POLITECNICO DI MILANO
Corso di Laurea Specialistica in Ingegneria Informatica
Dipartimento di Elettronica e Informazione

UNA METODOLOGIA PER LA VALUTAZIONE DEL SENTIMENT WEB BASATA SUL CONCETTO DI REPUTAZIONE

Relatore: Prof.ssa Chiara FRANCALANCI

Correlatore: Ing. Donato BARBAGALLO

Tesi di Laurea Specialistica di:

Luca Carlo CARMINATI

(Matricola 725193)

Anno accademico 2009/2010

Sommario

Il lavoro di tesi si propone di studiare la reputazione in ambienti Web 2.0 e si pone come obiettivo quello di dare delle metodologie di analisi che aiutino a trarre delle conclusioni, il più possibile dettagliate, sull'andamento del sentiment riguardante un brand o prodotto.

Lo studio riportato in questo elaborato vuole essere sia un contributo alla ricerca accademica del settore, attraverso gli algoritmi implementati e le idee innovative proposte per ottenere un risultato, che un aiuto alle aziende che intendono valutare l'immagine del marchio e dei propri prodotti online, arricchendo la Business Intelligence con una nuova fonte di informazione.

Ringraziamenti

Un sincero ringraziamento va alla Prof.ssa Chiara Francalanci e all'Ing. Donato Barbagallo, ottime guide per la realizzazione di questa tesi.

Grazie ai miei genitori e a mio fratello Roberto per avermi sostenuto negli studi e nella vita. Senza di loro non sarei arrivato sin qui.

Grazie a tutti i miei amici che hanno condiviso con me le mille avventure universitarie e della vita.

Grazie a Ulisse che mi ha aiutato a crescere sotto il profilo professionale e umano oltre ad essere un grande amico.

Grazie a Chiara, perché capace di sopportarmi, ascoltarmi e consigliarmi nei momenti di difficoltà, nelle gioie e nei dolori.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

Blog, forum e più in generale le pagine web, stanno diventando sempre più un luogo dove i consumatori commerciano, discutono e influenzano gli altri consumatori. L'analisi e l'estrazione delle opinioni sta attirando l'interesse di molte aziende che vedono nell'esplosione di questo fenomeno un'importante occasione di business.

Il lavoro di tesi propone uno studio empirico sull'analisi della reputation. In particolare questo elaborato ha come obiettivo quello di fornire degli indicatori per potenziare lo studio del sentiment ricavato dai siti di tipo "user generated content" come TripAdvisor e Lonely Planet. I temi principalmente trattati sono due: l'analisi dell'importanza delle sorgenti informative e la determinazione dell'importanza degli utenti che generano delle opinioni.

La tesi è strutturata nel modo seguente.

Il Capitolo 2 tratta in generale della sentiment analysis e delle teorie che stanno alla base dei tool che si pongono come obiettivo questo tipo di analisi. Vengono descritti gli strumenti di mercato, le teorie sull'analisi del linguaggio naturale, la teoria sulle reti sociali e sulla loro misurazione e infine viene trattato l'argomento della reputazione delle sorgenti informative.

Il Capitolo 3 descrive tutte le caratteristiche del tool in cui verrà posta la perte implementata utilizzando la teoria di questo lavoro di tesi.

Il Capitolo 4 presenta le metriche scelte per la valutazione dell'importanza sia degli utenti che delle fonti. In particolare si descrive ciò che è stato implementato nel cruscotto di analisi.

Il Capitolo 5 presenta i risultati di analisi con dei grafici e delle discussioni su ciò che è emerso dalle valutazioni empiriche.

Capitolo 2

Stato dell'arte

2.1 Introduzione

La continua crescita nell'utilizzo dei social media è un trend che si è consolidato nel corso degli ultimi anni. La grande quantità di utenti che ne fanno uso ha attirato l'attenzione del business delle grandi e piccole aziende. Molti artigiani e piccoli imprenditori utilizzando questo mezzo riescono a pubblicizzare la propria attività attraverso la pubblicazione di eventi, promozioni e altre attività volte ad accrescere il proprio parco clienti. Le grandi aziende, oltre alle esigenze di pubblicizzare i propri prodotti, hanno anche la necessità di tenere sotto controllo la reputazione del loro brand. Più cresce la dimensione dell'azienda più la sua reputazione deve essere misurata su un insieme di utenti sempre più grande, partendo dall'artigiano che si preoccupa dell'opinione dei clienti abituali fino a raggiungere le dimensioni delle multinazionali che devono mantenere una reputazione a livello globale. Più cresce il numero di utenti da monitorare, più si rende necessario sviluppare dei sistemi automatici che calcolino il sentiment degli utenti e forniscano degli indicatori ai manager che si occupano di marketing. In questo capitolo inizieremo con una breve descrizione dei tool attualmente sul mercato (Sezione 2.2), per poi passare ad una breve trattazione sullo studio del linguaggio naturale che sta alla base di questi strumenti e che ne abilita le analisi (Sezione 2.3). Successivamente si tratterà la teoria

delle reti sociali (Sezione 2.4), la creazioni di grafi in blog e forum (Sezione 2.5) e infine della reputazione delle fonti (Sezione 2.6), argomenti che sono alla base di questo lavoro di tesi.

2.2 Il mercato della reputation analysis

Tra gli strumenti che si dividono il mercato della reputation analysis si possono individuare due diverse tipologie di prodotti: i tool che si occupano di ricevere il feedback dagli utenti e che elaborano statistiche, e i tool che sono veri e propri CRM (Customer Relationship Management) e supportano i contatti con i clienti e aumentano l'efficacia delle campagne. Di seguito verranno presentati i principali strumenti presenti sul mercato; sia nazionali che internazionali.

2.2.1 Cogito Monitor

Cogito Monitor è un prodotto di Expert System, una delle più importanti software house a livello nazionale. Si occupa della gestione dei documenti con analisi linguistica e comprensione semantica. L'applicativo consente di analizzare in tempo reale i contenuti presenti in rete, rilevando in modo automatico le opinioni degli utenti di blog, forum e social network. Le principali funzionalità sono l'accessibilità tramite interfaccia Web in modalità SaaS (Software as a Service), il monitoraggio automatico delle fonti, sulla base di analisi semantiche esegue l'estrazione dei contenuti di potenziale interesse, analizza e classifica i contenuti, assegna automaticamente un valore di sentiment ad ogni commento e fornisce delle funzionalità di reportistica.

2.2.2 BuzzMetrics

BuzzMetrics è uno strumento presente sul mercato americano da una decina di anni e dal 2007 ha mosso i primi passi nel mercato europeo. Prodotto da Nielsen, è un applicativo che riscuote parecchio successo

nel mercato internazionale. La funzionalità principale è di indicizzare le fonti sulle quali i consumatori parlano, andando a individuare in tempo reale i commenti sul brand, il prodotto o l'argomento di interesse. I contenuti possono essere misurati quantitativamente e qualitativamente. Le piattaforme attingono da database di grandi dimensioni – oltre 103 milioni di blog e 100 mila forum indicizzati nel mondo. È uno strumento particolarmente utile per studiare la reputazione di un'azienda o prodotto, misurare l'efficacia di una campagna pubblicitaria o monitorare il lancio di un nuovo prodotto.

2.2.3 BuzzLogic

BuzzLogic è una compagnia che si occupa di media digitali. Il suo prodotto BuzzLogic Insight ottimizza le attività pubblicitarie sui più importanti blog presenti sul Web. Osservando le conversazioni e i link tra blog, crea una mappa delle conversazioni nella quale si possono rintracciare gli utenti che stanno parlando di un determinato argomento e gli utenti che esercitano una maggiore influenza sugli altri. Con l'utilizzo di questo strumento le aziende cercano di inserire la pubblicità all'interno di queste conversazioni.

2.2.4 Radian6

Radian6 fornisce alle organizzazioni una piattaforma per ascoltare, misurare e intervenire in conversazioni che si svolgono sui social media presenti sul Web. Viene usato da responsabili di pubbliche relazioni, marketing e customer service supportando i professionisti per comprendere e assistere meglio i propri clienti. Il tool viene utilizzato da più di 1400 clienti appartenenti a diverse industrie – tecnologiche, non-profit, manifatturiere, di consulenza e molti altri.

2.2.5 truCast

TruCast è un tool sviluppato da Visible Technologies che permette alle compagnie di monitorare le conversazioni che si svolgono su internet analizzando blogs, social network, forum di discussione, ecc.. per determinare cosa si sta dicendo sul web in un determinato momento. Molte aziende importanti come Dell, Microsoft, Panasonic e Hormel, usano TruCast per monitorare e rispondere alla propria clientela. Questo prodotto è stato eletto finalista alla MITX Technology Awards, evento organizzato da Massachusetts Institute of Technology (MIT), nella categoria social media.

2.2.6 SAS Teragram

Acquistata recentemente da SAS, Teragram è una software house che fa dell'analisi del linguaggio naturale il suo punto di forza. Per tale motivo è uno dei migliori tool nel calcolo della reputation in modo completamente automatico. Il prodotto è in grado di ottenere alti livelli di precisione in circa 30 lingue diverse. Una delle lacune di questo strumento riguarda le funzionalità di analisi che sono molto limitate.

2.3 Linguaggio naturale

L'elaborazione del linguaggio naturale (Natural Language Processsing (NLP)) è un campo della computer science e della linguistica che si occupa dell'interpretazione e dell'estrazione di informazioni da un testo scritto in linguaggio naturale tramite l'utilizzo di calcolatori elettronici. L'aggettivo "naturale" è utilizzato per distinguere tra il linguaggio umano e i linguaggi formali (come il linguaggio matematico o i linguaggi di programmazione). La nascita di software per il riconoscimento del linguaggio naturale è avvenuta in ambito militare durante la guerra fredda. Nel periodo compreso tra il 1971 e il 1982 nascono i seguenti paradigmi:

- **Simbolico**: studia il linguaggio naturale tramite regole e grammatiche. In verità la loro nascita avviene nel decennio precedente, ma in questo lasso di tempo vengono arricchite con algoritmi di riconoscimento, quali Hidden Markov Model (HMM);

- **Logic-based**: unifica le strutture in feature (interconnessioni tra le parti del discorso) analizzabili in modo più potente rispetto alle grammatiche context-free.

- **Natural language understanding**: comprende lo sviluppo del sistema SHRDLU, realizzato da Terry Allen Winograd nel 1972, il quale utilizza un meccanismo di comprensione basato su tre fasi di analisi: sintattica, semantica e deduttiva. Tramite uno schermo grafico l'utente osserva un ambiente virtuale costituito da una superficie piana, una scatola e una serie di oggetti di varie forme. L'utente può interagire in lingua inglese con un immaginario braccio meccanico per spostare gli oggetti. Il programma è in grado di risolvere molte ambiguità della lingua inglese e riuscire a capire, ad esempio, a quale tipo di oggetto ci si riferisce anche se è sottinteso. In Figura 2.1 viene mostrata una rappresentazione grafica dell'ambiente e degli oggetti utilizzati da questo sistema;

- **Discourse modelling**: branchia della linguistica che analizza sotto-strutture del linguaggio scritto e parlato.

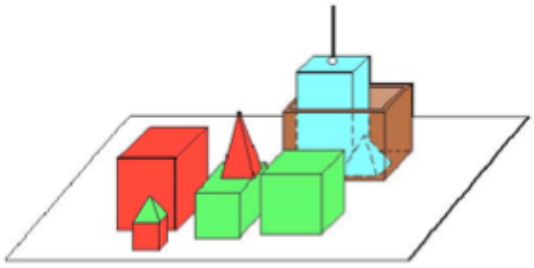

Figura 2.1: Ambiente virtuale del sistema SHRDLU

Negli ultimi anni si sta evidenziando un pesante uso di metodologie data-driven e modelli probabilistici.

In ambito universitario e di ricerca la Sentiment Analysis ha avuto un notevole interesse e ha portato alla nascita di diversi rami della stessa materia, ognuno dei quali concentrato su di uno specifico problema. Di seguito vengono elencati in modo sintetico, poiché una loro trattazione esula dagli obiettivi di questo lavoro di tesi.

Classificazione delle opinioni: La precisa classificazione delle opinioni è stata fin dal principio uno dei principali argomenti di discussione in ambito accademico [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. A tal fine la Sentiment Analysis viene trattata come un problema di classificazione di testi secondo due livelli di dettaglio. Il primo livello di classificazione consiste nell'assegnare una categoria positiva o negativa a un testo che esprime opinioni, la cosiddetta classificazione a livello di documento. Il secondo livello consiste invece nella classificazione di frasi dapprima in soggettive e oggettive e poi, per le soggettive, in positive, negative o neutrali. Si parla in questo caso di classificazione a livello di frase.

Sentiment Analysis basata su feature: La classificazione a livello di documento o di frase fornisce un'indicazione di massima sul giudizio degli utenti su un particolare aspetto di interesse e non riesce ad

esprimere informazioni dettagliate. Una frase con polarità negativa non è necessariamente indice di una visione negativa globale. Ad esempio sebbene la frase "The screen quality of this phone is not good" abbia una polarità negativa, questa non dovrebbe essere associata con il cellulare, ma solamente con lo schermo di questo. Il problema può essere scomposto in due fasi: identificare l'oggetto del discorso ("screen quality" nell'esempio) e determinare la polarità dell'opinione quella caratteristica. La teoria che tratta di questo problema si può trovare in [11, 12, 13, 14, 15, 16, 17].

Frasi comparative: Una frase comparativa esprime una relazione tra uno o più soggetti, generalmente con aggettivi o avverbi comparativi e superlativi. Necessitano di essere trattate diversamente rispetto alle altre frasi in quanto la loro semantica e la loro forma sintattica sono estremamente differenti. La letteratura si è dedicata allo studio di questo particolare problema [18, 19, 20] in modo da riconoscerle e assegnare loro una polarità.

False opinioni: L'utilizzo di messaggi automatici (spam) è un problema che si sta cercando di risolvere [21, 22]. L'identificazione di frasi pubblicitarie e false opinioni, cioè messaggi pubblicitari camuffati in modo da essere interpretati come vere e proprie opinioni, è una sfida per il futuro poiché attualmente non esistono algoritmi maturi per risolvere questo tipo problema.

2.4 Metriche di analisi delle reti sociali

L'analisi delle reti sociali è un ambito di ricerca nato già ad inizio secolo scorso e da sempre utilizzato per spiegare fenomeni pertinenti scienze quali la sociologia, la psicologia e l'economia. Solo negli ultimi anni lo studio delle comunità virtuali, tra cui in particolare quella dei social

media, sta prendendo in considerazione di affrontare il problema da un punto di vista delle reti sociali, riuscendo a collegare in questo modo risultati ottenuti in discipline molto diverse tra di loro, per meglio comprendere questo fenomeno di rapida ascesa e dal grande impatto sociale ed economico.

Gli studi sulle reti sociali si sono guadagnati dei riconoscimenti significativi negli ultimi anni sia in termini di avanzamenti teorici che di metodologie. Gli studi su questo argomento hanno impattato notevolmente su vari domini come il capitale sociale, la gestione della conoscenza e le teorie organizzative. Basate su costrutti teorici della sociologia e su quelli matematici propri della teoria dei grafi, grazie anche all'evoluzione di hardware e software degli ultimi anni, l'analisi delle reti sociali offre una metodologia che unifica la visualizzazione e l'investigazione delle strutture e delle relazioni sociali. Se, da un lato, per alcuni problemi, un sondaggio generale di carattere sociale può aiutare nello studio delle proprietà individuali in prima approssimazione, l'analisi delle reti sociali incorpora il contesto sociale per spiegare i comportamenti sia individuali che di gruppo. In questo modo le relazioni tra gli attori diventano di primaria importanza, mettendo in seconda posizione le caratteristiche proprie del singolo individuo. Lo sviluppo del campo delle reti sociali ha inizio negli anni Trenta tramite gli studi di diversi gruppi di ricerca che lavoravano in maniera indipendente. Grazie a Simmel cominciò ad emergere un approccio sistematico, costruendo una teoria che spiegava le cause dei fenomeni sociali e contribuì alla sociologia formale che può quindi essere considerata la progenitrice dell'analisi delle reti sociali. Nel 1934 Moreno fu il primo a operazionalizzare una rete sociale creando un sistema per rappresentarla come una combinazione di nodi e collegamenti tra di essi. Successivamente Cartwright e Harary [23, 24] continuarono il lavoro di Moreno cominciando ad applicare i concetti della teoria dei

grafi a quelli che allora erano ancora detti "sociogrammi". Grazie soprattutto all'introduzione dei collegamenti direzionati tra i nodi, nei loro studi sono stati capaci di spiegare pattern sociali di complessità molto maggiore rispetto ai risultati che si erano raggiunti sino a quel momento. Alla fine degli anni Trenta si sono formate due scuole separate: una, americana, formata dal gruppo di lavoro presso l'università di Harvard, l'altra, britannica, dagli antropologi dell'università di Manchester. La prima si concentrava soprattutto su tecniche per individuare e studiare sottogruppi di persone all'interno di gruppi originari più ampi con lo scopo di analizzare e comprendere i rapporti tra i sottogruppi. Da questa scuola, come illustrato brevemente in [25], si generò il cosiddetto *approccio sociocentrico*, che concerne lo studio quantitativo delle relazioni tra le persone all'interno di un determinato gruppo. L'obiettivo è la misura dei pattern strutturali di quelle interazioni e di come questi pattern riescano a spiegare i risultati. La seconda scuola, invece, si concentrò maggiormente sullo studio delle comunità dando origine all'*approccio egocentrico*. I ricercatori di questa scuola analizzarono le reti di relazioni sottostanti agli individui piuttosto che focalizzarsi sull'intera società. L'obiettivo finale è la generalizzazione delle caratteristiche trovate sulla rete personale all'intero gruppo. Successivamente, i loro lavori Barnes [26], Granovetter [27, 28] e Milgram [29] hanno formalizzato la teoria dell'analisi delle reti sociali tra gli anni '50 e '70. In particolare, Barres ha dato origine alla nozione di rete sociale come un risultato paradossalmente secondario del suo studio di una parrocchia su un'isola norvegese nel 1953. In seguito, Granovetter, ne sottolineò l'importanza come mezzo di diffusione della conoscenza e dell'informazione attraverso i suoi lavori sui legami deboli tra individui. Tali concetti vengono ripresi anche in studi relativi alle reti sociali su Internet, come *Source Forge.net* o le mailing list [30, 31] e le reti sociali di manager aziendali, come illustrato in [32, 33, 34]. Milgram, invece, introdusse il famoso concetto dei *six degrees of separtion* [29]

con cui tentava di dimostrare l'idea di un mondo piccolo (*small world phenomenom*). Mediante un approccio empirico supportò la sua tesi mostrando che tutte le persone negli Stati Uniti sembravano essere connesse tra di loro in media da sei legami di conoscenza, ipotizzando quindi che un cittadino statunitense è in grado di riuscire a conoscere qualsiasi altro connazionale attraverso le presentazioni di circa sei persone. Anche il concetto di "small world" viene applicato e studiato in numerose aree; ad esempio nel campo delle pubblicazioni scientifiche, due ricercatori sono collegati se sono coautori di uno stesso articolo [35, 36].

Contemporaneamente, importanti lavori indipendenti si svilupparono presso l'Università della California (Irvine) soprattutto intorno a Freeman, che si concentrò anche sulla definizione dei nodi importanti all'interno di una rete e delle relative metriche, come illustrato nel suo articolo [37] ed è autore, insieme a Borgatti e Everett, del software Ucinet [38], tra i più completi e diffusi per l'analisi direti sociali.

Oggi la teoria delle reti sociali è largamente utilizzata per studiare l'influenza della struttura della rete sociale all'interno di un'organizzazione o di un gruppo di lavoro e la gestione della conoscenza al suo interno. La gestione della conoscenza (*knowledge management*) comprende un insieme di pratiche volte a identificare, creare, rappresentare e ridistribuire la conoscenza all'interno delle organizzazioni.

La maggior parte degli studi si è focalizzata su due principali pratiche per gestire la conoscenza in maniera formale: l'applicazione di sistemi di supporto e le comunità. La prima soluzione punta sulla codifica e la distribuzione della conoscenza tramite l'*information and comunication technology,* come i database e Internet; la seconda, invece, promuove lo scambio di conoscenza tra persone con interessi e obiettivi comuni.

Queste ricerche hanno dimostrato che chiedere consiglio ai propri parigrado è un importante canale che favorisce quotidianamente uno scambio di conoscenza molto specifica. Per questo motivo una parte della ricerca in quest'ambito si sta focalizzando sulla relazione tra i benefici del *knowledge management* che derivano dalla struttura delle reti sociali corrispondenti come illustrato in [39, 40, 41].

Song et al. [42] affrontano uno studio della rete sociale in una grossa industria aerospaziale concentrandosi maggiormente sulle caratteristiche delle divisioni organizzative che individuali, concludendo che alcune misure strutturali delle reti di queste unità aziendali, come la centralità in termini di grado o i buchi strutturali, siano indicativi delle capacità creative dell'unità organizzativa stessa e più in generale della capacità di gestire la conoscenza. L'analisi delle reti sociali riguarda anche il campo degli studi economici soprattutto per capire il ruolo e l'importanza dell'individuo all'interno delle organizzazioni. In questo senso numerosi studi [43, 44, 45, 46, 47, 48, 49, 50] si concentrano sull'analisi delle reti di impiegati e manager di aziende di diverse dimensioni, costruendo tramite dei questionari le reti sociali individuali, misurandone le metriche relative e dimostrando correlazioni di questi valori con le performance individuali. I principali limiti di questi lavori possono essere trovati alcune volte nelle piccole dimensioni del campione e nel metodo utilizzato per costruire le relazioni basato su questionari individuali che lasciano troppo spazio alla soggettività.

2.4.1 Cenni teorici

L'idea alla base delle reti sociali è molto semplice: una rete è un insieme di attori (o punti, o nodi, o agenti) che hanno relazioni gli uni con gli altri. Dato un insieme finito U di elementi:

$$U = \{ X_1, X_2, \ldots, X_n \}$$

E un numero finito di relazioni *R*:

$$R_t \subseteq U \times U$$

con

$$t = 1, 2, \ldots, r$$

si definisce rete sociale *N* la *n*-upla composta dall'insieme finito di elementi *U* e da *(n - 1)* relazioni tra di essi:

$$N = (U, R_1, R_2, \ldots, R_r)$$

Le relazioni possono rappresentare qualunque tipo di legame: per esempio il legame tra padre e figlio oppure l'appartenenza ad uno stesso progetto da parte di due sviluppatori o i rapporti di amicizia tra le persone. Le relazioni possono godere delle seguenti proprietà:

1. Riflessiva: $\forall X \in U : XRX$
2. Irriflessiva: $\forall X \in U : \neg XRX$
3. Simmetrica: $\forall X, Y \in U : (XRY \Rightarrow YRX)$
4. Asimmetrica: $\forall X, Y \in U : \neg(XRY \wedge YRX)$
5. Antisimmetrica: $\forall X, Y \in U : (XRY \wedge YRX \Rightarrow X = Y)$
6. Transitiva: $\forall X, Y, Z \in U : (XRY \wedge YRZ \Rightarrow XRZ)$
7. Intransitiva: $\forall X, Y, Z \in U : (XRY \wedge YRZ \Rightarrow \neg XRZ)$

Inoltre una relazione è detta di ordinamento parziale se gode delle proprietà riflessiva, antisimmetrica e transitiva ed è utile per descrivere i rapporti gerarchici all'interno di una organizzazione, mentre si dice che è una relazione di equivalenza se gode delle proprietà riflessiva,

simmetrica e transitiva, che modella il caso in cui tutti gli elementi della rete sociale hanno lo stesso ruolo.

Una rete definita tramite una relazione R può essere rappresentata [51, 52, 53] in modi diversi:

Tramite la matrice binaria $R = [r_{i,j}]_{nxn}$, detta anche matrice di adiacenza, dove

$$R_{ij} = \begin{cases} 1 \text{ se } XiRXj \\ 0 \text{ altrimenti} \end{cases}$$

se gli archi sono pesati il valore r_{ij} può essere il numero reale che indica la "forza" del legame *R* tra X_i e X_j;

Tramite la lista dei vicini, specificando per ciascun nodo la lista degli altri nodi a cui è relazionato (è il formalismo adottato dai software di analisi di reti sociali);

Tramite un grafo G = (V, L), dove V è l'insieme dei vertici, che rappresentano le unità della rete, e $L = U\ L_i$ con i = 1….r è l'insieme degli archi che indicano le relazioni.

In base a quanto visto finora si possono elencare quattro tipi di reti sociali, che sono i più utilizzati in letteratura.

a. **Rete non direzionata**: la relazione R è simmetrica, ovvero tutti gli archi non hanno una direzione specifica, come per esempio nella relazione di matrimonio tra due persone;
b. **Rete direzionata**: la relazione R è asimmetrica, quindi tutti gli archi sono orientati, come ad esempio nella relazione di paternità tra padre e figlio;
c. **Rete mista**: nella stessa rete si possono avere archi sia direzionati che non direzionati, per esempio quando si

rappresentano due o più relazioni nella stessa rete, come quella di matrimonio e paternità;

d. **Rete a due modi**: è formato da due insiemi di unità $U = U_\alpha \cup U_e$, spesso definiti in letteratura come attori U_α ed eventi U_e e una relazione R che connette i due insiemi, per esempio l'appartamento di una persona ad una associazione.

Le proprietà e le relative misure sulle reti si basano su due differenti livelli di analisi: quelle relative alla rete considerata nella sua interezza come relazioni tra insiemi di attori e quelle relative al singolo individuo (o elemento).

2.4.2 Metriche per l'analisi di reti sociali

Di seguito vengono presentate le metriche per l'analisi delle reti sociali maggiormente utilizzate in letteratura. Inizialmente verranno illustrate le metriche che descrivono le caratteristiche della rete nella sua interezza, successivamente quelle pertinenti l'analisi dei singoli nodi.

Dimensione della rete

La *dimensione della rete,* ottenuta semplicemente contando il numero di nodi presenti e quindi definita come

$$SIZE(N) = |U|$$

è un parametro tanto semplice quanto importante de tenere in considerazione nello studio di un problema basato su reti sociali. Basta considerare, per esempio, un insieme di studenti che frequentano un dato corso: se il numero di studenti è pari a qualche decina si ha un'alta probabilità che questi si conoscano tutti tra di loro e quindi possano condividere appunti e quindi conoscenza, se questa stessa classe dovesse essere composta da qualche centinaio di studenti, la probabilità che questi si conoscano diminuisce, anzi l'esperienza suggerisce che

tenderanno a formarsi dei gruppi. Per questo motivo la dimensione è un aspetto che viene sempre analizzato preliminarmente in tutti i lavori, come in [54, 55, 56, 57, 58, 59].

Reti personali

Per meglio comprendere il comportamento dei singoli individui bisogna focalizzarsi sui loro intorni locali, cioè su quella porzione di rete immediatamente adiacente al nodo che si sta considerando. In questo senso si parla di reti degli ego, dove per ego si intende un nodo considerato nella sua individualità. Naturalmente una rete ha tanti ego quanti sono i nodi, e, generalizzando, un ego può essere sia un individuo che un gruppo, un'organizzazione o l'intera società. Si definisce *vicinato* l'insieme dell'ego e di tutti i nodi a cui l'ego è connesso tramite un percorso di una data lunghezza. Solitamente per *vicinato* si intende quello di distanza unitaria, quindi include solamente l'ego e i nodi direttamente adiacenti. È necessario sottolineare che il *vicinato* comprende non solo i collegamenti diretti tra l'ego e i nodi a lui "vicini", ma anche tutti i collegamenti tra gli attori che fanno parte del *vicinato*. Il *vicinato* a *N-passi* $V_N(n_i)$ del nodo n_i generalizza il concetto di *vicinato* includendo tutti i nodi con cui l'ego ha una connessione tramite un percorso di lunghezza al più *N* e, naturalmente, tutte le connessioni tra tutti gli attori in questo modo considerati. Formalmente:

$$V_N(n_i) = R_i^N$$

dove R_i^N rappresenta la posizione della rete R centrata nel nodo n_i e comprendente tutti i nodi *vicini* fino alla distanza *N*. un esempio è mostrato in Figura 2.2, dove la rete vicinato centrata nell'ego d_1 è riportato in azzurro.

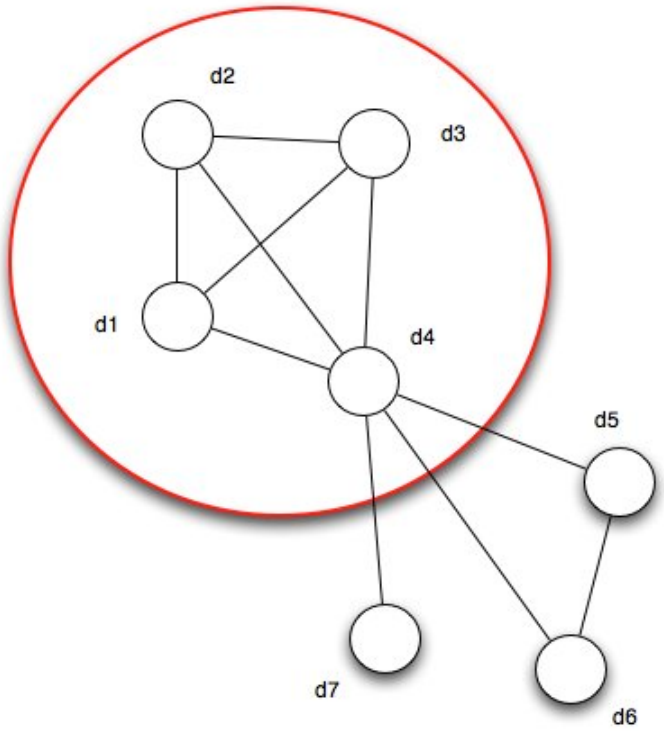

Figura 2.2: Esempio di rete vicinato ad un passo centrata sul nodo d_1

Quando si ha a che fare con reti direzionate si può fare la distinzione tra *vicinato entrante* e *uscente,* considerando rispettivamente solo i nodi da cui partono le relazioni verso l'ego o solo quelli a cui arrivano i collegamenti da parte dell'ego. Un esempio di rete *vicinato uscente* è riportato in Figura 2.3.

