

Estrazione e gestione di informazioni da collezioni testuali

Giuseppe Amodeo
Università degli studi dell'Aquila

Supervisor: Prof. Michele Flammini
Supervisor FUB: Gianni Amati, PhD



Ambito della ricerca

- Estrazione, recupero e classificazione di informazione da documenti testuali
- Applicazioni
 - Web/Intranet Search Engines
 - Enterprise Search
 - Blog Search
 - Opinion Retrieval
 - Question Answering
 - Patent and Legal Search, Bioinformatics
 - Expert Search
 - Integrazione DBMS e Search Engines



Problemi e Stato dell'arte

- **Indicizzazione**

- Dimensionalità delle collezioni
 - Sistemi efficienti di indicizzazione (Terrier SE distribuito e concorrente , versione open-source FUB)
 - Qualità delle collezioni (alta dimensionalità del lessico e molto rumore)
- Eterogeneità del formato dei documenti (XML, RSS, pdf, Excel, word, ecc.)
- Spamming, lingue,
- Eterogeneità dei contenuti (term-partitioning/document-partitioning degli indici)
- Integrazione effettiva DB e SE



Problemi e Stato dell'arte

- **Modelli**

- Tecniche di *Data Fusion* e *Learning to Rank* (*Data Mining*) per integrazione di sorgenti informative eterogenee
- *Feature selection*, per la riduzione della dimensionalità dello spazio degli indici
- Analisi della struttura dei documenti, come *title*, *body*, *anchor text*, *tags* (*Field Information Retrieval*)
- Ogni applicazione (*blog search*, *web search*, *expert search* ecc.) ha un modello specifico.



Text Mining

- Ambito di applicazione del Data Mining all'elaborazione del Testo (Text Mining)
 - Estrazione di informazione da testo
 - Entity and Concept recognition
 - Classificazione del testo e sue applicazioni
 - Sentiment Analysis, Topical Opinion Retrieval
- Tecniche analizzate
 - Support Vector Machine
 - Natural Language Processing, come catene di Markov



Opinion Retrieval

- Opinion Retrieval = Recuperare i documenti che esprimono un'opinione su X

“What do people think about X?”

- X = persona, luogo, organizzazione, prodotto, evento, tecnologia, ecc.
- Scoprire qual'è il sentimento pubblico su un particolare target X.



Opinion Retrieval

- Tecniche di recupero mediante l'uso di dizionari
 - Costruzione automatica da un training set di un dizionario sentimentale: modelli basati su funzioni di divergenza tra probabilità osservate e quelle a priori
 - Utilizzo dei dizionari in un processo di recupero a due fasi, ovvero basato su un processo di revisione del recupero iniziale (*reranking*)
- Sperimentazione condotta
 - Tramite *cross validation*: validazione sia del processo di generazione dei dizionari sia del modello di recupero



Opinion Retrieval

Sentimental Dictionary Excerpt

abide	0.0023	inaccurate	0.0064
abject	0.0031	inane	0.0009	wish	0.0060
absolute	0.0029	inappropriate	0.0028	wonder	0.0068
absurd	0.0076	incapable	0.0072	wonderful	0.0025
abusive	0.0047	incessant	0.0052	woo	0.0024
abyss	0.0008	inclin	0.0043	worri	0.0020
acclaim	0.0008	incoherent	0.0010	worse	0.0044
accuse	0.0012	incompetent	0.0012	worst	0.0041
activist	0.0023	incomprehensible	0.0018	worth	0.0018
actual	0.0069	inconvenient	0.0026	worthless	0.0097
admir	0.0024	incredible	0.0048	worthwhile	0.0016
admirable	0.0030	indefensible	0.0011	wound	0.0046
admire	0.0011	indicative	0.0017	wrath	0.0021
admit	0.0063	indifferent	0.0029	yeah	0.0070
.	...	indispensable	0.0033	yearn	0.0049



Risultati TREC 07

Group	Baseline	Re-rank by Opinion	% Increase
FUB	0.2727	0.3210	17.71%
Univ. of Glasgow	0.2817	0.3264	15.87%
Indiana Univ.	0.2537	0.2894	14.07%
Univ. of Arkansas LR	0.2554	0.2911	13.89%
Dalian Univ.	0.2890	0.3190	10.38%
Univ. of Waterloo	0.2486	0.2631	5.83%



Risultati TREC 08

	Precision
Baseline for all groups	0.3822

Best 5 Groups	Re-rank precision	% Increase
Korean University	0.4189	9.60
Illinois University	0.4067	6.41
Indiana University	0.4023	5.26
FUB	0.4007	4.81
Glasgow University	0.3964	3.72
Waterloo University	0.3381	-11.54



