

■■■
**COLLANA
INFORMATICA**

FrancoAngeli

Information Retrieval

Metodi e modelli
per i motori di ricerca

Massimo Melucci

Informazioni per il lettore

Questo file PDF è una versione gratuita di sole 20 pagine ed è leggibile con



La versione completa dell'e-book (a pagamento) è leggibile con Adobe Digital Editions. Per tutte le informazioni sulle condizioni dei nostri e-book (con quali dispositivi leggerli e quali funzioni sono consentite) consulta [cliccando qui](#) le nostre F.A.Q.



Collana di informatica – Nuova serie

diretta da Arrigo L. Frisiani

Comitato scientifico:

Giovanni Adorni (Università di Genova), Luigi Benedicenti (University of Regina), Maurelio Boari (Università di Bologna), Giacomo Bucci (Università di Firenze), Virginio Cantoni (Università di Pisa), Paolo Ciancarini (Università di Bologna), Gianni Conte (Università di Parma), Paolo Corsini (Università di Pisa), Fabio Crestani (Università della Svizzera Italiana), Rita Cucchiara (Università di Modena e Reggio Emilia), Valeria De Antonellis (Università di Brescia), Gianluca Foresti (Università di Udine), Alfonso Fuggetta (Politecnico di Milano), Andrea Fusiello (Università di Verona), Salvatore Gaglio (Università di Palermo), Marco Gori (Università di Siena), Enrico Grosso (Università di Sassari), Giovanni Guida (Università di Brescia), Giuseppe Iazeolla (Università di Roma "Tor Vergata"), Sebastiano Impedovo (Università di Bari), Pieter Kritzinger (University of Cape Town), Massimo Maresca (Università di Padova), Paolo Maresca (Università di Napoli Federico II), Giuseppe Mastronardi (Politecnico di Bari), Antonino Mazzeo (Università di Napoli Federico II), Massimo Melucci (Università di Padova), Marco Mezzalama (Politecnico di Torino), Stefano Mizzaro (Università di Udine), Alfredo Petrosino (Università di Napoli "Parthenope"), Antonio Puliafito (Università di Messina), Gabriella Sanniti di Baja (CNR - Istituto di Cibernetica), Nello Scarabottolo (Università di Milano), Fabrizio Sebastiani (CNR - Istituto di Scienza e Tecnologie dell'Informazione), Giovanni Semeraro (Università di Bari), Alberto Sillitti (Libera Università di Bolzano), Giancarlo Succi (Libera Università di Bolzano), Carlo Tasso (Università di Udine), Genoveffa Tortora (Università di Salerno), Marco Vanneschi (Università di Pisa), Mario Vento (Università di Salerno), Alessandro Verri (Università di Genova), Lorenzo Vita (Università di Catania), Renato Zaccaria (Università di Genova).

I lettori che desiderano informarsi sui libri e le riviste da noi pubblicati possono consultare il nostro sito Internet: www.francoangeli.it e iscriversi nella home page al servizio "Informatemi" per ricevere via e-mail le segnalazioni delle novità.

Information Retrieval

Metodi e modelli
per i motori di ricerca

Massimo Melucci

FrancoAngeli

 **COLLANA
INFORMATICA**

Grafica della copertina: *Alessandro Petrini*

Copyright © 2013 by FrancoAngeli s.r.l., Milano, Italy.

L'opera, comprese tutte le sue parti, è tutelata dalla legge sul diritto d'autore. L'Utente nel momento in cui effettua il download dell'opera accetta tutte le condizioni della licenza d'uso dell'opera previste e comunicate sul sito www.francoangeli.it.

Indice

Indice	5
Elenco delle figure	9
Elenco degli acronimi	15
Presentazione	21
1 Introduzione	25
1.1 Dal memex ai motori di ricerca	25
1.2 Information Retrieval	28
1.3 Motori di ricerca	37
1.4 Suggerimenti bibliografici	41
1.5 Quesiti	43
2 Metodi di indicizzazione	47
2.1 Introduzione	47
2.2 Indicizzazione ed efficacia	49
2.3 Analisi lessicale	52
2.4 Stemming	56
2.5 Costruzione dei termini	61
2.6 Controllo statistico dell'indice	62
2.7 Agente di ricerca	75
2.8 Suggerimenti bibliografici	90
2.9 Quesiti	92
3 Metodi di reperimento e ordinamento	95
3.1 Introduzione	95
3.2 Operatori logici	98
3.3 Livello di coordinamento	99

