POLITECNICO DI MILANO

FACOLTA' DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea Specialistica in Ingegneria Informatica

DEFINIZIONE E TESTING DI UNA METRICA PER LA VALUTAZIONE DELLA REPUTAZIONE DEI FORUM

Relatore: **Prof. Chiara Francalanci**

Correlatore: **Ing. Leonardo Bruni**

Tesi di laurea di:

Dario Poletti - 755604

Matteo Pettinaroli - 755901

Anno Accademico 2011 - 2012

Abstract

Web 2.0 has profoundly changed the way by which people express their ideas and opinions. It 'increasing the number of people who visit the Web looking for all kinds of information such as, for example, those in the medical field, and use the Forum as quickest way to collect, or distribute, ideas or information. For companies therefore a strategy is needed to understand fully automatically and precisely what people are saying about a particular product or directly on the company's reputation for so figure out where it is preferable to intervene. This thesis aims to define a metric designed to determine the importance of a forum. This metric, built using data automatically detectable, allowing you to build a ranking of significance of Forum based on the grade the student got. This metric is extended by adding the ability to consider the content of the post in order to understand if the arguments mentioned may change the ranking. The results obtained indicate how, using the metric extended, also a forum much smaller than others can be particularly significant.

Riassunto

Il Web 2.0 ha profondamente cambiato le modalità con tramite cui le persone esprimono le loro idee ed opinioni. E' sempre maggiore il numero di persone che consulta il Web alla ricerca di ogni genere di informazione come, ad esempio, quelle in ambito medico, e utilizza i Forum come mezzo più veloce per raccogliere, o diffondere, idee o informazioni. Per le aziende quindi risulta necessaria una strategia per comprendere in maniera automatica e precisa le opinioni della gente riguardo un determinato prodotto oppure direttamente sulla reputazione dell'azienda per così capire dove è preferibile intervenire. Questo lavoro di tesi si pone l'obiettivo di definire una metrica atta a determinare l'importanza di un Forum. Tale metrica, costruita utilizzando dati automaticamente estraibili, consente di costruire una classifica di rilevanza di Forum in base alla valutazione finale ottenuta. Tale metrica viene estesa aggiungendo la possibilità di considerare anche il contenuto dei post allo scopo di comprendere se gli argomenti di cui si parla possono modificare la classifica. I risultati ottenuti indicano come, utilizzando la metrica estesa, anche un forum molto più piccolo di altri può risultare particolarmente significativo.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

La nascita di Internet, ed in particolare del Web 2.0, ha cambiato drasticamente il modo di comunicare, di interagire e di scambiarsi informazioni delle persone. Queste utilizzano il Web per svariate ragioni, ma le principali sono sicuramente scambio di opinioni e la ricerca di informazioni. Le aziende conoscono tutto questo e, oltre ad utilizzare Internet per i loro scopi commerciali, potrebbero monitorare questo scambio di opinioni nei siti pubblici per cercare di trarre il maggior numero di informazioni possibili. Soprattutto nell'ultima decade vi è stato un enorme sviluppo di siti web che permettono alla popolazione di dirsi praticamente tutto riguardo qualsiasi argomento; ne sono una prova il boom commerciale recente di siti quali Facebook e Twitter, o di forum quali TripAdvisor. Un forum online è uno spazio virtuale dove persone unite da interessi particolari comuni condividono passioni ed esperienze, esprimono la propria opinione e leggono le opinioni degli altri, chiedono consigli o ne danno se possono. É come una piazza, si potrebbe dire. Una piazza dove un gruppo di amici si dà appuntamento la sera per discutere, in allegria, delle proprie passioni e per divertirsi o confidarsi. E, magari, per scambiare esperienze con altri gruppi di persone. Ma è pure più di una piazza, perché un forum è spesso anche un servizio, una fonte di informazioni utili, libere e gratuite. In questo senso, lo si può paragonare a un immenso mercato dove chi entra può prendere ciò che vuole, informarsi sui prodotti

migliori, valutare e sentire il parere dei clienti più informati. Nel caso una società fosse in grado di capire che la popolazione si sta facendo un'idea negativa su di sé potrebbe cercare di riparare in una moltitudine di modi come intervenire direttamente sul Web cercando di alzare la propria reputazione, oppure andare a correggere il proprio prodotto/servizio nei difetti che è riuscita a capire siano la fonte del malcontento. In molti casi se questo tipo di analisi si riuscisse a fare in un momento in cui l'azienda non fosse in perfetta salute, potrebbe significare la sopravvivenza o scomparsa dell'azienda stessa.

