.

BAYES-SWARM: UNO STRUMENTO OPEN SOURCE DI ANALISI DEI CONTENUTI WEB

Govoni, Riccardo, Associazione BayesFor, Via Stenone 3, 50139 Firenze, riccardo.govoni@bayesfor.eu

Zandi, Matteo, Associazione BayesFor, Via Stenone 3, 50139 Firenze, matteo.zandi@bayesfor

Bonazzi, Alessandro, Associazione BayesFor, Via Stenone 3, 50139 Firenze, alessandro.bonazzi@bayesfor.eu

Brunori, Paolo, Università di Bari e Associazione BayesFor, Via C. Rosalba 53, Bari, paolo.brunori@bayesfor.eu

Sommario

Il tradizionale mondo dell'informazione diviso tra giornali, televisione e radio trova oggi nel web un nuovo medium capace di raggiungere un sempre crescente numero di utenti a costi contenuti. La centralità dell'informazione online che si sta affermando, congiuntamente alla facilità con la quale il contenuto di pagine web può essere organizzato e conservato hanno incentivato lo sviluppo di sistemi automatici di analisi dei contenuti dell'informazione. Le finalità di queste analisi spaziano dall'analisi sociale e politica al marketing. Bayes-Swarm è uno strumento di analisi dei contenuti web con due caratteristiche peculiari: l'output è organizzato per una lettura cronologica; i contenuti delle pagine sono estratti e presentati attraverso una serie di strumenti che si basano su principi di analisi delle reti. Bayes-Swarm si basa su due passaggi fondamentali:

1. *Stoccaggio: i contenuti testuali di 100 siti web, provenienti dalle più autorevoli fonti di informazione italiane e straniere, sono salvati in un archivio digitale.*
2. *Indicizzazione: ogni parola contenuta in ciascun documento viene organizzata in un database strutturato che ne consente la fruizione in modo automatico. La posizione relativa nel testo di ogni parola rispetto alle altre viene utilizzata per determinare una misura di distanza, necessaria per la costruzione di un network nel quale le parole divengono nodi, e i loro legami sono calcolati come funzione pesata della frequenza congiunta e dell'inverso della distanza a cui mediamente si trovano.*

L'utente di Bayes-Swarm è in grado: di studiare l'andamento della frequenza di apparizione di una parola sui media, di risalire a serie storiche disaggregate per fonte e di recuperare i documenti che hanno trattato di un certo argomento nel tempo, e di interpretare i contenuti dei testi attraverso concetti di analisi dei network.

Alcune delle analisi più intuitive, come la visualizzazione delle serie storiche delle frequenze di una parola oppure della concentrazione di termini in periodi temporali, sono disponibili apertamente sul sito www.bayes-swarm.com.

Bayes-Swarm è sviluppato prevalentemente con i linguaggi di programmazione Ruby e Python e utilizza esclusivamente software open source. Sul sito di sviluppo è possibile trovare il codice sorgente, rilasciato secondo la licenza GPL e seguire l'evoluzione del progetto.

L'articolo presenta la struttura e il funzionamento del software ed è completato dall'esposizione di alcuni risultati ottenuti dall'analisi delle pagine web in concomitanza di alcuni dei più importanti avvenimenti politici dell'ultimo anno.

Parole Chiave: open source, spidering, ruby, phyton, mass media, network analysis.

.

1 INTRODUZIONE

1.1 Bayes-Swarm e l'analisi quantitativa di contenuti Web

Bayes-Swarm è uno strumento open source finalizzato all'indagine analitica e statistica di contenuti web. Nasce dalla volontà di fornire una visione quantitativa e imparziale del sempre crescente volume di informazioni oggigiorno disponibili sulla rete. Pur essendo strutturato per un approccio generico al problema, al momento si focalizza nel campo del “media monitoring” e dell'analisi del mondo dell'informazione online, un settore particolarmente dominato dall'analisi qualitativa ove un approccio più sistematico può fornire soluzioni e risultati innovativi.

E' un progetto dell'associazione no-profit BayesFor, formata da un gruppo di ricercatori e appassionati con lo scopo di promuovere la comprensione e la diffusione dell'analisi statistica, con particolare attenzione alle nuove tecnologie ed ai nuovi media.

Da un punto di vista tecnico Bayes-Swarm può essere considerato sotto diversi aspetti: è un software di spidering che estrae ed archivia i contenuti di pagine web ed altre fonti dati (quali, ad esempio, feed RSS). E' un motore di analisi che identifica ed estrae informazioni strutturate dal corpus di documenti estratti, basato sulle teorie dell'Information Retrieval e della Network Analysis. E' un insieme di strumenti per l'utente finale (sia web-based che client) per l'interazione con il set di dati raccolti e l'analisi dei risultati.

L'associazione BayesFor crede profondamente nell'importanza del codice libero. Bayes-Swarm si fonda completamente su software opensource ed è rilasciato sotto licenza GPL per permettere libero accesso ad altri sviluppatori. E' una scelta basata principalmente su due motivazioni: in qualità di utilizzatori di software libero, l'adozione di soluzioni open ci permette l'accesso ad un set di risorse, software, know-how e strumenti di supporto che sarebbe altrimenti impossibile in ottica closed-source, a meno di un significativo investimento sia in termini economici sia in termini di acquisizione del necessario know-how. Bayes-Swarm non sarebbe semplicemente stato possibile se la comunità opensource non avesse reso disponibili l'insieme di strumenti e librerie sulle quali si fonda.

