Un prototipo per la ricerca di opinioni sui blog dedicati alle trasmissioni televisive d'interesse nazionale

Giambattista Amati, Marco Bianchi, and Giuseppe Marcone

Fondazione Ugo Bordoni
Viale del Policlinico, 147
00161 Rome, Italy
gba@fub.it
mbianchi@fub.it
gmarcone@fub.it

Sommario In questo lavoro si riporta l'esperienza maturata durante la realizzazione di un prototipo per la ricerca delle opinioni pubblicate sui blog dedicati ai programmi televisivi trasmessi dalle emittenti italiane. Il contributo per la comunità scientifica italiana dell'Information Retrieval è duplice: da un lato si presenta il primo benchmark per il task dell'opinion finding applicato a piattaforme di blog in lingua italiana e si riporta la metodologia adottata per la sua creazione. In secondo luogo si descrive l'architettura di un sistema che implementa un algoritmo dictionary-based di comprovata efficacia utile ad affrontare il problema dell'opinion finding su testi in lingua italiana. Tale sistema, basato su componenti open-source, supporta la creazione di ulteriori benchmark a partire dai quali genera in modo automatico i dizionari necessari al funzionamento dell'algoritmo che implementa. Proprio quest'ultima funzionalità è da considerarsi strategica per la comunità scientifica vista la scarsa disponibilità di risorse linguistiche italiane e il costo necessario alla loro creazione e aggiornamento.

Keywords: Information Retrieval, Sentiment Analysis, Opinion Finding

1 Introduzione

La disciplina scientifica il cui fine è lo sviluppo di tecniche di estrazione della conoscenza da documenti contenenti opinioni è nota in letteratura con il nome di *sentiment analysis*. Oggigiorno le principali comunità scientifiche che si occupano di sentiment analysis sono due: la comunità dell'Intelligenza Artificiale, che utilizza prevalentemente tecniche di Processamento del Linguaggio Naturale (NLP) finalizzate alla classificazione automatica dei documenti e di estrazione puntuale di informazioni da documenti [10], e la comunità dell'Information Retrieval, che ha specializzato il problema al mondo del Web. La differenza fondamentale tra le ricerche effettuate dalla due comunità risiede nella tipologia di collezioni che prendono a riferimento.

Infatti la comunità dell'Intelligenza Artificiale basa i sui studi, nella maggior parte dei casi, su collezioni composte esclusivamente da documenti contenenti opinioni e strutturalmente omogenei. In questi casi il problema diventa quello di classificare, ad

esempio, i documenti contenenti opinioni positive da quelli contenenti opinioni negative, oppure quello di estrarre informazioni puntuali, come le caratteristiche più o meno apprezzate di un prodotto commerciale [4,11].

Diversamente numerosi studi della comunità dell'Information Retrieval sono basati su collezioni Web. Tali collezioni sono caratterizzate, tra l'altro, dalla presenza di documenti non contenti opinioni e dalla eterogeneità a livello strutturale delle pagine Web, quasi sempre scaricate da siti diversi. In questo scenario la sentiment analysis viene generalmente considerato un problema di re-rank a due fasi [7]: nella prima si cerca di individuare i documenti che sono rilevanti rispetto all'esigenza informativa dell'utente (topic), indipendentemente dalla presenza di opinioni rispetto al topic cercato; nella seconda l'insieme dei documenti individuati a valle di un processo di riordinato (re-rank) in funzione presenza o assenza di opinioni. L'intero processo di recupero, denominato *opinion-finding*[1], ha quindi l'obiettivo di recupere pagine Web contenenti opinioni rispetto ad un determinato topic.

Questo lavoro si inquadra nell'ambito delle attività finalizzate alla realizzazione del prototipo di un motore ricerca in grado di trovare le opinioni che i telespettatori di programmi televisivi riportano sui blog in lingua italiana. Tale motore può essere utile sia ai telespettatori che vogliono leggere, o scrivere, recensioni o commenti relativi ai loro programmi preferiti, sia alle emittenti televisive che intendono indagare l'opinione del popolo del Web in merito ai programmi trasmessi.