Questo lavoro di tesi si propone di determinare quali parametri sono fondamentali per la caratterizzazione di un Forum Web e nella valutazione della sua importanza strategica. Di creare un sistema automatizzato di estrazione di questi parametri dal maggior numero di fonti possibili e di creare un sistema di valutazione che sia in grado di determinare tramite un numero compreso in una scala l'indice di importanza che un Forum possiede sul Web. Questa valutazione sarà anche successivamente caratterizzata in base a determinati parametri non numerici scelti dall'utente.

La tesi è organizzata come segue. Il Capitolo 2 tratta una rassegna della letteratura che, dopo una breve discussione sull'evoluzione del Web, si concentra sempre più sui Forum, partendo dalla loro nascita fino alla caratterizzazione delle fonti. Vengono introdotti i principali lavori già svolti sui Forum e viene fatta una spiegazione dello stato attuale della medicina sui Forum Web. Il Capitolo 3 riporta la prima metrica, quella basata solamente sui numeri. Viene descritto lo strumento Crawler creato ed utilizzato ,vengono spiegati tutti i parametri che sono stati presi in considerazione e viene creata la prima metrica di valutazione. Il Capitolo 4 discute dell'estensione della metrica elaborata al punto precedente prendendo in considerazione anche il contenuto dei post, vengono trattate le motivazioni che hanno portato a questa estensione, lo strumento che permette di estrarre le informazioni e la metrica finale ottenuta. Nel Capitolo 5 vengono discussi e confrontati

tutti i risultati ottenuti facendo una precisa analisi sull'accuratezza delle metriche ottenute e su quali possono essere considerate più rilevanti/veritiere in determinati casi. Il Capitolo 6 riassume i risultati ottenuti da questo lavoro di tesi ed offre possibili idee per eventuali sviluppi futuri.

Capitolo 2

Stato dell'arte

2.1 L'evoluzione del web

La nascita del web partì dal Dipartimento della Difesa Americana con il progetto ARPANet nel 1958, con l'obiettivo di realizzare una rete di computer per semplificare la comunicazione tra gli utenti del proprio laboratorio di ricerca. A seguito dei miglioramenti ottenuti in campo di ricerca tramite l'utilizzo del sistema ARPANet, si diffuse questa nuova tecnologia anche ad altri paesi, rendendo necessaria la gestione della stessa, tramite una standardizzazione nella comunicazione. Dal 1990 al 1991 nacque ufficialmente l'HTML e il World Wide Web. Il Web però si diffuse al pubblico soprattutto per visualizzare documenti ipertestuali statici (linguaggio HTML) tale stadio può essere chiamato Web 1.0, comprende tutti i siti web che si relazionano all'utente in modo unidirezionale, senza possibilità di interazione con esso. La comunicazione era limitata allo scambio di e-mail e tramite la prima chat IRC (Internet Relay Chat) entrambe inizialmente disponibili e utilizzate dalle università. La crescita del web in ogni sua forma portò grazie all'integrazione con i database e all'utilizzo dei CMS (content management system) i primi siti dinamici (contenenti ad esempio forum, blog, ecc.) dando luogo al Web 1.5, aumentando di fatto le possibilità di comunicazione.