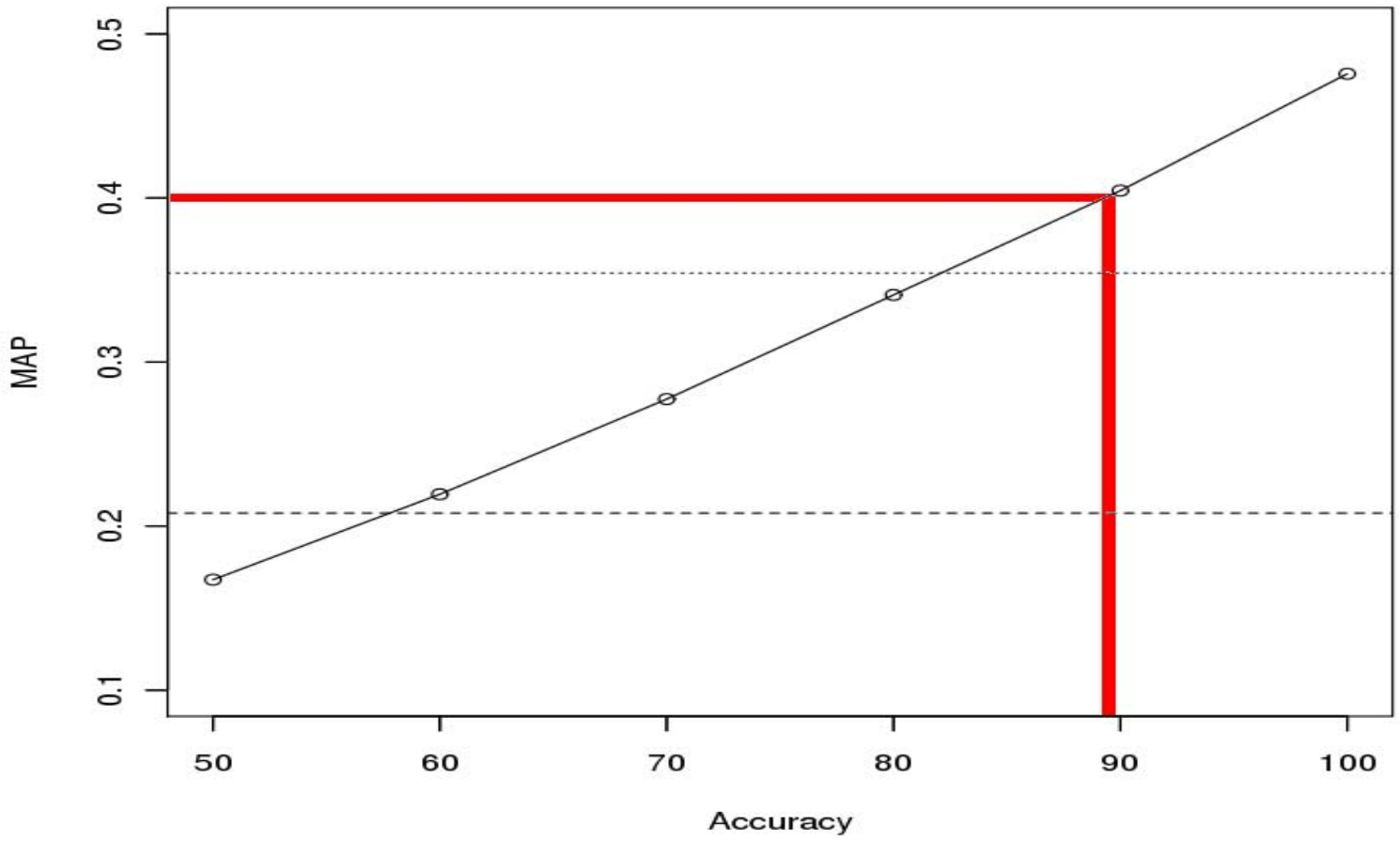
Valutazione di sistemi di OR

- La valutazione dei sistemi di Opinion Retrieval risulta essere particolarmente complessa
 - Discernere il contributo sui risultati della fase di recupero su topic e quella di analisi del contenuto d'opinione
- Come affrontare il problema della valutazione dei risultati?
 - Studio di tecniche di valutazione basate su analisi delle variazioni di precisione al variare della accuratezza dei classificatori di Opinion Detection



Valutazione di sistemi di OR

BL4



Bibliografia

- Gianni Amati and C.J. van Rijsbergen. Probabilistic Models of Information Retrieval Based on Measuring Divergence From Randomness, *ACM Transactions on Information Systems*, 20(4):357-389, 2002
- Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma, *Terrier: A High Performance and Scalable Information Retrieval Platform*, SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). Seattle, Washington, USA.
- G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, G. Gambosi “Automatic Construction of an Opinion-Term Vocabulary for Ad Hoc Retrieval”, 30th European Conference on IR Research, ECIR 2008, LNCS 4956, Springer, pp.89-100
- G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, G. Gambosi, *FUB, IASI-CNR and University of Tor Vergata at Trec 2008 Blog Track* - Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, 2008.
- G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, G. Gambosi “A uniform theoretic approach to opinion and information retrieval” chapter of book “Intelligent Information Access” – Springer. To appear