A partire dall'analisi dei requisi è stato eseguito uno studio dello stato dell'arte che ha confermato quanto già riportato in [7] e cioè che le principali tecniche di opinion-finding possono essere classificate in due principali categorie: strategie basate su classificatori (classification-based) e strategie basate su dizionari (lexicon-based). Considerata l'efficacia dimostrata da quest'ultima classe di tecniche, si è deciso di applicare la strategia lexicon-based presentata in [1]. Tale tecnica è di particolare interesse non solo perchè si è dimostrata tra le più performanti nelle varie edizioni della TREC, ma anche perchè permette la creazione *automatica* di dizionari di termini "portatori" di opinione (opinion-bearing terms). Considerato che non esistono, ad oggi, dizionari italiani per la sentiment analysis si ritiene che la realizzazione di un prototipo che supporti la generazione automatica di dizionari italiani sia da considerarsi un valore aggiunto per l'intera comunità scientifica italiana. Il basso costo di generazione, e quindi di aggiornamento, del dizionario va infatti incontro a quel requisito di economicità che dovrebbe contraddistinguere la voce "costo di manutenzione" di ogni sistema software. È tuttavia necessario precisare che a fronte di un risparmio in termini di impiego di risorse umane, si è costretti ad accettare la presenza, all'interno del dizionario, di termini "intrusi", ovvero termini che, almeno in apparenza, non sono portatori di opinione.

Il prototipo, presentato nella Sezione 4, è caratterizzato dall'originale integrazione tra la catena di tool nutch-solr [3,12], lo standard di fatto della comunità dell'open-source per la realizzazione di motori di ricerca, e il framework Terrier [8], strumento di Information Retrieval estremamente diffuso nella comunità scientifica e necessario per l'implementazione della tecnica presentata in [1]. Grazie a tale integrazione è possibile soddisfare anche due importanti requisiti non funzionali aventi come obiettivo la realizzazione di un motore di ricerca che sia:

[1] Nomenclatura introdotta nell'ambito della Blog Track di TREC 2006 [9,13].

1. in grado di scalare alle dimensioni tipiche del Web, proprio per rendere possibile il monitoraggio di una porzione significativa della blogosfera italiana;
2. caratterizzato da un basso costo di realizzazione e manutenzione.

Il prototipo supporta anche il processo per la realizzazione di generici benchmark per l'Information retrieval. Proprio grazie a tale supporto, è stato creato il primo benchmark per la sperimentazione di soluzioni al problema dell'opinion finding su testi italiani, come riportato nella Sezione 3.

Tra i contributi del lavoro si evidenzia la descrizione di due diverse strategie per l'acquisizione dei contenuti pubblicati su blog con relativi vantaggi e svantaggi. Le considerazioni relative ai due approcci, oggetto della Sezione 2, sono generalizzabili a tutti i contesti in cui i contenuti su cui effettuare le ricerche sono fortemente condizionati da eventi esterni alla Rete quali, appunto, la trasmissione di un programma televisivo o la diffusione di notizie.

Infine la Sezione 5 conclude il lavoro.

2 Metodologie per l'acquisizione dei contenuti di un blog

Al fine di rendere più chiara la presentazione delle metodologie per l'acquisizione dei contenuti pubblicati su un blog, le componenti logiche di una piattaforma di blogging da tenere in considerazione sono:

- *permalink*, o link permanente: URL relativo a una pagina Web che contiene un post e i relativi commenti. Il contenuto testuale raggiungibile con permalink è l'obiettivo finale dell'attività di acquisizione;
- *homepage*: pagina dinamica sulla quale vengono riportati gli ultimi post pubblicati e i relativi permalink;
- *navigatore*: strumento che implementa una tecnica di "navigazione a faccette" (faceted search) al fine di semplificare la ricerca dei post di interesse. In genere su tutte le pagine di un blog sono presenti un numero significativo di navigatori;
- *pagina di aggregazione*: pagine dinamiche che contengono tutti i post che soddisfano un criterio di navigazione a faccette (ad esempio tutti i post pubblicati in un determinato mese);
- *RSS feed*: file in formato RSS (Really Simple Syndication)[2] sul quale vengono periodicamente riportati i permalink degli ultimi post pubblicati.

A partire dalle componenti logiche appena elencate, l'acquisizione dei contenuti pubblicati su una piattaforma di blog può avvenire adottando due diverse strategie a seconda delle esigenze.

La prima strategia consiste nell'effettuare una sorta di "fotografia" dei contenuti presenti sull'intero blog mediante attività di *crawling*. In questo caso l'idea è quella di fornire al crawler la URL della homepage e lasciare a quest'ultimo la responsabilità di navigare (in modo automatico) sul blog al fine di scaricarne i contenuti. Il vantaggio di questa tecnica è dato dalla completezza del risultato (alta recall): ciò significa che al

[2] Per le specifiche del procollo RSS far riferimento al sito `http://www.rssboard.org/`

termine dell'attività di crawling tutti i contenuti indirizzati da permalink saranno stati scaricati. Lo svantaggio principale sarà dato dalla bassa precisione (precision) poichè un crawler non è in grado di distinguere un permalink da altre URL (a meno dello sviluppo di filtri di URL specializzati per le singole piattaforme di blog, operazione che però ha controindicazioni in termini di costo di scalabilità e manutenzione del sistema). Di conseguenza il crawler scaricherà anche l'homepage e, soprattutto, le pagine di aggregazione di post. Queste ultime sono da considerarsi "rumore", in quanto replicano il testo dei post già raggiungibile seguendo i permalink. Vale la pena evidenziare che il numero di pagine di aggregazione è proporzionale al numero di navigatori e che può, di conseguenza, anche essere consistente.