Oggigiorno l'utilizzo di linguaggi di programmazione come JavaScript, degli elementi dinamici e dei CSS (fogli di stile) per gli aspetti grafici è possibile creare delle vere e proprie "applicazioni web" che si allontanano dal vecchio concetto di ipertesto ed anzi assomigliano sempre più alle tradizionali applicazioni che siamo abituati a vedere nei personal computer. Da notare è che dal punto di vista tecnologico non vi è alcuna differenza tra il Web1.0 ed il Web2.0, poiché l'infrastruttura di rete continua ad essere la medesima, con i soliti protocolli TCP/IP e HTTP e l'ipertesto è ancora un concetto base nella relazione tra i contenuti. La vera differenza si trova nell'approccio con il quale gli utenti si rivolgono al Web, che non è più la semplice consultazione (anche se agevolata da mezzi più potenti), ma l'interazione con essi e con il Web stesso, potendo così aggiungervi il proprio contributo attraverso contenuti personali.
Il Web 2.0 ha permesso di sfruttare gli utenti quali produttori di nuove informazioni e alimentare l'interesse e la fidelizzazione al proprio contenuto, infatti possiamo notare come i siti nati dopo l'avvento del web 2.0, tipo Facebook, Twitter, TripAdvisor, Youtube, Flickr ecc. che macinano milioni di contenuti ogni giorno, prodotti da altrettanti utenti, realizzano milioni di visite giornaliere, accrescendo il valore del proprio marchio.
Si suddividono le specie di contenuti della rete in gruppi: Social Network, Forum, Blog, altro. Un social network identifica una struttura internet incentrata sull'utente e la sua rete di contatti, basando le interazioni ad un sottogruppo di iscritti che generano contenuti. Un forum è una struttura informatica localizzata sul web che permette agli utenti iscritti di scambiarsi informazioni ed esperienze relative all'argomento di discussione, le interazioni tra utenti avvengono con scambio di messaggi regolati da supervisori che tutelano il regolamento del sito web. Un blog è un sito web gestito da una persona, un'ente o un'azienda, in cui l'autore pubblica periodicamente con contenuti testuali o multimediali; il blog inoltre permette l'interazione con gli utenti, tramite commenti, guestbook, ecc.. Altri siti che non rientrano nelle categorie precedenti che permettono l'interazione con gli utenti sono

ad esempio siti di vendita e-commerce dove è possibile integrare all'interno del listino una valutazione e/o una descrizione dei prodotti già acquistati e provati dai clienti.

La possibilità di confrontarsi con altre persone all'interno dei siti web 2.0 permette una visione globale sui prodotti e sulle informazioni scambiate, si possono citare alcuni esempi: se una persona decidesse di partire per un viaggio lontano, in un ambiente completamente sconosciuto; potrebbe chiedere all'agenzia di viaggi, supponiamo affidabile al 100%, i luoghi più affascinanti e l'abbigliamento per essere il più comodo possibile. D'altra parte potrebbe scegliere nella vastità di internet un sito, un forum o un social network che risponda alle sue domande e rimanere soddisfatta dalle risposte ottenute, contando sull'opinione ed esperienza di una vasta quantità di utenti raccogliendo ogni informazione "cum grano salis". Dallo sviluppo del web2.0 nel 2004 possiamo registrare un incremento notevole di navigatori rendendo necessario il paragone ad una popolazione, diventando soggetto di studi relativi al comportamento in varie discipline sociologia, psicologia, economia, antropologia, geografia, scienza dell'organizzazione, sociolinguistica e biologia.

2.2 Reti Sociali

Si può definire rete sociale una struttura che lega ogni individuo ad altre persone con rapporti di lavoro, legami di parentela, amicizia o conoscenza. La rappresentazione astratta di una rete sociale avviene tramite modellizzazioni matematiche delle teorie dei grafi, in cui i nodi raffigurano le persone mentre gli archi simboleggiano le relazioni sociali tra gli individui. Gli studi dell'evoluzione dei legami sono molto più chiari nei social network rispetto ad altre forme di comunicazione, poiché le interazioni tra gli utenti son ben definite e rintracciabili. La sociometria è la scienza che si occupa dell'analisi delle relazioni interpersonali. La SNA trova ora applicazione in diverse scien-

ze sociali, ed è stata utilmente impiegata nello studio di diversi fenomeni, come il commercio internazionale, la diffusione dell'informazione e lo studio delle istituzioni.