D'altro canto, in qualità di fornitori di servizi, rilasciare il nostro codice in formato libero garantisce la trasparenza dei nostri risultati: chiunque é libero di verificare, riprodurre, commentare e migliorare le tecniche che adottiamo ed i risultati che presentiamo. Questo aspetto ha un'importanza rilevante specialmente nel campo del “media monitoring” in cui Bayes-Swarm opera, ove l'imparzialità dell'informazione è un elemento cruciale e purtroppo non sempre garantito e dimostrabile.

1.2 La struttura di Bayes-Swarm

Bayes-Swarm divide il processo di data mining e analisi in 3 fasi separate: estrazione, analisi e presentazione. Durante la prima fase un processo di spidering, eseguito a cadenza periodica, estrae pagine web e feed RSS dalle fonti oggetto di analisi e salva i contenuti grezzi su filesystem locale. La fase successiva processa i dati archiviati al fine di estrarre informazioni strutturate e preparare i dati per consultazione e indagine statistica. Diversi motori di analisi possono essere utilizzati e facilmente aggiunti al set attualmente definito. La fase di presentazione si occupa dell'interfacciamento con l'utente finale e della predisposizione degli strumenti necessari alla fruizione dei risultati.

La chiara separazione tra le fasi di elaborazione permette una migliore organizzazione del codice e porta ad una naturale divisione in componenti che dialogano tra loro secondo interfacce definite, ed una migliore schedulazione dei processi, gestione degli imprevisti e rielaborazione dei dati a fronte di cambiamenti e bugfixes.

La figura 1 mostra la struttura logica secondo cui Bayes-Swarm è organizzato.

.

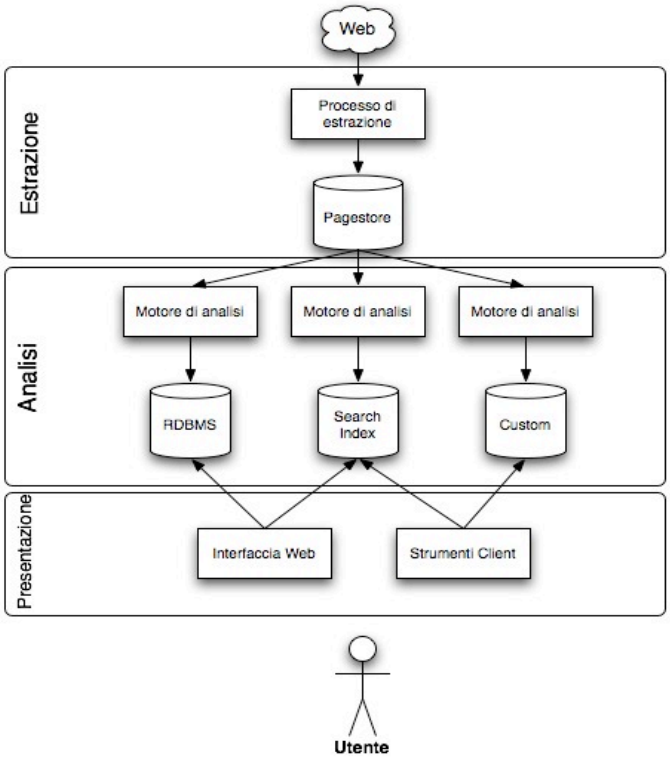

Figura 1. Struttura logica delle fasi di elaborazione e dei componenti di Bayes-Swarm.

Bayes-Swarm utilizza principalmente i linguaggi Ruby, Python (ed il linguaggio per analisi statistica R in misura minore) per le diverse componenti, e si appoggia a diversi strumenti open, tra cui rientrano:

•Xapian, come motore di indicizzazione e ricerca all'interno dei contenuti estratti.

•MySql per l'archiviazione strutturata dei dati estratti al fine della loro presentazione tramite interfaccia web.

•Rails, integrato da diverse librerie di visualizzazione grafica (e.g. OpenFlashChart) per la fase di presentazione web; GTK+ e PyGTK per gli strumenti client di presentazione e analisi.

•Varie librerie per web mining (Hpricot, Mechanize), text mining (Ferret, Lucene) e network analysis (igraph et al.).

Sulla base di tali tecnologie, Bayes-Swarm offre due strumenti principali per l'analisi dei dati:

•**Pulsar**: gestisce l'attività di spidering e Information Retrieval di primo livello: archivia pagine web, effettua operazioni di pulizia e normalizzazione del testo e archivia i risultati in un database MySql ottimizzato per lo studio di serie cronologiche e la presentazione dei contenuti tramite interfaccia web.

•**MeanMachine**: libreria di Network Analysis per lo studio degli argomenti trattati e delle relazioni tra essi in base a grafi di relazioni tra i termini lessicali che appaiono nel testo, accessibile tramite client GTK+.