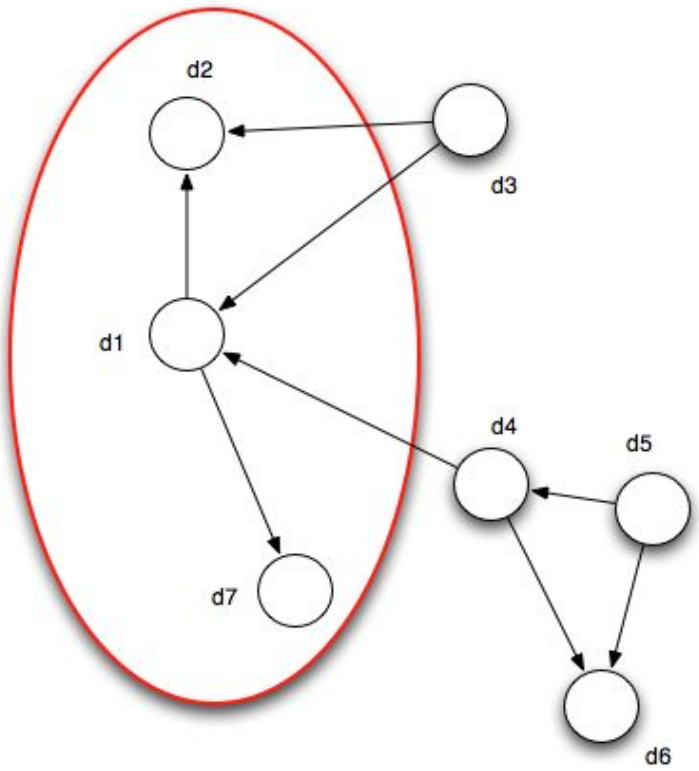

Figura 2.3: Esempio di rete vicinato uscente ad un passo centrata sul nodo d_1

Si parla anche di *vicinato a legami forti* e a *legami deboli* quando le relazioni tra i nodi sono pesati e non basati solo su dati binari. In quest'ultimo caso, infatti, è semplice stabilire se un nodo appartiene al *vicinato* dell'ego, in quanto questi sono connessi oppure non lo sono. Se invece si hanno misure della forza del legame tra due attori allora il problema si sposta verso quello della scelta di un valore di soglia per discriminare chi appartiene e chi no al vicinato dell'ego. La Figura 2.4 mostra un esempio di rete sociale con archi, i cui pesi sono dei valori reali. In questo caso la rete vicinato basata su legami forti, evidenziata in azzurro è centrata sull'ego d_1 ed è composta solo dai nodi d_2 e d_3, il cui legame con d_1, è superiore al valore di soglia unitario.

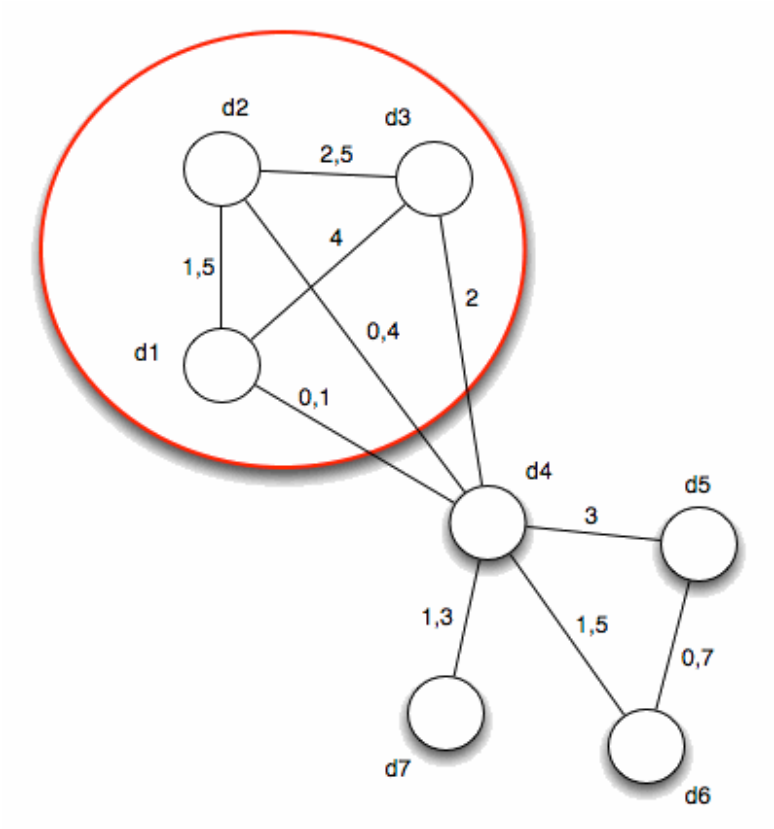

Figura 2.4: Esempio di rete vicinato a legami forti ad un passo centrata sul nodo d_1

Metriche di centralità

Se da un lato i sociologi sono d'accordo che l'importanza di un singolo nodo all'interno della rete sia una proprietà fondamentale da analizzare all'interno delle strutture sociali per le ripercussioni che può avere sull'intero insieme di relazioni, c'è molto meno accordo su come questa possa essere definita e quindi successivamente descritta insieme con le sue cause e le sue conseguenze. L'analisi delle reti sociali ha portato ad una migliore comprensione di quello che può significare potere sociale. Fondamentalmente, l'approccio di questo tipo di analisi porta a enfatizzare il fatto che il potere sia strettamente legato alle relazioni. Un

individuo non ha potere in senso astratto perché il potere è dato dal dominio nei confronti di altri individui e quindi dalle relazioni con gli altri nodi della rete. Poiché il potere è una conseguenza dei pattern delle relazioni, l'ammontare totale del potere può variare, infatti se un sistema è scarsamente accoppiato (quindi la rete ha una densità bassa) non può essere esercitato un potere molto grande, mentre in sistemi di dimensione maggiore e a maggiore densità c'è la possibilità di esercitare un'influenza maggiore. Per quanto detto il potere può essere quindi considerato sia una proprietà a livello di sistema (macro), che a livello di relazione (micro). L'insieme del potere in un sistema e la sua distribuzione tra gli individui sono correlati ma non possono essere confusi, infatti dato un ammontare totale di potere, ci sono reti in cui si hanno pochi nodi molto importanti che concentrano su di loro la maggior parte del potere e reti in cui questo è distribuito in maniera eguale tra tutti i nodi. Spesso la descrizione di un attore da parte di un analista viene fatta attraverso una serie di vincoli e opportunità che derivano dalla posizione in cui si trova all'interno della rete. Se un individuo si trova in una posizione più favorevole avrà meno vincoli e maggiori opportunità, cosa che si traduce, in generale, in una ,maggiore influenza e quindi in maggior successo rispetto a quelli che si trovano in posizione più svantaggiate. Se quindi il concetto di potere risulta difficile da spiegare, questo è stato declinato molto precisamente in una serie di definizioni matematiche precise che hanno dato origine a una serie di metriche, come illustrato in [37, 60, 61, 62, 63, 64, 65, 66, 67, 46, 68, 69, 70]. È possibile raggruppare le suddette metriche in tre concetti di centralità di un autore:

- Centralità in termini di grado o degree centralità;
- Centralità intermedia per singola unità o betweenness centrality;

- centralità in termini di vicinanza closeness centrality, eccentricity centrality e proximity centrality.

Centralità in termini di grado

Questa metrica basa la sua ragion d'essere sul presupposto che quanto più un attore è centrale nella rete, nel senso di avere molte connessioni, tanto più tende ad avere una posizione più favorevole e quindi maggior potere o influenza. Il vantaggio sta nel fatto di disporre di un maggior numero di alternative per soddisfare i propri bisogni ed essere meno dipendenti da altri individui; inoltre, avendo un maggior numero di relazioni, gli attori più importanti possono aver accesso anche a maggiori risorse presenti nella rete, soprattutto di tipo informativo, aumentando la loro probabilità di fungere da intermediari tra altri (concetto tenuto conto maggiormente dalla centralità intermedia). In definitiva, può essere considerata la metrica di centralità più semplice ed intuitiva, anche se certamente non esaustiva. L'approccio principale alla misura di questa metrica è quello di Freeman [37], il quale suggerisce di basare la misura semplicemente contando il numero di relazioni con il vicinato nel senso stretto del termine, quindi con un solo passo. Da questa affermazione deriva la definizione della centralità in termini di grado $C_d(a_i)$ dell'attore a_i come:

$$C_d(a_i) = SIZE\ (R_l) - 1$$

La misura finale può essere normalizzata o meno, in base alle esigenze, rispetto al numero di nodi della rete escluso uno, che è quello che si sta prendendo in considerazione. Il grande vantaggio di questo tipo di misura è anche la bassa complessità temporale del suo algoritmo di calcolo che risulta $O(n)$.

Bonacich [71] ha proposto una modifica alla metrica di Freeman. Bonacich, pur apprezzando l'efficacia della metrica di Freeman, contesta il fatto che avere uno stesso numero di connessioni non rende necessariamente due attori importanti in maniera uguale, ma che anzi è necessario guardare e confrontare la situazione del vicinato. Intuitivamente è possibile considerare che un individuo con pochissime connessioni può essere molto avvantaggiato se uno dei pochi attori cui è connesso ha a sua volta moltissime relazioni, perché in questo modo basterebbe inviare un solo messaggio per raggiungere un gruppo molto vasto. Da un altro punto di vista, però, l'attore in questa condizione non esercita alcuna forma di influenza sui suoi vicini potenti. Viceversa è da considerare che se gli individui a cui un attore è connesso non sono altrettanto ben connessi questi risultano altamente dipendenti da quest'ultimo. Lo stesso Bonacich argomentò che essere connessi ad attori a loro volta ben connessi rende un attore centrale ma non potente, invece essere connessi ad attori che non sono molto connessi rende potenti ma non centrali. Per comprendere meglio l'idea alla base della metrica di Bonacich si faccia riferimento alle Figure 2.5 e 2.6. Secondo Freeman il nodo d_1 in Figura 2.5 risulta più centrale del nodo d_6 in Figura 2.6 perché ha un numero maggiore di legami diretti, se si considera la sua rete vicinato ad un passo. Secondo l'approccio di Bonacich, invece, il nodo d_1, avendo dei vicini non connessi a loro volta si trova in posizione più svantaggiata rispetto a d_6, che, pur avendo un numero minore di legami diretti, può vantare un vicino, d_8, a sua volta meglio connesso.

Nel caso in cui gli archi della rete sono orientati, è possibile distinguere ulteriormente tra due tipi di misure: la centralità in ingresso e quella in uscita, dove la prima tende ad assumere la sfumatura di prestigio, perché vuol dire che l'attore in questione è "cercato" da molti altri, mentre la seconda denota l'influenza di un determinato attore, poiché

avere molte relazioni di verso uscente indica la capacità di scambiare con molti altri.

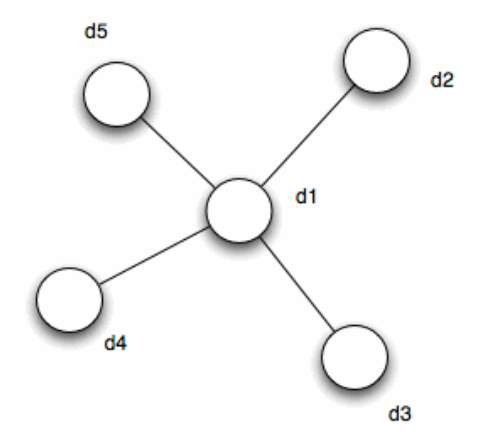

Figura 2.5: Esempio di rete in cui il nodo d_1 risulta molto centrale secondo Freeman

La metrica risultante da questo ragionamento è basata su un algoritmo iterativo (che raggiunge facilmente la convergenza solo nei casi particolari di reti simmetriche) in cui la misura di un attore corrisponde in partenza a quella di centralità secondo Freeman più una funzione pesata delle centralità degli attori a cui si è connessi. Queste iterazioni portano la complessità temporale dell'algoritmo di calcolo a $O(n^3)$. Quest'ultima metrica spesso viene considerata superiore alla precedente, anche se la prima è quella ad essere più spesso utilizzata sia per la sua semplicità interpretativa che implementativa, soprattutto considerando il fatto che, per reti di grandi dimensioni entrambe tendono a dare risultati molto simili.

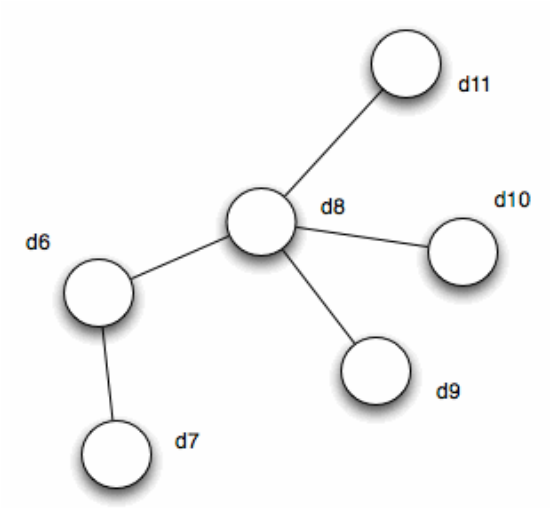

Figura 2.6: Esempio di rete in cui il nodo d_6 risulta molto centrale secondo Bonacich

Centralità intermedia

L'idea che sta alla base di questa metrica è se un attore a_i sta in mezzo ad un cammino tra un'altra coppia di attori a_j e a_k, a_i avrà molto probabilmente il ruolo di intermediario tra a_j e a_k. L'intermediario a_i, come si può intuire, ha sia la capacità di mettere in contatto gli attori a_j e a_k, ma anche eventualmente il potere di bloccare la comunicazione, oltre che di controllare il flusso di informazioni che passa attraverso i due individui a_j e a_k per i quali a_i fa da ponte.

La prima metrica per questo tipo di centralità ricalca la definizione appena data. Introdotta da Freeman [37], è calcolata sommando il numero di volte in cui un attore a_i fa parte di un cammino minimo presente tra tutte le coppie di nodi della rete $g_{ik}(n_i)$ rapportato al numero totale di cammini minimi tra tutte le coppie possibili g_{ik}. Quindi $C_b(a_i)$ è:

$$C_b(a_i) = \sum_{j<k} \frac{g_{jk}(a_i)}{g_{ik}}$$

La misura può essere normalizzata rispetto al numero massimo di cammini minimi possibili per tutti i nodi.

La complessità temporale dell'algoritmo di calcolo di questa metrica è $O(n^3)$.

Una variante di questo primo metodo è quella di considerare tutti i cammini tra i nodi, non solo quelli minimi e tra questi sceglierne uno casualmente. Questa tecnica, descritta in [72], ha il principale vantaggio nel diminuire la complessità temporale dell'algoritmo di calcolare portandola a $O(nm)$ dove n è il numero di nodi e m il numero di relazioni. In questo caso la centralità intermedia $C_b(a_i)$ è definita come:

$$C_b(a_i) = \sum_{j<k} \frac{r_{jk}(a_i)}{l_{ik}}$$

dove l_{jk} è il numero totale di cammini tra il nodo a_j e il nodo a_k, mentre $r_{jk}(a_i)$ rappresenta il numero di volte in cui il nodo a_i si trova in mezzo a un cammino tra tutti quelli che collegano i nodi a_j e a_k.

Centralità in termini di vicinanza

Un ulteriore motivo per il quale un attore può essere considerato importante all'interno di una rete è il fatto di essere più "vicino" agli altri attori e quindi più facilmente raggiungibili o poter raggiungere la restante parte del gruppo. Questo concetto può essere tradotto in termini di potere in quanto attori con valori bassi di questo tipo di centralità si troveranno nella periferia della rete, mentre al contrario valori alti implicano l'appartenenza ad un gruppo centrale ed influente. Queste metriche cercano di colmare una lacuna lasciata dalla centralità in termini di grado, la quale teneva in conto solamente le relazioni dirette tra gli attori o al limite quelle dei vicini, ma non i legami indiretti con tutti gli altri. Riprendendo l'idea di Bonacich, se un attore ha molte relazioni

con altri attori che ne hanno poche, questo risulta molto centrale, ma solo nel suo vicinato locale. Il concetto basilare per poter calcolare questo tipo di centralità è la distanza tra un nodo e tutti gli altri. Si ottiene una misura intermedia detta lontananza, che è la somma delle distanze tra l'ego e tutti gli altri nodi della rete. La lontananza è successivamente trasformata in centralità in termini di vicinanza facendone l'inverso ed eventualmente normalizzando moltiplicando per il numero di nodi della rete. La centralità in termini di vicinanza $C_c(a_i)$ per l'attore a_i è:

$$C_c = \frac{1}{\sum_{j=1}^{L} d(a_i, a_j)}$$

dove $d(a_i, a_j)$ rappresenta la distanza del cammino minimo tra i nodi a_i e a_j.

Nel caso in cui la rete fosse direzionata potremmo distinguere anche in questo caso in centralità entrante e uscente secondo le stesse interpretazioni viste per le altre metriche. Anche questa metrica è stata introdotta da Freeman [37], il suo algoritmo di calcolo ha una complessità temporale pari a $O(n^3)$ e presenta il problema di non poter essere applicata a reti disconnesse, se non evitando di considerare i nodi irraggiungibili.

Un altro modo per declinare il concetto di vicinanza è quello di conoscere la sottorete che un dato attore può raggiungere in un dato numero di passi: questa metrica è definita come raggiungibilità $C_r(a_i)$:

$$C_r(a_i) = k \sum_{j=1}^{T} \frac{1}{p(a_i, a_j)}$$

Il numero *T* di nodi a_j raggiunti dal nodo a_i viene pesato in base alla loro distanza $p(a_i, a_j)$, in modo tale che gli attori raggiunti in passi via via maggiori abbiano un peso minore moltiplicato per il coefficiente *k*. Alla fine i valori vengono normalizzati rispetto al valore più alto osservato di raggiungibilità all'interno della rete.

Un'altra metrica di centralità in termini di vicinanza è la centralità calcolata con gli autovettori della matrice di adiacenza (*eigenvector centrality*) definita da Bonacich [71]. La misura si ottiene calcolando l'autovettore principale della matrice di adiacenza dell'intera rete, ovvero l'autovettore corrispondente all'autovalore di modulo massimo. La sua stessa definizione porta all'interpretazione secondo cui un nodo che ha un alto valore di questo tipo di centralità è adiacente ad altri nodi dal punteggio alto. La centralità basata su autovettori $C_e(n_i)$ per l'attore a_i è definita ricorsivamente come:

$$C_e(a_i) = \frac{1}{\lambda} \sum_{j=1}^{N} A_{ij} C_e(a_j)$$

dove A rappresenta la matrice di adiacenza della rete sociale e λ l'autovalore di modulo massimo. Gli approcci basati sul calcolo degli autovettori e sulla distanza geodesica hanno il difetto di considerare solo il cammino più breve e dunque il più "efficiente". Secondo altre teorie potere e influenza dovrebbero invece essere espressi tenendo conto di tutti i cammini che connettono gli ego a ogni altro nodo della rete di cui fa parte. Secondo la tecnica di Hubbell [73] e Katz [74] è necessario contare tutte le connessioni tra gli attori, anche se la rete è direzionata, considerando insieme i percorsi entranti e uscenti, pesando ciascun percorso secondo la sua lunghezza: maggiore sarà quest'ultima, più debole sarà la connessione; l'indebolimento con l'aumentare della distanza è un fattore di attenuazione che può variare in base alla rete. I due approcci si differenziano per il solo fatto che Katz include una

matrice identità nel calcolo considerando come connessione più pesante quella di ogni ego con se stesso, cosa che invece Hubbel non fa. La centralità secondo Katz $C_k(a_i)$ per l'attore a_i è definita come:

$$C_k(a_i) = \sum_{j=1}^{N} W_{ij}$$

dove, posta $R = A^T$, trasposta della matrice di adiacenza A, si ha:

$$W = \alpha R + \alpha R^2 + \alpha R^3 + \ldots + \alpha R^N$$

con α un fattore di attenuazione. In maniera simile è definita la metrica di Hubbell $C_h(a_i)$:

$$C_h(a_i) = \sum_{j \neq i \wedge j=1}^{N} W_{ij}$$

Queste due metriche hanno lo svantaggio di essere realmente significative solamente nelle reti non direzionate poiché, per quanto spiegato sopra, fanno perdere di fatto l'informazione sul verso non considerandolo.

Un ultimo approccio alla misura della centralità in termini di vicinanza è quello di Stephenson e Zelen [75], anche detto harmonic closeness, perché riassume la centralità di ogni attore con la media armonica di tutte le sue distanze dagli altri. Questa metrica $C_{hc}(a_i)$ è quindi definita per l'attore a_i come:

$$C_{hc}(a_i) = \left(\frac{N}{\sum_{j=1}^{N} \frac{1}{p(a_i, a_j)}} \right)^{-1}$$

dove $p(a_i, a_j)$ rappresenta la distanza tra i nodi a_i e a_j calcolata secondo una teoria definita a priori: lunghezza del cammino minimo, o di un cammino casuale, o del cammino medio.

Le complessità temporali degli algoritmi per calcolare queste ultime metriche sono $O(n^3)$.

L'importanza del valore delle metriche di centralità per il successo sia individuale che di gruppo è stato confermato da numerosi studi in letteratura, soprattutto riguardanti le reti sociali di persone all'interno delle unità organizzative aziendali. In particolar modo vengono analizzati tre tipi di reti: la cosiddetta *advice network,* ovvero la rete che permette il flusso di conoscenza, *friendship network* la rete delle amicizie e la *hindrance network*, ovvero la rete delle inimicizie [76]. Uno studio condotto su 190 impiegati di cinque grandi aziende dimostra che la performance lavorativa individuale è correlata positivamente alla centralità in termini di grado nell'*advice network*, mentre è correlata negativamente alla centralità nella *hindrance network*, che inoltre a conseguenze negative anche sulle performance dell'intero gruppo di lavoro.

Analogamente in [77] viene condotto uno studio presso un'azienda americana di pubbliche relazioni che evidenzia come la centralità in termini di grado nella variante proposta da Bonacich nella *advice network* e la raggiungibilità nella *friendship network* siano positivamente correlate rispettivamente ai contributi individuali e alle posizioni formali all'interno dell'organizzazione.

Il lavoro di Carroll e Teo [78] approfondisce le questioni legate alle reti sociali dei manager confrontandole con quelle degli impiegati. I tipi di rete utilizzate sono due: quella della *membership organizzativa*, cioè i legami dei manager con le aziende presso cui lavorano, e quelle delle discussioni importanti (*core discussion*), ovvero la rete delle relazioni interpersonali relative a questioni strettamente private. Analizzando le dimensioni delle reti e il grado di centralità nel senso della vicinanza è emerso come i manager abbiano in generale delle reti di dimensione

maggiore e che siano strettamente tra di loro (grazie anche all'appartenenza ad associazioni esterne come quelle di beneficenza); inoltre dirigenti con un numero più elevato di collaboratori appartenenti alla loro rete personale sono associati ad un numero maggiore di premi sul lavoro e stipendi più elevati.

2.5 Costruzione di reti sociali in blog e forum

Contrariamente ai social network, blog e forum non hanno una vera e propria rete sociale su cui calcolare le metriche di Sezione 2.3. In [79] viene presentato un metodo per analizzare le discussioni che avvengono in questi tipi di social media. Il paper si focalizza su reti di professionisti ma i metodi e gli algoritmi introdotti possono essere applicati ad ogni social network dove la struttura delle discussioni segue le regole tipiche di un forum con post e repliche. All'interno di una discussione si può distinguere intuitivamente tra due tipi di utenti: "*information providers*" e "*information consumers*". Specialmente nelle discussioni on-line si può facilmente identificare i soggetti che forniscono informazioni tenendo traccia dei post che essi pubblicano. In contrasto, determinare i lettori di una determinata informazione è tutt'altro che banale. Non potremo mai essere sicuri che un utente legga un commento in un forum, anche se esso dovesse visualizzarlo nel suo browser. Diversamente, se qualcuno replica ad un post in particolare, allora possiamo essere sicuri che l'utente in questione abbia letto il messaggio e lo abbia giudicato interessante e degno di replica. L'utente che replica può essere visto come un consumatore di informazione ma allo stesso tempo anche come un information provider. Nell'approccio che si utilizzerà in seguito tracceremo esattamente questo comportamento all'interno di una discussione, identificando un'interazione tra due utenti quando uno di essi replica al post dell'altro.

In contrasto con gli approcci più comuni, non si utilizzeranno ne il numero di post prodotti (almeno non direttamente), ne il feedback degli altri utenti ma piuttosto si cercherà di estrarre le interazione che riguardano le relazioni tra coppie di utenti. Interpreteremo un thread simile ad un gruppo di discussione e si stabiliranno relazioni tra i partecipanti che postano in una catena. La Figura 2.7 mostra la struttura di una discussione a thread dove ogni quadrato rappresenta un commento fornito dal partecipante indicato. Per il commento evidenziato, scritto da *w*, le frecce mostrano le iterazioni tra gli utenti. L'utente che scrive il commento *w* da un voto ai post scritti dall'utente *v* e *u*, guadagnando il contributo di *k*. Il peso è unitario per la particolare struttura dei forum che andremo ad analizzare (Capitolo 5). Per strutture più complesse si rimanda a [79].

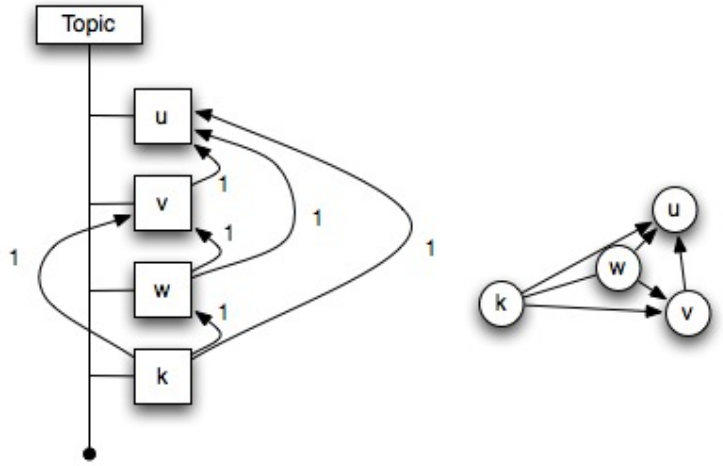

Figura 2.7: Costruzione grafo

Tutte le interazioni tra due partecipanti sono aggregate e vengono inserite in un grafo rappresentando delle interazioni dirette e pesate come mostrato in Figura 2.7.

2.6 Importanza delle sorgenti informative

Il concetto di reputazione di una fonte prende spunto dalla letteratura sul data quality. Si parte da una classificazione delle dimensioni di reputation fornita da [80]. L'articolo spiega come *accuracy*, *completeness* e *time* rappresentino le dimensioni fondamentali di data quality in molti contesti. *Interpretability, authority* e *dependability* rappresentano dimensioni aggiuntive che dovrebbero essere considerate quando si valuta la reputation, specialmente per fonti informative semi-strutturate o non strutturate. Gli esperimenti per valutare queste metriche sono stati eseguiti su blogs e forums e per questo trova applicazione nel contesto della Sentiment Analysis. Si sono identificati quattro aspetti per valutare la reputazione di blog e forum:

- **Traffic**: il volume totale di informazione prodotta e scambiata in un determinato intervallo di tempo;
- **Breadth of contributions**: la totalità dei temi trattati all'interno della fonte;
- **Relevance**: il grado di specializzazione di una fonte in un dato dominio (es. turismo);
- **Liveliness**: rapidità di reazione a nuove discussioni ed eveni.

La Tabella 2.1 riporta sulle colonne le metriche per il calcolo della reputazione, mentre sulle righe le diverse dimensioni di data quality. Tra parentesi sono riportate anche le sorgenti da cui si ricavano le metriche, dove "crawling" è un'ispezione della fonte che può essere effettuata manualmente oppure in modo automatizzato. Nel nostro caso l'estrazione delle informazioni viene fatta in modo automatico da dei crawler specifici per ogni sito. Alcune delle metriche sono fornite da Alexa (www.alexa.com), un'azienda molto conosciuta nel fornire informazioni statistiche su accessi, pagine visualizzate, ecc.. di siti presenti su internet. Si può anche notare che non tutte le dimensioni sono applicabili a tutte le variabili (in tabella indicate con n.a.).

	Traffic	Breath of contributions	Relevance	Liveliness
Accuracy	n.a.	Number of comment to selected post (crawling)	Number of feed subscriptions (feedburner tool)	n.a.
Completeness	n.a.	Number of open discussions (crawling)	Number of open discussions compared to largest web blog/foru (crawling)	Number of comments per user (crawling)
Time	Traffic rank (www.alexa.com)	Age of source (crawling)	n.a.	Number of new discussions opened per day (www.alexa.com)
Interpretability	n.a.	Number of distinct tags (crawling)	n.a.	n.a.
Authority	-daily visitors (www.alexa .com) -daily page views (www.alexia.com) average time spent on site (www.alexa.com)	n.a.	Number of inbound links (www.alexa.com)	Number of daily page views per daily visitor (www.alexa.com)
Dependability	n.a.	Number of comments per discussion (crawling)	Bounce rate (www.alexa.com)	Average number of comments to post provided within 24 hours (crawling)

Tabella 2.1: Metriche di reputation

Per validare questi tipi di metriche si sono eseguite 100 queries [Reputation-based selection] che riguardano il turismo sul motore di ricerca Google. Si è scelto tale dominio considerando che più del 60% degli utenti web eseguano ricerche relative a viaggi e turismo. Riferirsi a uno specifico dominio, aiuta a selezionare un insieme di queries che rispecchi il più possibile il modello di ricerca dello specifico dominio. In questo esperimento ci si riferisce al "Anholt-GfK Roper Nations Brand Index" [81]. Questo indice definisce sei dimensioni fondamentali lungo le

quali si possono identificare le variabili di decisione dei potenziali turisti. Queste dimensioni sono: presence, place, prerequisites, people, pulse e potential. Lungo queste dimensioni si sono identificate dieci variabili:

1. Weather and environment
2. Transportation
3. Low fares and tickets
4. Fashion and shopping
5. Food and drinks
6. Arts and culture
7. Events and sport
8. Life and entertainment
9. Night and music
10. Service and schools

La discussione del modello e delle variabili che guidano la scelta degli utenti va al di là dei nostri scopi e può essere trovata in [81]. Ciò che è più importante è verificare se questo concetto di reputation può essere utile per identificare delle fonti informative importanti per tutte le variabili o almeno per alcune di esse. Sono state definite 10 queries per ogni variabile. Esse sono ottenute sulla base delle 5 queries descritte in Tabella 2.2 aggiungendo "London" e "New York" a ogni query.