Il web 2.0 avvicina gli individui nascondendo le barriere geografiche, sviluppando un potenziale illimitato di relazioni di amicizia e di corrispondenza. L'avvento dei siti interattivi ha certamente incuriosito e avvicinato molti utenti "scollegati" a far parte della comunità di internet, non solo limitandosi al social network più in voga del momento, ma anche ad esplorare gli angoli più storici del web[1]. Aiutati dai motori di ricerca i nuovi utenti imparano a trovare soluzioni alle proprie curiosità e scoprono siti specializzati che offrono uno spazio al dibattito e condivisione di idee inerenti un argomento, i forum. Presenti fin dal web1.5 han mantenuto la loro funzionalità, lo scambio di messaggi pubblici immediati, la condivisione di idee e punti di vista e la semplicità di ridurre le informazioni all'argomento di discussione. Il web 2.0 ha semplificato la nascita di forum, forgiati e strutturati per cooperare con i social network e siti sociali (blog), sapendo sfruttare della crescente quantità di utenti rendendoli subito capaci di intervenire senza completare le burocratiche registrazioni al forum.

2.3 Forum

Un internet forum o una bacheca di messaggi, è un luogo di discussione online, dove le persone possono intrattenere conversazioni nella forma di messaggi pubblici. Differiscono dalle room chat dall'archiviazione temporale dei messaggi. I dati pubblicati, dipendono anche dal tipo di accesso che un utente possiede all'interno del forum, un messaggio potrebbe passare sotto controllo di un moderatore prima di essere pubblicato e reso visibile.
I Forum hanno sviluppato una serie di forme gergali per esprimere le attività sul sito; e.g. una singola conversazione è chiamata "thread".

Una discussione è sviluppata su una struttura gerarchica ad albero: un forum può contenere molte discussioni, ognuna delle quali può incorporare molti argomenti, chiamati topic. All'interno di un topic, ogni nuova discussione viene chiamata thread e può essere articolato da qualsiasi utente.
Dipendentemente dalla politica del forum, gli utenti possono agire anonimamente oppure iscriversi e accedere tramite log in per scrivere messaggi. Mentre nella maggior parte dei forum non è necessario effettuare un log in per leggere i messaggi. Anche le informazioni degli utenti sono soggette alle limitazioni del forum che solitamente rendono pubblici i dati relativi ai caratteri generali e l'esperienza sul forum: numero di post, data di iscrizione.
Negli ultimi anni sono stati implementati alcuni strumenti in grado di condividere una discussione o un commento anche sui social network, rimandando gli utenti alla visita del forum.

2.3.1 Storia dei Forum

I forum moderni hanno origini dalla nascita del BBS, "Bulletin Board System", nato nel 1977 da due studenti dell'Università di Chicago, Ward Christensen e Randy Suess che scrissero un programma battezzato MODEM, che permetteva il trasferimento di file tra i loro pc e nel 1978 misero a punto anche il Computer Bulletin Board System con il sistema BBcode, che consentiva al PC di trasmettere e archiviare messaggi.
I limiti dell'interesse di un BBS frequentato solo da poche persone di una ristretta area geografica vennero superati con la nascita di reti di BBS. I BBS aderenti alla stessa rete scambiavano fra loro (la notte, quando le tariffe telefoniche erano inferiori) tutti i messaggi scritti dagli utenti. In questo modo l'utente aveva l'impressione di usare un solo grande BBS diffuso in tutto il pianeta, e con moltissimi più utenti di qualunque singolo BBS. La prima rete del genere fu Fidonet, che arrivò ad avere decine di migliaia di nodi. Venne imitata da altre reti più piccole ma specializzate su temi specifici.