A questi si aggiungono altri componenti custom, sviluppati a seconda delle necessità, quali ad esempio analisi statistiche basate su R.

Quanto descritto offre una discreta scalabilità. A titolo di esempio, alcuni dati relativi processati dall'installazione di Bayes-Swarm trattata in questo paper:

•ca. 200 fonti (pagine web, feed RSS) elaborati giornalmente, equamente suddivisi tra 100 tra i più' noti portali di informazione italiani e stranieri.

.

- 40.000 termini lessicali tracciati giornalmente.
- 20M di termini identificati negli ultimi 5 mesi.
- ca. 50Mb di dati giornalieri, equivalenti a 60 libri di 300 pagine l'uno
- ca. 25Gb di dati accumulati da fine 2007 ad oggi, equivalenti ad una biblioteca di 40.000 volumi.

.

2 LE COMPONENTI DI BAYES-SWARM

2.1 Pulsar

Pulsar è una libreria che si occupa del processo di spidering, archiviazione dei documenti estratti e normalizzazione dei testi. Per ogni pagina web o contenuto oggetto di analisi, Pulsar effettua le seguenti operazioni:

•Spidering: estrazione a cadenza periodica della risorsa (pagina web o altro) oggetto di analisi.

•Archival: archiviazione storica del contenuto grezzo e indicizzazione su filesystem per data e fonte.

•Cleaning: Rimozione della formattazione indesiderata (nel caso di documenti HTML). Pulizia del codice HTML e della punteggiatura.

•Filtering: Separazione delle stop-word per ogni lingua analizzata.

•Stemming: Identificazione delle radici lessicali per ogni termine estratto e raggruppamento.

•Weighting: Pesatura dei termini in base a posizione e prominenza all'interno delle pagine web.

•Storage: Salvataggio su RDBMS con struttura ottimizzata per analisi cronologiche e lo studio di serie storiche.

Le fasi descritte sono fondamentali per l'analisi dei contenuti a causa dell'elevato rapporto rumore/contenuti tipico di ogni contenuto testuale e soprattutto presente oggigiorno sul web. A titolo di esempio, la Figura 2 mostra la distribuzione di frequenza dei primi 13000 termini lessicali analizzati per la giornata del 1-Maggio-2009, in perfetta corrispondenza con la distribuzione di Zipf che descrive l'andamento classico della frequenza di termini lessicali all'interno di un corpus testuale.

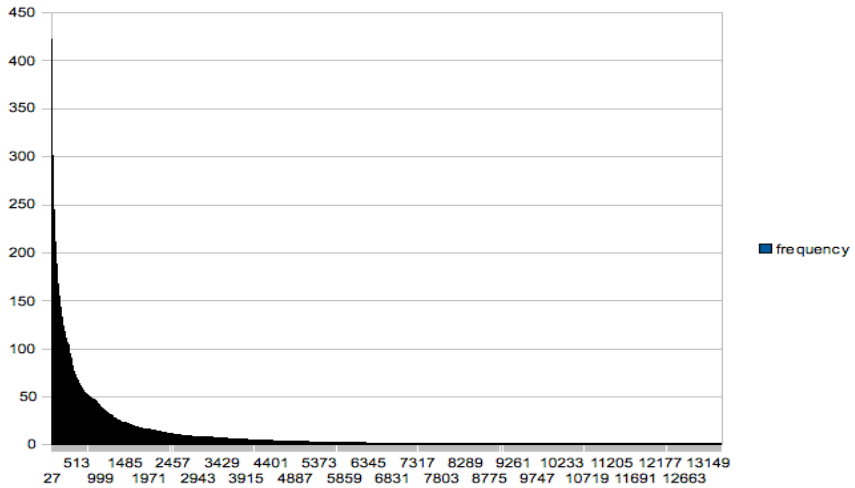

Figura 2. Distribuzione di Zipf della frequenza dei termini lessicali analizzati da Bayes-Swarm per la giornata dell 1-Maggio-2009.

.

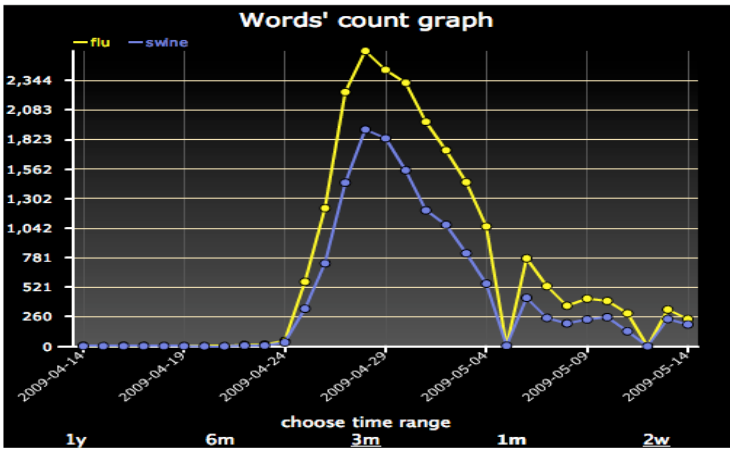

Figura 3. Serie storica del caso 'Influenza suina' come osservato tramite l'interfaccia web di Bayes-Swarm.