Decision making variable	***Tags for five basic queries***
Weather and environment	level of pollution, congestion charge, sustainable tourism, weather, air quality
Transportation	underground, rail, airport, traffic jam, street
Low fares and tickets	low-cost flights, cost of living, discounts and reductions, student fare, tickets discount
Fashion and shopping	shopping, fashion, department store, second hand, vintage
Food and drinks	pub, wine, beer, pizza, good cooking
Arts and culture	museums, monuments, parks, festivals, art
Events and sport	sport, tennis courts, city marathon, NBA, football
Life and entertainment	cinema, restaurants, clubs&bars, theaters, theme parks
Night and music	nightlife, music, theaters, party, jazz
Services and schools	public transports, accommodation, university, utilities, healthcare

Tabella 2.2: Query base

Per fare in modo che google si limiti a riportare i risultati ottenuti da blog e forum tutte le query sono della forma: <"tag" [London or New York] "tag" [blog or forum]>. La figura 2.8 mostra i risultati ottenuti da google utilizzando una query riguardo il cinema a Londra.

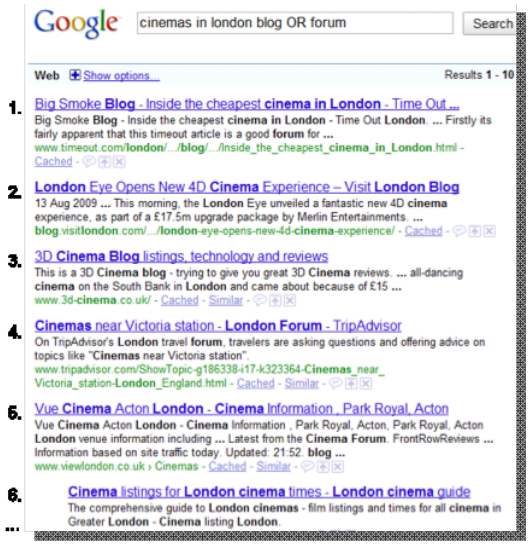

Figura 2.8: Risultato della query

Per ogni sito restituito, sono state calcolate le metriche di Tabella 2.1 ed è stato eseguito un riordinamento, ottenendo più di 1000 posizioni scambiate. Dagli studi eseguiti [82] emerge che Google non si focalizza sulla qualità delle informazioni contenute nei siti che ritorna. Il modello presentato tiene conto di maggiori fattori ma questa completezza introduce una complessità data dalla grande quantità di metriche considerate. Per ridurre questa complessità è stata eseguita una "Principal Component Analysis" (PCA) [82]. Questo tipo di analisi è usato per ridurre le metriche iniziali a un più piccolo insieme di gruppi correlati. In Tabella 2.3 presentiamo il risultato di questa analisi che si compone di tre costrutti:

- **Traffic**: costrutto che raggruppa tutte quelle metriche che sono direttamente o indirettamente coinvolte nel traffico generato dall'autorità del sito;

- **Perticipation**: sono contenute tutte le metriche che misurano il contributo degli utenti che scrivono messaggi, repliche e che tengono aggiornato il contenuto del sito;
- **Time**: costrutto che è un indice dell'interesse mostrato dagli utenti e colleziona le metriche che tengono traccia del tempo speso sul sito.

Variabile	**Costrutto**
Traffic rank	**Traffic**
Daily visitors	
Daily page views	
Number of inbound links	
Number of open discussions compared to largest Web blog/forum	
Number of new discussions opened per day	**Participation**
Number of comments per discussion	
Average number of comments to post provided within 24 hours	
Average time spent on site	**Time**
Bounce rate	

Tabella 2.3: Risultato PCA

Gli esperimenti mostrano che si ottengono diversi risultati applicando le varie metriche di reputation e ometterle può produrre una valutazione che non prende in considerazione alcune caratteristiche delle sorgenti.

La classificazione delle sorgenti informative aiuta gli utenti a selezionare le fonti più autoritarie e può essere utile per la Sentiment Analysis.

Capitolo 3

Architettura

3.1 Introduzione

Questo lavoro di tesi trova collocazione come parte di un progetto di ricerca il cui fine e quello di definire una metodologia e i relativi strumenti per valutare la qualità dei dati prodotti da un'analisi automatica di Web reputation. Da un punto di vista di business, il concetto di Web reputation può essere definito come l'opinione espressa sul Web riguardo un prodotto di un'azienda o sul suo brand.

Nel presente capitolo verranno illustrati dapprima gli obiettivi del progetto di ricerca nel suo complesso (Sezione 3.2), per poi proseguire con la descrizione dell'architettura dello strumento di analisi sviluppato, ponendo particolare attenzione ad evidenziare i problemi di integrazione tra componenti e le soluzioni adottate (Sezione 3.3).

3.2 Obiettivi del progetto

Come già accennato, obiettivo del progetto di ricerca, di cui questo lavoro di tesi fa parte, è quello di definire una metodologia e i relativi strumenti per valutare la Web reputation, ovvero dell'opinione espressa su fonti online relativamente ad un brand di un'azienda o di un suo prodotto. Questa opinione può derivare dalla percezione degli utenti, da recensioni ufficiali o da attacchi di concorrenti. Qualunque sia il caso, monitorare le fonti dove le opinioni sono pubblicate e gli effetti che

queste hanno sui clienti, sia attuali che potenziali, sta diventando un obiettivo di marketing necessario.

Molte aziende concordano che il Web è divenuto una preziosa fonte di informazioni per le operazioni di marketing, in quanto viene visto come una enorme e ricca base di conoscenza costantemente aggiornata. Monitorare il Web è visto come un'alternativa, in tempo reale, ai costosi sondaggi tradizionali. Sfortunatamente, i manager sono ancora convinti che gli strumenti attualmente presenti sul mercato siano immaturi e, poiché le decisioni critiche dovrebbero essere prese in relazione alle informazioni derivanti da questi strumenti, si mantengono ancora molto cauti sul loro impiego [83]. In particolare, è richiesta una valutazione oggettiva della qualità dei dati prodotti da un'analisi automatica della Web reputation e di prove concrete di come questa possa essere la base di informazioni su cui prendere delle decisioni [83]. Allo stato attuale, la letteratura non è in grado di soddisfare queste necessità a causa delle difficoltà tecniche di un'analisi della data quality in questo contesto.

Riepilogando, la valutazione della qualità di analisi automatiche di Web reputation richiede algoritmi in grado di valutare la qualità dei dati forniti come input, una conoscenza approfondita dell'architettura degli strumenti software che supportano tali analisi e, idealmente, uno strumento di riferimento che fornisca output di così alta qualità da poter essere considerati come benchmark per i confronti. Tuttavia, gli strumenti software necessari al supporto di tali analisi sono molto complessi, di conseguenza la costruzione di uno strumento di riferimento rappresenta tuttora un problema di ricerca aperto.

L'obiettivo del progetto è quindi quello di fornire una metodologia per costruire uno strumento di riferimento e di implementare i necessari

componenti software che sono responsabili della scarsa qualità dei dati. In dettaglio, gli obiettivi posti sono i seguenti:

- Analizzare lo stato dell'arte degli strumenti automatici di analisi della Web reputation già esistenti;
- Identificare e classificare le variabili tecniche che differenziano gli strumenti esistenti e che impattano sulla qualità dei risultati;
- Identificare i criteri più rilevanti e i relativi algoritmi da impiegare per la valutazione della qualità dei risultati;
- Definire e implementare strumenti di misura per valutare le variabili tecniche identificate e il loro impatto sulla qualità dei risultati;
- Costruire un corpus di test che sia scientificamente e statisticamente valido;
- Analizzare e confrontare i dati al fine di inferire delle linee guida di una metodologia per l'implementazione di uno strumento innovativo di analisi della Web reputation.

È necessario osservare come la data quality sia un aspetto chiave per tali strumenti, in quanto questa tipologia di analisi è soggetta ad errori derivanti dalla natura non strutturata o semi-strutturata dei dati presenti sul Web. Nello specifico, l'estrazione di un'opinione da un testo e la sua valutazione del grado di positività o negatività, sono operazioni soggette ad errori a causa dell'incertezza dovuta all'impiego di tecniche di elaborazione del linguaggio naturale.

3.3 Architettura dello strumento

Gli strumenti esistenti a supporto di analisi di Web reputation possono essere divisi in due categorie: semantici e non semantici. I primi basano

le loro valutazioni dei contenuti Web sull'interpretazione semantica del linguaggio naturale. I secondi basano la loro competitività sull'enorme quantità di informazioni che analizzano (alto numero di fonti Web), fornendo però un'interpretazione grossolana della reputation, ottenuta senza comprendere la semantica del contenuto.

I testi che derivano da fonti di informazione ufficiali, quali giornali online, possono essere assunti come affidabili, generalmente ben scritti e semplici da interpretare. Ma se derivano da siti Web 2.0, quali siti di microblogging, devono essere necessariamente trattati con operazioni di data cleaning prima di poter essere interpretati. Questo componente di data cleaning rappresenta un primo componente critico mancante in molti degli strumenti esistenti ma che è stato invece inserito e testato nell'architettura proposta.

Inoltre, come discusso in uno speciale report pubblicato dall'Economist il 27 Febbraio 2010 [84], una delle sfide maggiori per il futuro dell'ICT è la gestione dell'overload informativo dovuto all'aumento della disponibilità di acquisizione e scambio di dati. Come conseguenza, diviene cruciale la definizione di un metodo capace di valutare la qualità dell'informazione nonché di valutare la sua rilevanza per compiti specifici. Il software necessario ad abilitare l'interpretazione semantica del linguaggio naturale e a eseguire i relativi test di qualità rappresenta il secondo componente attualmente mancante in letteratura ma che è stato inserito e testato nell'architettura proposta [85].

Un terzo componente mancante sia in letteratura che negli strumenti esistenti è relativo alla valutazione dell'impatto della reputazione delle fonti Web sulla qualità dei risultati delle analisi. Secondo la nostra conoscenza, in letteratura non sono presenti lavori in cui vengono discusse funzioni per il bilanciamento di credibilità e reputazione di differenti fonti Web.

Infine, sulla base dell'architettura definita e sull'analisi degli strumenti esistenti, sono stati definiti degli indicatori di qualità e i relativi algoritmi necessari per la loro misurazione. Tali indicatori sono quindi inclusi in una dashboard. Le dimensioni degli indicatori di qualità, quali confidenza, completezza e accuratezza, sono definite tramite ben affermate tecniche, provenienti da ambiti di sociologia e microeconomia, per la definizione di un insieme coerente di indicatori al fine di ridurre gli errori dovuti alla eterogeneità delle analisi. In particolare, la valutazione della qualità dei risultati viene correlata alla qualità e all'affidabilità delle fonti, fornendo così un contributo innovativo in materia di data quality.

In Figura 3.1 è riportata l'architettura software dello strumento proposto, dove, con linee continue, è indicato il flusso dei dati mentre, con linee puntate, sono mostrate le interazioni tra componenti software o tra componenti e sorgenti dati. Per chiarezza, ad interrompere il flusso di elaborazione sono riportate fittizie basi di dati allo scopo di indicare qualè l'informazione che andrà ad arricchire quella già presente nella base di dati reale.

Nella rimanente parte del capitolo verranno discussi i singoli moduli software che compongono l'architettura proposta, ponendo particolare attenzione ad evidenziare i problemi di integrazione e le soluzioni adottate.

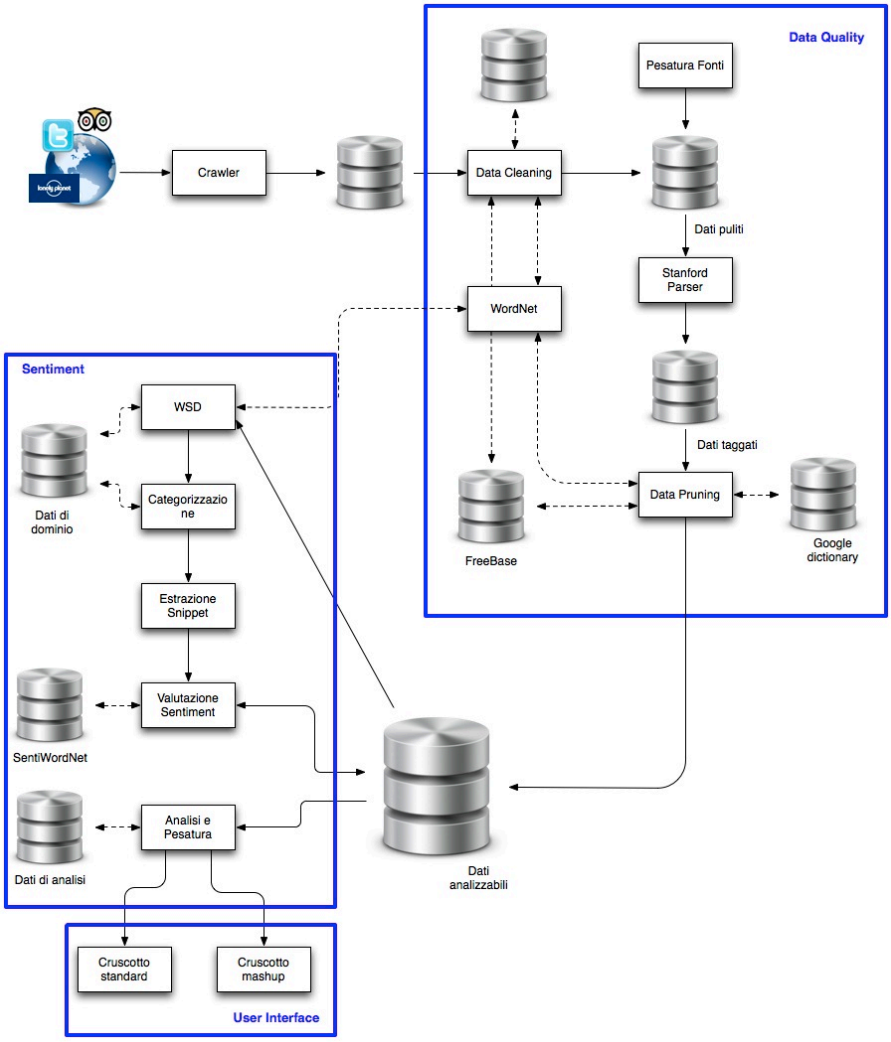

Figura 3.1: Architettura generale

3.3.1 Crawler

Un Web crawler [86], anche noto come Web spider o bot, è un componente software il cui fine è quello di navigare il Web, in maniera metodologica e automatica, tipicamente per reperire informazioni. I motori di ricerca, ad esempio, utilizzano dei crawler per creare una copia di tutte le pagine visitate in previsione di una loro successiva

indicizzazione per permettere un'esecuzione più rapida delle ricerche. I crawler, se utilizzati per raccogliere specifiche informazioni dalle pagine Web, possono essere impiegati con successo anche a supporto di attività quali lo spam, caso in cui l'interesse è rivolto alla raccolta di indirizzi e–mail.

L'algoritmo più generale eseguito da un Web crawler necessita di una lista di URL come input e, in maniera ripetitiva, esegue i seguenti passi: rimuove un URL dalla lista, determina l'indirizzo IP del suo host, esegue il download del corrispondente documento ed estrae ogni link che vi è contenuto aggiungendolo, se non è stato già visitato in precedenza, alla lista. Se necessario, dopo aver eseguito il download del documento, ne esegue l'elaborazione in una qualche maniera (ad esempio, indicizzandone il contenuto).

I Web crawler si collocano all'interno dell'architettura proposta come componenti abilitanti dell'intero processo di analisi, in quanto il loro compito è quello di reperire i dati che verranno analizzati. Obiettivo dei Web crawler è infatti quello di estrarre dalle fonti Web di interesse, post e commenti relativi a determinati argomenti. Quali sono le fonti da monitorare e quali gli argomenti di cui reperire informazioni devono essere necessariamente definiti in base al dominio di analisi. Attualmente, il caso di studio è incentrato sull'argomento turismo, in particolare sulle città di Milano, Madrid e Londra. Le fonti selezionate per reperire informazioni su questo specifico argomento sono Twitter, TripAdvisor e Lonely Planet, di cui sono stati implementati i rispettivi Web crawler chiaramente evidenziati nelle architetture di Figura 3.1.

3.3.2 Pesatura delle fonti

Classificare le fonti è un argomento che in passato ha ricevuto notevoli attenzioni. I motori di ricerca fanno della classificazione di siti web un

differenziale competitivo come ad esempio l'algoritmo di Page Rank [106] utilizzato da Google. Data la natura general purpose dei motori di ricerca, si è reso necessario sviluppare un nuovo metodo di ranking che tenga in considerazione molti aspetti tipici dei blog e delle community. Le metriche di importanza per una fonte di questo genere, possono essere sintetizzate in tre costrutti [82].

Traffic: prende in considerazione i volumi di traffico che attraversano il sito. Si basa su delle metriche comuni come ad esempio il numero di visitatori giornalieri, le pagine visualizzate, numero di discussioni aperte ecc..

Participation: la partecipazione è una misura che occupa una posizione importante per il nostro lavoro. La valutazione del sentiment deve tenere conto di quanto un sito sia vivo e dinamico poiché alti livelli di partecipazione possono favorire il diffondersi di opinioni positive o negative riguardanti un brand. Il numero di discussioni aperte ogni giorno, la media dei post per discussione e la rapidità di risposta degli utenti a nuovi topic sono gli indicatori di partecipation presi in considerazione.

Time: il tempo speso sul sito può essere un indice di importanza e di gradimento dei contenuti. Poiché i volumi delle visite al sito possono essere influenzate da un posizionamento favorevole nei motori di ricerca oppure dalla presenza di molti link sparsi per il web. Misurare la media del tempo speso sul sito e la percentuale di bounce rate ricopre un ruolo fondamentale per la valutazione di una fonte.

Nella valutazione del sentiment la pesatura per fonte riceve un ruolo molto importante. Ogni social media ha dei livelli di traffico, partecipazione e di tempo diversi tra di loro. Considerare in modo uguale una fonte con livelli di traffico notevolmente superiori rispetto a un'altra, introdurrebbe una distorsione nei risultati e di conseguenza un

calcolo del sentiment non veritiero. Ricordando che l'obiettivo di queste analisi ha come scopo principale supportare il marketing di un'azienda, differenziare l'opinione espressa su una fonte con dei livelli di utilizzo elevato rispetto a un'altra meno importante è fondamentale per la buona riuscita del business.

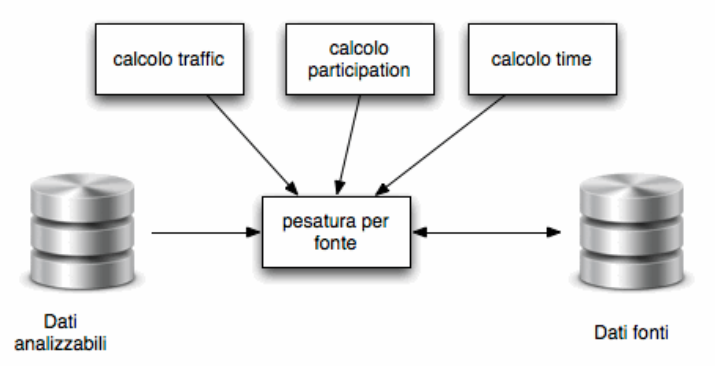

Figura 3.2: Architettura pesatura fonti

3.3.3 WordNet

WordNet [87] è un database lessicale gratuitamente disponibile per la lingua inglese, il quale organizza nomi, verbi, aggettivi e avverbi in distinte gerarchie di synonym set o synset. Ciascun synset raggruppa parole aventi uno stesso significato ed è provvisto di una glossa che descrive brevemente il concetto che rappresenta. Ad esempio, il synset composto dalle parole {apartment, flat} rappresenta il concetto definito dalla glossa "a suite of rooms usually on one floor of an apartment house". Molte delle glosse sono estese con l'aggiunta di alcuni esempi di utilizzo dei concetti che descrivono.

WordNet è organizzato come rete di concetti collegati gli uni agli altri per mezzo di relazioni semantiche, creando così delle vere e proprie gerarchie di significato [88]. Tali relazioni semantiche sono legate ad

una particolare parte del discorso, creando quindi differenti e separati sottografi per nomi, verbi, aggettivi e avverbi.

In Figura 3.3 è mostrato un frammento della rete semantica di WordNet relativo al concetto di animal. Le linee solide rappresentano relazioni di tipo is–a, le linee tratteggiate rappresentano relazioni di tipo has–a/part–of e infine linee puntate rappresentano una serie di relazioni di tipo is–a che sono state omesse dalla figura per compattezza.

WordNet è attualmente giunto alla versione 3.0, versione utilizzata in questo lavoro, la quale conta 147278 parole uniche tra nomi, verbi, aggettivi e avverbi [89]. Benché molte parole sono uniche all'interno di una stessa categoria sintattica, sono comunque presenti in più di una categoria. In Tabella 3.1 sono riportati i dati relativi al numero di parole uniche per ognuna delle categorie sintattiche.

A molte delle parole contenute in WordNet è associato più di un significato (polisemia), ovvero una medesima parola appartiene a più di un synset. In WordNet i concetti seguono un ordinamento decrescente rispetto alla loro frequenza di utilizzo basata sull'analisi di corpus taggati semanticamente, ovvero i sensi più frequenti sono indicati da un più basso numero ordinale. Nelle Tabelle 3.2 e 3.3 sono riportate informazioni dettagliate riguardo il grado di polisemia delle parole contenute in WordNet.

Negli anni, WordNet si è affermato come risorsa di riferimento per l'elaborazione del linguaggio naturale e, in modo particolare, per i compiti di Word Sense Disambiguation.

Si può notare come, sebbene la descrizione appena fornita di WordNet faccia pensare ad un suo utilizzo unicamente come fonte dati, sia rappresentato in architettura come un componente software. Il motivo è che allo stato

POS	Stringhe uniche	Synset	Totale coppie parola-senso
Nomi	117798	82115	146312
Verbi	11529	13767	25047
Aggettivi	21479	18156	30002
Avverbi	4481	3621	5580
Totale	155287	117659	206941

Tabella 3.1: Numero di parole, sanse e sensi in WordNet 3.0

POS	Parole e sensi monosemici	Parole polisemiche	Sensi polisemici
Nomi	101863	82115	146312
Verbi	6277	13767	25047
Aggettivi	16503	18156	30002
Avverbi	3748	3621	5580
Totale	128391	117659	206941

Tabella 3.2: Numero di parole monosemiche e polisemiche presenti in WordNet 3.0

POS	Polisemia media (incl. parole monosemiche)	Polisemia media (escl. parole monosemiche)
Nomi	1,24	2,79
Verbi	2,17	3,57
Aggettivi	1,40	2,71
Avverbi	1,25	2,50

Tabella 3.3: Grado di polisemia medio delle varie categorie sintattiche presenti in WordNet 3.0

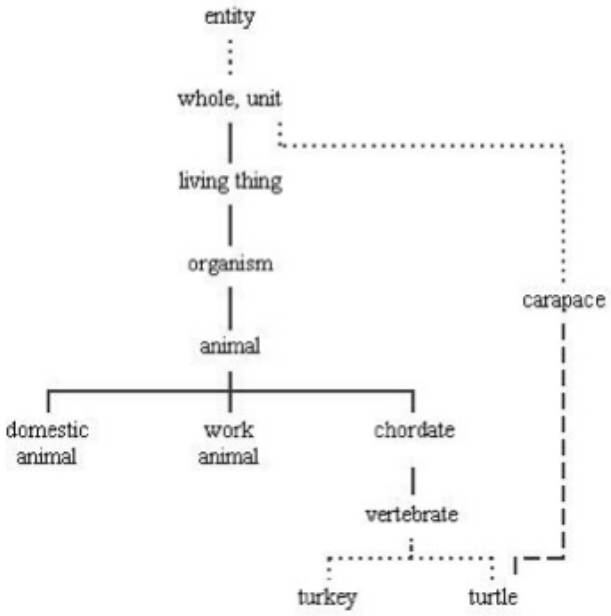

Figura 3.3: Frammento di WordNet

attuale è in fase di implementazione un modulo che permetta la personalizzazione di WordNet abilitando operazioni quali l'aggiunta e la modifica di synset e di relazioni semantiche. Lo scopo è quello di riuscire ad integrare in WordNet dati di dominio, al fine di incrementare le performance dei moduli in cui è coinvolto.

3.3.4 Data cleaning

I testi che derivano da fonti di informazione ufficiali, quali giornali online, possono essere assunti come generalmente ben scritti e semplici da interpretare. Ma se derivano da siti Web 2.0 devono essere necessariamente trattati con operazioni di data cleaning prima di poter essere interpretati in quanto nelle frasi saranno frequenti errori ortografici, abbreviazioni ad–hoc ed utilizzi impropri delle regole grammaticali. Tutto ciò è accentuato in siti di microblogging quale

Twitter a causa del numero limitato di caratteri utilizzabili per un singolo post. Il componente di cleaning è stato sviluppato principalmente per trattare con i tweet, in quanto questi sollevano il maggior numero di problemi.

Il componente svolge dapprima un'attività di entity recognition per identificare URL, indirizzi e–mail e smile al fine di rimuoverli dal testo in quanto responsabili di peggiori performance della successiva fase di parsing. La rimozione è giustificata dalla mancanza di utilità di questo genere di informazione relativamente alle attività di Sentiment Analysis. Successivamente, il testo necessita di essere correttamente formattato. Questo implica la conversione del testo in una codifica standard, la rimozione di punteggiatura multipla o la sua aggiunta dove è necessaria e l'unificazione di formati quali date o quote. Infine viene svolta la correzione ortografica delle parole. Quest'ultima è l'attività di maggiore criticità in quanto i tweet contengono molti acronimi o parole derivanti da slang. Data una parola presumibilmente errata, obiettivo della correzione ortografica è quello di scegliere in maniera automatica la correzione più plausibile. Certamente non potrà essere noto con certezza quale sia la giusta correzione per un certo errore, ad esempio, non possiamo dichiarare con certezza se la parola errata "lates" debba essere corretta in "late" o "latest" o addirittura un'altra ancora. Ciò suggerisce quindi l'uso della probabilità.

3.3.5. Stanford Parser

Il parser di Stanford [90] è l'analizzatore sintattico scelto per essere inserito nell'architettura. Un parser per il linguaggio naturale è un programma che opera a livello della struttura grammaticale della frase, ad esempio su gruppi di parole associate o su relazioni tra parole quali verbo-soggetto o verbo-complemento oggetto. In particolare, un parser di tipo probabilistico sfrutta la conoscenza del linguaggio, ottenuta

tramite l'elaborazione di un insieme di frasi parsate manualmente, in modo tale da cercare di produrre l'analisi più probabile per una nuova frase. Il parser di Stanford è un'implementazione, scritta in linguaggio Java, di un parser probabilistico per il linguaggio naturale.

La scelta del parser da includere nell'architettura è stata effettuata sulla base dei risultati ottenuti da un'analisi preliminare di differenti analizzatori sintattici. Scopo di questa analisi è stato quello di valutare la capacità dei differenti strumenti di analizzare correttamente un testo, verificando non soltanto la capacità di identificare in maniera esatta i soggetti delle frasi, ma anche la correttezza dell'albero sintattico creato e delle dipendenze identificate. In entrambi i test, il parser di Stanford ha dimostrato essere il più accurato [91].

Il risultato dell'analisi del parser di Stanford consiste sia nella struttura ad albero della frase che in una lista di dipendenze tipizzate, anche note come relazioni grammaticali. Un esempio della struttura ad albero generata dal parser è mostrato in Figura 3.4 con riferimento alla frase "The themes for the Hearings will be based on the comprehensive report of the Secretary–General, contained in document A/59/2005, and the clusters defined therein.".

La fase di integrazione tra i componenti software dell'architettura e il parser di Stanford ha richiesto la modifica di alcuni aspetti del comportamento di quest'ultimo [91]. In particolare, le modifiche necessarie sono dovute al fatto che il parser di Stanford, riconoscendo e analizzando in maniera corretta le contrazioni normalmente utilizzate nella lingua inglese ('s, 'm, . . .), mantiene anche nell'analisi fornita come output la forma contratta della parola, ovvero non effettua operazioni di normalizzazione del testo. Ciò influisce in maniera negativa sul processo di integrazione con WordNet, in quanto, nel caso questo sia interrogato con una contrazione non è in grado di risalire alla

forma non contratta della parola e, conseguentemente, non è in grado recuperare né il lemma né i synset corrispondenti.

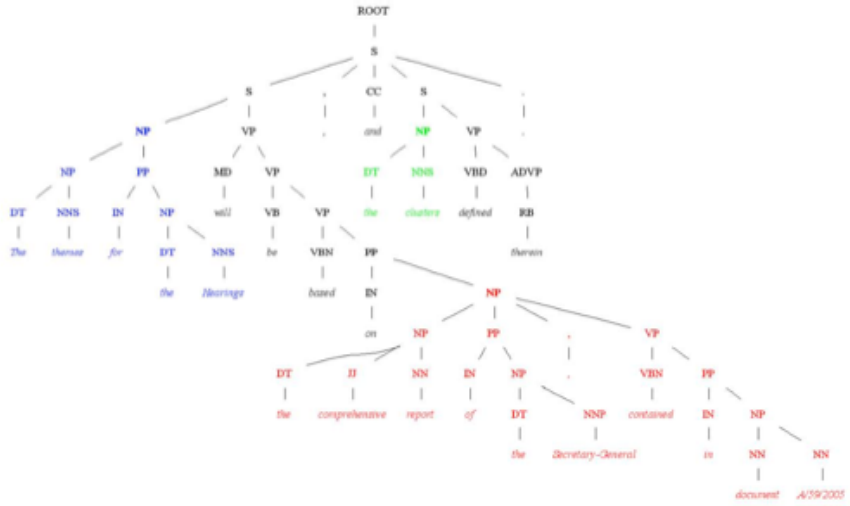

Figura 3.4: Esempio della struttura ad albero di una frase creata dal parser di Stanford

3.3.6 Data pruning

La precisione di un qualsiasi analizzatore sintattico difficilmente potrà raggiungere un valore del 100%. Una precedente fase di cleaning del testo può indubbiamente migliorare i risultati, tuttavia alcuni margini di errore continueranno a permanere. Di conseguenza un ulteriore modulo dovrebbe essere posizionato dopo l'analizzatore sintattico in modo da identificare e possibilmente eliminare dalle successive elaborazioni le frasi la cui analisi sintattica è errata. Faremo riferimento a questo tipo di operazioni con il termine "data pruning".