Con la crescita di Internet della metà/fine degli anni novanta, la popola-

rità dei BBS calò rapidamente. Attualmente, molti di essi sono connessi a Internet e possono essere letti tramite un comune browser web, prendendo il nome di Forum.
Si può descrivere l'utilizzo dei forum come una versione web di una "mailing list" elettronica; permettendo alle persone di scrivere messaggi e commenti su altri messaggi. Gli sviluppi di ulteriori gruppi e liste di interesse hanno portato la creazione di nuovi forum, mantenendo una suddivisione per argomenti ordinata.
I forum internet sono presenti in molti Paesi sviluppati. Gli argomenti su cui discutere sono ampi: tecnologia, video games, sport, musica, moda, religione, e politica, ma sono presenti anche forum che raccolgono più argomenti contemporaneamente.

Sono disponibili pacchetti software per la realizzazione dei forum in internet in vari linguaggi di programmazione, come PHP, Perl, Java e ASP. Le configurazioni e i testi dei post possono essere memorizzate in file di testo o in database. Ogni pacchetto differisce dal numero e tipo di feature, dalla forma base che permette il post di semplici messaggi di testo, ai pacchetti più avanzati, che offrono supporto multimediale e un metodo di formattazione. Molti pacchetti possono essere integrati nei siti esistenti per permettere ai visitatori di lasciare post e commenti negli articoli.
La differenza tra ieri ed oggi sta nel numero dei forum e sui contenuti, ora si può trovare di tutto. Alcuni esempi: Forum dedicato alla carne alla griglia e forum dedicato ai capelli.

2.4 L'importanza delle fonti

Prima di compiere un acquisto o prendere una decisione ci affidiamo automaticamente ad un processo di selezione sulle informazioni che abbiamo acquisito. Le informazioni su cui il pensiero ragiona sono acquisite dall'esperienza della nostra vita, dalla lettura di un testo o dal consiglio delle persone

che vivono attorno a noi. Come è possibile discriminare le notizie inaffidabili da quelle affidabili?
Non possiamo a priori definire l'attendibilità di una notizia, ma se sappiamo da quale fonte proviene è più facile dare l'importanza che merita. Diverso risalto prendono le affermazioni dette da un medico circa le condizioni di salute di una persona, rispetto ai discorsi da salotto uditi dal parrucchiere. Purtroppo non sempre le fonti sono disponibili e l'interlocutore o il mezzo che riporta le informazioni diventa la nostra fonte, riponendo in esso la nostra fiducia. Ognuno è in grado di separare le fonti attendibili dalle altre assegnando ad esse importanza in base al riscontro che hanno le informazioni nella realtà. Quando però non è possibile determinare la natura delle notizie e rimaniamo con il dubbio della verità. Quali sono i fattori che influiscono sulle nostre decisioni? Ad esempio, lo stile cognitivo, la maggiore o minore propensione alla fiducia o al sospetto, così come le precedenti esperienze con una certa fonte sono tutti elementi che influenzano la percezione di credibilità della fonte[2]. Quando abbiamo stabilito un rapporto di fiducia con le fonti, le informazioni apprese da esse aumentano di credibilità. La dipendenza dalle fonti di informazione è seguita molto da vicino dall'economia e dal marketing che veicolando la pubblicità in esse sfruttano la fidelizzazione delle persone[3].

Dal punto di vista economico, il numero di persone che una fonte riesce a coinvolgere è di particolare interesse, per poter massimizzare il profitto derivato da pubblicità e diffusione delle idee. Le aziende che si affacciano a questa opportunità di divulgazione devono saper scegliere il mezzo comunicativo più adatto. Prima dell'avvento di internet i metodi più efficaci, in grado di raggiungere la quasi totalità della popolazione erano mezzi stampa, Tv ed emittenti radiofoniche. Oggi le attenzioni si spostano sul web, un bacino di utenza in costante aumento e diverse possibilità di veicolare le informazioni: blog, siti web, social network, forum, advertise, motori di ricerca ecc.; ognuno di essi è un candidato adatto per la diffusione di notizie.