L'output finale prodotto dalla libreria Pulsar, acceduto tramite interfaccia web, permette di effettuare analisi cronologiche sui dati, serie storiche e aggregazioni di vario livello, inclusa la consultazione storica dei documenti oggetto di analisi (funzione TimeMachine). La figura 3 mostra un esempio per un caso di recente interesse mediatico.

2.2 Mean machine

La Mean Machine costruisce reti di parole elaborando i documenti ottenuti nel processo di spidering. I nodi delle reti sono costituiti dalle parole di interesse mentre i legami sono funzione della distanza alla quale in media le parole si trovano nei tesi analizzati.

Le pagine salvate da Pulsar passano attraverso un processo di pulizia che consente di eliminare il codice html, restituendo esclusivamente il contenuto testuale delle pagine web. Successivamente il testo è indicizzato attraverso la libreria Xapian, la quale si occupa di rimuovere la punteggiatura e di associare ad ogni parola un numero crescente. Questo consente di poter ricercare i documenti che contengono frasi esatte costituite da più parole oppure che presentano due parole a meno di una certa distanza. Ad Esempio, consideriamo un database che contiene un solo documento composto esclusivamente dalle frasi:

"*Obama is explaining his dream about creating a new country. Obama has an unfeasible dream, however people tend to trust him.*"

Mean Machine associa le seguenti posizioni nel testo:

Obama	is	explaining	his	dream	about	creating	a	new	country	
1	2	3	4	5	6	7	8	9	10	
Obama	has	an	unfeasible	dream	however	people	tend	to	trust	him
11	12	13	14	15	16	17	18	19	20	21

Questo documento sarà restituito quindi dalle query:

- "Obama" (presenza della parola)
- "Obama is explaining" (frase esatta)
- "Obama NEAR country" (presenza di entrambe le parole con distanza inferiore a 11)

.

Le reti di parole. Affinché la Mean Machine possa creare una rete di parole, è necessario specificare una query di ricerca che limiti l'ambito di documenti da analizzare. L'obiettivo della query è quello di filtrare i documenti disponibili e considerare solo quelli che l'utente ritiene siano interessanti.
La determinazione dei documenti che rientrano nella rete è affidata a Xapian, secondo i passaggi:

1. Costituzione del *Matching Set*: il set di documenti completo viene filtrato sulla base di parametri di ricerca inseriti dall'utente. Fanno parte di questi parametri: la lingua, il range di date, il set di fonti e soprattutto le parole chiave. L'utente può specificare ad esempio "Obama AND health", in questo caso entreranno a far parte del Matching Set, solo i documenti che contengono sia la parola Obama che health. I documenti che verificano la condizione sono ordinati per importanza decrescente e vengono considerati solo i primi K documenti.
2. Determinazione dell' *Expansion Set*: nell'ambito dei documenti che appartengono Matching Set, vengono individuate le parole che hanno peso maggiore (*vertex size*).
3. Calcolo della Distanza (*edge weight*): per ciascuna coppia di parole facenti parte del Expansion Set, si ottiene la loro posizione nell'ambito di tutti i documenti che appartengono al Matching Set, al fine di ottenere una misura di distanza.

Attraverso una interfaccia grafica è possibile selezionare:
- periodo temporale di interesse (T);
- sottoinsieme di fonti da includere (S);
- massimo numero di documenti da selezionare (K);
- massimo numero di parole da includere (N), in alternativa le parole da utilizzare possono essere quelle provenienti da una lista prescelta;
- vertex size: soglia di rilevanza al di sotto della quale le parole non vengono incluse nella rete;
- edge weight: soglia di forza al di sotto della quale le parole non vengono incluse nella rete.

Vertex size. Il network creato dalla Mean Machine potrà essere infittito o diradato, includendo o escludendo parole, sulla base della loro rilevanza (che nella rete è un attributo del vertice chiamato size). Il concetto di rilevanza (alla base della costruzione delle reti si basa sul numero di documenti nei quali si trova la parola dei K che costituiscono il Matching Set e sulla frequenza del termine nel database. La formula si basa su alcuni lavori di Robertson e altri (Robertson et al., 1992).

La parola più rilevante, non necessariamente l'oggetto della ricerca, ha la rilevanza massima (r_{MAX}), tutte le altre parole hanno una rilevanza r_i (con $r_{MAX} > r_i$) fino alla parola che ha rilevanza minima (r_{MIN}). Nella visualizzazione grafica, la grandezza del nodo (vertex size) è ottenuto attraverso una normalizzazione della rilevanza:

$$r_i = \frac{r_{MAX} - r_i}{r_{MAX} - r_{MIN}} \quad (1)$$

Dalla rete possono essere escluse parole alzando il livello minimo di vertex size (che varia da 0 a 1) man mano che si innalza il livello minimo le parole meno rilevanti dai testi vengono escluse. Abbassare il valore vertex size implica includere parole progressivamente meno frequenti nei documenti. Se vertex size = 0 allora le parole incluse nella rete sono N.

