2.2 La web como biblioteca digital 517
16.2.3 El papel de la recuperación de la información en las bibliotecas digitales 519
16.2.4 Ejemplos de bibliotecas digitales 520
16.3 RECURSOS DIGITALES. ALMACENAMIENTO Y RECUPERACIÓN MULTIMEDIA 522
16.3.1 Audio 522
16.3.2 Vídeo e imagen 523
16.3.3 Búsqueda en imágenes 524
16.3.4 Búsqueda en vídeos 524
16.4 METADATOS PARA BIBLIOTECAS DIGITALES 525
16.5 INTEROPERABILIDAD. ESTÁNDARES Y PROTOCOLOS 529
16.5.1 Introducción 529
16.5.2 Evolución histórica 530
16.5.3 Algunos elementos de interés respecto a OAI-PMH 532
16.6 MODELOS (5S, DELOS) 535
16.6.1 5S 536
16.6.2 DELOS 538
16.7 RETOS DE FUTURO 543

CAPÍTULO 17. RECUPERACIÓN DE IMAGEN 547
Antonio Mosquera González, María José Carreira Nouche y Manuel Francisco González Penedo

17.1 INTRODUCCIÓN 548
17.2 SISTEMAS DE RECUPERACIÓN DE IMAGEN BASADOS EN CONTENIDO (SRIBC) 549
17.2.1 Consulta basada en contenido 552
17.2.2 Representación del contenido 553
17.2.3 Organización y acceso 553

17.2.4 Medición de la similitud 553
17.3 PROCESADO DIGITAL DE IMÁGENES 555
17.3.1 Imagen digital 555
17.3.2 Color 557
17.3.3 Textura 561
17.3.4 Forma 562
17.3.5 Descriptores 563
17.4 RECUPERACIÓN DE IMAGEN 564
17.4.1 Extracción y representación de propiedades 564
17.4.2 Medidas de similitud 573
17.5 OTRAS DIMENSIONES 578
17.6 EL DOMINIO DE LA IMAGEN Y EL GAP SEMÁNTICO 579
17.7 EJEMPLOS DE SRIBC 580

CAPÍTULO 18. ASPECTOS AVANZADOS DE LA IMPLEMENTACIÓN DE MOTORES DE BÚSQUEDA 583

Fidel Cacheda Seijo, Diego Fernández Iglesias, Vreixo Formoso López y Rafael López García

18.1 INTRODUCCIÓN 584
18.2 *CRAWLING* DISTRIBUIDO 584
18.2.1 Asignación estática 587
18.2.2 Otros factores 591
18.3 ÍNDICE DISTRIBUIDO 592
18.3.1 Partición por términos 598
18.3.2 Partición por documentos 602
18.4 *CACHING* 607

CAPÍTULO 19. MINERÍA DE DATOS EN LA WEB 613

Marcelo Mendoza Rocha

19.1 INTRODUCCIÓN 614
19.2 MINERÍA DE CONTENIDO DE LA WEB 617
19.2.1 Extracción de información 618
19.2.2 Agrupamiento de documentos 621
19.2.3 Categorización de documentos 623
19.2.4 Minería de opiniones en la web 624
19.3 MINERÍA DE LA ESTRUCTURA DE LA WEB 625
19.3.1 Estructura microscópica de la web 626

19.3.2 Estructura macroscópica de la web 627
19.3.3 Estructura mesoscópica de la web 628
19.4 MINERÍA DE USO DE LA WEB 630
19.4.1 Archivos de *logs* 631
19.4.2 Sesiones de usuarios 632
19.4.3 Análisis de tráfico 635
19.4.4 Agrupamiento en minería de uso de la web 636
19.4.5 Reglas de asociación en minería de uso de la web 637
19.4.6 Patrones secuenciales en minería de uso de la web 640
19.4.7 Categorización en minería de uso de la web 640
19.4.8 Privacidad de datos en minería de uso de la web 641

CAPÍTULO 20. TECNOLOGÍAS WEB SEMÁNTICA Y RECUPERACIÓN DE INFORMACIÓN 649

Eduardo Peis Redondo, José Manuel Morales del Castillo y Enrique Herrera Viedma

20.1 INTRODUCCIÓN 650
20.2 UN NUEVO MODELO DE WEB 650
20.2.1 El modelo multicapa 651
20.2.2 Recuperación de información semántica: semejanzas y diferencias con los modelos tradicionales 657
20.2.3 Plataformas de desarrollo 659
20.3 VOCABULARIOS SEMÁNTICOS PARA LA RECUPERACIÓN DE INFORMACIÓN 659
20.3.1 RDF (*Resource Description Framework*) 660
20.4 SPARQL: RECUPERACIÓN DE INFORMACIÓN CON TECNOLOGÍAS SEMÁNTICAS 667
20.4.1 SPARQL: Un lenguaje para la recuperación de información semántica 668