2.5 Lavori Correlati

Attualmente gli studi del web analizzano le motivazioni che spingono gli utenti alla ricerca delle informazioni online, come fare per evitare false notizie, si analizzano gli strumenti presenti sul web che classificano le notizie in base alla qualità e alla popolarità, affidando agli utenti una votazione per potersi orientare verso la verità[4].
Uno studio dell'Università del Nebraska è stato condotto sugli internauti per valutare come i tempi di caricamento delle pagine influiscono sull'umore delle persone; i risultati dimostrano già una prima selezione di preferenze tra siti web, indipendentemente dal contenuto; le pagine web con tempi di caricamento minori sono i più apprezzati. É stato affrontato un'ulteriore approfondimento del campione per valutare il riscontro dei metodi che possano ridurre l'impatto negativo dell'attesa come: l'aggiunta di informazioni nella barra di caricamento, la visualizzazione parziale del sito che si compone in modo incrementale. Nonostante i cambiamenti introdotti, il test ha smentito l'ipotesi che un utente con maggior controllo sul caricamento abbia tolleranza migliore rispetto al tempo di caricamento[5].

Per quanto riguarda la qualità delle informazioni in internet non ci sono evidenze tra gli studi che identificano un metodo generico che analizzi i contenuti. Gli utenti presuppongono che il motore di ricerca effettui una selezione automaticamente e che mostri i risultati in ordine di importanza e qualità; successivamente altri parametri influiscono sul giudizio della fonte, come la struttura ordinata del sito, la velocità con cui si trovano le informazioni e la contestualizzazione delle notizie [6].
Come è possibile riconoscere i siti affidabili? Janet E. Alexander e Marsha Ann Tate nel loro libro spiegano quali parametri deve rispettare un sito web per sembrare affidabile. La struttura è importante e rispecchia il messaggio che si vuole esprimere, che si tratti di un sito di informazione o di intrattenimento ogni fine deve essere ben chiaro nell'impostazione del sito web; il

libro ci fornisce gli strumenti per scegliere i siti con la struttura più seria e adatta a rappresentare le informazioni, ma insegna anche ai web designer come devono relazionarsi con l'utente [7] [8].

Una ricerca condotta dallo studio congiunto di due Università, Victoria University of Wellington e University of Bath, del settore di School of Information Management ha adattato gli strumenti presenti nel web per la classificazione dei siti governativi di vari paesi. Lo strumento utilizzato è WebQual, inizialmente creato per confrontare siti di e-commerce è stato ampliato per il confronto tra siti della stessa categoria oppure per valutare l'effetto di un cambiamento nel aspetto strutturale. Lo strumento è stato utilizzato in diverse fasi temporali, per raccogliere una valutazione iniziale e quindi poter evidenziare un cambiamento. WebQual ha raccolto le impressioni degli utenti somministrando loro un questionario che richiedeva di valutare i siti in esame su una scala di 7 valori, le domande erano volte a evidenziare un punteggio su 3 aree di interesse: qualità di informazione, qualità dei servizi e interazione, usabilità (Human-computer interaction). Il campione di utenti è stato preso dai visitatori abituali del sito FSMKE su cui è stata compiuta la valutazione e comprendeva navigatori di diverse regioni culturali, in maggior parte dal Giappone, UK, Canada e Australia, mentre il periodo di osservazione è stato da Aprile a Maggio. Dopo aver modificato l'aspetto e la struttura del sito, è stato somministrato lo stesso questionario per un periodo di tempo che va da Luglio a Settembre, registrando una simile partecipazione degli utenti, suddivisi in gran parte come nel primo test, cioè per nazionalità, età e genere. Lo strumento ha evidenziato un aumento notevole rispetto le aree: design, usabilità, informazione e servizi; oltre al questionario sono stati analizzati i commenti degli utenti per verificare l'effettivo miglioramento dell'interazione con il sito. Gli autori del test hanno osservato che l'interazione con l'utente e la conoscenza delle loro opinioni sia importante per sviluppare i servizi giusti e offrirli nella maniera più adeguata. Inoltre è un modo per mantenere alto il ranking del sito e indurre l'utente a tornare

nella visita [9].