Edges weight. L'edges weight è un indice di prossimità, i legami fra nodi in questo database sono la distanza in parole fra i termini in un testo. Questo fa si che virtualmente tutte le parole abbiano legami fra di loro se sono presenti almeno una volta entrambe nello stesso documento, tutto dipende dalla soglia di distanza che definiamo come minima per l'esistenza di un legame. Il peso del legame è ottenuto dallinverso della distanza media fra due parole.

Il legame tra nodo *i* e *j* è calcolato come:

.

$$eweight_{i,j} = \sum_{n \in N} \frac{\sum_{c \in C} \frac{1}{x_i - x_j}}{f_{MAX}} \quad (2)$$

C è ottenuto come segue: partendo da tutte le possibili coppie tra i e j con i < j , consideriamo solo le coppie che hanno distanza minima. In questo modo il legame fra due parole è compreso fra 0 e 1, è pari a 1 se i e j sono le due parole più frequenti e sono sempre a distanza 1'una dall'altra. Riprendiamo l'esempio precedente:

Obama	is	explaining	his	dream	about	creating	a	new	country	
1	2	3	4	5	6	7	8	9	10	
Obama	has	an	unfeasible	dream	however	people	tend	to	trust	him
11	12	13	14	15	16	17	18	19	20	21

Supponiamo che l'utente ricerchi la parola "obama" e che quindi l'unico documento nel database rientri nel Matching Set sia questo. Supponiamo inoltre che le parole "obama" e "dream" facciano parte del Expansion Set. L'insieme di tutte le possibili coppie delle due parole determina il seguente insieme di coppie di posizioni (combinazioni senza ripetizione): {(1,5), (1,15), (5, 11), (11, 15)} che corrisponde alle seguenti distanze: {4, 14, 6, 4} . L'insieme C delle distanza considerate sarà: {4, 4} con numerosità pari a 2 (frequenza della parola obama). Per cui la distanza tra le parole obama e dream equivale a (1/4+1/4)/2 = 0,25[1].

[1]Allo stesso modo sarà possibile calcolare il legame fra "obama" e "people", e fra "dream" e "people", si noti che la parola con frequenza massima ha frequenza pari a 2 e che il numero di documenti è pari a 1.

$$ew_{OBAMA,PEOPLE} = \frac{(\frac{1}{6} + \frac{1}{16})}{2} = 0{,}114583$$

$$ew_{DREAM,PEOPLE} = \frac{(\frac{1}{12} + \frac{1}{2})}{2} = 0{,}291667$$

.

3 I RISULTATI

3.1 Bayes-Swarm all'opera.

Il progetto di sviluppo del software Bayes-Swarm è stato accompagnato da un'attività di analisi dei dati raccolti che è tesa ha dimostrare la validità scientifica del software. L'analisi dei contenuti web ha un elevato numero di possibili applicazioni. I motori di ricerca tradizionali rappresentano un esempio emblematico ma negli ultimi anni un grande numero di software si sono specializzati nell'analisi dei contenuti online finalizzata ad interessi specifici. Sin dall'Ottobre del 2007 la nostra attività di spidering si è concentrata su alcune fonti di informazione online. Alcune decine di pagine web dei maggiori canali di informazione sono stati monitorati giornalmente e le occorrenze delle parole sono state organizzate in un database mysql. Una parte della validazione del software ha perciò riguardato l'analisi dell'informazione in rete, così come presentata dalle principali fonti ufficiali di informazione (giornali, agenzie di stampa, aggregatori di notizie, etc.)[2]. L'analisi delle fonti di informazione è storicamente pertinenza di esperti in marketing e scienziati della comunicazione, ci è apparso opportuno quindi confrontare i nostri risultati con la letteratura esistente a riguardo. In particolare, la letteratura ha mostrato un grande interesse per il comportamento dei media, ed i suoi effetti sui fruitori dei loro servizi, in concomitanza con eventi politici di particolare rilievo, come le scadenze elettorali (si vedano ad esempio Kern, (2001), Mullainathan s. e Shleifer A (2002) o Puglisi, 2004).

Per analizzare il comportamento dei media italiani durate le elezioni primarie per la scelta del segretario del Partito Democratico dell'autunno 2007 sono state monitorate 31 fonti di informazione in lingua italiana. Come prima approssimazione si è voluto verificare se la presenza dei candidati, misurata come numero totale delle apparizioni del loro nome, fosse in qualche modo legata al numero di preferenze raccolte. Questa relazione è discussa in letteratura ed è stata misurata da altri autori in passato anche nel nostro paese ((Pira et al., 2001). La percentuale relativa della presenza di ciascun candidato sui media è ottenuta sommando tutti le apparizioni dei principali candidati nel mese e dividendo per il totale. Questa misura rappresenta, con una certa approssimazione, che percentuale di visibilità ciascun candidato ha avuto sul totale dello spazio dedicato alle primarie del PD. La figura sotto riportata rappresenta la percentuale delle apparizioni medie per ciascun candidato (a destra) e la percentuale dei voti ottenuti (a sinistra), come si nota queste percentuali sono molto simili[3].