Per consentire una facile lettura della discussione seguente è utile descrivere brevemente le basi di conoscenza impiegate dal modulo: WordNet, FreeBase e Google Dictionary.

WordNet [87] è un database lessicale gratuitamente disponibile per la lingua inglese, già precedentemente descritto nella Sezione 3.2.3.

FreeBase è un repository aperto e sotto licenza Creative Commons di dati strutturati che contiene, al momento attuale, circa dodici milioni di entità connesse tra loro a formare un grafo. Un'entità può essere una singola persona, un luogo o una cosa. Più in generale un'entità è un concetto esistente nel mondo reale. In FreeBase ogni entità è rappresentata come un topic node nel grafo, ognuno dei quali viene identificato per mezzo di un id univoco utile per distinguere tra multiple entità con uno stesso nome. Ad esempio, "Henry Ford" l'industriale e "Henry Ford" il campione di football sono indubbiamente due entità distinte aventi però lo stesso nome. I topic sono associati con almeno un tipo e possono avere ulteriori proprietà come ad esempio latitudine e longitudine per un luogo. I tipi sono una sorta di categorie in cui i topic sono raggruppati come persone, luoghi, libri o film. FreeBase è inoltre una community di migliaia di persone che contribuiscono a migliorare e incrementare i dati esistenti.

Google Dictionary è un servizio gratuito introdotto da Google nel dicembre 2009. Supportando oltre 25 differenti lingue, permette di ricercare facilmente la definizione di una specifica parola oltre che garantire l'accesso ad un insieme di informazioni addizionali quali sinonimi, espressioni correlate o definizioni alternative reperite in siti ritenuti affidabili sul web. In aggiunta fornisce un servizio di traduzione automatica di parole.

In [91] sono stati eseguiti esperimenti su un campione di 86 frasi selezionate dalle tre fonti di cui abbiamo implementato i crawler: 56 frasi sono state estratte da Twitter, 15 da Lonely Planet e 15 da TripAdvisor. La scelta delle frasi è stata fatta in modo da selezionare non soltanto esempi sintatticamente corretti ma anche frasi con alcuni errori grammaticali e strane combinazioni di simboli. La dimensione del campione è limitata dall'elevata quantità di tempo necessaria a valutare manualmente il dettagliato risultato dell'analisi sintattica effettuato dal

parser di Stanford per ciascuna frase. I risultati ottenuti sono riportati in Tabella 3.4. Si può notare facilmente come i risultati ottenuti siano promettenti. Infatti il componente ha dimostrato di eseguire la corretta azione su 75 frasi (87%) del campione totale, di cui 60 sono state tenute per le successive analisi ed effettivamente corrette e 15 sono state scartate perché realmente errate da un punto di vista della struttura grammaticale. L'errore commesso sulle rimanenti 11 frasi (13%) è generalmente dovuto a inesattezze nella struttura dell'albero o nelle dipendenze di Stanford e non a errori di catalogazione delle singole parole.

	Struttura frase corretta	**Struttura frase errata**
Frase tenuta	60	10
Frase scartata	1	15

Tabella 3.4: Performance del componente di Data Pruning

3.3.7 Word sense disambiguation

L'espressione Word Sense Disambiguation (WSD) si riferisce alla capacità di identificare il senso delle parole in una data frase. Rappresenta un argomento di ricerca fondamentale nel campo dell'elaborazione del linguaggio naturale con molte applicazioni pratiche, quali information retrieval, motori di ricerca e sentiment analysis.

Nello strumento di Sentiment Analysis proposto, il componente di WSD ha il ruolo di disambiguare i nomi presenti nelle frasi che sono stati identificati nella precedente fase di analisi sintattica al fine di rimuovere dall'analisi quelle frasi che trattano argomenti non di interesse per il dominio in esame. In relazione all'attuale caso di studio sul turismo è necessario, ad esempio, in frasi che contengono la parola "Milan", riuscire a disambiguare se questa vuole significare la città di Milano oppure la sua squadra di calcio. Il problema permane anche nel caso in

cui la parola sia scritta in lingua italiana, “Milano”, poiché è comunque necessario riuscire a comprendere se l’autore si riferisce alla città, all’attrice Alyssa Milano oppure ad una famosa marca di biscotti americana, Milano Cookies. È evidente come il processo di WSD sia di cruciale importanza sia per la precisione che per la recall dell’intero strumento di Sentiment Analysis.

3.3.8 Categorizzazione

Lo scopo della categorizzazione del testo è la classificazione dei documenti, nel caso in esame post, in un numero fissato e predefinito di categorie. In generale ogni documento può appartenere ad una categoria, a più categorie o a nessuna categoria. Il problema della categorizzazione viene generalmente visto come un caso particolare del problema più generale dell’apprendimento, ovvero sull’idea che l’esperienza possa migliorare la capacità di un agente automatico di agire in futuro. In letteratura, il classificatore Naïve Bayes è uno dei metodi più usati per la classificazione del testo [93]. Tuttavia non è l’unico approccio, infatti, recentemente è emerso un differente orientamento che fa riferimento alle Support Vector Machine (SVM) [94], oppure è stata impiegata una tecnica più vecchia chiamata TF–IDF [95]. Una grossa barriera all’adozione di questi algoritmi è quella di richiedere un elevato numero, spesso proibitivo, di dati etichettati per poter essere addestrati ed ottenere risultati accurati [96]. L’etichettatura dei dati è un compito che deve essere tipicamente eseguito da un essere umano, risultando quindi un processo lungo e costoso. Il bisogno di grandi quantità di dati per ottenere una buona accuratezza e la difficoltà di ottenere dei dati etichettati ha portato, durante lo sviluppo del componente da inserire nell’architettura dello strumento di Sentiment Analysis, alla ricerca di un metodo alternativo che non richiedesse dati già etichettati.

Da quanto illustrato emerge la necessità di dover definire una tassonomia o un modello che rappresenti al meglio gli aspetti della realtà che si intende analizzare. Tale modello risulta necessariamente essere dipendente dal contesto: un modello sviluppato per analizzare il turismo è diverso da uno definito per l'industria automobilistica. Sebbene la discussione sulla metodologia di definizione di un tale modello esuli dagli obiettivi di questo lavoro di tesi, alcune categorie (o tag) molto generali per l'ambito del turismo possono essere Services and Transport, Arts and Culture, Food and Drink, Fares and Ticket, Events and Sport. Tali categorie possono essere ulteriormente dettagliate per mezzo di relazioni di tipo is–a a formare una struttura non necessariamente gerarchica, quale ad esempio una struttura a matrice.

Nell'approccio per la categorizzazione sviluppato per l'utilizzo nella nostra architettura, l'attività manuale, sebbene molto ridotta rispetto alla creazione di dati etichettati per l'addestramento, non viene tuttavia completamente eliminata.

La discussione di queste tematiche esula da questo lavoro di tesi e per ulteriori approfondimenti e dettagli si rimanda ad altri elaborati.

3.3.9 Estrazione degli snippet

Prima dell'avvento del Web 2.0 vi era l'assunzione che ogni frase esprimesse al più una e una sola opinione su uno e un solo oggetto. Con la diffusione di blog e forum questa assunzione è venuta meno, o almeno non risulta più applicabile in contesti di social network, in quanto in un medesimo post un utente può esprimere più opinioni su più oggetti. Sebbene l'opinione complessiva espressa in un commento sia indubbiamente utile, questa è solo una parte dell'informazione di interesse. La valutazione del sentiment a livello di frase, o addirittura di documento, non è in grado di rilevare l'opinione espressa a proposito di

specifici aspetti del commento. Ad esempio, sebbene un individuo possa essere complessivamente soddisfatto del suo telefono cellulare, potrebbe essere insoddisfatto della durata della sua batteria. Al produttore, conoscere queste debolezze o punti di forza individuali è ugualmente importante, se non addirittura di maggior valore, del livello di soddisfazione generale dei clienti.

Alla luce di quanto appena discusso, è evidente l'importanza di introdurre in architettura un componente il cui fine sia quello di estrarre da un post tutte le opinioni, relative a differenti soggetti, che lo compongono, ovvero di estrarre i cosiddetti snippet. Uno snippet è una piccola porzione di testo centrata su un aspetto di interesse, nel caso specifico, sulle opinioni. Il processo di snippettizzazione è di cruciale importanza, in quanto permette di focalizzare l'analisi solo sulle parti rilevanti del testo.

Allo stato attuale, è stata implementata una prima versione di test del componente basata sull'estrazione di coppie verbo-complemento oggetto dal testo, che tuttavia non fornisce ancora risultati soddisfacenti [91]. Sono necessari ulteriori studi per poter utilizzare con successo il componente nel processo di analisi.

3.3.10 Valutazione sentiment

Il modulo di valutazione del sentiment, denominato SentiEngine e sviluppato in un differente lavoro di tesi [97], ha l'obiettivo di valutare il grado di positività o negatività di un testo. Rispetto all'attuale letteratura e agli strumenti attualmente sul mercato, l'aspetto innovativo di questo modulo risiede nelle diverse metriche implementate per ricavare una polarità prima a livello di frase e poi a livello di insieme di frasi analizzate, per ottenere un valore di sentiment univoco. Inoltre, è stato osservato come le fonti possano contenere disturbi o distorsioni a livello

di sito o di singolo post ed è stato studiato un metodo per bilanciare questo fattore per ottenere una valutazione oggettiva del sentiment.

SentiEngine impiega SentiWordNet come risorsa lessicale per la valutazione della polarità delle singole parole. SentiWordNet (SWN) [98] (attualmente alla versione 3.0) è un'estensione di WordNet 3.0 ai cui synset sono stati associati tre valori numerici a indicare quanto i termini compresi in un dato synset siano oggettivi, positivi o negativi. Ognuno dei tre punteggi varia tra zero e uno, inoltre, la loro somma è unitaria. SentiWordNet è costruito in maniera automatica mediante l'utilizzo di otto classificatori ternari caratterizzati da un livello di accuratezza simile ma da un differente criterio di classificazione. Indicando con P(s) la positività, N(s) la negatività e O(s) l'oggettività, ad esempio, in riferimento all'aggettivo "estimable" compreso in tre synset, SentiWordNet fornisce le seguenti informazioni: al synset {estimable#1} definito da "deserving of respect or high regard" è associato P(s) = 0.75, N(s) = 0.00 e O(s) = 0.25; al synset {respectable#2, honorable#4, good#4, estimable#2} definito da "deserving of esteem and respect" è associato P(s) = 0.75, N(s) = 0.00 e O(s) = 0.25; al synset {estimable#3, computable#1} definito da "may be computed or estimated" è associato P(s) = 0.00, N(s) = 0.00 e O(s) = 1.00.

L'operazione di valutazione del sentiment avviene dapprima estraendo da SentiWordNet i synset dei vari lemmi identificati dalla precedente analisi sintattica del testo. Successivamente viene valutata la polarità a livello di parola, aggregata infine a livello di frase. Le differenti operazioni sono intramezzate dall'applicazione di filtri atti a perfezionare l'analisi.

Il calcolo della polarità di una singola parola viene effettuato mediante il calcolo della media aritmetica dei valori di polarità di tutti i synset corrispondenti al lemma della parola restituiti da SentiWordNet. Ovvero,

utilizzando la notazione introdotta precedentemente e indicando con n il numero totale di synset per quel lemma,

$$pol_{+}(lemma) = \sum_{k=1}^{n} \frac{P(syn_k)}{n}$$

$$pol_{-}(lemma) = \sum_{k=1}^{n} \frac{N(syn_k)}{n}$$

$$pol_{o}(lemma) = \sum_{k=1}^{n} \frac{O(syn_k)}{n}$$

Facendo ancora riferimento all'aggettivo "estimable" come esempio, si ottiene pol_{+}(estimable) = 0.50, pol_{-}(estimable) = 0.00 e pol_{o}(estimable) = 0.50.

Il calcolo della polarità a livello di frase viene a sua volta ottenuto come media aritmetica dei valori di polarità delle singola parole, escludendo dal calcolo quelle con polarità nulla.

$$pol_{+}(frase) = \sum_{w \in W} \frac{pol_{+}(w)}{size(w)}$$

$$W = \{lemma | pol_{+}(lemma) \neq 0 \vee pol_{-}(lemma) \neq 0\}$$

Le formule per il calcolo della polarità negativa e dell'oggettività sono analoghe. La polarità risultante viene infine discretizzata su una scala a cinque livelli.

3.3.11 Analisi e pesatura dei dati

Una volta che i dati estratti dai crawler sono stati elaborati dai moduli precedentemente discussi fino ad ottenere dei valori di sentiment opportunamente categorizzati, si ha la necessità di analizzarli per dare l'opportunità ai decision maker di elaborare delle strategie di marketing

che siano più efficaci possibili. L'applicativo sviluppato per le analisi è l'obiettivo di questa tesi e si compone di diverse parti rappresentate in Figura 3.5.

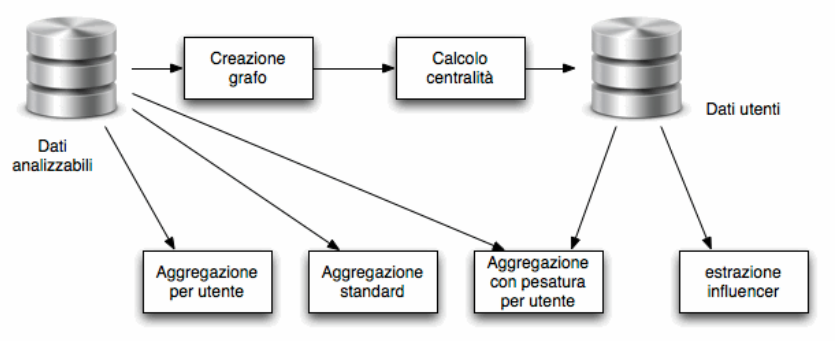

Figura 3.5: Architettura pesatura utente e aggregazione

Le opinioni, come accade in qualsiasi società, non sono tutte uguali. Esistono delle persone che, grazie a particolari facoltà, riescono a occupare posizioni importanti e di prestigio che gli consentono di influenzare il parere comune. Uno dei compiti più importanti del modulo di analisi è di identificare questi utenti, ricostruendo i grafici delle interazioni sociali utilizzando la teoria ampiamente discussa nel Capitolo 2. Analizzando la ricostruzione grafica delle community virtuali si ha la possibilità di calcolare metriche di importanza, quali:

Indegree: numero di archi che puntano verso un utente;

Proximity: misura del grado di prossimità di tutti gli utenti della community rispetto all'utente analizzato;

Outdegree: numero di archi che partono da un utente;

Eccentricity: la distanza dell'utente più lontano della rete;

Closeness: molto simile alla proximity da una misura qualitativa della vicinanza di tutti gli altri utenti della rete;

Betweenness: misura la posizione dell'utente e in particolare quanto si trova sui percorsi di comunicazione tra gli altri utenti.

Tali metriche, che verranno illustrate in detteglio nel Capitolo 4, ci consentono di determinare quali siano gli influencer e sviluppare delle analisi su di essi.

Infine il sentiment viene calcolato in tre modi:

Aggregazione standard: è la media di tutti i voti su un particolare prodotto o brand.

Aggregazione per utente: valuta il sentiment degli individui indipendentemente dal volume di post prodotto;

Aggregazione con pesatura utente: utilizzando le metriche di centralità elencate si dà un maggiore peso all'opinione degli utenti più importanti.

Nei tool sul mercato il calcolo del sentiment viene fatto utilizzando l'aggregazione standard mentre sono inesistenti le altre due tipologie. Per questo un'analisi di questo tipo può fornire informazioni preziose sul sentiment e sul suo sviluppo futuro.

3.3.12 User interface

Sono state implementate due differenti tipologie di interfacce grafiche per mezzo delle quali gli utenti possono fruire dei risultati delle analisi: standard e mashup.

L'interfaccia standard fornisce un insieme di strumenti di analisi predefiniti, tipicamente grafici, che mostrano i risultati attraverso differenti modalità di visualizzazione interattive. Ogni tipologia di visualizzazione mostra i dati in maniera incrementale al fine di permettere all'utente una semplice interpretazione dei risultati senza

essere disturbato da un eccessivo carico di informazioni. Ad esempio, un grafico a torta di alto livello può essere facilmente espanso in grafici più dettagliati. Le analisi possono essere effettuate sulla base delle fonti da cui provengono i dati, su particolari aspetti di interesse del modello di analisi e su base temporale.

L'interfaccia mashup [100] permette agli utenti di crearsi un ambiente di analisi totalmente personalizzato permettendo l'aggiunta di nuovi servizi di analisi in maniera semplice ed immediata senza che siano necessarie conoscenze tecniche sull'applicazione. Oltre ai servizi di base resi disponibili anche attraverso l'interfaccia standard, sono state importate diverse API pubbliche (quali Google Charts, Google Maps, Flickr, Tag Clouds) a supporto delle analisi. La Figura 3.6 mostra un esempio di utilizzo dell'interfaccia mashup.

Figura 3.6: Esempio di composizione dell'interfaccia mashup

Capitolo 4

Metriche di valutazione della qualità delle fonti informative Web

4.1 Introduzione

Dopo aver presentato tutta la teoria riguardo alle misure di centralità nelle reti sociali e l'importanza delle sorgenti informative (Capitolo 2), in questo capitolo vengono discusse tutte le metodologie scelte e le analisi che si è deciso di implementare per inserirle poi nell'architettura presentata nel Capitolo 3. Si inizierà con una presentazione generale dei tipi di metriche considerate (Sezione 4.2), si passerà alla presentazione dei tipi di aggregazione sulle singole fonti (Sezione 4.3), per poi passare alla descrizione dei metodi per l'aggregazione di più fonti (Sezione 4.4). Infine il capitolo si chiude con degli esempi di analisi (Sezione 4.5) che verranno poi svolte nel Capitolo 5.

4.2 Tassonomia delle metriche

La necessità di monitorare i social media sta diventando un argomento di notevole attualità e interesse. Sia le piccole che le grandi imprese hanno a disposizione delle sorgenti informative mai avute fino ad ora. Twitter ha espanso i suoi confini fino al Giappone, Indonesia, Messico ed altre regione, diventando una rete sociale in grado di raggiungere i

suoi utenti in ogni angolo della terra. Questa rapida espansione ha sollevato la necessità di migliorare le capacità delle aziende di monitorare ed analizzare queste sorgenti informative al fine di costruire o mantenere la propria reputazione. Se da un lato è facile intravedere l'utilità di queste analisi, dall'altro l'applicazione e lo sviluppo di tool in grado di farlo è tutt'altro che banale a causa della forte eterogeneità delle fonti. Ogni social media ha caratteristiche proprie che lo differenzia da tutti gli altri, quali la struttura della rete sociale, le tipologie di interazione tra gli utenti oltre che agli argomenti trattati al suo interno. Inoltre l'informazione estratta da queste sorgenti può essere analizzata in varie dimensioni diverse ed ognuna di esse fornisce indicazioni utili a seconda del contesto operativo. Perché una misura sia effettiva, deve essere allineata direttamente con l'obiettivo che si vuole raggiungere. Questa deve essere specifica, misurabile, realistica e calcolabile in un tempo accettabile. Una misurazione effettiva raramente risiede nei confini di una singola metrica ma piuttosto in una loro combinazione che aiuti a illustrare gli andamenti e l'evoluzione verso un obiettivo finale. Analizzando i social media e le loro caratteristiche si è cercato di trovare delle metriche che aiutassero a capire l'opinione degli utenti, i motivi e le modalità della propagazione dell'informazione e i meccanismi con cui si forma ed evolve la reputazione all'interno di queste entità. Per svolgere un'analisi approfondita che pur entrando il più possibile nel dettaglio di ogni social media, mantenga un alto grado di flessibilità si sono trovate due tipi di aggregazioni:

- Metriche per l'aggregazione del sentiment sulle singole fonti;
- Metriche per aggregare più fonti eterogenee.

Figura 4.1: The modern sales funnel

Di seguito vengono descritte nel dettaglio le metriche di analisi che vengono applicate alle singole fonti (sezione 4.2.1) per proseguire con la descrizione di quelle che, prendendo in considerazione le caratteristiche di ogni social media, ci consentiranno di effettuare un'analisi su più fonti eterogenee (sezione 4.2.2).

4.2.1 Metriche per l'aggregazione del sentiment sulle singole fonti

Esistono diversi modi per aggregare il sentiment fornito dal sentiEngine. In Figura 4.2 vengono riportati alcuni metodi che verranno analizzati nel proseguo del capitolo.

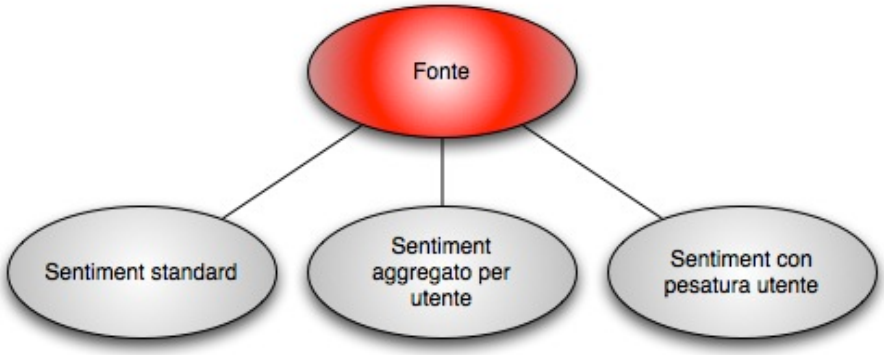

Figura 4.2: Schema aggregazioni

Partendo da sinistra verso destra, le varie aggregazioni aumentano di complessità, introducendo un'analisi sempre più potente ed efficace che prende in considerazione le caratteristiche strutturali della fonte e dell'interazione tra gli utenti che ne fanno parte.

4.2.1.1 Sentiment Standard

L'aggregazione standard è la più semplice fra le tre aggregazioni; consiste nel calcolo della media del sentiment di tutti i post diviso per il numero di post.

$$SS = \frac{\sum_{i=1}^{n} x_i}{n}$$

dove:

x_i – valore del sentiment per il post i-esimo;

n – numero totale di post su un dato argomento.

4.2.1.2 Sentiment aggregato per utente

Lo scopo dell'aggregazione per utente è quello di democratizzare il voto delle persone nel calcolo del sentiment. L'obiettivo che ci si pone è

quello di evitare che l'opinione di un ristretto gruppo di persone abbia un peso maggiore rispetto alle altre sulla base del volume di post prodotto. L'aggregazione viene eseguita in due passi:

- prima si calcola il sentiment per ogni persona (SU) sulla base dei post che l'utente genera;

$$SU_j = \frac{\sum_{i \in P_j} x_i}{|P_j|}$$

indicando con:

x_i – il valore del sentiment per il post i-esimo;

P_j – l'insieme dei post scritti dall'utente j-esimo su un dato argomento;

$|P_j|$ – la cardinalità dell'insieme P.

- successivamente si aggrega il sentiment facendone la media rispetto al numero di utenti che partecipa alle conversazioni sul tema analizzato. Il sentiment con aggregazione per utente (SAU) è infine calcolato con la formula:

$$SAU = \frac{\sum_{j=1}^{n} SU_j}{n}$$

dove:

SU_j – il sentiment calcolato per l'utente j-esimo;

n – numero di utenti che discute del tema trattato.

L'esempio di Figura 4.3 utilizza un campione di cinque utenti che producono un diverso volume di post. Per ognuno di essi si calcola SU ottenendo i valori SU_1=4, SU_2=2, SU_3=3, SU_4=2, SU_5=2. Una volta in

possesso di tali valori si procede con il calcolo del sentiment aggregato per utente (SUA) che in Figura 4.3 ha un valore di 2,6.

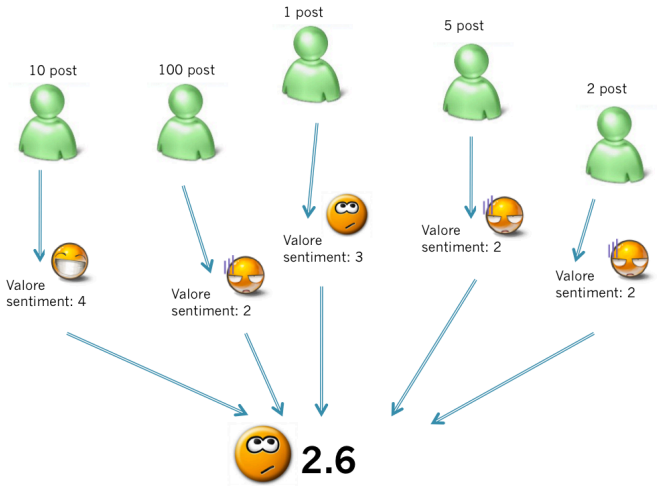

Figura 4.3: Esempio aggregazione per utente

Uno studio di questo genere ha lo scopo di risolvere due problemi principali non risolvibili con l'aggregazione standard:

- Poche persone che producono molti post possono influenzare l'analisi;
- Si possono registrare variazioni dei risultati di analisi non dovuti ad un cambio di opinione da parte degli utenti ma a una riduzione dei volumi dei post. Per esemplificare questo concetto ci riportiamo alla Figura 4.3. Se l'utente che scrive 100 post, nel mese successivo diminuisse il volume a soli 10 post, e tutti gli altri dati rimanessero invariati, l'aggregazione standard registrerebbe un aumento della reputation anche se l'opinione degli utenti non dovesse cambiare (1 utente con sentiment positivo, 3 con sentiment negativo e 1 neutro).

Riferendoci al primo punto, supponiamo di avere cinque persone interessate a un prodotto. Come si può vedere in Figura 4.3, gli utenti

non parlano con la stessa frequenza; esiste una notevole differenza nel numero dei post prodotti da ogni singola persona. Se l'azienda produttrice non volesse sapere quanto si parla bene o male su internet del propri prodotti ma fosse interessata a rilevare la soddisfazione media dei propri clienti, o potenziali clienti, il metodo standard di aggregazione non sarebbe più adatto poiché l'opinione dell'utente che produce cento post avrebbe un peso che è cento volte tanto l'opinione del cliente che produce un solo post. Utilizzando l'aggregazione a due livelli che è presentata in questa sezione, si calcola il sentiment per utente in modo che ogni persona fisica abbia lo stesso peso indipendentemente dal volume di post che esso produce. Riportiamo due situazioni molto diverse per fare ulteriore chiarezza:

- 100 utenti producono un post. Supponiamo di avere 50 post con sentiment positivo e gli altri 50 negativi;
- 50 utenti producono un post e uno ne genera 50. L'utente che produce 50 post ha un'opinione negativa e gli altri 50 al contrario, hanno un opinione positiva.

Usando l'aggregazione standard in entrambi i casi si avrebbe lo stesso scenario: 50 post positivi e 50 negativi. "Ma siamo sicuri di essere di fronte alla stessa situazione?" Se ci mettiamo nei panni dell'azienda che studia questi dati ci si può rendere conto che la situazione è profondamente diversa. Nel primo caso si deve effettuare un intervento di custumer care su un insieme di 50 persone, mentre nel secondo caso solo un utente ha un'opinione negativa e di conseguenza anche le spese di gestione della clientela saranno diverse ed in particolari minori rispetto al caso precedente.

La diversa partecipazione a discussioni e quindi il diverso volume di post prodotto da ogni utente, potrebbe alterare l'analisi della reputazione misurata su un periodo più o meno lungo. Se si monitora l'andamento

del sentiment su un social network si potrebbe registrare un'alterazione sostanziale del sentiment. Ora è lecito chiedersi se tale variazione sia dovuta a un diverso volume dei post oppure a un cambio di opinione del prodotto o servizio offerto da un'azienda. Nell'esempio riportato in Figura 4.3 se l'utente che ha pubblicato 100 post era particolarmente attivo in quel periodo a seguito di un acquisto più o meno fortunato di un prodotto, è probabile che in futuro non avrà la stessa partecipazione. Con il metodo standard, a seguito di un calo di volume da parte di un utente particolarmente attivo si ottiene un cambio del sentiment complessivo alterando significativamente i risultati. Da questi esempi si evince l'utilità di avere entrambi i tipi di aggregazione in modo da fare chiarezza sulle cause di determinati andamenti del sentiment.

4.2.1.3 Sentiment con pesatura utente

Una delle questioni più cruciali nell'analisi dei social network è quello di estrarre i membri più importanti di una comunità. In questa circostanza valutare l'importanza di un utente significa valutarne la posizione all'interno del network rispetto agli altri membri. Diversamente dai metodi precedenti, la pesatura dell'utente complica un po' le cose. L'analisi delle reti offre molte metriche di centralità che sono utilizzate con successo nel processo di valutazione dei profili di utenti all'interno dei social network [105]. In questa trattazione sono considerate solo le metriche che possono essere usate con l'ausilio di un grafo che ricostruisca la rete, dove i nodi rappresentano gli utenti e gli archi le loro interazioni. Le relazioni tra gli utenti danno due tipi di posizioni:

- Posizione di Prestigio;
- Posizione di Centralità.

Prestigio

Un utente può essere considerato di prestigio se, osservando il grafo, sono presenti numerosi legami che partono da altri utenti e si dirigono verso il soggetto analizzato. Esistono varie metriche per il calcolo del prestigio di cui ne riportiamo due:

- Indegree centrality;
- Proximity prestige;

nel seguito verranno esaminate nel dettaglio una alla volta.

Indegree centrality

Indegree centrality, chiamato anche degree prestige, è basato sul numero di collegamenti verso un utente. Preso un grafo della rete, si tengono in considerazione i membri più adiacenti a una persona: colegati con un arco diretto verso di essa [105]. In altre parole un utente sarà più interessante se riceve più "nomination" da altri membri della community. L'indegree centrality si calcola con la seguente formula:

$$IDC(x) = \frac{i(x)}{m-1}$$

dove:

i(x) – è il numero di persone adiacenti a x nella community considerando solo il primo livello di prossimità;

m – il numero totale di utenti del social network.

Come è facilmente osservabile questa misura è di tipo locale: non va in profondità nella struttura della rete ma si limita a considerarne il primo livello.