La seconda strategia consiste nell'individuare e scaricare i permalink dei nuovi post mediante il *monitoraggio* degli RSS feed. Questa tecnica, adottata per la realizzazione di due benchmark internazionali utilizzati nell'ambito delle gare TREC [6], ha il vantaggio di produrre una elenco composto esclusivamente da permalink. Purtroppo però il monitoraggio degli RSS non permette lo scaricamento dei vecchi permalink, ossia delle URL che sono già state eliminate dall'RSS perché la pubblicazione del post da loro riferito non rappresenta più una "novità" per gli utenti del blog.

Indipendentemete dai vantaggi e dagli svantaggi, l'adozione della prima strategia è obbligatoria quando si vuole includere nella collezione post poco recenti, o quando non si è nella condizione di aspettare il tempo necessario per eseguire il monitoraggio degli RSS. Nel caso dell'acquisizione dei contenuti pubblicati su blog che trattano trasmissioni televisive, la prima strategia è da considerarsi una scelta obbligata anche quando le trasmissioni di interesse sono già andate in onda.

3 Un benchmark per l'opinion finding task in lingua italiana

In genere un tipico benchmark di Information Retrieval è composto da:

1. *un insieme di topic*, ovvero un elenco di esigenze informative, esprimibili come query, definite da esperti di dominio in modo tale da essere rappresentative dell'utenza reale;
2. una *collezione di documenti*;
3. un *insieme di valutazioni*, ottenute grazie all'apporto degli esperti di dominio, nel quale a ogni topic viene associato un sottoinsieme di documenti della collezione rilevante rispetto a tale topic.

Coerentemente con quanto appena riportato, la creazione del benchmark per l'opinion finding task applicato al dominio delle trasmissioni televisive trasmesse da emittenti TV ha richiesto le seguenti attività:

- definizione di un elenco di trasmissioni televisive su cui eseguire ricerche (topics) e che, almeno sulla carta, suscitino dibattito tra gli utenti del Web. Nel caso specifico sono state individuate 65 trasmissioni televisive di vario genere (es. attualità, reality, fiction, satira, ecc.);
- individuazione delle piattaforme di blog dalle quali acquisire i contenuti: grazie al coinvolgimento di esperti di dominio è stato stilato un elenco di 100 URL (seeds) relative a piattaforme di blog tematiche;

– conduzione dell'attività di crawling. Considerato che molti dei programmi selezionati non andavano in onda durante il periodo dedicato all'acquisizione dei contenuti (periodo che va dai primi di novembre 2010 e alla prima metà di dicembre 2010), si è deciso di adottare la strategia del crawling dei blog. Al termine dell'attività di crawling la collezione risulta composta da 6.067.494 pagine HTML, tra le quali sono compresi permalink, homepage, pagine di aggregazione e duplicati (per lo più ottenuti a causa dell'utilizzo di pagine dinamiche da parte delle piattaforme di blog);
– rimozione dei duplicati. Successivamente alla fase di crawling, le pagine duplicate sono state rimosse a seguito di un controllo sul valore MD5 e riducendo il numero di documenti della collezione a 1.531.837 pagine Web;
– creazione dell'insieme delle valutazioni. Dopo aver indicizzato l'intera collezione con Lucene, sono state selezionate 30 trasmissioni televisive tra le 65 precedentemente individuate e, per ognuna di queste, è stato eseguito un recupero di 200 risultati utilizzando il nome della trasmissione come query e il search handler di seguito riportato[3]:

```
<requestHandler name="/topicSearch"
  class="solr.SearchHandler">
    <lst name="defaults">
      <str name="defType">dismax</str>
      <str name="echoParams">explicit</str>
      <float name="tie">0.01</float>
      <str name="qf">content^1.0</str>
      <str name="pf">anchor^1.0 title^0.1</str>
      <int name="ps">3</int>
      <str name="fl">url</str>
      <bool name="hl">false</bool>
    </lst>
</requestHandler>
```

Per ognuna delle 6.000 URL così individuate, un esperto di dominio ha registrato (mediante un'applicazione Web appositamente sviluppata) il proprio parere in merito a:

- la pertinenza della pagina recuperata rispetto al topic. Più precisamente la domanda di riferimento per i valutatori è stata: "La pagina Web associata alla URL è pertinente rispetto alla trasmissione televisiva in oggetto?" con possibili valori di risposta: *rilevante* e *non rilevante*.
- la tipologia di pagina Web. In questo caso la domanda di riferimento per i valutatori è stata: "La pagina Web associata alla URL è una home-page, una pagina di aggregazione o un permalink?", con possibili valori di risposta: *homepage*, *pagina di aggregazione post*, *permalink* o *altro*.
- la presenza di opinioni nella pagina Web, con la seguente domanda di riferimento: "La pagina Web associata alla URL contiene opinioni positive, opinioni negative, opinioni miste o nessuna opinione?" con possibili valori di risposta: *nessuna opinione*, *opinioni positive*, *opinioni negative* o *opinioni miste*.

[3] Per approfondimenti sui parametri del request handler si veda `http://wiki.apache.org/solr/DisMaxQParserPlugin`

4 Un prototipo per la ricerca di opinioni sui blog

Nell'ambito del progetto TV++ condotto dalla Fondazione Ugo Bordoni e dall'Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione è stato realizzato un prototipo per l'applicazione della tecnica dell'opinion finding al dominio dei blog dedicati ai programmi televisivi trasmessi dalle emittenti italiane.

Come già anticipato nella Sezione 1, il prototipo implementa la tecnica dictionary-based presentata in [1]. In estrema sintesi tale tecnica prevede due passi principali:

1. Costruzione automatica di un dizionario composto da termini che caratterizzano i documenti in cui vengono espresse opinioni (termini *opinion-bearing*). A ogni termine del dizionario è associato un peso che fornisce, informalmente parlando, una misura del suo grado di soggettività, ove per termine soggettivo s'intende un termine che usualmente compare in una frase soggettiva. Ad esempio i termini "credo" e "penso" si suppone che abbiano un grado di soggettività alto in quanto spesso usati in frasi che esprimono opinioni. La costruzione automatica del dizionario avviene adottando un approccio di tipo statistico-probabilistico basato su modelli della famiglia *Divergence from Randomness* (DFR) [2].
2. Esecuzione di un algoritmo di opinion retrieval che, sfruttando le informazioni presenti nel dizionario appena descritto, assegni ad ogni documento uno score funzione sia della rilevanza rispetto alla query, sia della presenza di opinioni nel testo. L'algoritmo tenderà pertanto a far emergere nelle prime posizioni i documenti rilevanti e contenenti opinioni, a discapito dei documenti solo rilevanti o contenenti solo opinioni.

A livello realizzativo, l'implementazione della metodologia appena richiamata rende necessario l'utilizzo di Terrier [8], l'unico framework per l'IR che, ad oggi, supporta nativamente i modelli di recupero DFR. Terrier è quindi indispensabile sia per la creazione del dizionario che per l'implementazione dell'algoritmo di re-rank. D'altro canto la comunità dell'open-source di Apache Software Foundation[4] sta, già da qualche anno, concentrando le energie sullo sviluppo del crawler Nutch [3] e del framework per motori di ricerca Solr [12]. Tale impegno si concretizza nel frequente rilascio di versioni sempre più stabili e di funzionalità sempre più avanzate. Inoltre, se l'uso di componenti open-source va incontro al requisito non funzionale di economicità dichiarato nella Sezione 1, si evidenzia come la scelta di Nutch permetta di soddisfare anche il requisito di scalabità grazie al suo supporto nativo verso la piattaforma Hadoop [15].

La Figura 1 riporta uno schema architetturale a partire dal quale è possibile descrivere sia le modalità di creazione del dizionario (linee piene etichettate con numeri), sia quelle relative al suo utilizzo (linee tratteggiate contrassegnate da lettere).

Per quanto riguarda la creazione del dizionario, Nutch viene utilizzato per eseguire la strategia di crawling (1) descritta nella Sezione 2. La collezione così prodotta viene indicizzata da Solr (2). Successivamente l'indice viene ripulito (3) per mezzo delle funzionalità di rimozione dei duplicati offerte dalle librerie Lucene [5] . A partire dal contenuto dell'indice viene generato il benchmark (4), secondo le modalità descritte nella Sezione 3, ed esportata una collezione in formato TREC (5) grazie alla quale risulta

[4] http://www.apache.org/

semplice generare un indice Terrier i cui documenti condividono un identificativo comune con i documenti presenti nell'indice della piattaforma Solr. A partire dall'indice Terrier viene infine generato il dizionario (6).