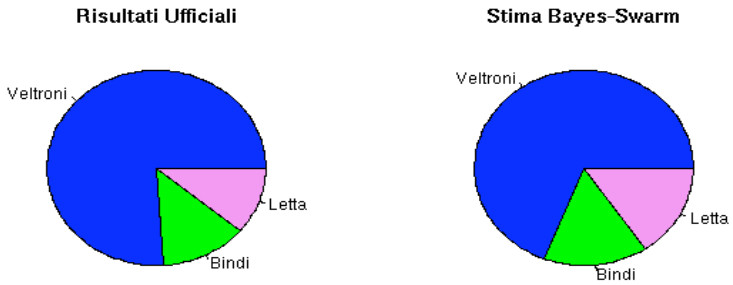

Figura 4. *Le percentuali stimate da Bayes-Swarm, considerando le nostre osservazioni come estrazioni casuali da una popolazione sconosciuta, sono comprese, con un livello di confidenza del 95% fra i valori Veltroni: 64,03% – 73,86%, Letta: 11.16% - 19.96%, Bindi 12.06% - 18.90%; i risultati hanno visto un'affermazione di Veltroni con 75,81% seguito da Bindi con il 12,88 % e Letta 11,07%.*

[2]L'elenco completo delle pagine analizzate è riportato in appendice.
[3] L'analisi delle primarie del PD dell'Ottobre 2007 sono state pubblicate nell'articolo di Brunori et al (2008).

.

L'interessante regolarità che sembra legare visibilità online e successo elettorale è un dilemma stimolante: sono i media che esercitano un'influenza sugli elettori? O sono i clienti dei mezzi di informazione che con le loro "preferenze" inducono il c. d. "media bias"? Una parte della letteratura protende per la prima idea (si veda ad esempio McCombs (2002)) sottolineando il ruolo di "agenda setting" che i media sarebbero in grado di esercitare. Un'altra parte della letteratura considera invece che a prevalere sia un effetto di segno opposto, i media sono "di parte" in quanto devono vendere ad un pubblico che preferisce acquistare giornali in linea con il proprio pensiero (fra gli altri si veda Gentzkow e Shapiro (2005)). Come sottolineato recentemente da alcuni autori entrambi gli effetti possono convivere e rinforzarsi a vicenda (Slater, 2007), in ogni caso questo tipo di regolarità ci ha spinto a ripetere l'analisi nell'imminenza della scadenza elettorale del 2008.

Le elezioni rappresentano un fenomeno di più difficile monitoraggio. Il solo nome dei candidati rappresenta difficilmente una buona approssimazione della visibilità di un partito o di una coalizione, che spesso sono l'oggetto della scelta. Per questo motivo si è scelto di descrivere la visibilità di una coalizione sommando le apparizioni del nome del candidato premier e delle sigle dei partiti. La visibilità della coalizione che appoggiava Veltroni è stata ricavata come somma delle occorrenze delle parole: "veltroni", "idv" e "pd". Anche in questo caso abbiamo deciso di allungare il periodo di monitoraggio da 4 a 6 settimane. L'analisi oltre a confrontare visibilità e voti ha preso in considerazione anche una serie di sondaggi i cui risultati erano disponibili in rete.[4]

Di seguito riportiamo la tabella riassuntiva dei risultati ottenuti, si noti come in la capacità predittiva dei sondaggi sia molto simile a quella di una semplice osservazione della visibilità online.

Coalizione	**voti**	**Visibilità on line**	**Sondaggi**
Berlusconi	47.07%	45.23%	44.43%
Veltroni	37.78%	39.54%	36.47%
Casini	5.59%	9.92%	6.58%
Bertinotti	3.13%	2.45%	7.3%

Tabella 5. Voti, visibilità e sondaggio nelle elezioni 2008, fonte Bonazzi et al. (2008).

Come si è detto questo legame, così regolare, fra visibilità e preferenze elettorali rappresenta uno spunto interessate e suggerisce l'esistenza di un legame estremamente forte fra l'opinione pubblica e i contenuti dei media, anche nel caso dei media online. Per questo motivo è stato predisposto un database contenente 38 fonti di informazione statunitensi che sono stati monitorati durante l'ultima compagna elettorale per le elezioni presidenziali del 2008. La figura 6 mostra l'andamento delle visibilità di Obama e McCain e dei sondaggi pubblicati nello stesso mese. Si tratta del periodo nel quale hanno avuto luogo le due *convention* nelle quali Obama e McCain sono stati ufficialmente nominati candidati dei rispettivi partiti per le elezioni presidenziali. È interessante notare come a seguito di due fasi di alta visibilità sia del candidato democratico a fine Agosto, sia di quello repubblicano, a inizio Settembre, si registri un incremento delle preferenze nei sondaggi, si noti che a metà Settembre, i due momenti di maggior visibilità di McCain precedono le uniche due settimane nelle quali le sorti delle elezioni sono sembrate incerte, con il candidato repubblicano in recupero di consensi.

[4] L'analisi delle elezioni 2008 sono state presentate alla Confereza E-Democracy 2008 che si è svolta all'Università di Krems in Austria (si veda Bonazzi, et al. (2008)).

.