Proximity Prestige

Proximity prestige PP(x) misura la vicinanza di tutti gli altri utenti all'interno della community rispetto al membro x. Questa metrica si basa sulla distanza geodesica, indicata con d(x, y_i), la quale indica la distanza di tutti gli utenti y_i dall'utente x [105]. La formula per calcolare la Proximity Prestige è:

$$PP(x) = \frac{p(x)^2}{(m-1)\sum_{i=1}^{p(x)} d(x, y_i)}$$

dove:

p(x) – numero di tutti gli utenti y_i nella rete che possono raggiungere x (esiste un percorso da y_i verso un dato utente x);

m – numero di nodi della rete.

Centralità

Le metriche di centralità ci permettono di trovare gli utenti che instaurano una grande quantità di relazioni con gli altri iscritti alla rete. Per misurare la centralità di un utente è sufficiente costruire un grafo delle relazioni vhe non deve essere necessariamente diretto. Le metriche che considereremo per misurare la centralità sono:

- Outdegree centrality;
- Eccentricity centrality;
- Closeness centrality;
- Betweenness centrality.

Come abbiamo già fatto con le metriche di prestigio, procediamo con la spiegazione nel dettaglio di ogni metrica.

Outdegree centrality

Misura il numero di archi che vanno dal nodo x verso gli altri nodi [105]. La formula per calcolare l'Outdegree centrality (ODC) è:

$$ODC(x) = \frac{O(x)}{m-1}$$

dove:

o(x) – il numero degli utenti direttamente connessi a x considerando la struttura della rete in modo locale;

m – numero totale di utenti iscritti al social network.

Gli utenti che comunicano con il più grande numero di persone ottengono un più elevato valore di questa metrica. ODC e IDC sono le metriche per l'analisi dei social network più semplici e più intuitive che si possano calcolare.

Eccentricity centrality

Utilizzando l'Eccentricity centrality (EC) si cerca di trovare il nodo più centrale della rete. L'utente che minimizza la massima distanza rispetto a ogni altro nodo dalla rete ottiene il valore più alto di EC [105]. La formula per calcolare questa metrica è:

$$EC(x) = \frac{p(x)}{\max\{d(x,y) : y \in M\} * (m-1)}$$

dove:

p(x) - numero di tutti gli utenti y_i nella rete che possono raggiungere x

d(x,y) – è la lunghezza del percorso più breve che va dal nodo x al nodo y;

M – l'insieme di tutti gli utenti appartenenti al social network;

m – numero dei nodi della rete.

Closeness centrality

La closeness centrality (CC), in contrasto con la proximity prestige, esprime la vicinanza di un utente rispetto a tutti gli altri facenti parte della rete. L'idea che sta alla base, è che un utente occupa una posizione centrale se può raggiungere in modo veloce gli altri utenti della rete. Questa metrica invece di misurare la quantità di relazioni, misura la qualità della posizione all'interno della community. Un utente con alti valori di CC è un buon propagatore di idee ed informazioni [105]. Anche questa misura dipende dalla ormai ben nota distanza geodesica ed è calcolata nel modo seguente:

$$CC = \frac{\sum_{y \neq x, y \in M} c(x,y)}{m-1}$$

dove:

c(x,y) – è una funzione che descrive la distanza tra i nodi x e y (es. max, min, mean, median);

M – l'insieme di nodi della rete;

m – numero dei nodi della rete.

Betweenness centrality

Questa metrica misura la centralità di un utente, non sulla base di particolare interazione, ma sulla base di una particolare strutturazione della rete. Gli utenti con un elevato valore di Betweenness centrality (BC) sono molto importanti per la propagazione dell'informazione all'interno della rete. I componenti della rete per riuscire a condividere le informazioni molto probabilmente devono passare per questi nodi. BC è calcolata dividendo il numero di percorsi più brevi che vanno da y a z

rispetto al numero di quelli che passano attraverso x [105]. La formula per calcolare la BC è:

$$BC = \frac{\sum_{i \neq x \neq j; i,j \in M} b_{ij}(x)}{b_{ij}}$$

dove:

$b_{ij}(x)$ – il numero dei percorsi più corti che vanno da i a j e passanti per x;

b_{ij} – numero dei percorsi più brevi che vanno da i a j (tutti i percorsi minimi all'interno della rete);

M – Insieme di nodi della rete. Se un utente ottiene un alto valore di BC, allora significa che senza tale utente il social media sarebbe diviso in sottoreti.

4.2.2 Metriche per aggregare più fonti

In questa sezione discuteremo le varie metriche utilizzate per confrontare i vari social media presenti su internet. Queste sono importanti al fine di dare un risultato aggregato sull'evoluzione del sentiment proveniente dalla totalità delle sorgenti analizzate. Le varie fonti si differenziano in molte caratteristiche tra cui il numero di potenziali lettori, i temi trattati all'interno della piattaforma e molte altre metriche che descriveremo con maggiore dettaglio in seguito. Dividiamo le metriche trattate in due categorie:

- Metriche che misurano la reputazione di una fonte;
- Metriche che misurano la specializzazione di una fonte intesa come il grado con cui la sorgente si focalizza su un argomento.

4.2.2.1 Metriche che misurano la reputazione di una fonte

Le metriche per il calcolo della reputazione di una fonte sono tratte dall'articolo [82]. Il concetto di reputation ed il suo calcolo operativo derivano direttamente dalla letteratura sul data quality. In particolare, la classificazione delle dimensioni a essa relativa sono fornite da [80]. L'articolo spiega come accuracy, completeness e time rappresentano le dimensioni fondamentali di data quality in molti contesti. Interpretability, authority e dependability rappresentano dimensioni aggiuntive che dovrebbero essere considerate quando si valuta la reputation, specialmente per fonti informative semi-strutturate o non strutturate. Gli esperimenti per valutare queste metriche sono stati eseguiti su blog e forum ma la teoria può facilmente essere applicata su qualsiasi tipo di social media. Dopo una fase di valutazione e validazione delle varie metriche trattate nell'articolo [82], esso prosegue con una loro riduzione in modo tale da semplificare il modello di valutazione. Le metriche a cui siamo interessati sono quelle che risultano da questa fase si scrematura. Esse possono essere ricondotte a tre costrutti:

- **Traffic**: raggruppa tutte quelle metriche che, direttamente o indirettamente, hanno a che fare con il traffico generato sul sito e ne qualificano in un certo senso l'autorità;
- **Participation**: si misura il contributo degli utenti che scrivono messaggi, replicano a discussioni e mantengono aggiornate le informazioni;
- **Time**: è un indice del livello di coinvolgimento e di interesse degli utenti. Contiene tutte quelle misure che considerano il tempo speso sul sito.

Di seguito, per ognuno dei costrutti precedenti, si indicano le metriche contenute seguite da una breve spiegazione riguardo a cosa misurano e la fonte da cui sono ricavate.

Traffic

Come già riportato in precedenza, queste metriche misurano l'autorità di un sito sulla base del traffico prodotto all'interno delle sue pagine. Le metriche che sono contenute in questo costrutto sono:

- **Traffic rank**: è un indice ottenuto dal sito (www.alexa.com), specializzato nell'effettuare analisi di traffico dei principali siti presenti su internet. Questo indice viene aggiornato con cadenza giornaliera e misura il grado di popolarità della fonte selezionata. Il rank è calcolato utilizzando una combinazione tra la media dei visitatori che ogni giorno visitano il sito e le pagine visitate negli ultimi tre mesi. Il sito con la più alta combinazione dei due valori ottiene il rank #1;

- **Daily visitors**: Misura il numero di utenti che visitano un sito (inteso come la totalità delle sue pagine) nell'arco di un giorno;

- **Daily page views**: questa metrica conta il numero di pagine, appartenenti a un dato sito, visitate ogni giorno;

- **Number of inbound links**: conta il numero totale di link provenienti da altre fonti. Questo tipo di metrica è ampiamente utilizzato dai motori di ricerca per determinare l'autorità di un sito. Ogni link proveniente da altre fonti, diretto verso il sito di interesse, è visto come una sorta di voto sulla sua popolarità. È da sottolineare come link multipli provenienti dalla stessa sorgente siano contati una sola volta;

- **Number of open discussions compared to largest web blog/forum**: questa metrica non è fornita da un sito specializzato ma è necessario effettuare un crawling delle pagine per calcolare il numero di discussioni aperte in un certo intervallo di tempo. Per confrontare i dati ottenuti dal sito di cui si sta svolgendo l'analisi con i più importanti blog/forum si utilizzano dei benchmarks ottenuti da www.technorati.com.

Se le prime quattro metriche sono fornite da Alexa e non richiedono molti calcoli, l'ultima necessita di una fase di crawling. Il costrutto traffic oltre a fornire una dimensione di valutazione delle fonti, può fornire informazioni interessanti sui potenziali lettori di un sito. Questi sono un elemento fondamentale per il calcolo della reputation. Un'opinione espressa su un sito con un alto valore di traffico raggiungerà un maggiore numero di persone rispetto a quello che si avrebbe su un sito che mostra livelli inferiori di tale misura. Si ricorda che si sta parlando di potenziali lettori poiché il traffico è l'unica misura attualmente disponibile e che più si avvicina ad una valutazione di questo genere. Purtroppo le fonti a disposizione non danno i livelli di visualizzazione di un post il che rende impossibile un calcolo dettagliato del volume di utenti che legge un'opinione espressa.

Participation

Queste metriche valutano il coinvolgimento degli utenti nel creare contenuti. Sono state selezionate tre metriche considerate più rilevanti:

- **Number of new discussion opened per day**: indica l'attitudine di una comunità a reagire ad eventi esterni come può essere una manifestazione, un dibattito politico o un evento sportivo.
- **Number of comments per discussion**: A differenza della precedente metrica che è fornita dal sito Alexa, per calcolare il

numero di commenti per discussione è necessario effettuare un crawling della fonte. La metrica misura il coinvolgimento verso discussioni aperte sul sito ed è importante per valutare la partecipazione degli utenti, e quanto il dibattito sia vivo;

- **Average number of comments to post provided within 24 hours**: anche questa metrica deve essere effettuata utilizzando un crawler che analizzi le pagine. Tale misura indica la reattività del social media a nuovi post. Siti con una reattività bassa potrebbero avere un basso grado di coinvolgimento, di partecipazione oppure semplicemente una scarsa attitudine alla condivisione dei contenuti.

Time

In questo costrutto si analizza il tempo speso sul sito con l'utilizzo di due metriche principali:

- **Average time spent on site**: misura la media del tempo che gli utenti spendono sul sito. Maggiore è il tempo passato sul sito e maggiore è la probabilità che i contenuti raggiungano un maggior numero di utenti e quindi rendono le fonti con un valore più alto di questa metrica più appetibili per il propagarsi della reputazione;

- **Bounce rate**: questa metrica tiene traccia della percentuale di utenti che abbandonano il sito dopo aver visualizzato la prima pagina. È utile tenere traccia di questa misura poiché spesso tali utenti hanno effettuato una ricerca sui motori tradizionali (es. google o yahoo) ed effettuano un accesso al sito che dura solo pochi istanti e di conseguenza non andrebbero calcolati all'interno dei volumi.

Queste due metriche sono fornite da Alexa e non necessitano di calcoli complessi.

4.2.2.2 Metriche per valutare il grado di specializzazione di una fonte

Con il termine specializzazione s'intende il grado con cui una fonte si focalizza rispetto a un determinato dominio come ad esempio può essere il turismo oppure il cinema, la tecnologia, la cucina, ecc.. Il nostro obiettivo è automatizzare l'identificazione delle sorgenti informative definite generaliste da quelle specializzate o focalizzate su un particolare dominio. Se si effettua una ricerca nel campo del turismo ci si aspetta che siti come TripAdvisor e LonelyPlanet finiscano nella categoria delle fonti focalizzate mentre siti come Twitter e Facebook tra i generalisti. Le due metriche che si sono individuate per perseguire tale scopo sono:

- **Post per dominio**: misura la percentuale dei post appartenenti a uno specifico dominio rispetto al volume totale dei post. Di seguito viene riportata la formula di facile comprensione dove con F si indica il grado di focalizzazione.

$$F = \frac{\#\,post\_del\_dominio\_x}{\#\,di\_post\_totali}$$

- **Domini per post**: indica in media quanti domini vengono identificati in un post. Questa metrica più che essere utile per l'analisi è utilizzata per verificare la bontà della metrica precedente. Se ogni post contiene un numero elevato di domini, la metrica precedente può essere ricavata con meno precisione. Con P si indica la precisione nel rilevare il dominio che è calcolata nel modo seguente.

$$P = \frac{\sum_{1}^{n} x_n}{n}$$

in cui:

- n – numero di post
- x_n – numero di domini appartenenti al post n

Anche se precedentemente abbiamo catalogato le fonti come Facebook o Twitter tra le fonti generaliste, la classificazione a cui siamo interessati non ha lo scopo di porre dei paletti sulla base dell'opinione che tutti hanno di queste sorgenti, ma ha lo scopo di mantenersi il più flessibile possibile. Può capitare, in determinate circostanze, che una fonte generalista, come ad esempio twitter, si focalizzi su un argomento di pubblico interesse. In questo caso Twitter potrebbe passare nell'insieme delle fonti focalizzate poiché il volume e la visibilità dei commenti su tale argomento eguagliano le caratteristiche tipiche dei siti come TripAdvisor o altri. È da sottolineare che non si sono trovati ne studi ne esperimenti al riguardo. Di conseguenza è necessario effettuare dei test per analizzare dove porre la soglia che fa passare i siti da generalisti a focalizzati.

4.3 Calcolo della reputazione di una fonte

La necessità di avere un valore unico della reputazione di un'azienda o prodotto su Internet ci spinge a fare un ulteriore passo di aggregazione. Dopo aver aggregato il sentiment in ognuno dei media a nostra disposizione ci si chiede come unire i risultati tenendo in considerazione le diverse caratteristiche delle sorgenti informative sulla base dei tre costrutti Traffic, Participation e Time. Questa aggregazione è composta

da vari calcoli in cascata (Figura 4.4), in cui si distinguono 5 passi fondamentali:

- **Calcolo della metrica**: alcune metriche sono fornite da aziende specializzate come Alexa mentre per altre è necessario effettuare un crawling delle pagine. Nella Sezione 4.2.2, dedicata alle metriche di valutazione delle fonti, si può trovare un'ampia descrizione su come ricavarle;

- **Normalizzazione**: Una volta ottenuto il valore della metrica lo si riconduce all'intervallo [0,1]. Nella Sezione 4.3.2 verranno discussi alcuni metodi per normalizzare le misure;

- **Aggregazione per costrutto**: in questa fase si aggregano le varie metriche per costrutto in modo da ottenere un valore/livello di Traffic, Partecipation e Time per ogni fonte;

- **Sentiment pesato**: a questo punto siamo in possesso di tutte le informazioni per effettuare l'aggregazione delle diverse fonti ottenendo tre valori: 1) sentiment pesato per il Traffic ($S_{traffic}$); 2) sentiment pesato per Participation (S_{part}); 3) sentiment pesato per Time (S_{time});

- **Pesatura dei costrutti**: i tre valori di sentiment ricavati nel passo precedente vengono aggregati per ottenere un unico valore. In questa fase si è deciso di lasciare l'assegnazione dei pesi all'utente in modo da mantenere un certo grado di flessibile d'analisi.

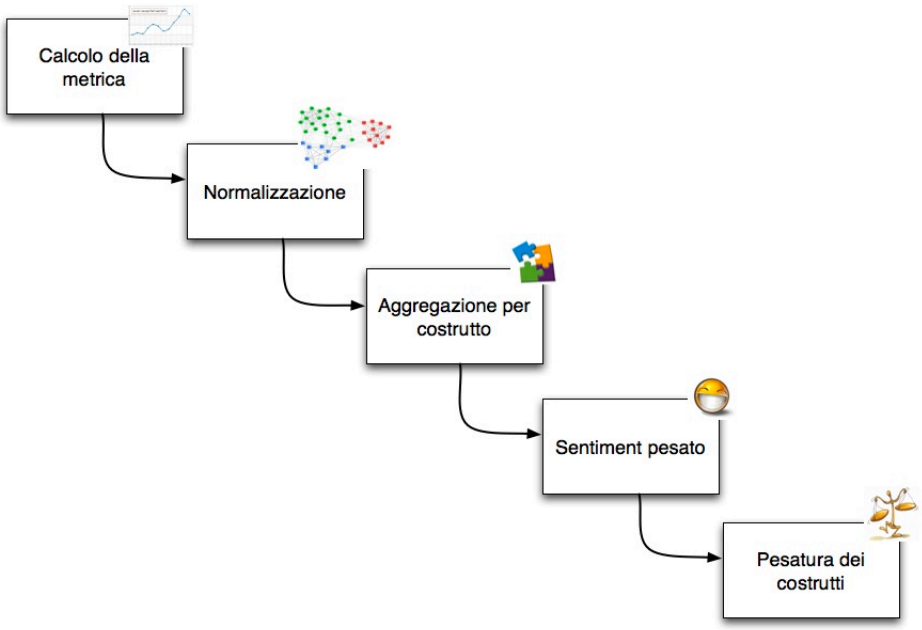

Figura 4.4: Schema del calcolo reputazione fonte

Per fare chiarezza su come si svolgono le fasi sopra descritte, si è deciso di utilizzare un esempio operativo che ripercorre l'intera cascata.

4.3.1 Esempio

Supponiamo di avere tre fonti e di considerare per le prime tre fasi solo le metriche del costrutto Time. In Tabella 4.1 sono riportati i valori.

	Average time spent on site (min)	**Bounce Rate (%)**
Fonte 1	10	5
Fonte 2	12	3
Fonte 3	5	9

Tabella 4.1: Metriche di esempio

Dopo aver ottenuto i valori delle due metriche per ogni fonte si passa alla fase di normalizzazione. Per assegnare un valore normalizzato ad ogni misura ci sono diverse soluzioni che verranno discusse nella

Sezione 4.3.2. I valori delle due metriche normalizzati nell'intervallo [0,1] sono riportati in Tabella 4.2.

	Average time spent on site (Norm)	**Bounce Rate (Norm)**
Fonte 1	1	0,5
Fonte 2	1	0,2
Fonte 3	0,5	1

Tabella 4.2: Metriche di esempio normalizzate

Il passo successivo è l'aggregazione per costrutto. Per ottenere il valore di Time si effettua la media di tutte le metriche che ne fanno parte. Il valore di Time della fonte 1 (T_1) del nostro esempio si calcola come:

$$T_1 = \frac{Avg + Br}{n} = \frac{1 + 0,5}{2} = 0,75$$

Dove:

T_1 – è il valore di Time della sorgente 1;

Avg – il valore della metrica Average time spent on site (normalizzata);

Br – il valore della metrica Bounce Rate (normalizzata);

n – numero totale di metriche del costrutto.

Una volta calcolato il valore T per ogni fonte (T_1=0,75; T_2=0,6; T_3=0,75), si procede con il calcolo del sentiment aggregato pesato per il costrutto Time (S_{time}) nel modo seguente:

$$S_{time} = \frac{\sum_{i=1}^{m} s_i T_i}{\sum_{i=1}^{m} T_i}$$

dove:

m – numero di fonti;

s_i – sentiment della fonte i-esima;

T_i – Valore del costrutto time per la fonte i-esima.

Supponiamo di avere i seguenti valori di sentiment per le rispettive fonti: s_1=4, s_2=3, s_3=2. Otteniamo un valore S_{time}=3.

Prima di procedere all'ultima fase di pesatura dei costrutti (Figura 4.4), è necessario procedere con gli stessi calcoli anche per gli altri due insiemi di metriche in modo da avere il sentiment aggregato per Traffic ($S_{traffic}$) e per Participation (S_{part}). A questo punto si passa alla fase di pesatura. Dopo diverse riflessioni si è pensato di lasciare all'utente la possibilità di settare a suo piacimento i vari pesi. Una tale scelta permette un'analisi più flessibile e consente di valutare il sentiment sulla base di diverse caratteristiche delle sorgenti informative. La formula generale è la seguente:

$$S = w_1 S_{traffic} + w_2 S_{Part} + w_3 S_{time}$$

dove i pesi w_1, w_2, w_3 sono impostati dall'utente e inoltre devono soddisfare le seguenti condizioni:

- Devono appartenere all'intervallo [0,1];
- $w_1 + w_2 + w_3 = 1$.

Alla fine di questa sequenza di calcoli siamo in grado di presentare all'utente un unico valore aggregato del sentiment.

4.3.2 Normalizzazione

Per assegnare un valore normalizzato a una metrica abbiamo analizzato tre metodi:

- Normalizzazione continua;

- Normalizzazione con livelli di soglia percentuali;
- Normalizzazione rispetto al massimo valore.

Normalizzazione continua

Supponiamo di avere dieci fonti con i seguenti valori della metrica A:

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
20	10	15	13	2	5	20	8	4	9

Tabella 4.3: Valori di esempio metrica M

Per normalizzare i valori della metrica non si fa altro che costruire un intervallo prendendo come limite superiore il massimo valore misurato e come valore minimo lo 0. Sulla base di questo intervallo si ricavano, tramite proporzioni, i valori delle metriche normalizzate nell'intervallo [0,1]. La Figura 4.5 esemplifica la normalizzazione dei valori di Tabella 4.4.

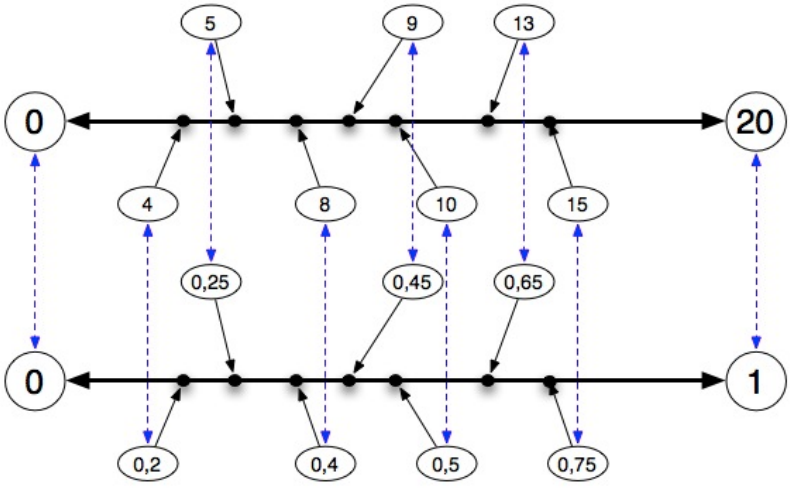

Figura 4.5: Normalizzazione continua

Normalizzazione con livelli di soglia percentuali

A differenza del metodo precedente che utilizza valori continui, questo metodo definisce delle soglie A^k per ogni metrica. Supponiamo di utilizzare i valori di Tabella 4.4. Definiamo:

- $A^{0,2}$ come il valore di soglia sopra il quale si trova il 20% delle misurazioni effettuate;
- $A^{0,5}$ come il valore di soglia sopra il quale si trova il 50% delle misurazioni effettuate;
- $A^{0,8}$ come il valore di soglia sopra il quale si trova il 80% delle misurazioni effettuate.

Si è scelto di utilizzare le soglie 0,2, 0,5 e 0,8 poiché sono quelle che di norma sono utilizzate nella documentazione in nostro possesso.

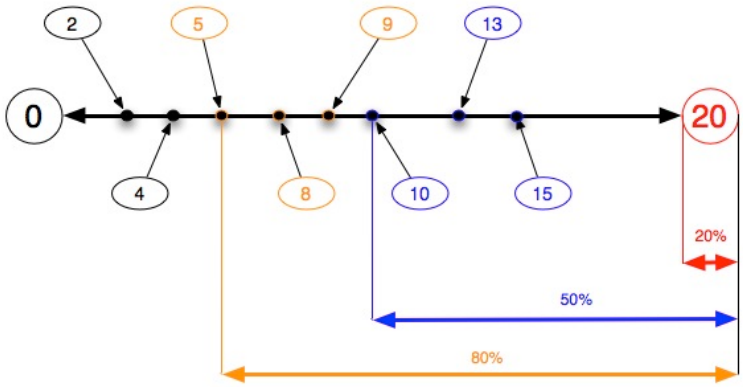

Figura 4.6: Normalizzazione con livelli di soglia percentuali

Dalla Figura 4.6 si osservano le seguenti soglie: $A^{0,2}$=20, $A^{0,5}$=10 e $A^{0,8}$=5. Una volta determinati questi valori si definisce la funzione di normalizzazione nel modo seguente:

$$A_{norm\_i} = \begin{cases} 1 & \text{if } A_i \geq A^{0,2} \\ 3/4 & \text{if } A^{0,5} \leq A_i < A^{0,2} \\ 1/2 & \text{if } A^{0,8} \leq A_i < A^{0,5} \\ 1/4 & \text{altrimenti} \end{cases}$$

Si sconsiglia l'utilizzo del metodo di normalizzazione con livelli di soglia percentuali a causa di un problema discusso con un ulteriore esempio. Supponiamo di ottenere i valori della metrica "Daily visitors" per 10 fonti (Tabella 4.4).

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
2000	104	156	135	2	5	205	8	4	9

Tabella 4.4: Valori di esempio metrica "Daily visitors"

Calcoliamo i livelli di soglia come abbiamo fatto nell'esempio precedente (Figura 4.7).

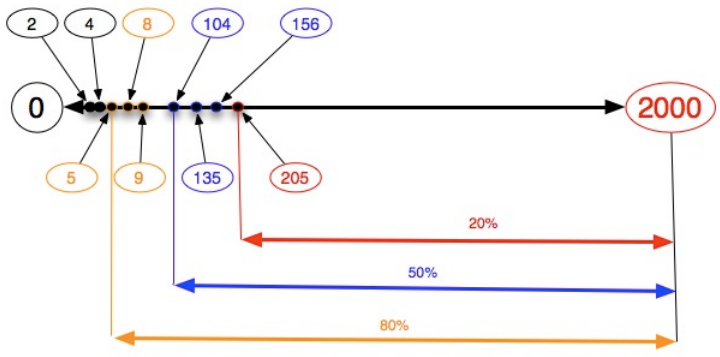

Figura 4.7: Normalizzazione con livelli di soglia percentuali (caso particolare)

Se si implementa la funzione di normalizzazione precedente si darebbe lo stesso valore alla fonte A_1=2000 e A_7=205. Sicuramente siamo di

fronte a un errore non banale. La fonte 1 ha 10 volte i visitatori della fonte 7. È auspicabile avere due pesi diversi a fronte di un'ampia differenza delle due misure.

Normalizzazione rispetto al massimo valore

Questo ultimo metodo prende in ingresso il massimo valore della metrica che nell'esempio di Tabella 4.5 è A_1=2000 e si costruiscono le soglie su percentuali del valore massimo. Come nell'esempio di Figura 4.7 si calcola:

- $A^{0,2}$: il massimo valore misurato per la metrica A ridotta del suo 20%;
- $A^{0,5}$: il massimo valore misurato per la metrica A ridotta del suo 50%;
- $A^{0,8}$: il massimo valore misurato per la metrica A ridotta del suo 80%.

Utilizzando i valori di Tabella 4.5 abbiamo $A^{0,2}$=1600, $A^{0,5}$=1000 e $A^{0,8}$=400. In Figura 4.8 riportiamo la schematizzazione utilizzata anche per gli altri metodi.

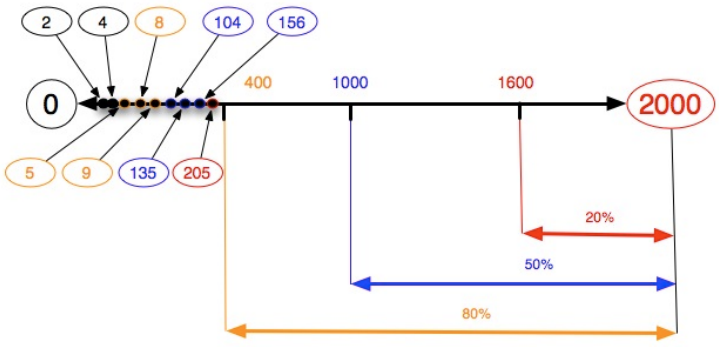

Figura 4.8: Normalizzazione rispetto al massimo valore

4.4 Tipi di aggregazione e obiettivi di analisi

Precedentemente abbiamo discusso tutto le metriche per la valutazione sia delle fonti che degli utenti e introdotto delle metodologie di aggregazione utili per l'analisi dei social media. In questa sezione vengono analizzate particolari aggregazioni volte a misurare caratteristiche utili per la sentiment analysis.

In Tabella 4.5 viene richiamata la lista delle metriche finora trovate distinguendo quelle applicate alla singola fonte rispetto a quelle su più fonti.

SINGOLA FONTE	**PIU' FONTI**
Sentiment standard	Traffic
Sentiment aggregato per utente	Participation
Sentiment con pesatura utente	Time
	Specializzazione fonte

Tabella 4.5: Categorizzazione metriche

Sentiment con pesatura delle fonti

Dopo aver calcolato il sentiment per ogni social media utilizzando il metodo standard, cioè tralasciando ogni informazione sugli utenti che hanno scritto i post, si procede con la pesatura per fonte. Ad ogni valore di sentiment ottenuto vengono applicati dei pesi sulla base delle metriche di traffic, participation e time. L'utente è libero di prendere in considerazione ogni combinazione di queste, e decidere a suo piacimento se analizzarle separatamente o tutte e tre contemporaneamente.

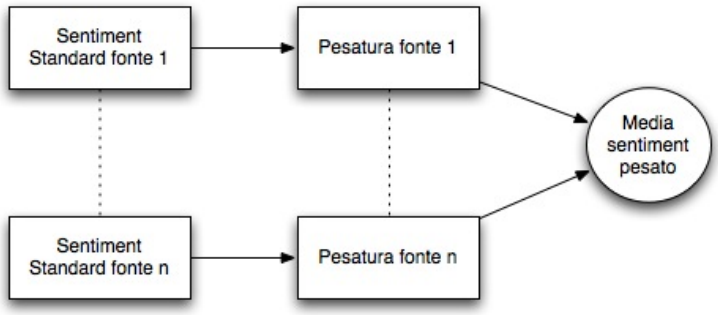

Figura 4.9: Pesatura Fonte

Aggregazione per utente con pesatura fonte

Rispetto all'analisi precedente, si prende in considerazione l'utente ed in particolare si cerca di eliminare il bias del volume dei post per fare largo ad un'analisi più democratica del sentiment. La fase di pesatura della fonte è uguale a quella precedente mentre per il calcolo del sentiment ci si basa sull'aggregazione per utente .