CAPÍTULO 21. TÉCNICAS AVANZADAS DE RECUPERACIÓN DE INFORMACIÓN I 681

Enrique Alfonseca

21.1 INTRODUCCIÓN 681
21.2 TERMINOLOGÍA 682
21.3 APLICACIÓN DEL PROCESAMIENTO DE LENGUAJE NATURAL A LA RECUPERACIÓN DE INFORMACIÓN 683
21.3.1 Morfología 686
21.3.2 Desambiguación de sentidos 686
21.3.3 Similitud léxica computacional 687

21.3.4 Corrección ortográfica 689
21.4 APLICACIÓN DE LA RECUPERACIÓN DE INFORMACIÓN AL PROCESAMIENTO DE LENGUAJE NATURAL 691
21.5 SISTEMAS DE BÚSQUEDA DE RESPUESTAS 694
21.5.1 Caso de uso: búsqueda de respuestas factuales para un buscador web 696
21.6 RECUPERACIÓN DE INFORMACIÓN PATROCINADA 698
21.6.1 Visión general 698
21.6.2 Principales diferencias con respecto a la recuperación de información documental 699
21.6.3 Arquitectura de un sistema de recuperación de información patrocinada 705
21.7 RESUMEN 706

CAPÍTULO 22. TÉCNICAS AVANZADAS DE RECUPERACIÓN DE INFORMACIÓN II 711

Rafael Berlanga Llavori, Aurora Pons Porrata, David E. Losada y Ronald T. Fernández

22.1 INTRODUCCIÓN 712
22.2 DETECCIÓN Y SEGUIMIENTO DE TEMAS DE ACTUALIDAD 712
22.2.1 Conceptos preliminares 713
22.2.2 Estructura general de un sistema TDT 715
22.2.3 Principales aproximaciones a la detección de temas de actualidad 717
22.2.4 Aproximaciones al seguimiento de temas de actualidad 719
22.2.5 Evaluación de los sistemas de TDT 720
22.3 DETECCIÓN DE NOVEDAD 722
22.3.1 Más allá de la independencia en la relevancia 722
22.3.2 Detección de novedad 723
22.3.3 Novedad y diversidad 724
22.3.4 Aproximaciones a nivel de documento 726
22.3.5 Aproximaciones basadas en pasajes 727
22.3.6 Bancos de prueba para la detección de novedad 730
22.3.7 Aplicaciones de la detección de novedad 732
22.4 CONSTRUCCIÓN AUTOMÁTICA DE RESÚMENES DE DOCUMENTOS TEXTUALES 733
22.4.1 Aproximaciones para la construcción de extractos 735
22.4.2 Aproximaciones para la construcción de resúmenes 738
22.4.3 Evaluación 739

CAPÍTULO 23. PROBLEMAS EMERGENTES EN RECUPERACIÓN DE INFORMACIÓN E INVESTIGACIÓN 745
Craig MacDonald (Traducido por Nahir Seijo Saavedra)

23.1 INTRODUCCIÓN 745
23.2 INVESTIGACIÓN EMPÍRICA 746
23.2.1 Método científico 746
23.2.2 Plataformas de recuperación de información 748
23.2.3 Tareas de recuperación de información con evaluación 749
23.2.4 Publicar tu investigación 752
23.2.5 Ejemplo de investigación empírica 753
23.3 INVESTIGACIÓN DIRIGIDA POR LOS DATOS 755
23.3.1 Aprender a ordenar 756
23.3.2 Aprendiendo reformulaciones de consulta 758
23.3.3 Herramientas para la investigación con gran volumen de datos 760
23.4 LA PUBLICIDAD EN RECUPERACIÓN DE INFORMACIÓN 763
23.4.1 Búsqueda patrocinada 763
23.4.2 Anuncios asociados al contexto 764
23.5 MEDIOS SOCIALES Y BÚSQUEDA SOCIAL 765
23.5.1 La búsqueda en los medios sociales 765
23.5.2 Búsqueda social 770
23.6 CONCLUSIONES 771

BIBLIOGRAFÍA 773

MATERIAL ADICIONAL 805

ÍNDICE ALFABÉTICO 807