3.4	Implementazione del reperimento	102
3.5	Suggerimenti bibliografici	109
3.6	Quesiti	110
4	Metodi di valutazione	115
4.1	Introduzione	115
4.2	Base di partenza e gruppo di controllo	119
4.3	Misure di valutazione	122
4.4	Collezione sperimentale	136
4.5	Iniziative di valutazione	140
4.6	Suggerimenti bibliografici	145
4.7	Quesiti	146
5	Modelli di indicizzazione e reperimento	149
5.1	Introduzione	149
5.2	Modello logico	151
5.3	Modello vettoriale	153
5.4	Modello probabilistico	169
5.5	Modello linguistico	186
5.6	Suggerimenti bibliografici	193
5.7	Quesiti	194
6	Metodi avanzati di indicizzazione	197
6.1	Introduzione	197
6.2	Valutazione e indicizzazione	198
6.3	Àncore dei link	200
6.4	Autorevolezza delle pagine	201
6.5	Pagine duplicate	209
6.6	Compressione dei dati	210
6.7	Collocazioni e termini	214
6.8	Analisi della semantica latente	216
6.9	Indicizzazione collaborativa	225
6.10	Suggerimenti bibliografici	226
6.11	Quesiti	227

7	Metodi avanzati di reperimento e ordinamento	229
7.1	Introduzione	229
7.2	Sistemi paralleli e distribuiti	232
7.3	Ordinamento per autorevolezza	238
7.4	Reperimento per semantica latente	249
7.5	Espansione delle interrogazioni	256
7.6	Retroazione di rilevanza	261
7.7	Suggerimenti bibliografici	276
7.8	Quesiti	277
8	Metodi di Machine Learning	279
8.1	Introduzione	279
8.2	Separabilità dei punti	281
8.3	Classificazione	285
8.4	Ordinamento dei documenti	299
8.5	Raggruppamento	301
8.6	Suggerimenti bibliografici	305
8.7	Quesiti	306
9	Contesti applicativi	309
9.1	Introduzione	309
9.2	Immagini, suoni e musica	311
9.3	Reti sociali e blog	318
9.4	Personalizzazione	324
9.5	Pubblicità digitale	330
9.6	Analisi dei dati	333
9.7	Suggerimenti bibliografici	337
9.8	Quesiti	337
	Suggerimenti bibliografici	341
	Indice analitico	359

Elenco delle figure

1.1	Il memex immaginato da Bush (1945)	26
1.2	Distribuzione delle lingue	33
1.3	Architettura funzionale di un Information Retrieval System	34
1.4	Information Retrieval System	35
1.5	Collezione di documenti e relativo indice	36
1.6	Data center	38
1.7	World Wide Web (web) visto, visibile e invisibile	39
1.8	Profondo web	40
2.1	Sorgente della home-page di un quotidiano	50
2.2	Indicizzazione esaustiva e indicizzazione specifica	52
2.3	Analisi lessicale di alcune parole	53
2.4	Stop-list per l'inglese	54
2.5	Stop-list per l'italiano	55
2.6	Frammento iniziale dell'algoritmo di Porter per lo stemming	57
2.7	Riduzione di parole italiane a radici comuni	57
2.8	Lista di affissi	59
2.9	Perdita d'informazione dopo rimozione delle stop word e stemming	60
2.10	Calcolo dei lemmi	60
2.11	Etichettatura delle parti del discorso (Part-of-Speech (POS) tagging)	62
2.12	Distribuzione di frequenza di in-link della collezione sperimentale WT10G	64
2.13	Le cinquanta parole delle query più frequentemente inviate a un motore di ricerca nel 2002 (riportate da Jansen e Spink (2006))	64
2.14	Distribuzione di frequenza delle parole della collezione sperimentale CACM	64