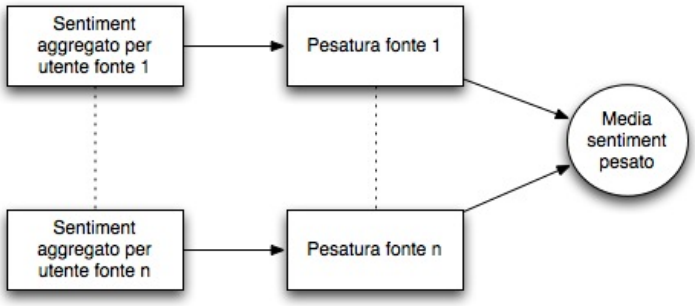

Figura 4.10: Aggregazione utente + Pesatura fonte

Pesatura per utente e per fonte

Spesso si ha la necessità di differenziare ogni utente sulla base della sua posizione all'interno della rete. L'opinione degli utenti che occupano le posizioni più centrali o più prestigiose hanno una maggiore leva sull'evoluzione del sentiment globale del social network. Grazie alla

posizione che occupano, questi utenti guidano i trend e le opinioni propagando le proprie idee e ostacolando quelle contrastanti la sua presa di posizione.

In Figura 4.11 è riportato lo schema di analisi il quale aggiunge un livello di pesatura dell'utente allo schema precedente.

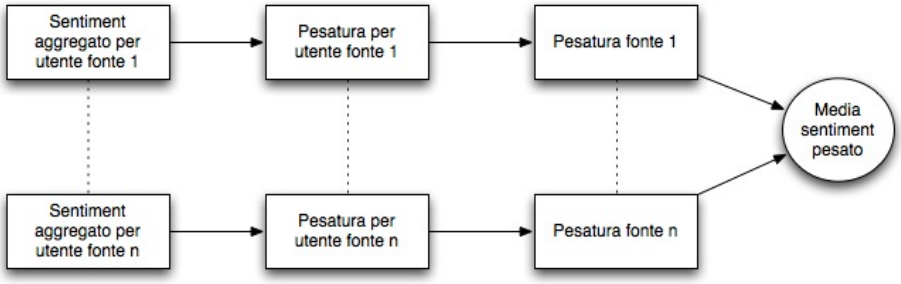

Figura 4.11: Pesatura per utente e per fonte

Capitolo 5

Analisi dei dati e presentazione dei risultati

5.1 Introduzione

In questo capitolo verrà svolta una valutazione empirica del cruscotto di indicatori presentati nel Capitolo 4. I dati utilizzati sono stati estratti dai siti di TripAdvisor e Lonely Planet tramite dei crawler dedicati, i quali hanno riempito dei database con le informazioni sui post e sugli utenti iscritti ai forum delle due sorgenti. Il capitolo è composto da una Sezione 5.2 che presenta le fonti analizzate nel resto del capitolo; la Sezione 5.3 fa un'analisi di correlazione tra le metriche di centralità della sorgente TripAdvisor. Nella Sezione 5.4 viene presentata l'analisi di correlazione tra le metriche ottenute da TripAdvisor e i destination expert, utenti promossi dal sito come esperti di particolari località. L'analisi del sentiment è stata svolta nella Sezione 5.4 che tratta nuovamente della correlazione tra le metriche utilizzando però i dati di sentiment (Sezione 5.4.1) per poi passare ad altre analisi quali la variabilità delle aggregazioni rispetto alla deviazione standard dei voti (Sezione 5.4.2), la valutazione per categorie (Sezione 5.4.3), il sentiment degli opinion leader (Sezione 5.4.4) e per finire la valutazione del silenzio (Sezione 5.4.5). In seguito nella Sezione 5.5 viene trattato l'argomento della pesatura delle fonti ed infine (Sezione 5.6) vengono riportate le conclusioni dei risultati ottenuti..

5.2 Fonti analizzate

Per testare la metodologia presentata nel Capitolo 4 si è scelto di analizzare il sentiment di tre importanti città europee come Milano, Madrid e Londra. Tra i siti specializzati in viaggi e turismo si è scelto di creare il dataset utilizzando le discussioni presenti sulle fonti TripAdvisor e Lonely Planet. Dopo aver implementato dei crawler specializzati nell'estrarre i dati da ogni fonte, si è passati alla fase di estrazione dei post accompagnati da tutti i dati rilevanti per le analisi come ad esempio il thread di appartenenza, il numero sequenziale all'interno del thread e alcune informazioni sull'utente che lo ha scritto. Queste informazioni sono fondamentali per costruire il grafo di analisi utilizzando la teoria in [79].

Lonely Planet

Lonely Planet è specializzato nella pubblicazione di Travel Book a livello mondiale. Dal 2007, la compagnia è stata controllata dalla BBC Worldwide la quale ne possiede il 75%, mentre ai fondatori Maureen e Tony Wheeler rimane il restante 25%. Originariamente chiamata Lonely Planet Publications, la compagnia ha cambiato nome nel luglio 2009 in Lonely Planet. Nel 2010 pubblica circa 500 titoli in 8 lingue cosi come programmi televisivi, magazine, applicazioni mobile e website. La sede è posizionata in Footscray, una cittadina di Melbourne, Australia, con uffici affiliati con sedi in Londra e Oakland, California. La comunità online di Lonely Planet (Thorn Tree) è usata da più di 600000 viaggiatori per suggerimenti e pubblicità. Il sito è stato aggiornato nel 2009 e si muove verso l'integrazione con funzionalità di social networking e contenuti generati dall'utente: nuove caratteristiche includono BlogSherpa blogs, connessioni a Facebook, l'aggiunta di gruppi, la possibilità di votare e scrivere recensioni su luoghi e ristoranti per poi salvarli in una lista di favoriti.

Figura 5.1: Lonely Planet

Il forum è strutturato in 5 categorie di discussione:

1. Departure Lounge: sono presenti discussioni che riguardano la condivisione d'informazioni su viaggi, suggerimenti e notizie. Questa categoria è a sua volta suddivisa in zone geografiche in modo che gli utenti interessati ad aprire un topic sul Kenya non dovranno fare altro che selezionare la categoria Departure Lounge, entrare in Africa e creare il topic di discussione;

2. The lobby: qui si trovano informazioni su budget, salute, vacanze tematiche ecc..;

3. Check in Lonely Planet: per parlare direttamente a Lonely Planet, segnalare bug al sito, condividere un feedback per i guide book e conoscere le novità;

4. Tree House: qui gli utenti possono parlare di qualsiasi topic non appartenente alle altre categorie , come la fotografia, la musica, i film, i libri, ecc…;

5. Sell, Swap & Meet up: in questa sezione si trovano compagni di viaggio per condividere esperienze, consigli per rendere i propri viaggi meno costosi e più abbordabili facendo conoscenza con persone che raccontano le proprie esperienze di viaggio agli altri utenti.

TripAdvisor

TripAdvisor.com è una guida di viaggi completamente gratuita che assiste gli utenti nel reperire informazioni su viaggi, pubblicare opinioni e interagire nel forum. TripAdvisor.com fa parte di TripAdvisor Media Network; è un esempio di "consumer generated media". I servizi del sito sono gratuiti per gli utenti i quali forniscono la maggior parte dei contenuti ed è supportato da un modello di business basato sulla pubblicità.

TripAdvisor Media Group è composto da 14 brand molto popolari come TripAdvisor, Virtual Tourist, Cruise Critic, Seat Guru, TravelPod, Airfare Watchdog, Booking Buddy, FlipKey, Holiday Watchdog, Independent Traveler, Onetime, Smarter Travel, Frequent Flier e Travel Library. TripAdvisor opera in diverse nazioni con siti in Canada, UK, Francia, Germania, Giappone, India, Italia e Spagna e la sede principale è posizionata in Newton, Massachusetts. TripAdvisor è stato fondato nel Febbraio del 2000 da Stephen Kaufer. I primi finanziamenti provenivano da Flagship Ventures, Bollard Group e investitori privati. Nel 2004 la

compagnia è stata comprata da InterActive Corporation (IAC) e dal 2004 al 2009 ha continuato ad espandersi comprando tutti gli altri brand che ne fanno parte. Il sito conta oltre 10 milioni di utenti iscritti, 25 milioni di visitatori e quasi 2 milioni di foto inviate dagli utenti rappresenta un sito di riferimento per persone intenzionate a prenotare un hotel o fare un viaggio. Non mancano le critiche ricevute da fonti autorevoli, come *The Times* [101] e *The Guardian* [102], principalmente per tre motivi:

- Ospiti che non hanno mai soggiornato in un hotel possono lasciare una finta recensione positiva o negativa;
- Strutture ricettive con un basso livello di gradimento da parte degli utenti, possono far salire la propria reputazione in poche ore, scrivendo finte recensioni positive;
- Molti albergatori cercano di danneggiare la reputazione della concorrenza, innescando una sorta di competizione che genera la creazione di recensioni false.

Di contro però, TripAdvisor dispone di uno staff che controlla manualmente i commenti [103] e di un algoritmo che previene eventuali abusi [104].

Figura 5.2: TripAdvisor

Il sito presenta molte caratteristiche tra cui:

1. Reviews: le classifiche di hotel, ristoranti e attività si basano soprattutto su recensioni fornite da utenti registrati;

2. Traveler Articles: è una struttura in stile Wiki in cui gli utenti possono aggiungere informazioni ad un articolo comune scritto a più mani. Includono informazioni sui trasporti, attività e guide sul cibo;

3. Photo & video: gli utenti possono aggiungere fotografie e video delle mete turistiche che recensiscono in modo da documentare ulteriormente i luoghi visitati;

4. Hotel popularity index: fornisce un ranking dinamico degli hotel di tutto il mondo misurato tenendo conto della quantità e della qualità dei commenti che ricevono;

5. Maps: un mashup tra TripAdvisor e Google maps per fornire informazioni geografiche delle posizioni degli hotel;

6. Forum: dove gli utenti creano discussioni per rispondere a dubbi, domande e condividere informazioni utili e importanti.

Per quanto riguarda i test che condurremo in questa sezione ci limiteremo ad analizzare la parte dei forum sia per costruire la rete di utenti che per valutare il sentiment delle città Milano, Londra e Madrid.

5.3 Analisi di correlazione

Il primo passo per valutare l'importanza degli utenti di una comunità virtuale è la costruzione del grafo delle relazioni. Utilizzando la struttura delle discussioni come illustrato in Figura 5.3 siamo stati in grado di ricostruire il grafo degli utenti che hanno partecipato alle discussioni su Milano, Londra e Madrid nei forum di Lonely Planet e TripAdvisor.

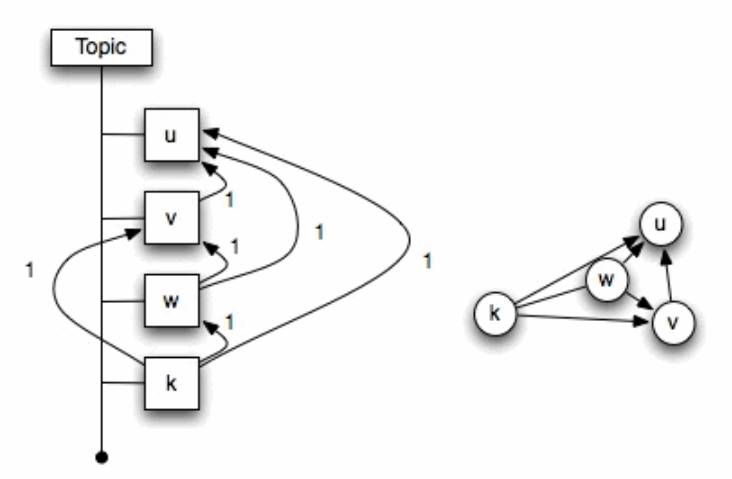

Figura 5.3: Costruzione grafo

Per esempio l'utente *u,* che ha scritto il commento prima degli altri, sarà collegato tramite archi entranti agli utenti *v*,*w* e *k* poiché si assume che

questi abbiano letto il commento di *u* prima di intervenire nella discussione.

Una volta costruito il grafo, sono state calcolate le metriche mostrate nel Capitolo 4 per determinare la centralità di ogni utente al fine di considerarne l'importanza. Il campione di analisi si compone di 2423 utenti per Lonely Planet e 2936 utenti per TripAdvisor. Su tali metriche abbiamo eseguito due studi di correlazione, il primo (sezione 5.3.1) per valutare la correlazione tra le diverse metriche e il secondo (sezione 5.3.2) per analizzare la correlazione tra le metriche degli utenti di TripAdvisor rispetto ai "Destination Expert", particolari utenti iscritti al sito e da questo riconosciuti come utenti leader nel fornire informazioni riguardanti determinate località.

5.3.1 Correlazioni tra metriche

In Tabella 5.1 sono riportati i valori di correlazione tra le varie metriche calcolate sugli utenti di TripAdvisor. Queste presentano una buona correlazione le une con le altre; in particolare la *indegree*, la *outdegree* e la *betweenness* hanno dei valori molto elevati, indicazione del fatto che gli utenti che attirano più commenti attraverso i loro post sono anche quelli più propensi a commentare quelli degli altri; inoltre, elevati livelli di *indegree* e *outdegree* favoriscono anche la centralità calcolata attraverso la *betweenness*. La *proximity* e la *closeness* hanno un indice di correlazione pari a 1. Sul campione di 2941 utenti le due metriche producono lo stesso ranking suggerendo così di mantenere solo una delle due in modo da ridurre sensibilmente i calcoli di analisi. Anche l'*eccentricity* ha una forte correlazione con queste due metriche ma, a differenza delle altre due, fornisce informazioni aggiuntive e quindi è bene mantenerla nel nostro cruscotto.

		indegree	proximity	outdegree	eccentricity	closeness	between.
indegree	**Person Correlation**	1	,334 (**)	,945 (**)	,278(**)	,347(**)	,880(**)
	Sig. (2-tailed)		,000	,000	,000	,000	,000
	N	2941	2941	2941	2941	2941	2941
proximity	**Parson Correlation**	,334 (**)	1	,274 (**)	,989 (**)	1,000 (**)	,141 (**)
	Sig. (2-tailed)	,000		,000	,000	,000	,000
	N	2941	2941	2941	2941	2941	2941
outdegree	**Parson Correlation**	,945 (**)	,274 (**)	1	,220 (**)	,286 (**)	,884 (**)
	Sig. (2-tailed)	,000	,000		,000	,000	,000
	N	2941	2941	2941	2941	2941	2941
eccentricity	**Parson Correlation**	,278 (**)	,989 (**)	,220 (**)	1	,986 (**)	,101 (**)
	Sig. (2-tailed)	,000	,000	,000		,000	,000
	N	2941	2941	2941	2941	2941	2941
closeness	**Parson Correlation**	,347 (**)	1,000 (**)	,286 (**)	,986 (**)	1	,151 (**)
	Sig. (2-tailed)	,000	,000	,000	,000		,000
	N	2941	2941	2941	2941	2941	2941
betweenness	**Parson Correlation**	,880 (**)	,141 (**)	,884 (**)	,101 (**)	,151 (**)	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000
	N	2941	2941	2941	2941	2941	2941

Tabella 5.1: Correlazione tra metriche

5.3.2 Correlazioni con Destination Expert

I Destination Expert sono utenti eletti dallo staff di TripAdvisor come esperti di una determinata destinazione turistica. Per diventare un esperto, l'utente deve partecipare frequentemente alle discussioni, fornire consigli utili e segnalare eventuali post non corretti aiutando il sito a mantenere elevati livelli di credibilità. In Tabella 5.2 sono riportati i valori di correlazione tra le metriche di centralità e lo status di un utente. Essendo la variabile *expert* dicotomica (vale 1 se l'utente è un esperto, 0 se non lo è), il test utilizzato per testare la correlazione è una point-biserial.

		indegree	proximity	outdegree	eccentricity	closeness	between.
Expert	**Point biserial Cor.**	,422(**)	,032	,369(**)	,0,15	,036(*)	,413(**)
	Sig. (2-tailed)	,000	,087	,000	,423	,0,49	,000
	N	2941	2941	2941	2941	2941	2941

Tabella 5.2: Correlazione Destination Expert

I livelli di correlazione sono elevati per le metriche *indegree*, *outdegree* e *betweenness*, mentre si possono considerare non correlate con le altre tre. È interessante notare come la correlazione sia elevata con le metriche locali. Una tale correlazione suggerisce che nella scelta dei destination expert si faccia particolare attenzione alla quantità dei post pubblicati e non si tenga traccia delle interazioni con gli altri utenti. Un risultato importante è dato dalla forte correlazione con la *beetweenness*, secondo Freeman [37] la metrica che sintetizza meglio la centralità degli utenti. Si ricorda che le metriche locali come l'*indegree* e l'*outdegree* sono facili da calcolare rispetto alle altre; inoltre da questi risultati tali metriche dimostrano avere anche una forte espressività.

5.4 Aggregazioni del sentiment

Dopo aver calcolato tutte le metriche di centralità utilizzando il tool di analisi, diamo una breve presentazione dei dati utilizzati nel corso del capitolo per lo studio della reputation. Com'è già stato detto in precedenza, gli utenti coinvolti nelle analisi sono 2423 per Lonely Planet e 2936 per TripAdvisor. Ricordando che il tool di analisi si appoggia su un'architettura molto più complessa ancora in fase di costruzione, i dati utilizzati provengono da diversi database, di conseguenza il numero di utenti con la disponibilità dei dati sia di sentiment che per calcolare le metriche, potrebbe essere inferiore rispetto ai numeri presentati in

precedenza. In Tabella 5.3 si riporta il numero di utenti e quello dei post per le tre località analizzate: Londra, Madrid e Milano.

Città	TripAdvisor		Lonely Planet	
	n° utenti	n° post	n° utenti	n° post
Londra	419	3952	327	1488
Madrid	37	557	51	150
Milano	45	416	81	234

Tabella 5.3: Dati di analisi

I valori di sentiment sono forniti dal database di Expert System contenente i post scritti nell'intervallo di tempo che va da Dicembre 2002 a Giugno 2010, ottenuti con una sessione di crawling svoltasi nel primo semestre del 2010. Il sentiment è espresso su una scala di 5 valori, dove:

- Valore 1: sentiment molto negativo;
- Valore 2: sentiment negativo;
- Valore 3: sentiment neutro;
- Valore 4: sentiment positivo;
- Valore 5: sentiment molto positivo.

Utilizzando le metriche di centralità calcolate, si procederà con la valutazione del sentiment nei vari modi presentati in Sezione 4.2.1. Per maggiore chiarezza riportiamo le tre tipologie di aggregazione anche di seguito:

- Aggregazione standard: media dei valori dei sentiment relativi ai post che recensiscono un determinato brand;
- Aggregazione per utente: media del valore di sentiment per ogni utente e successiva aggregazione di tutti gli utenti;

- Aggregazione con pesatura utente: le metriche ottenute con l'utilizzo dell'applicativo sono usate per calcolare il sentiment prendendo in considerazione l'importanza dell'utente che recensisce un brand.

5.4.1 Verifica della correlazione tramite l'analisi di sentiment

La prima città ad essere analizzata è Londra e in Tabella 5.4 sono riportati i risultati delle varie aggregazioni sia per la sorgente TripAdvisor che per Lonely Planet.

Londra			
Aggregazione		**Lonely Planet**	**TripAdvisor**
Standard		3,67	3,78
Per Utente		3,67	3,90
Pesatura utente	**Indegree**	3,58	3,84
	Proximity	3,63	3,90
	Outdegree	3,64	3,83
	Eccentricity	3,63	3,90
	Closeness	3,63	3,90
	Betweenness	3,61	3,78

Tabella 5.4: Aggregazioni Londra

Per una lettura più immediata in Figura 5.4 viene riportato il grafico con i valori messi a confronto tra le due fonti.

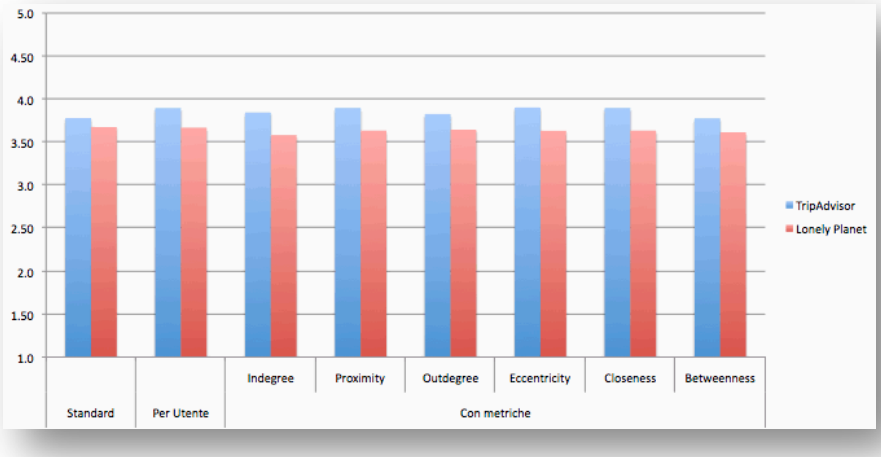

Figura 5.4: Grafico aggregazioni Londra

Evidentemente TripAdvisor ha una maggiore positività rispetto a Lonely Planet. Tale differenza è mantenuta su tutte le tipologie di aggregazione indicando che l'opinione più positiva, è riscontrata sia dagli utenti comuni che da quelli con elevati livelli di centralità.

Nella sezione 5.3.1 abbiamo osservato che le metriche *proximity* e *closeness* presentano un valore di correlazione uguale a uno. Analizzando i valori di aggregazione che si basano su queste due metriche (Figura 5.4) gli scostamenti sono pressoché nulli. Come riprova della forte correlazione in Tabella 5.5 e 5.6 riportiamo rispettivamente gli utenti con valore di *proximity* e *closeness* più elevati.

Proximity		
Pos	**Username**	**Valore metrica**
1	TravellerPlus	0,460
2	adamhornets	0,448
3	Glyn1	0,420
4	Mr_Cellophane	0,416
5	joeintheuk	0,408
6	Alanrow	0,407
7	offtolondon	0,406
8	GreenwichNick	0,405
9	bob007	0,405
10	Kayb95	0,402

Tabella 5.5: Ranking Proximity

Closeness		
Pos	**Username**	**Valore metrica**
1	TravellerPlus	0,521
2	adamhornets	0,505
3	Glyn1	0,461
4	Mr_Cellophane	0,455
5	Alanrow	0,447
6	joeintheuk	0,446
7	GreenwichNick	0,440
8	offtolondon	0,439
9	bob007	0,437
10	Kayb95	0,433

Tabella 5.6: Ranking Closeness

Dalle tabelle si osserva che gli utenti che compaiono nelle prime 10 posizioni sono gli stessi. L'unica differenza sta nell'ordinamento che, se pur diverso, rimane molto simile con un paio di posizioni scambiate.

Passiamo ora ad analizzare la città di Madrid che come riportato in Tabella 5.3 presenta dei volumi inferiori rispetto a Londra. In Tabella 5.7 e in Figura 5.5 riportiamo i risultati delle aggregazioni.

Madrid			
Aggregazione		**Lonely Planet**	**TripAdvisor**
Standard		3,68	3,88
Per Utente		3,74	3,95
Pesatura utente	**Indegree**	3,58	4,16
	Proximity	3,66	3,96
	Outdegree	3,61	4,27
	Eccentricity	3,66	3,98
	Closeness	3,65	3,96
	Betweenness	3,55	4,36

Tabella 5.7: Aggregazioni Madrid

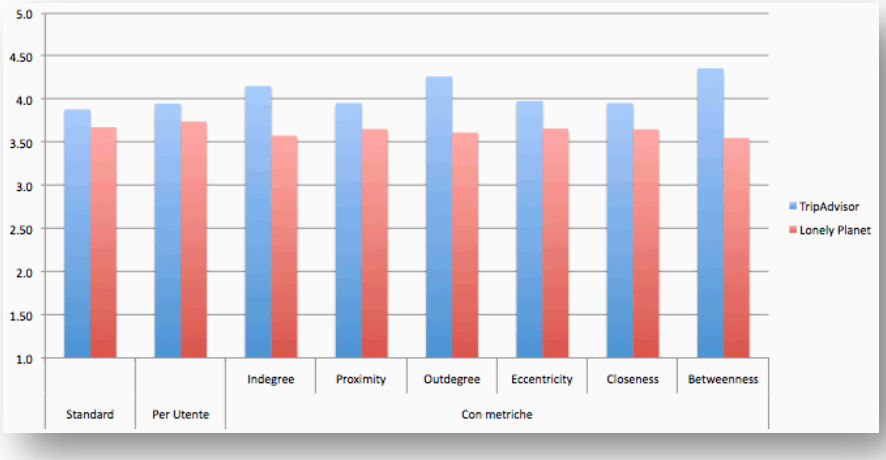

Figura 5.5: Grafico aggregazioni Madrid

Gli utenti di TripAdvisor, come era accaduto per Londra, continuano a mantenere un sentiment più positivo rispetto a Lonely Planet. Dal grafico gli utenti con il sentiment più alto sono opinion leader secondo le

metriche *indegree*, *outdegree* o *betweenness*. Anche in questo caso proviamo a verificare la correlazione trovata nella sezione 5.3.1 tra queste tre metriche (Tabella 5.8).

Indegree		**Outdegree**		**Betweenness**	
Pos	**utente**	**Pos**	**utente**	**Pos**	**utente**
1	TravellerPlus	1	TravellerPlus	1	TravellerPlus
2	dicanio10	2	dicanio10	2	dicanio10
3	Nurioopta	3	RonaldoC	3	Nurioopta
4	RonaldoC	4	Nurioopta	4	perilizia
5	pintpont	5	littlerussell	5	RonaldoC
6	littlerussell	6	amcsocal	6	Scarpenter4
7	girlmanc	7	RomanCitizen	7	pintpont
8	Scarpenter4	8	pintpont	8	littlerussell
9	perilizia	9	Scarpenter4	9	RomanCitizen
10	gambl_R	10	girlmanc	10	gambl_R

Tabella 5.8: Ranking Indegree, Outdegree e Betweenness

Tra i primi 10 utenti con valori più alti di *betweenness* ben 7 compaiono nella top 10 sia del ranking costruito con l'*indegree* che quello con l'*outdegree* mentre le altre tre compaiono solo in una di esse. Se si pensa al significato delle tre metriche si può giungere facilmente alla conclusione che gli utenti con un numero elevato di archi entranti e uscenti, hanno una maggiore probabilità di trovarsi sui cammini minimi che collegano gli altri nodi, ottenendo così valori elevati di *betweenness*.

Concludendo l'analisi di correlazione, è emerso che i valori di *proximity* e di *closeness* degli utenti producono lo stesso ranking e inoltre, osservando i risultati delle aggregazioni basate su queste due metriche ci si accorge che sono pressoché identiche presentando delle differenze dell'ordine del centesimo. Da questi dati si potrebbe pensare di utilizzarne solo una in modo da alleggerire il carico computazionale.

5.4.2 Deviazione standard dei voti

Tornando ad analizzare la Figura 5.4 è possibile notare come i valori siano tutti compresi nell'intervallo (3,5; 4) mostrando solo piccole variazioni introdotte dai diversi metodi di aggregazione. La causa di questi risultati può essere spiegata solo con un'analisi sulla distribuzione del valore di sentiment dei 327 utenti che hanno fornito dei post su Londra nel forum Thorn Tree di Lonaly Planet (Figura 5.6).

Figura 5.6: Distribuzione dell'opinione utenti

Una grande quantità di utenti ha un giudizio della metropoli britannica compreso tra 3 e 5: tra un parere neutrale e un parere molto positivo. Per l'esattezza il 29% dei voti tra 3 e 4 e il 57% tra un giudizio positivo (4) e uno molto positivo (5). A questo punto ci si chiede se la scarsa variabilità dei voti influisca in qualche modo nei risultati di Figura 5.4 e se le aggregazioni illustrate possano essere di un qualche aiuto. Per fare questo è necessario valutare la distribuzione dei voti per poi calcolarne la deviazione standard e determinare se le nostre analisi

applicate a un dataset con voti più distribuiti possa fare la differenza e cambiare la polarità delle analisi. In Figura 5.7 mostriamo la distribuzione dei voti di tutti i post su Londra.

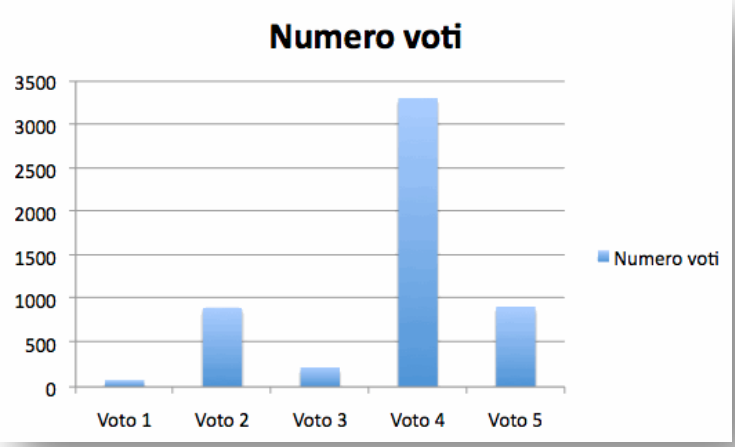

Figura 5.7: Distribuzione voti Londra

La Tabella 5.9 riporta i valori numerici con il calcolo della deviazione standard e la media di tutti i voti.

	Numero voti
Voto 1	78
Voto 2	907
Voto 3	223
Voto 4	3312
Voto 5	919
Deviazione standard	0,97
Media voti	3,75

Tabella 5.9: Deviazione standard Londra

I valori di Tabella 5.4 si discostano in modo più marcato per le analisi su TripAdvisor. Utilizzando prima l'*eccentricity* e poi la *betweenness* abbiamo una differenza di 0,13 equivalente al 13% della deviazione

standard. Con una deviazione standard di 0,97, è probabile che molti degli opinion leader abbiano anch'essi un valore di sentiment molto vicino alla media dei voti. In Tabella 5.10 si mostra il sentiment per i 20 utenti appartenenti a Lonely Planet con livelli di *indegree* più elevati.