L'algoritmo di re-rank entra in gioco durante la fase di recupero. Più precisamente a fronte di una query eseguita dall'utente (a), il sistema Solr esegue un primo recupero sul proprio indice e inoltra il risultato a Terrier (b). Quest'ultimo esegue l'algoritmo di re-rank (c) e restituisce i risultati a Solr (d) che si incarica di farli visualizzare all'utente (e).

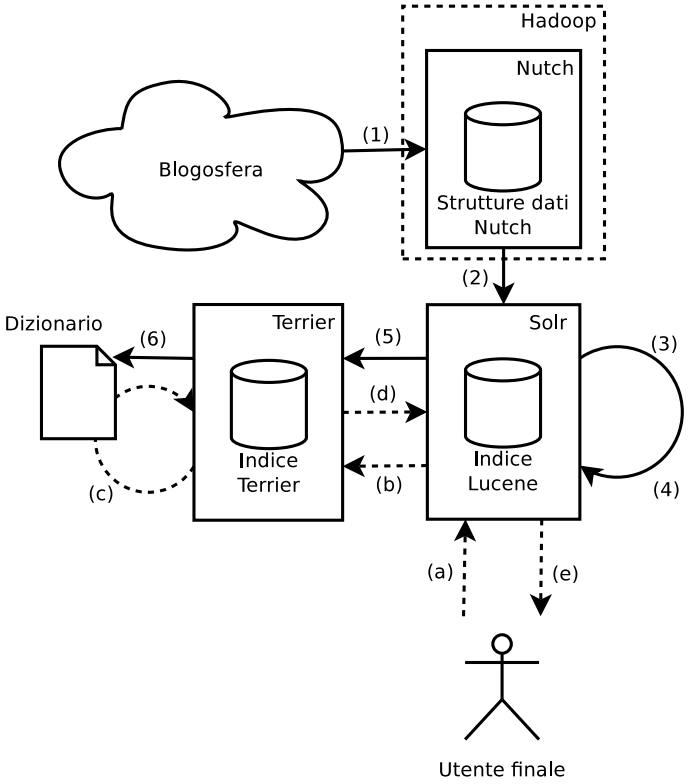

Figura 1. Schema architetturale del prototipo realizzato. Le frecce piene, etichettate con numeri, delineano il processo di creazione del dizionario. Le frecce tratteggiate, etichettare con lettere, mostrano il processo di interrogazione del sistema.

5 Conclusioni e sviluppi futuri

In questo lavoro si riporta l'esperienza maturata nell'applicazione di tecniche di opinion finding al dominio dei blog dedicati ai programmi televisivi trasmessi dalle emittenti italiane. Le attività hanno condotto alla realizzazione di un prototipo, basato su

componenti open-source, in grado non solo di fornire una risposta al problema in questione, ma anche di supportare la creazione di benchmark per l'Information retrieval e la creazione automatica di dizionari italiani composti da termini "opinion-bearing". In tal senso l'intera piattaforma può essere riutilizzata in altri domini applicativi, favorendo sia la realizzazione di nuovi benchmark che la creazione di nuovi dizionari specializzati per i singoli domini.

Riferimenti bibliografici

1. Giambattista Amati, Edgardo Ambrosi, Marco Bianchi, Carlo Gaibisso, and Giorgio Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 89–100. Springer, 2008.
2. Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20:357–389, October 2002.
3. Mike Cafarella and Doug Cutting. Building nutch: Open source search. *Queue*, 2:54–61, April 2004.
4. Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM.
5. Lucene. The Lucene search engine, 2005.
6. Craig Macdonald and Iadh Ounis. The TREC Blog06 collection: Creating and analysing a blog test collection. Technical report, Department of Computing Science, University of Glasgow, Scotland, United Kingdom, 2006.
7. Craig Macdonald, Rodrygo L. T. Santos, Iadh Ounis, and Ian Soboroff. Blog track research at TREC. *SIGIR Forum*, 44(1):57–74, 2010.
8. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.
9. Iadh Ounis, Craig Macdonald, Maarten de Rijke, Gilad Mishne, and Ian Soboroff. Overview of the trec 2006 blog track. In Voorhees and Buckland [14].
10. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.
11. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.
12. David Smiley and Eric Pugh. *Solr 1.4 Enterprise Search Server*. Packt Publishing, 2009.
13. Ellen M. Voorhees. Overview of the trec 2006. In Voorhees and Buckland [14].
14. Ellen M. Voorhees and Lori P. Buckland, editors. *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, November 14-17, 2006*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.
15. Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, original edition, June 2009.