2.15	Legge di Heap	65
2.16	Indice di una collezione	69
2.17	Architettura di un indice	71
2.18	Semplice algoritmo d'indicizzazione	72
2.19	Distribuzione delle trenta parole più frequenti del manuale di MySQL	72
2.20	Struttura di un URL e dialogo in HTTP	75
2.21	Funzionamento di un agente di ricerca	76
2.22	Agente di ricerca e DNS server	76
2.23	Dialogo con un server web	77
2.24	Illustrazione di un metodo per determinare punti di partenza	80
2.25	Esplorazione in ampiezza	80
2.26	Algoritmo d'esplorazione in ampiezza	81
2.27	Esplorazione in profondità	81
2.28	Algoritmo d'esplorazione in profondità	82
2.29	Evoluzione di una coda ed evoluzione di una pila	83
2.30	Architettura di un sistema di Information Retrieval (IR) con un agente di ricerca	84
2.31	Canonizzazione degli Uniform Resource Locator (URL)	85
2.32	Sitemap	89
2.33	Robots.txt	89
2.34	Really Simple Syndication (RSS)	90
3.1	Rapporti di grandezza tra documenti reperiti, rilevanti e non reperiti	96
3.2	Operatori logici	98
3.3	Descrittori e insiemi di documenti	98
3.4	Operazioni con FONDI e POTA	103
3.5	Indice risultante dalla fig. 3.3	104
3.6	Elaborazione di un'interrogazione booleana	104
3.7	Schema di un algoritmo di reperimento	104
3.8	Term-At-A-Time (TAAT) e Document-At-A-Time (DAAT)	105
3.9	Algoritmo di reperimento TAAT con operatori logici	106

3.10	Reperimento DAAT	107
3.11	Algoritmo di reperimento TAAT con terminazione anticipata	110
4.1	Quadrato latino	121
4.2	Quadrato greco-latino	122
4.3	Tabella di contingenza per il calcolo di richiamo e precisione	124
4.4	Run di venti documenti	127
4.5	Due configurazioni a confronto	129
4.6	Misura E	129
4.7	Calcolo di richiamo e precisione per due ranking	130
4.8	Interpolazione di richiamo e precisione	132
4.9	Caso peggiore, normale e ideale di Cumulative Gain	134
4.10	Numero di scambi e τ di Kendall	135
4.11	Metodi di costruzione di una collezione sperimentale	138
4.12	Pooling method	139
4.13	Fase di addestramento	139
4.14	Misurazione e test statistico delle ipotesi	140
4.15	Fase di test sperimentale di una configurazione	140
4.16	Prime collezioni sperimentali	141
4.17	Composizione della collezione TIPSTER	143
4.18	Topic TREC n. 301	143
4.19	Documento TREC rilevante al topic n. 301	144
5.1	Teoria, modelli, esperimenti e realtà fisica	150
5.2	Vettori di documenti e interrogazioni nello spazio vettoriale a due dimensioni	154
5.3	Vettori di documenti e interrogazioni nello spazio vettoriale a tre dimensioni	154
5.4	Modello vettoriale e multimedia	155
5.5	Cluster Hypothesis	157
5.6	Coseno tra vettori nello spazio vettoriale a due dimensioni	161
5.7	Coseno tra vettori nello spazio vettoriale a tre dimensioni	161
5.8	Nozione di normalizzazione a perno	163

5.9	Coseno tra vettori e ruolo delle relazioni tra descrittori	164
5.10	Funzione di IR con il modello vettoriale	166
5.11	Decisione secondo il modello probabilistico	169
5.12	Stima del modello probabilistico	170
5.13	Spazio probabilistico	171
5.14	Tabella dei costi di decisione nel modello probabilistico	172
5.15	Costi di imprecisione e perdita	173
5.16	Curse of dimensionality	178
5.17	Indipendenza stocastica e indipendenza stocastica condizionata	179
5.18	Tabella di contingenza per il modello probabilistico	183
5.19	Language Model	187
5.20	Bigrammi e probabilità	187
5.21	Metafora di un LM	188
5.22	Mistura di LM	190
5.23	Lisciamento di LM	191
6.1	Àncore	201
6.2	Pagine e link web che non formano una catena di Markov ergodica	203
6.3	Pagine e link web che formano una catena di Markov persistente, ma periodica	204
6.4	Algoritmo PageRank	208
6.5	Rilevamento di pagine duplicate mediante hashing e fingerprinting	210
6.6	Compressione e codifica dei caratteri	211
6.7	Albero binario di compressione	212
6.8	Metodo di compressione γ	213
6.9	Metodo di compressione δ	214
6.10	Metodo di compressione basato sulle differenze	214
6.11	Due basi per lo stesso documento	218
6.12	Quattro vettori nello spazio a due dimensioni	225
7.1	Relevance Feedback	231