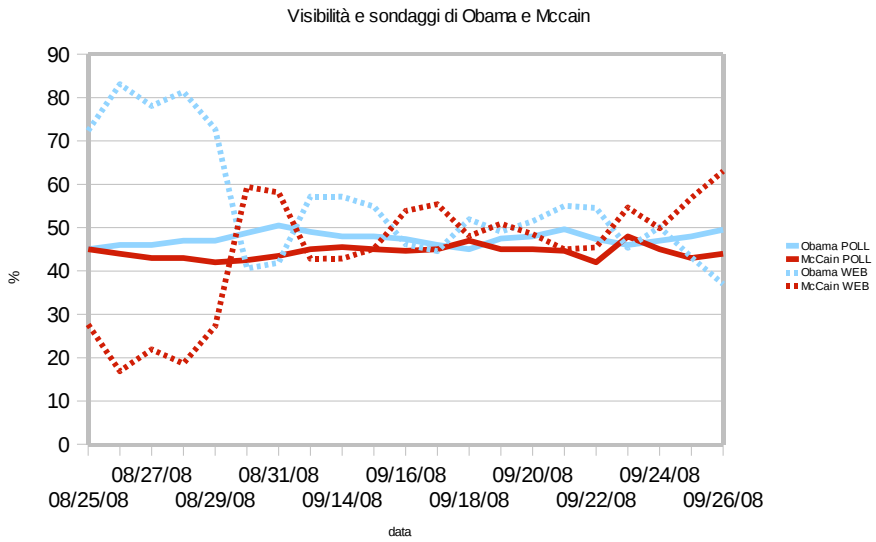

Figura 6. L'andamento di sondaggi (linea continua) e visibilità online (linea tratteggiata) nelle 4 settimane dall'inizio delle convention[5].

3.2 Mean Machine all'opera.

Le elezioni presidenziali statunitensi dello scorso anno hanno costituito anche il banco di prova per la Mean Machine. Il pagestore creato da pulsar per il periodo cruciale della campagna elettorale rappresentava un'ottima base sulla quale testare l'algoritmo alla base della costruzione delle reti di parole. La tabella 7 mostra la forza dei legami fra la parola "Obama", due parole chiave della campagna elettorale del candidato democratico: "change", "dream" e la parola "president". Assumiamo che un media interessato a sostenere un candidato dovrebbe tendere ad associare il suo nome ai termini chiave della sua campagna, e in ogni caso al termine "presidente". Questa assunzione è verificata dai dati della tabella 7.

Parola chiave: obama	**CNN**	**Fox**	**Google news**	**Yahoo news**
change	0,040942	0,004164	0,050787	0,117352
dream	0,001288	0,000765	0.097836	0,018918
president	0,108492	0,003915	0,079294	0,15223

Tabella 7. *I legami fra "obama" e tre parole chiave.*

Il sito di informazione dell'emittente Fox, tradizionalmente legata ai conservatori del partito repubblicano mostra legami assai blandi (circa un ordine di grandezza) rispetto a quanto non faccia il sito della CNN. Allo stesso modo

[5] L'analisi della visibilità on line dei candidati alle elezioni presidenziali negli Stati Uniti nel 2008 è in parte contenuta nel lavoro di presentato a San Diego alla conferencza annuale di SUNBELT (Zandi et al., 2009), una parte dell'analisi è stata svolta da Martina De Siervo nella sua tesi di Laurea "I temi di politica estera e le strategie comunicative nelle campagne presidenziali di Barack Obama e John McCain" dell'Università di Firenze.

.

La tabella 8 si riferisce ai quotidiani on line nel mese delle due convention (dal 25 Agosto). In questo caso si mettono a confronto le distanze fra la parola “president” e il nome dei due candidati e fra qusti e le loro due parole chiave “experience” e “change”. Si noti che malgrado il legame fra “change” e “Obama” sia sempre più forte rispetto a “McCain” e “experience”, un solo giornale inverte questa relazione, il Washington Times, che è stato fra i pochissimi giornali statunitensi a schierarsi ufficialmente al fianco del candidato repubblicano..

	Obama - change	**McCain - experience**
Boston Globe	0,07896'	0,008789
New York Times	0,116736	0,050459
Atlanta J. Const.	0,295286	0,002572
Washington Times	0,008199	0,070864
WSJ	0,149151	0,061636

Tabella 8. *I legami fra “obama” e tre parole chiave.*

Presentiamo infine una due network ottenuti con la MeanMachine. Riguardano ancora la campagna elettorale statunitense nel mese delle convention. Le due reti sono ottenute con la ricerca “Obama AND McCain” includendo solo i termini che oltre ad essere presenti nelle pagine on line sono stati anche utilizzati dai candidati nei loro discorsi alle convention. A colpo d'occhio si nota l'orientamento politico differente dei due giornali. In particolare la parola chiave “experience” di McCain, ben lontana da Obama e strettamente legata a Mccain sul Washington Times, si trova in posizione quasi simmetrica su USA Today. La simmetria nella rete di USA Today è chiara anche per “state”, “care”, “promise”, “save”, “going” e “Bush”. Sempre sulla sinistra della rete di USA Today, si legge, in senso antiorario un premonitore: “Obama”- “can” - “win”!