Pos	Username	Valore metrica	Sentiment su Londra
1	Tony_b	0,26	2,93
2	Badger1492	0,24	3,50
3	Belsa	0,19	3,50
4	alanR	0,09	3,60
5	MTL	0,06	4,00
6	stevegerms	0,05	4,50
7	Voyager_2002	0,05	3,30
8	timothysolberg	0,04	3,90
9	newone	0,04	3,73
10	qwovadis	0,02	4,44
11	Ria	0,02	3,68
12	scaryant	0,02	3,54
13	RealHerrBert	0,02	4,00
14	greencelery	0,02	3,53
15	Sugoi	0,02	3,55
16	mr_rush	0,02	3,78
17	Nerb	0,02	3,21
18	njarratt	0,02	3,20
19	RAK	0,02	3,70
20	Man_in_Seat_61	0,02	3,50
		Media	3,65

Tabella 5.10: Opinion leader (Indegree)

Le caselle evidenziate indicano che l'opinione dell'utente cade vicino alla media dei voti di Tabella 5.9. Tra i primi 20 utenti con più alto valore della metrica *indegree*, ben 14 hanno un sentiment compreso tra 3,5 e 4. Si spiega così il motivo per cui le varie tipologie di aggregazione hanno valori simili tra loro nonostante si prendano in considerazione, ogni volta, caratteristiche diverse della rete e più in particolare degli

utenti. La stessa analisi può essere svolta per ognuna delle metriche utilizzate.

Ora torniamo a valutare la distribuzione dei voti anche per Madrid e il calcolo della deviazione standard così come abbiamo fatto per Londra. In Figura 5.8 e Tabella 5.11 sono riportati i risultati per quanto riguarda Madrid.

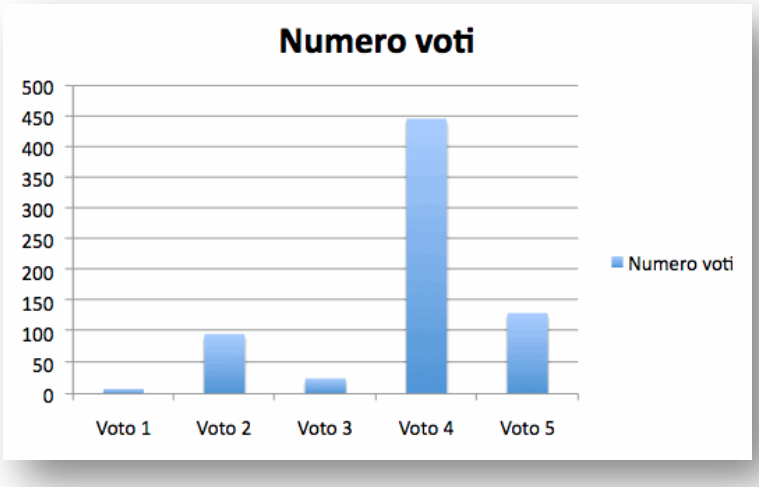

Figura 5.8: Distribuzione voti Madrid

	Numero voti
Voto 1	8
Voto 2	97
Voto 3	25
Voto 4	446
Voto 5	131
Deviazione standard	0,92
Media voti	3,84

Tabella 5.11: Deviazione standard Madrid

Confrontando i risultati con quelli di Figura 5.5, si può concludere che in questo caso, i vari tipi di aggregazione hanno un peso decisamente

rilevante nella determinazione del sentiment. Madrid presenta una differenza massima tra i valori di aggregazione pari a 52% della deviazione standard (0,48). Infine, in Tabella 5.12 e Figura 5.9 mostriamo i risultati anche per Milano così come abbiamo fatto in precedenza per le altre due città.

Milano			
Tipo Aggregazione		**Lonely Planet**	**TripAdvisor**
Standard		3,65	3,66
Per Utente		3,80	3,60
Pesatura utente	**Indegree**	3,87	3,68
	Proximity	3,82	3,59
	Outdegree	3,89	3,67
	Eccentricity	3,83	3,60
	Closeness	3,83	3,60
	Betweenness	3,97	3,85

Tabella 5.12: Aggregazioni Milano

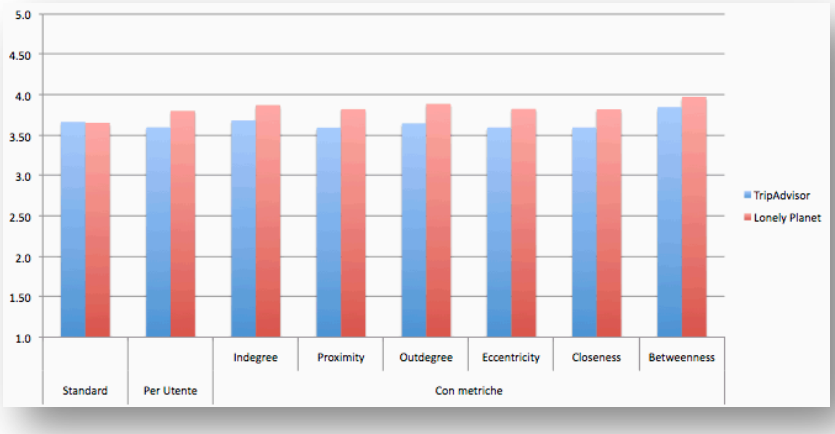

Figura 5.9: Grafico aggregazioni Milano

Diversamente dalle altre analisi, Lonely Planet mostra una maggiore positività rispetto a TripAdvisor anche se le differenze in questo caso sono molto piccole. Il grafo della distribuzione dei voti è riportato in Figura 5.10 mentre i dati delle deviazioni standard sono riportati in Tabella 5.13.

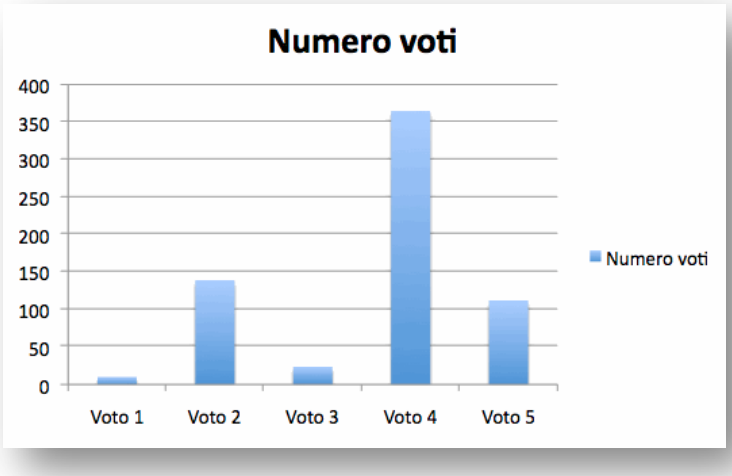

Figura 5.10: Distribuzione voti Milano

	Numero voti
Voto 1	10
Voto 2	139
Voto 3	24
Voto 4	365
Voto 5	112
Deviazione standard	1,04
Media voti	3,66

Tabella 5.13: Deviazione standard Milano

Per tutte le città analizzate si sono registrati degli scostamenti piccolissimi applicando le varie aggregazioni, ma con uno studio approfondito considerando la variabilità intrinseca del fenomeno e il bias di positività, è emerso che per il dataset di voti a nostra disposizione le

variazioni riscontrate dalla nostra analisi, anche se piccole erano rilevanti. In sintesi riportiamo in Tabella 5.14 i valori percentuali rispetto alla deviazione standard dei nostri scostamenti.

Città	% Dev. Standard
Londra	13%
Madrid	52%
Milano	31%

Tabella 5.14: Scostamenti rispetto alla deviazione standard

In conclusione, applicando una delle aggregazioni presentate, il valore di sentiment può cambiare sensibilmente, in relazione al fenomeno. È necessario prestare molta attenzione a ciò che ognuna di esse significa per non incorrere in errori di valutazioni. Un'altra considerazione da fare riguarda la distribuzione dei voti. Se si osservano i grafici relativi a tali distribuzioni si vede che hanno un andamento molto simile e presentano una grande quantità di voti positivi. La maggior parte degli utenti preferisce raccontare online di fatti, avvenimenti positivi, quali aver trascorso una bella giornata o essere contenti dell'acquisto effettuato, perché in fondo tutti fanno "marketing di se stessi", tutti cercano di dare agli altri una buona immagine di sé. Questo tipo di comportamento non fa altro che rendere più difficile le analisi sulle varie fonti e quindi sviluppi futuri di queste analisi deve tenere in considerazione questo fatto per cercare di correggerlo e rendere l'analisi del sentiment più efficace.

5.4.2 Categorizzazioni

Nella Sezione 5.4.1 sono stati presentati i risultati di sentiment per le città di Londra, Madrid e Milano utilizzando i metodi di aggregazione basati sugli utenti. Gli scostamenti dei risultati erano limitati a causa della forte positività espressa dagli utenti registrando gli stessi livelli di reputation. Per avere degli scostamenti più elevati è necessario scendere a una granularità più fine di dettaglio, analizzando per ogni

città le diverse categorie riportate in Tabella 5.15. Il punto debole di questa analisi è che molte categorie contengono un numero di post limitato. La categorizzazione utilizzata per queste analisi non è riuscita a categorizzare una grande quantità di post assegnandoli così alla categoria General. In Tabella 5.15 riportiamo i numeri di post e di utenti per tutte le categorie analizzate.

Città	Categoria	TripAdvisor		Lonely Planet	
		n° utenti	n° post	n° utenti	n° post
Londra	**Arts&Culture**	153	448	63	118
	Events&Sports	16	18	5	5
	Life&Entertainment	11	15	6	6
	Services&Transport	280	1692	161	506
	Weather&Environmental	107	280	53	78
	Fashion&Shopping	32	53	15	21
	Food&Drink	135	363	64	159
	Night&Music	15	21	12	16
	Ticket	94	252	53	119
Madrid	**Arts&Culture**	12	123	10	18
	Events&Sports	9	17	1	1
	Life&Entertainment	7	28	1	1
	Services&Transport	25	138	22	47
	Weather&Environmental	7	27	7	15
	Fashion&Shopping	24	64	4	4
	Food&Drink	9	91	8	11
	Night&Music	7	22	1	1
	Ticket	53	119	5	6
Milano	**Arts&Culture**	19	54	25	40
	Events&Sports	5	9	1	1
	Life&Entertainment	4	4	0	0
	Services&Transport	25	172	38	82
	Weather&Environmental	4	15	5	11
	Fashion&Shopping	7	15	5	11
	Food&Drink	10	41	10	14
	Night&Music	0	0	0	0
	Ticket	6	31	8	8

Tabella 5.15: Volumi per categorie

Alcune categorie presentano un numero di post molto piccolo mentre altre sono ben commentate. Ciò che ci interessa è trovare, utilizzando una granularità più fine, dei risultati più marcati e consistenti, causando dei cambi di sentiment tra i diversi metodi di calcolo.

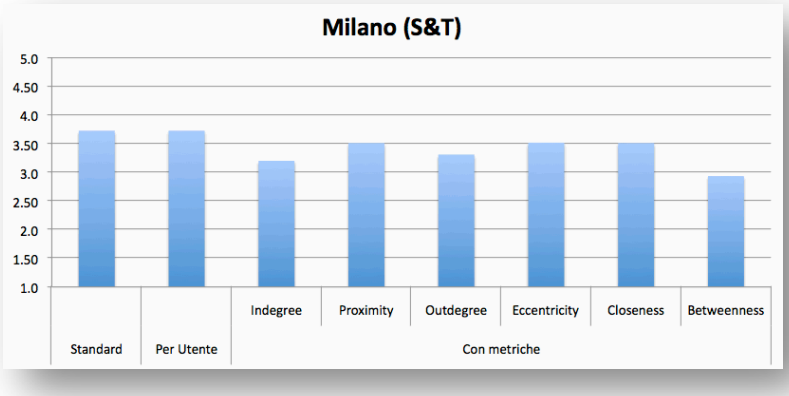

Figura 5.11: Aggregazioni Milano (S&T)

In Figura 5.11 si vede che, cambiando la metodologia di calcolo del sentiment, possiamo ottenere un valore positivo (valore 4) per le aggregazioni standard e per utente, mentre utilizzando le pesature della *betweenness* otteniamo un sentiment neutrale (valore 3). La Figura 5.12 riporta delle differenze ancora più marcate e il valore di sentiment passa da un'opinione molto positiva (valore 5) utilizzando la *betweenness* a un valore neutro con l'aggregazione Standard.

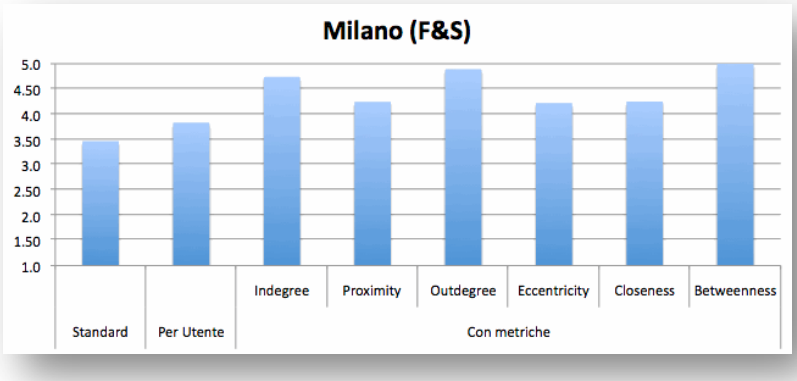

Figura 5.12: Aggregazioni Milano (F&S)

I due grafici però mostrano un dato importante. Osservando la Tabella 5.15, per Milano S&T il numero dei post è 172 prodotti da 25 utenti, mentre per Milano F&S le aggregazioni sono state calcolate su un campione molto più ristretto composto da 15 post appartenenti a 7 utenti. Il primo risultato (Figura 5.11) offre una valutazione più attendibile poiché il secondo è calcolato su un campione troppo piccolo e il minimo errore di calcolo può variare notevolmente i valori rendendo il risultato inesatto.

5.4.3 Opinion Leader

In questa sezione si vuole condurre uno studio sugli opinion leader e trarre delle conclusioni sul loro giudizio: se si discosta dal resto degli utenti oppure è in linea con esso. Per l'analisi utilizzeremo la *betweenness*, una delle metriche più importanti secondo il parere di molti. Nelle Figure 5.13 e 5.14 riportiamo le analisi fatte su Madrid e Milano.

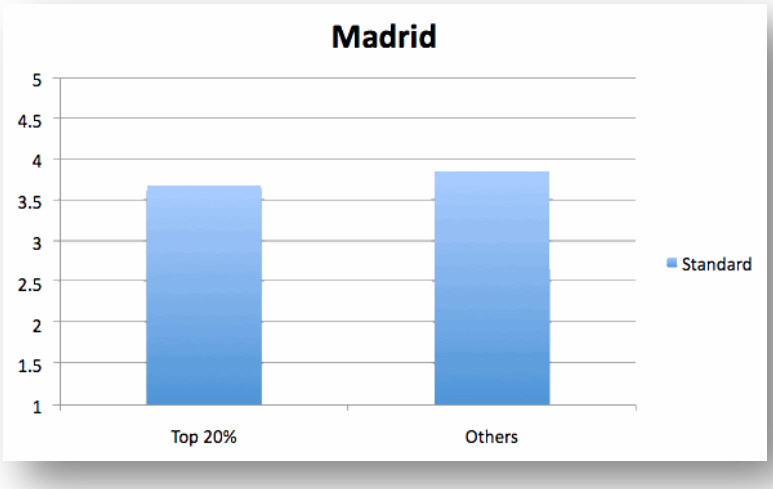

Figura 5.13: Influencer Madrid

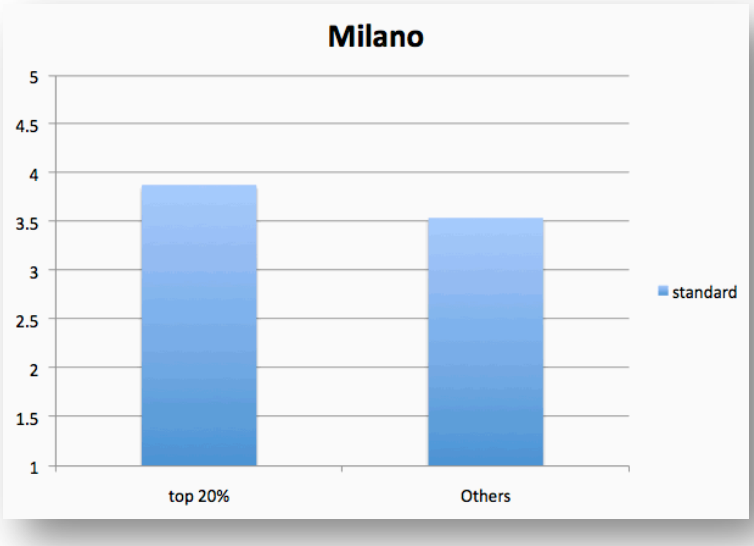

Figura 5.14: Influencer Milano

Un utente è considerato un opinion leader se rientra nel 20% degli utenti con valori di *betweenness* più elevati. Il risultato conferma l'allineamento tra i due gruppi di utenti e con buona approssimazione possiamo

affermare che analizzando gli utenti più significativi si hanno delle buone informazioni sull'opinione dell'intera community.

5.4.4 Valutazione del silenzio

Una delle analisi più importanti e che può suscitare un notevole interesse è la valutazione del silenzio. Un tale studio può essere un ottimo termine di paragone tra due o più brand oppure, nel nostro caso specifico, tre città. Scoprire quanto si parla o non si parla di una particolare caratteristica di un prodotto o città, è un'informazione che arricchisce la valutazione del sentiment. Per una città, le varie caratteristiche che un utente può pensare di analizzare sono le categorie di Tabella 5.15. Il primo passo da compiere è quello di calcolare quanto si parla di una piuttosto che dell'altra città sulle fonti a nostra disposizione (Tabella 5.16).

Città	n° post TA	n° post LP
Londra	3952	1488
Madrid	557	150
Milano	416	234

Tabella 5.16: Volumi di post per città

Le città di Milano e Madrid hanno dei volumi che si aggirano intorno al 10% rispetto a quelli di Londra. Se si vuole valutare i punti di forza o di debolezza di una città sulla base dei volumi prodotti, un confronto senza tenere conto delle proporzioni di Tabella 5.16 sarebbe improprio e nettamente sbilanciato in favore della capitale inglese. Valutare il silenzio significa studiare se ci sono delle categorie in cui, per qualche ragione, le proporzioni dei volumi vengono a mancare in favore di una o l'altra città. In altre parole se di Londra si parla 10 volte più di Madrid è plausibile che la stessa proporzione si mantenga all'incirca anche per le categorie di Tabella 5.15.

In Figura 5.15 è riportato il grafico di valutazione del silenzio prendendo come riferimento Londra. Come si può vedere per molte categorie la proporzione iniziale non viene mantenuta e le città di Milano e Madrid sembrano ottenere un maggior interesse da parte degli utenti.

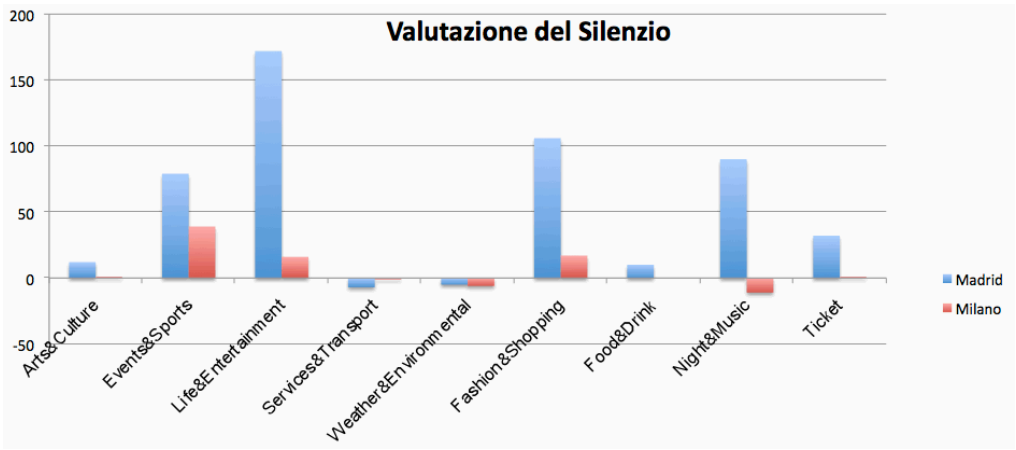

Figura 5.15: Valutazione del silenzio

Questa analisi fa riferimento unicamente ai volumi e di conseguenza il fatto che si parli più o meno di qualcosa può essere positivo o negativo. La valutazione del silenzio fine a se stessa non dice nulla; è necessario affiancare tale analisi ai risultati di sentiment per valutare se i volumi espressi manifestano delle lamentele oppure esperienze positive. Dalle analisi discusse nella Sezione 5.4.1 capita spesso che il sentiment abbia lo stesso andamento da città a città. In questo caso la valutazione del silenzio può fornire un nuovo indice di valutazione e di comparazione molto importante.

Purtroppo l'assegnazione delle categorie ai post presenta notevoli carenze e fino a quando non si riuscirà ad aumentare la precisione di questa operazione l'analisi presentata risulta pressoché inutilizzabile. La Figura 5.15 omette una colonna molto importante, che è la categoria General. All'interno di questa categoria finiscono tutti i commenti di cui

non si è riusciti a determinare la categoria di appartenenza. Una grande quantità dei post su Londra sono stati messi in questa categoria, di conseguenza i risultati hanno una distorsione intrinseca che favorisce le altre due città.

5.5 Pesatura delle fonti

Dopo aver calcolato tutte le aggregazioni non resta che applicare l'ultimo passo del nostro metodo di analisi: la pesatura delle fonti. Come abbiamo presentato nel Capitolo 4, la pesatura delle fonti prende in considerazione tre caratteristiche:

- Traffic;
- Participation;
- Time.

Ognuna di queste comprende diverse metriche i cui valori sono ricavati sia da siti specializzati nel fornire statistiche come Alexa, sia attraverso delle query complesse su i database a nostra disposizione. Una volta ottenute tutte le metriche è necessario normalizzarle su un intervallo compreso tra 0 e 1. Nella Sezione 4.3.2 sono stati presentati tre metodi di normalizzazione. In questa trattazione useremo la normalizzazione continua, anche se va sottolineata la validità delle altre due. Per maggiore chiarezza riportiamo l'esempio già trattato nella Sezione 4.3.2. Supponiamo di avere dieci fonti con i seguenti valori della metrica A:

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
20	10	15	13	2	5	20	8	4	9

Tabella 5.17: Valori di esempio metrica M

Per normalizzare i valori non si fa altro che costruire un intervallo prendendo come limite superiore il massimo valore misurato, nel nostro

esempio 20, e come valore minimo lo 0. Con delle proporzioni ricaviamo il valore delle restanti metriche così come riportato nella Figura 5.16.

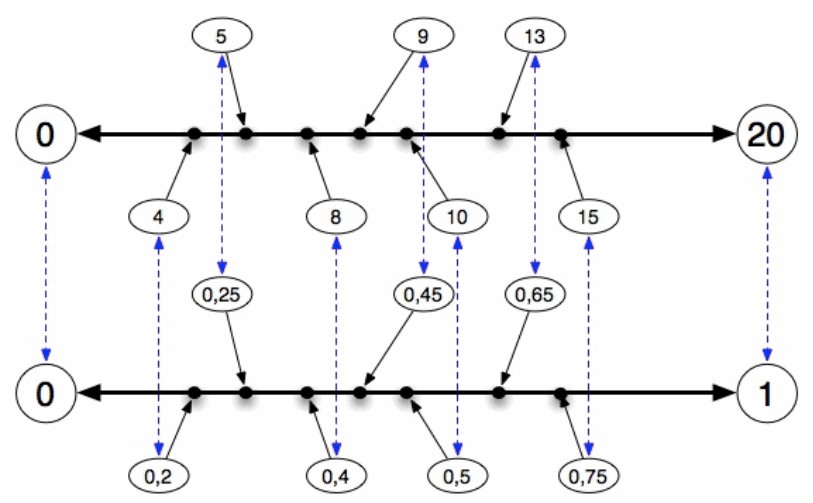

Figura 5.16: Normalizzazione continua

Va sottolineato che questa procedura non è valida su tutte le metriche. Ad esempio quando la metrica è espressa attraverso delle percentuali, come ad esempio il Bounce Rate, il 100% equivale allo 0 e lo 0% a 1; viceversa nel caso dei "Daily visitors" al 100% è associato il valore 1 e allo 0% il valore 0.

Prima di applicare la pesatura al sentiment già calcolato presentiamo i valori delle metriche divisi per i tre costrutti appena elencati.

Traffic

Il *traffic* viene calcolato utilizzando 5 metriche opportunamente normalizzate come descritto in precedenza. In Tabella 5.18 sono riportati sia i valori reali che quelli già normalizzati per le fonti Lonely Planet e TripAdvisor.

Metrica	Lonely Planet		Trip Advisor	
	r.v	Norm	r.v.	Norm
Traffic rank	1840	0,82	268	0,97
Daily Visitors	0,076	0,2	0,376	1
Daily page views	0,0025	0,15	0,0174	1
Number of inbound links	13698	0,79	17343	1
Number of open discussion	627865	0,19	3257353	1

Tabella 5.18: Valori di Traffic

Elenchiamo di seguito il significato dei valori delle metriche e il modo in cui sono state calcolate:

- **Traffic Rank**: è un indice calcolato dal sito di Alexa e non è altro che un loro rank interno che prende in considerazione diverse caratteristiche di traffico opportunamente combinate che ne determina la posizione rispetto agli altri siti;
- **Daily Visitors**: è la percentuale degli utenti presenti su internet a livello globale che visita i siti in questione;
- **Daily page views**: percentuale delle pagine visitate sulle fonti analizzate, rispetto alle pagine viste a livello globale;
- **Number of inbound links**: semplicemente il numero dei link diretti verso la fonte;
- **Number of open discussion**: numero di discussioni aperte sulla fonte ottenuto attraverso il crawling delle pagine e calcolato tramite delle query sui database.

In Figura 5.17 riportiamo il grafico delle metriche normalizzate comparando i due siti di interesse.

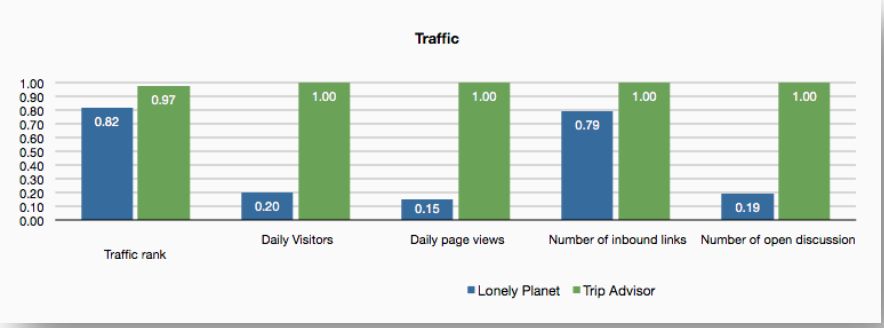

Figura 5.17: Grafico valori Traffic

Participation

La *participation*, a differenza del *traffic*, è calcolata utilizzando tre metriche. Il calcolo delle metriche è basato su query complesse eseguite su dati ottenuti tramite una procedura di crawling. In Tabella 5.19 riportiamo i valori.

Metrica	**Lonely Planet**		**Trip Advisor**	
	r.v	**Norm**	**r.v.**	**Norm**
Number of new discussion opened per day	7,27	0,91	8,04	1
Number of comments per discussion	2,76	0,67	4,1	1
Average number of comments to post provided within 24 H	2,02	0,74	2,73	1

Tabella 5.19: Valori di Partecipation

Dove le metriche indicano:

- **Number of new discussion opened per day**: è la frequenza con cui gli utenti iniziano nuove discussioni sui siti di interesse. Per ottenere tale valore si è reso necessario eseguire delle query

sui database, ricostruendo le discussioni e valutandone la data di inizio. Per fare in modo che la metrica indicasse un valore attuale si è scelto di effettuare la media delle discussioni aperte ogni giorno solo per i primi sei mesi del 2010;

- **Number of comments per discussion**: il numero di commenti per discussioni è stato calcolato come la metrica precedente, attraverso query ai database e limitando il calcolo della media alle discussioni della prima metà del 2010;
- **Average number of comments to post provided within 24H**: questa metrica indica la reattività della community a nuovi topic. Come per le altre metriche, ricostruendo le discussioni e tenendo traccia della data di inizio, sempre tramite query, siamo stati in grado di determinare quanti commenti sono stati forniti nell'arco delle prime 24 H. Anche in questo caso abbiamo ristretto il calcolo alle discussioni iniziate nei primi 6 mesi del 2010.

In Figura 5.18 viene riportato il grafico per le metriche che compongono la *participation*.

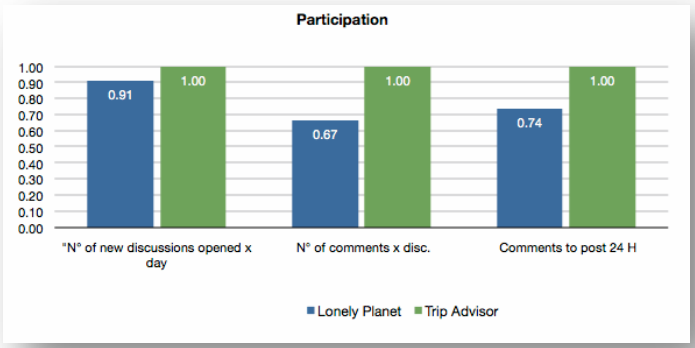

Figura 5.18: Grafico valori di Partecipation

Time

Il tempo speso sul sito è sicuramente un indice di importanza e di gradimento dei contenuti. Poiché i volumi delle visite sono spesso influenzate da un posizionamento favorevole nei motori di ricerca, oppure dalla presenza di molti link sparsi per il web, il bounce rate e il tempo medio trascorso sul sito possono limitare questi elementi di disturbo. In Tabella 5.20 e Figura 5.19 presentiamo i valori delle metriche di Time.

Metrica	Lonely Planet		Trip Advisor	
	r.v	Norm	r.v.	Norm
Average time spent on site	4,03	0,91	4,42	1
Bounce rate	52,2	0,48	41,8	0,58

Tabella 5.20: Valori di Time

Dove:

- **Average time spent on site**: è la media del tempo, espresso in minuti, che ogni utente trascorre sul sito una volta che ha effettuato l'accesso;
- **Bounce rate**: è la percentuale degli utenti che lascia il sito dopo aver visualizzato la prima pagina.

Entrambe le metriche di questo costrutto sono fornite dal sito di statistiche Alexa.