7.2	IRS parallelo	232
7.3	Ripartizione di una matrice di occorrenza di un Information Retrieval System (IRS) parallelo	233
7.4	IRS distribuito	235
7.5	Mutuo rinforzo	239
7.6	Algoritmo Hyperlinked Induced Topic Search (HITS)	240
7.7	Grafo di sette nodi per calcolare autorevolezza e centralità	244
7.8	Popolarità, conformismo, centralità e autorevolezza	245
7.9	Un grafo per HITS	246
7.10	Insieme radice	247
7.11	Insieme radice, insieme base e HITS	248
7.12	Confronto tra HITS e PageRank	249
7.13	Piccola collezione usata per illustrare il modello di analisi latente (tratta dall'articolo di Deerwester e altri (1990))	252
7.14	Disposizione di documenti e interrogazione dopo LSA	254
7.15	Decomposizione a Valori Singolari	256
7.16	Tipi di Relevance Feedback	262
7.17	Stato iniziale della retroazione	263
7.18	Reperimento prima della retroazione	263
7.19	Addestramento durante la retroazione	263
7.20	Assegnazione del giudizio di rilevanza durante la retroazione	263
7.21	Reperimento dopo la retroazione	264
7.22	Cluster Hypothesis verificata e non verificata	264
7.23	Relevance Feedback (RF) vettoriale positivo e negativo	266
7.24	Algoritmi di RF	272
8.1	Machine Learning e spazi	280
8.2	Separabilità di un insieme di punti	283
8.3	Trasformazione dei dati di un training set	285
8.4	Separabilità perfetta di un insieme di punti	290
8.5	Iperpiano di una Support Vector Machine (SVM)	291
8.6	Calcolo dell'iperpiano ottimo	295
8.7	Kernel trick	296

8.8	Non separabilità dei punti	298
8.9	Raggruppamento di otto punti in tre gruppi in \mathbb{R}^2	302
8.10	Algoritmo k-Means	303
8.11	Rappresentazione del metodo Scatter and Gather	304
8.12	Punti distribuiti in tre gruppi e tre classi	305
8.13	Misura F	305
9.1	Frammento di testo apparso sullo schermo e poi riconosciuto mediante Optical Character Recognition (OCR)	312
9.2	Struttura di un'immagine	314
9.3	Notazione della melodia di un brano musicale segmentata in profili melodici	316
9.4	Profili melodici di un brano ottenuti con uno stemming per la melodia	317
9.5	Struttura di un blog	322

Elenco degli acronimi

ASCII	American Standard Code for Information Interchange
ASK	Anomalous State of Knowledge
CGI	Common Gateway Interface
CG	Cumulative Gain
CG	Discounted Cumulative Gain
CLEF	Cross-Language Evaluation Forum
CLIR	Cross-Language IR
CMS	Content Management System
CPA	Cost Per Action
CPC	Cost Per Click
CPM	Cost Per Mille
CTD	Click-Through Data
DAAT	Document-At-A-Time
DARPA	Defense Advanced Research Program Agency
DNS	Domain Name System
DVS	Decomposizione a Valori Singolari
DV	Term Discrimination Value
EM	Expectation Maximization
FIFO	First-In First-Out
FND	Forma Normale Congiuntiva
GPS	Global Positioning System
HITS	Hyperlinked Induced Topic Search
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IDF	Inverse Term Frequency
IE	Information Extraction
IP	Internet Protocol
IRF	Implicit RF
IRS	Information Retrieval System

IR	Information Retrieval
LIFO	Last-In First-Out
LM	Language Model
LR	Learning to Rank
LSA	Latent Semantic Analysis
MAP	Mean Average Precision
ML	Machine Learning
MMIR	IR multimediale
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
NIST	National Institute of Standard and Technology
NLS	oN Line System
NTCIR	NII-NACSIS Test Collection for IR Systems
OLTP	Online Transaction Processing
OCR	Optical Character Recognition
OPAC	Online Public Access Catalogue
POS	Part-of-Speech
PRP	Probability Ranking Principle
QA	Question Answering
QE	Query Expansion
QLM	Query-Likelihood Model
RF	Relevance Feedback
RGB	Red Green Blue
RSS	Really Simple Syndication
SGBD	Sistema di Gestione di Basi di Dati
SML	Statistical Machine Learning
SMS	Short Message System
SQL	Structured Query Language
SVM	Support Vector Machine
TAAT	Term-At-A-Time
TFIDF	Term Frequency \times Inverse Document Frequency
TF	Term Frequency
TREC	Text REtrieval Conference

TRW	Term Relevance Weight
URL	Uniform Resource Locator
VSM	Vector Space Model
WWW	World Wide Web (web)
XML	eXtensible Markup Language

per mia moglie, Alessandra
e
per mio figlio, Oleg