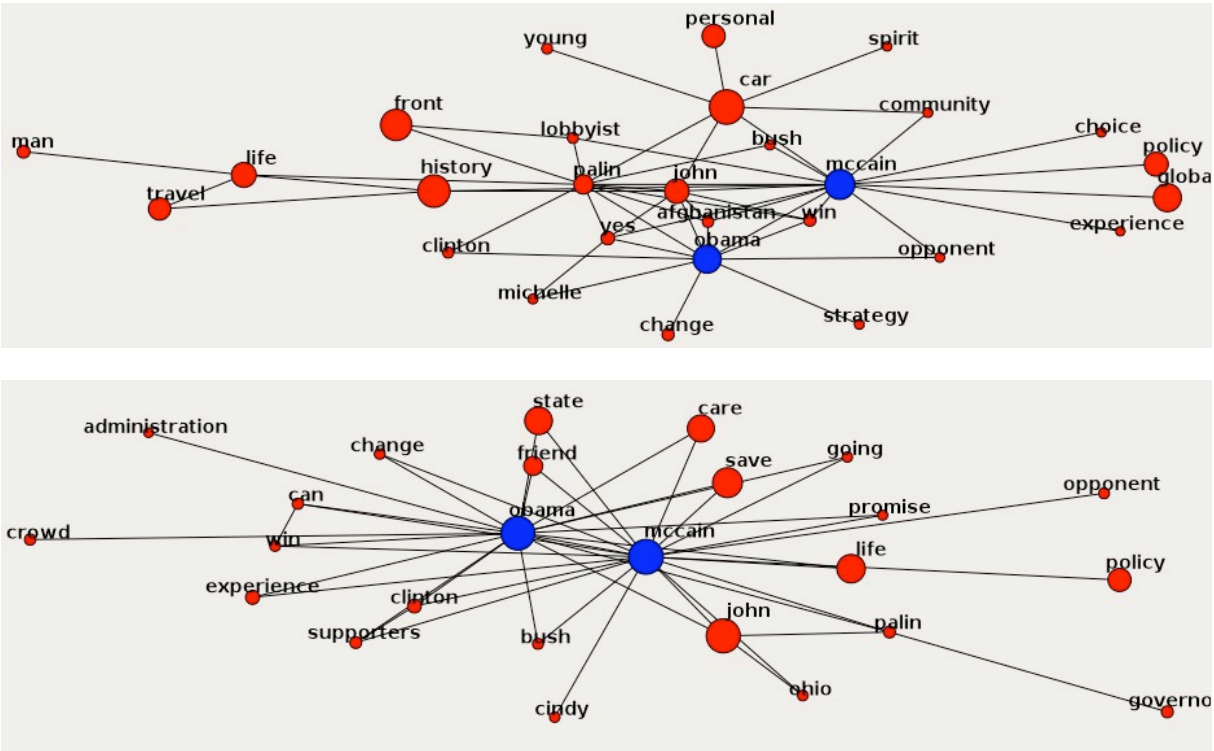

Figura 9. Due network per la ricerca “Obama AND McCain”, in alto il Washington Times, in basso USA Today, giornale tradizionalmente neutrale.

.

Bibliografia

Bonazzi A., Brunori P., Govoni R., Lampronti G.I. e Zandi M. (2008). Italy 2008 Polls, Web Visibility and Election Results, EDem2008 E-Democracy Conference proceedings, Danube University, Krems. Austria.

Brunori P., Zandi M., Bonazzi A., Govoni R. and Lampronti G. (2008). Visibilità mediatica dei candidati al le primarie del Partito Democratico. Analisi con i dati di Bayes- Swarm. Il Politico, 217 (1).

Gentzkow M. e Shapiro J. M. (2005) Media Bias and Reputation, NBER Working Paper No. 11664.

Kern M. (2001). Disadvantage Al Gore in Election 2000: Coverage of Issue and Candidate attributes, including the Candidate as Campaigner, on Newspaper and Television News Web Sites, American Behavioral Scientist, N. 44, 2125-2139.

McCombs M. E. (2002) The Agenda-Setting Role of the Mass Media in the Shaping of Public Opinion, Mass Media Economics 2002 Conference, LSE, London, UK.

Mullainathan s. e Shleifer A (2002). Media bias. Working Paper 9295, National Bureau of Economic Research, October 2002.

Pira F., Gaudiano L. e Manzo C. (2001). Monitoraggio dei mezzi di informazione locali durante la campagna elettorale per le elezioni amministrative in Friuli-Venezia Giulia, CORECOM.

Puglisi R. (2004) "Being the New York Times: the Political Behaviour of a Newspape", Paper presented at the annual meeting of the American Political Science Association, Hilton Chicago and the Palmer House Hilton, Chicago, www.allacademic.com/meta/ p59266_index.html.

Robertson S.E. , Walker S., Hancock-Beaulieu M., Gull A., e Lau M., (1992). Okapi at TREC, om Text Retrieval Conference, pp. 21–30.

Zandi M., Grippa F, Bazarnick T., Brunori P., Frongia D., Govoni R. e Bonazzi A., (2009). Media Behavior During 2008 Electoral Campaign: a Web Content Analysis, SUNBELT Annual Conference, San Diego USA.