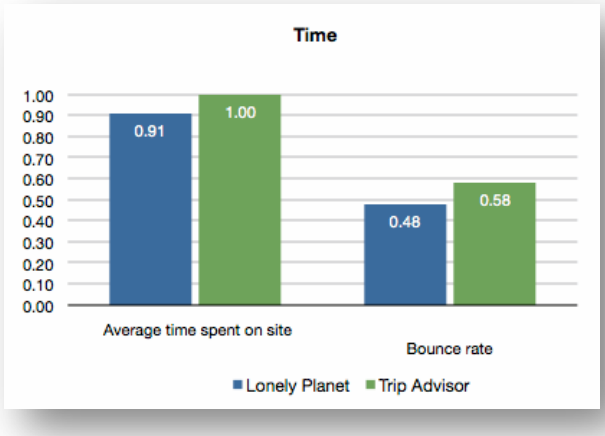

Figura 5.19: Grafico valori di Time

5.5.1 Aggregazioni con pesatura utenti delle fonti

Come mostrato in Figura 4.4 del capitolo precedente l'aggregazione con pesatura della fonte si compone di 5 passi: calcolo della metrica, normalizzazione, aggregazione per costrutto, sentiment pesato e pesatura costrutti.

In precedenza sono stati riportati i risultati dei primi due passaggi mentre ora passiamo ad aggregare le metriche per costrutto in modo da avere un unico valore che descriva i livelli di *Traffic*, *Participation* e *Time* per ogni fonte. Il calcolo consiste nella media delle metriche che appartengono ad un costrutto. La Tabella 5.20 riporta i valori per entrambe le fonti.

Fonte	Traffic	Participation	Time
Lonely Planet	0,43	0,77	0,69
Trip Advisor	0,994	1	0,79

Tabella 5.21: Valori delle metriche delle fonti

Dai risultati si osserva che TripAdvisor ottiene risultati migliori rispetto a Lonely Planet su tutti e tre i costrutti; il traffico di utenti è più elevato e questi partecipano maggiormente alle discussioni e in maniera più tempestiva, passando maggior tempo sul sito rispetto a quelli iscritti al forum di Lonely Planet.

Il passo successivo è quello del calcolo del sentiment per costrutto. Utilizzando i valori di sentiment calcolati nella Sezione 5.4 e i valori dei costrutti appena calcolati, si procede con il calcolo del sentiment con pesatura delle fonti per alcune delle analisi fatte in precedenza.

Anche con la pesatura per fonte non si hanno grandi variazioni nel sentiment, sia perché le opinioni provenienti dai due siti sono simili, sia perché Lonely Planet e TripAdvisor sono dei brand forti e raggiungono dei livelli di importanza che non si discostano molto tra loro. Confrontando i risultati con la deviazione standard dei voti si ha una variazione che è vicina al 10%. In Figura 5.20, 5.21, e 5.22 sono riportate rispettivamente le analisi per Madrid, Milano e Londra. Come in precedenza, per avere dei cambiamenti significativi è necessario aumentare la granularità e passare ad analizzare le singole categorie in modo separato. In Figura 5.23 per completezza riportiamo un esempio della categoria Fashion&Shopping su Milano. Come si può vedere i valori di sentiment sono molto più vicini agli andamenti di TripAdvisor, soprattutto per quanto riguarda la pesatura del *traffic* poiché come si osserva dalla Tabella 5.21 i livelli di traffico sul sito di TripAdvisor sono

molto più alti rispetto a quelli di Lonely Planet mentre per le altre metriche le due fonti si assomigliano.

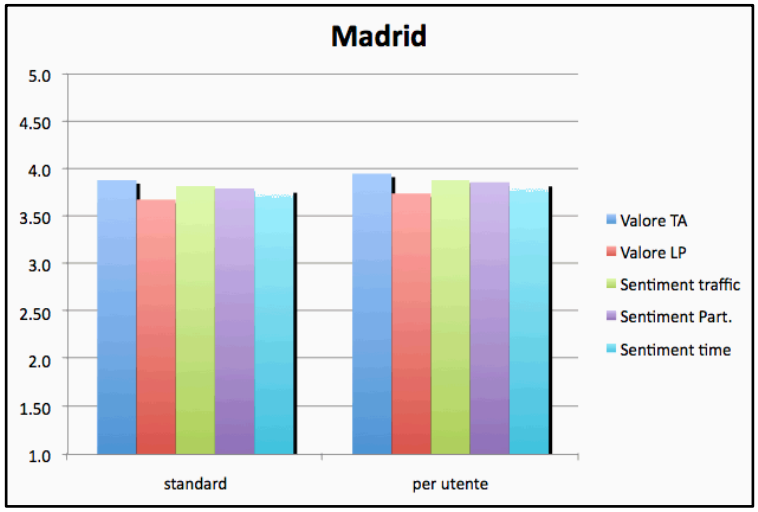

Figura 5.20: Pesatura fonti Madrid

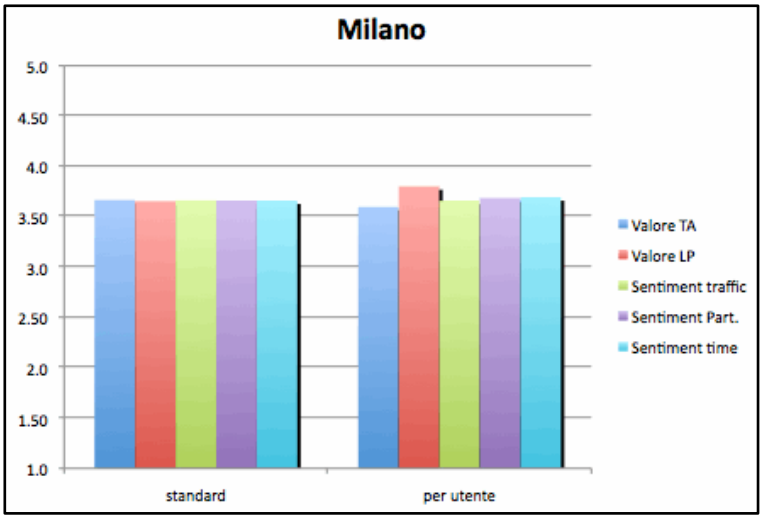

Figura 5.21: Pesatura fonti Milano

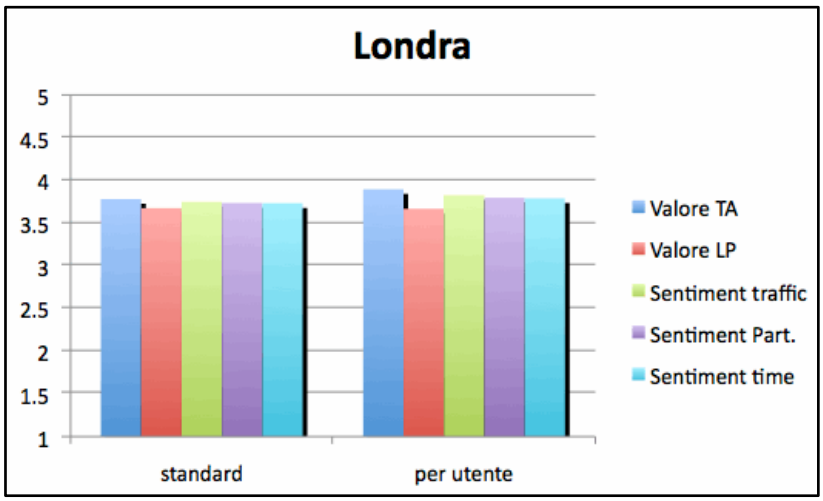

Figura 5.22: Pesatura fonti Londra

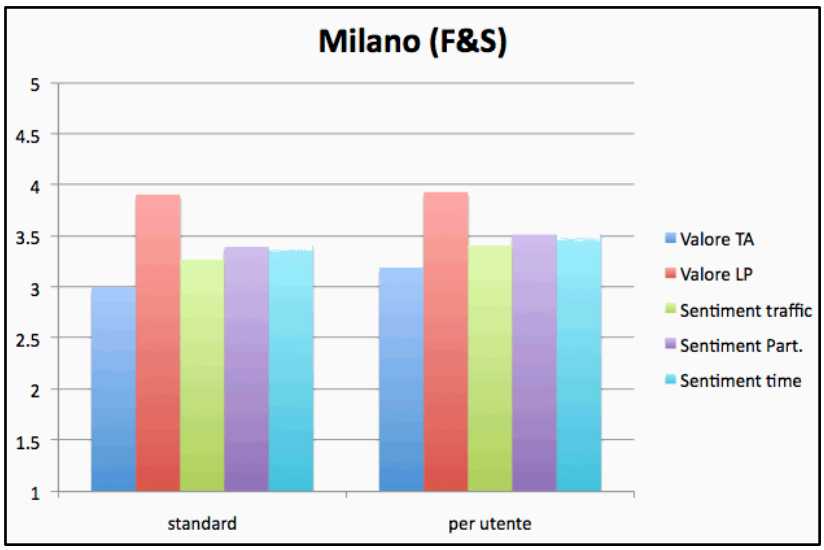

Figura 5.23: Pesatura fonti Milano (F&S)

Per dimostrare quanto la variazione dei valori di queste metriche possa essere significativo quando si tratta di pesare diverse sorgenti informative, proviamo a valutare due diversi siti web che vengono ritornati da Google quanto si inseriscono le parole di ricerca

"monuments in Milan blog OR forum": TripAdvisor e AboutMilan. La Tabella 5.22 riporta i valori normalizzati per le due fonti. Per semplicità non tratteremo il problema di pesare in modo diverso i tre costrutti, anche se è chiaro che un decisore, che è più interessato nella *partecipation* piuttosto che il *traffic*, può decidere di pesarli in modo diverso a seconda delle sue preferenze.

Metric	TripAdvisor	AboutMilan
Traffic	0,99	0,13
Participation	0,99	0,04
Time	0,79	0,47
Overall	0,93	0,21

Tabella 5.22: Valori costrutti di TA e AM

Nei tool presenti sul mercato, il sentiment è calcolato come il valore medio di fonti diverse, attraverso la seguente espressione.

$$Sentiment_{std} = \frac{Sent_{TA} + Sent_{AM}}{2}$$

dove $Sent_{TA}$ e $Sent_{AM}$ sono i due valori di sentiment osservati in TripAdvisor e AboutMilan rispettivamente. Usando i risultati di Tabella 5.22, è possibile applicare una media pesata:

$$Sentiment_{ranked} = \frac{w_{TA} Sent_{TA} + w_{AM} Sent_{AM}}{w_{TA} + w_{AM}}$$

dove $Sent_{TA}$ e $Sent_{AM}$ sono i due valori di sentiment osservati in TripAdvisor e AboutMilan, mentre w_{TA} e w_{AM} sono le due metriche aggregate per TripAdvisor e AboutMilan.

Consideriamo come esempio di un insieme di frasi ottenute dalle due sorgenti, s_{TA} con un valore di sentiment uguale a 3 e s_{AM} con un sentiment medio pari a 4.1. Applicando le formule otterremo i seguenti risultati:

- $Sentiment_{standard}$ = 3,6 (arrotondato a 4);
- $Sentiment_{rande}$ = 3,2 (arrotondato a 3).

Questo esempio mostra come, applicando un sistema di pesatura i risultati possano cambiare in modo sostanziale.

5.6 Discussione dei risultati

Gli studi svolti in questo capitolo ci hanno permesso di trarre delle conclusioni sulla valutazione del sentiment e sulle varie tipologie di analisi che possono essere svolte. Il primo risultato nasce dall'analisi di correlazione tra le metriche in cui si sono evidenziati due gruppi altamente correlati. Il primo gruppo è formato dalla *proximity*, dalla *closeness* e dall'*eccentricity,* in cui le prime due sembrano esprimere lo stesso concetto di importanza. Questi risultati suggeriscono l'utilizzo soltanto di una delle due metriche in modo da ridurre il carico computazionale. Il secondo gruppo di metriche formato da *indegree*, *outdegree* e *betweenness,* oltre ad avere una forte correlazione tra di loro, presentano una correlate anche con i destination expert, utenti ritenuti importanti dal sito perché mostrano elevati livelli di partecipazione e coinvolgimento. Secondo Freeman [37] la *betweenness* è la metrica per eccellenza che descrive l'importanza di un utente all'interno di una rete ed è interessante che questa trovi elevati livelli di correlazione con le metriche locali come la *indegree* e la *outdegree* che sono per altro le più facili da calcolare soprattutto le più facili da esportare nell'analisi di altri tipi di social media come micro-blogging, ecc..

Successivamente abbiamo valutato i metodi di aggregazione del sentiment, registrando delle piccole variazioni tra i risultati ottenuti. Dopo una valutazione della deviazione standard, tali scostamenti, seppur limitati, si sono dimostrati significativi. Da questi risultati è emerso che

applicando le diverse tipologie di aggregazione il risultato di sentiment può cambiare sensibilmente per questo è necessario fare attenzione al significato di ogni aggregazione e a cosa si vuole misurare.

Utilizzando i valori delle metriche e i risultati di sentiment, abbiamo comparato l'opinione degli utenti più centrali con quelli meno centrali osservando che i due gruppi sono allineati e analizzando il 20% degli utenti più centrali si ha una buona approssimazione sull'opinione dell'intera community.

Anche la valutazione del silenzio, seppur non utilizzabile per i motivi discussi nella Sezione 5.4.4, si è dimostrata importante per aggiungere maggiore espressività all'analisi della reputation e in futuro potrà ricoprire un ruolo fondamentale negli applicativi di mercato.

Infine, anche la pesatura delle fonti si è dimostrata essere un fattore molto importante nello studio della reputation. Dai risultati ottenuti nella Sezione 5.5 abbiamo dimostrato la rilevanza dell'utilizzo delle metriche di importanza nel calcolo del sentiment, specialmente quando le fonti hanno valori non uniformi come nell'esempio di TripAdvisor e AboutMilan.

Capitolo 6

Conclusioni

L'obiettivo di questo lavoro di tesi è di fornire una metodologia per lo studio della reputazione sul web tenendo conto di diversi fattori come la centralità degli utenti e l'importanza di una sorgente informativa.

Nel Capitolo 2 sono stati presentati diversi tool che implementano la sentiment analysis sottolineando l'interesse del mercato a questo tipo di analisi. Va però detto che nessuno di questi tool raggiunge un perfetto livello di accuratezza e spesso necessitano ancora dell'intervento umano.

Il lavoro di tesi ha introdotto due principali contributi innovativi: (i) un framework di valutazione delle sorgenti informative e degli utenti specifici; (ii) un tool, che prendendo le informazioni crawlate, come post e informazioni sugli utenti, ricostruisce la rete sociale e calcola tutte le metriche di reputazione per ogni utente. Queste informazioni ci permettono di calcolare i valori di sentiment secondo diverse topologie di aggregazione.

I risultati ottenuti ci hanno permesso di trarre delle conclusioni sul sentiment proveniente dalle varie sorgenti informative e sulla variazione di questo, prendendo in considerazione le varie metriche di reputazione. In particolare nella Sezione 5.5 abbiamo dimostrato la rilevanza nell'utilizzo delle metriche d'importanza nel calcolo del sentiment, specialmente quando le fonti hanno valori non uniformi come

nell'esempio di TripAdvisor e AboutMilan, in cui il valore di sentiment passa da una valutazione neutrale a una valutazione positiva utilizzando la metodologia proposta.

Inoltre, il calcolo della centralità di ogni utente ci ha permesso di paragonare l'opinione degli influencer con quella degli utenti più comuni, osservando che i due gruppi sono allineati, e gli utenti più centrali sintetizzano bene l'opinione dell'intera community.

Uno delle conclusioni più importanti deriva dallo studio dei metodi di aggregazione del sentiment. Applicando tali metodi, si sono registrate delle variazioni tra i risultati ottenuti. Dopo una valutazione della deviazione standard, tali scostamenti, seppur limitati in senso assoluto, si sono dimostrati significativi in relazione alla variabilità dei fenomeni osservati. Da questi risultati è emerso che applicando le diverse tipologie di aggregazione il risultato di sentiment può cambiare sensibilmente per questo è necessario fare attenzione al significato di ogni aggregazione e a cosa si vuole misurare.

La limitazione di questa tesi è di non aver compiuto un testing estensivo. Uno sviluppo futuro potrebbe essere quello di estendere i casi di test del framework aggiungendo altre sorgenti informative e nuove analisi sul grado di specializzazione di una fonte.

Un possibile studio futuro potrebbe riguardare l'evoluzione del sentiment tenendo conto anche della variabile tempo. Valutare se e in che modo gli utenti più importanti riescono a modificare l'opinione generale delle community; verificando se una differenza di opinione degli influencer rispetto al resto degli utenti, porti ad una evoluzione del sentiment generale che, con il passare del tempo, si avvicini sempre più a quello degli influencer.

Bibliografia

[1] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of RANLP*. Citeseer, 2005.

[2] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440, 2007.

[3] E. Breck, Y. Choi, and C. Cardie. Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[4] S. R. Das and M. Y. Chen. Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9): 1375-1388, 2007.

[5] M. Gamon. Sentiment classification on custode feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International conference on Computational Linguistics*, page 841. Association for Computational Linguistics, 2004.

[6] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. Citesser, 2007.

[7] A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2): 110-125, 2006.

[8] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International conference on Knowledge capture*, page 77. ACM, 2003.

[9] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327-335. Association for Computational Linguistics, 2006.

[10] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational linguistics*, 30(3): 277-308, 2004.

[11] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceeding of EMNLP*, volume 4, pages 412-418, 2004.

[12] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM International conference on Information and knowledge management, page 631. ACM, 2005.

[13] Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International conference on World Wide Web*, page 180. ACM, 2007.

[14] Y. Liu, X. Huang, A. An, and X. Yu. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual International ACM SIGIR conference on Reach and development in information retrival*, page 614. ACM, 2007.

[15] Q. Su, X. Xu, H. Guo, Z. Guo, X. Zhang, B. Swen, and Z. Su. Hidden sentiment association in chinese web opinion mining. In *Proceeding of the 17th International conference on World Wide Web*, pages 959-968. ACM, 2008.

[16] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Association for Computational Linguistics (ACL-08: HLT),* pages 308-316, Columbus, Ohio, USA, June 2008.

[17] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In ICDM-03, *the IEEE International Conference on Data Mining*, 2003.

[18] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. Advances in Intelligent Data Analysis VI, pages 121-132, 2005.

[19] N. Jindal and B. Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual International ACM SIGIR conference on Research and development in information retrieval*, pages 244-251. ACM, 2006.

[20] N. Jindal and B. Liu. Mining comparative sentences and relations. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1331. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[21] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, page 1190. ACM, 2007.

[22] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219-230. ACM, 2008.

[23] F. Harary, R.Z. Norman, and D. Cartwright. Structural Models: An introduction to the theory of directed graph. Wiley, New York, 1965.

[24] D. Cartwright. Studies in social power. University of Michigan, 1959.

[25] K. K. S. Chung, L. Hossain, and J. Davies. Exploring sociometric and egocentric approaches for social network analysis. In *Proceedings of International Conference on Knowledge Management Asia Pacific*, pages 1-8, 2005.

[26] J. Barnes. Class and committees in a norwegian Island parish. *Human Relations*, 7: 39-58, 1954.

[27] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78: 1360-1380, 1973.

[28] M. S. Granovetter. The strength of weak ties: A Network theory revised. *Sociological Theory*, 1: 201-233, 1983.

[29] S. Milgram. The small world problem. *Psychology Today*, 2: 60-67, 1967.

[30] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *Proceedings of CEAS-1*, 2004.

[31] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. Mining email social networks. In MSR '06: *Proceedings of the 2006 international workshop on Mining software repositories*, pages 137–143, New York, NY, USA, 2006. ACM Press.

[32] S. E. Seibert, M. L. Kraimer, and R. C. Liden. A social capital theory of career success. *The Academy of Management Journal*, 44:219–237, Aprile 2001.

[33] J. Xu, S. Christley, and G. Madey. *Application of Social Network Analysis to the Study of Open Source Software*. Elsevier Press, 2006.

[34] Y. Gao. Topology and evolution of the open source software community. Master's thesis, University of Notre Dame, Notre Dame, IN, 2003.

[35] D. B. Horn, T. A. Finholt, J. P. Birnholtz, D. Motwani, and S. Jayaraman. Six degrees of jonathan grudin: a social network analysis of the evolution and impact of cscw research. In *CSCW '04*: *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 582–591, New York, NY, USA, 2004. ACM Press.

[36] L. Licamele, M. Bilgic, L. Getoor, and N. Roussopoulos. Capital and benefit in social networks. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 44–51. ACM Press, 2005.

[37] L. C. Freeman. Centrality in social networks. *Social Networks*, 1:215– 239, 1979.

[38] S. P. Borgatti, M. G. Everett, and L. C. Freeman. Ucinet for windows: Software for social network analysis. Harvard, MA: Analytic Technologies, 2002.

[39] M. Ohira, T. Ohoka, T. Kakimoto, N. Ohsugi, and K. Matsumo- to. Supporting knowledge collaboration using social networks in a large-scale online community of software development projects. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, 2005.

[40] M. Ohira, K. Nakakoji, and K. Matsumoto. D-sns: A knowled- ge exchange *mechanism using social network density among mega-community users. In Proceedings of Supporting the Social Side of Large Scale Software Development (CSCW2006 Workshop)*, pages 39–42, Novembre 2006.

[41] M. Ohira, N. Ohsugi, T. Ohoka, and K. Matsumoto. Accelerating crossproject knowledge collaboration using collaborative filtering and social networks. In *Proceedings of the International Conference on Software Engineering (ICSE'05)*, 2005.

[42] Seokwoo Song, Sridhar Nerur, and James T.C. Teng. An exploratory study on the roles of network structure and knowledge processing orientation in work unit knowledge management. *SIGMIS Database*, 38(2):8–26, 2007.

[43] Ray Reagans and Bill McEvily. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly*, 48:240–267, Giugno 2003.

[44] N. M. Tichy, M. L. Tushman, and C. Fombrun. Social network analysis for organizations. *The Academy of Management Review*, 4:507–519, Ottobre 1979.

[45] D. L. Rulke and J. Galaskiewicz. Distribution of knowledge, group network structure and group performance. *Management Science*, 46:612–625, Maggio 2000.

[46] L. Hossain, A. Wu, and K. K. S. Chung. Actor centrality correlates to project based coordination. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 363–372, New York, NY, USA, 2006. ACM Press.

[47] W. B. Stevenson and D. Greenberg. Agency and social networks: Strategies of action in a social structure of position, opposition, and opportunity. *Administrative Science Quarterly*, 45:651–678, Dicembre 2000.

[48] A. Mehra, M. Kilduff, and D. J. Brass. The social networks of high and low self-monitors: Implications for workplace performance. *Administrative Science Quarterly*, 46:121–146, Marzo 2001.

[49] W. Tsai. Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *The Academy of Management Journal*, 44:996–1004, Ottobre 2001.

[50] M. K. Ahuja and K. M. Carley. Network structure in virtual organizations. *Organization Science Special Issue: Communication Processes for Virtual Organizations*, 10:741–757, Novembre-Dicembre 1999.

[51] L. C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1, 2000.

[52] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis 2005),* 2005.

[53] S. Zeqian, M. Kwan-Liu, and T. Eliassi-Rad. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE Transactions on Visualization and Computer Graphics*, 12:1427 – 1439, 2006.

[54] Y. Gao, Y. Huang, and G. Madey. Data mining project history in open source software communities. In *Proceedings of NAACSOS 2004*, 2004.

[55] G. Valetto, M. Helander, K. Ehrlich, S. Chulani, M. Wegman, and C. Williams. Using software repositories to investigate socio-technical congruence in development projects. In *MSR '07: Proceedings of the Fourth International Workshop on Mining Software Repositories*, page 25, Washington, DC, USA, 2007. IEEE Computer Society.

[56] S. Abrams. Two-mode social network analysis as exploratory tool for cscw: Technology adoption and use.

[57] S. D. Burks. Social networks and its uses in collaborative strategies. Master's thesis, Georgia Institute of Technology, 2004.

[58] Y. Gao and G. Madey. Network analysis of the sourceforge.net community. In *Proceedings of The Third International Conference on Open Source Systems (OSS 2007)*, 2007.

[59] Y. Gao and G. Madey. Towards understanding: A study of the sourceforge.net community using modeling and simulation. In *Proceedings of Agent-Directed Simulation (ADS'07)*, 2007.

[60] C. Kadushin. Review of linton c. freeman. the development of social network analysis: A study in the sociology of science. Vancouver, Canada: Booksurge publishing, 2004, 205 pp.. *Journal of Social Structure*, 6, 2005.

[61] M. Jamali and H. Abolhassani. Different aspects of social network analysis. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 66–72. IEEE Computer Society, 2006.

[62] F. Yiweil, C. Wentian, and Y. Huahai. Identification of important actors in edge-weight social networks. In *Proceedings of International Conference on Service Systems and Service Management,* 2006, pages 1643–1647, 2006.

[63] M. Tubi, R. Puzis, and Y. Elovici. Deployment of dnids in social networks. In *Intelligence and Security Informatics, 2007 IEEE*, pages 59–65, 2007.

[64] E. E. Santos, L. Pan, D. Arendt, and M. Pittkin. An effective anytime anywhere parallel approach for centrality measurements in social network analysis. *In Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2006. ICSMC '06.*, pages 4693–4698, 2006.

[65] K. Faust. Centrality in affiliation networks. *Social Networks*, 19:157-191, Aprile 1997.

[66] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. *Cambridge University Press*, 1994.

[67] R. D. Nolker and L. Zhou. Social computing and weighting to identify member roles in online communities. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference*

on Web Intelligence, pages 87–93, Washington, DC, USA, 2005. IEEE Computer Society.

[68] R. Memoli. L'analisi delle reti sociali.

[69] S. P. Borgatti. Centrality and aids. *Connections*, 18(1):112–115, 1995.

[70] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27:55–71, 2005.

[71] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120, 1972.

[72] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 2003.

[73] C. H. Hubbell. An input-output approach to clique identification. *Sociometry*, 28:377–399, 1965.

[74] L. Katz. A new status index derived from sociometric data analysis. *Psychometrika*, 18:34–43, 1953.

[75] K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11:1–37, 1989.

[76] R. T. Sparrowe, R. C. Liden, S. J. Wayne, and M. L. Kraimer. Social networks and the performance of individuals and groups. *The Academy of Management Journal*, 44:316–325, Aprile 2001.

[77] H. Ibarra and S. B. Andrews. Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions. *Administrative Science Quarterly*, 38:277–303, Giugno 1993.

[78] G. R. Carroll and A. C. Teo. On the social networks of managers. *The Academy of Management Journal*, 39:421–440, Aprile 1996.

[79] F. Skopik, H. L. Truong, S. Dustdar. Trust and Reputation Mining in Professional Virtual Communities. *Distributed Systems Group, Vienna Uneversity of Technology*. Argentinierstr. 8/184-1, A-1040 Vienna, Austria.

[80] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):1-52, 2009.

[81] S. Anholt. Competitive Identity: the new brand management for nations, cities and regions. Palgrave Macmillan, 2006.

[82] D. Barbagallo, C. Cappiello, C. Francalanci and M. Matera. Reputation-Based selection of web information sources. Politecnico di Milano, Dipartimento di Elettronica e Informazione. 2010.

[83] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[84] K. Cukier. Data, data everywhere: A special report on managing information. *The Economist*, 27 February 2010.

[85] Donato Barbagallo, Leonardo Bruni, and Chiara Francalanci. Exploi- ting WordNet glosses to disambiguate nouns through verbs. In *The Fourth International Conference on Advances in Semantic Processing*, Florence, October 2010.

[86] Web crawler. Wikipedia, The Free Encyclopedia. Last access: 08/23/2010.

[87] C. Fellbaum et al. WordNet: An electronic lexical database. *MIT press Cambridge*, MA, 1998.

[88] B. Magnini and C. Strapparava. Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet. *Linguaggio e Cognizione. Bulzoni, Palermo, Italy*, 1997.

[89] George A. Miller. WordNet – About Us, 2009. http://wordnet. princeton.edu.

[90] The Stanford Parser: A statistical parser. http://nlp.stanford.edu/ software/lex-parser.shtml.

[91] Angela Adragna. Messa a punto e verifica empirica di una metodologia per l'utilizzo dell'analisi sintattica dei testi per il data clening. Master's thesis, Politecnico di Milano, Italy, July 2010.

[92] P. Resnik and D. Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural language engineering*, 5(2):113–133, 1999.

[93] A.K. McCallum and K. Nigam. A Comparison of Event Models for Naïve Bayes Text Classification. In *AAAI/ICML–98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[94] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *ECML–98, Tenth European Conference on Machine Learning*, pages 137–142, 1998.

[95] T. Joachims. A Probalistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning ICML97*, pages 143–151, 1997.

[96] A.K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the Construction of Internet Portals with Machine Learning. *Information Retrivial*, December 1999.

[97] Lorenzo Radice and Saverio Bruno. Analisi della reputazione in am- bienti Web 2.0: dallo studio alla realizzazione di un tool per il Senti- ment Analysis. Master's thesis, Politecnico di Milano, Italy, December 2009.

[98] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC, volume 6*, pages 417–422. Citeseer, 2006.

[99] B. Magnini and G. Cavagli. Integrating subject field codes into WordNet. In *LREC–2000 Second International Conference on Language Resources and Evaluation*, pages 1413–1418, Athens, Greece, June 2000.

[100] M. Picozzi, G. Sprega, D. Barbagallo, C. Cappiello, C. Francalanci, and M. Matera. DashMash: a Mashup Environment for End-User Programming. Technical report, DEI – Politecnico di Milano, 2010.

[101] Gareth Walsh and Steven Swinford. Hotel review websites: a five-star scam. *The Sunday Times*, November 2006.

[102] Sean Dodson. Bestof the net: TripAdvisor – The great divide. *The Guardian*, July 2007.

[103] Joe Sharkey. Online Reviews of Hotel and Restaurant Flourish. *New York Times*, January 2008.

[104] Ginny McGrath and Steve Keenan. We're clean' pledges TripAdvisor. *The Sunday Times*. May 2007.

[105] K. Musial, P. Kazienko, P. Bròdka. User Position Measures in Social Networks. The *3rd SNA-KDD Workshop '09 (SNA-KDD'09)*, June 28, 2009.

[106] S.Brin and L. Page. The anatomy of large-scale hipertextual web search engine . *Computer Network*, ISDN Syst., 30(1-7):107-117, 1998.