Dal punto di vista della social information (es. Blog) le notizie prendono il nome dell'autore che riporta l'informazione. Diventa importante studiare come discriminare le notizie vere da quelle false, un esempio è quello di valutare un individuo in base alla sua importanza e al numero di utenti al seguito. Raymond Morin, un consulente del WEB-marketing ha ipotizzato i parametri in grado di valutare i social web influencer, cioè coloro in grado di raggiungere molte persone con i loro siti e blog. Con la sua esperienza di 15 anni nell'era digitale e media, ha estratto una lista di indicatori in grado di evidenziare l'impatto degli influencer in base allo scopo da raggiungere; per ogni scopo viene valutato, il valore numerico delle visite e le qualità dei contenuti; la reputazione online dell'influencer, ossia la sua presenza online con pubblicazione e condivisione di contenuti; la credibilità, in parte correlata alla reputazione, riguarda la qualità delle persone con cui si relaziona e dai riscontri nella realtà dei contenuti pubblicati; la quantità di notizie divulgate dal proprio pubblico; oltre all'esperienza online viene valutata la carriera dell'influencer nella realtà in base alla competenza, con titoli acquisiti e carriera lavorativa; anche il contesto sul quale lavora ha rilevanza nella valutazione. Ad aggiungersi alla misurazione sociale calcolata seguendo un algoritmo, Raymond Morin, identifica due valori legati ai risultati economici che seguono le pubblicazioni degli influencer; la fiducia acquisita si rispecchia nella credibilità che gli utenti ripongono nell'influencer; l'impegno, cioè il numero di condivisioni e di citazioni riprodotte [10].

I parametri individuati da Morin lasciano agli studiosi l'onere di estrarre da essi una valutazione parametrica dell'influenza degli individui online. L'analisi delle fonti può essere ampliato indagando sulle opinioni della comunità del web, a questo proposito il dipartimento CSE della regione indiana "Tamil Nadu" ha lavorato sul data mining. Si tratta di un'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di un sapere o di una co-

noscenza a partire da grandi quantità di dati. La ricerca è stata condotta osservando i forum relativi ad alcuni prodotti commerciali; analizzando il sentiment dei post è stato sviluppato uno strumento in grado di aiutare i clienti nelle decisioni[11].

Un'altra università, la Peking University in Cina, ha analizzato i contenuti prodotti dai forum, con lo scopo di valutare l'umore dei topic, tramite sentiment analysis. Non è stato specificata l'area di studio dei topic, ma tramite l'individuazione di parole chiave i thread che trattano dello stesso argomento sono stati raggruppati in cluster. Lo scopo della ricerca è stato quello di verificare nuovi metodi di analisi, software e statistiche in grado di dare maggior precisione circa il sentiment di forum [12]. In aiuto al team cinese può essere utile parlare del lavoro svolto in Svezia nella Lund University nel 2011, il quale ha sottoposto i commenti relativi a diversi forum a 9 algoritmi per la valutazione del sentiment, estraendo infine le caratteristiche e differenze di ognuno[13].

Nel 2009 l'università del Maryland ha condotto una ricerca sul social network Facebook [14], riguardo la somministrazione di pubblicità agli utenti. Non è segreto che la pubblicità su internet è una grande fonte di guadagno, Google ne è un esempio formidabile e a tal proposito, la possibilità di raggiungere milioni di persone ha permesso a Facebook di avere un introito notevole. Per questo motivo lo studio è rivolto a massimizzare e migliorare il servizio di divulgazione della pubblicità analizzando l'influenza che ogni individuo ha sui propri amici. Diversamente di quanto accade per l'analisi dei siti internet la cui valutazione avviene tramite somministrazione di questionari, su Facebook e nei Social Network in genere le interazioni vengono divulgate e memorizzate, rendendo più facile la fase di raccolta dei dati. Dopo aver analizzato i dati e le interazioni tra gli utenti è stata determinata una influenza eterogenea tra gli utenti, rilevando lo stesso riscontro nelle interazioni reali, cioè che ogni utente viene influenzato da pochi utenti e a sua volta non ha il

potere di raggiungere molte persone. La potenza della ricerca si ferma nella determinazione degli influencer del social network tramite la valutazione delle interazioni e del tempo di collegamento degli utenti, tralasciando molti dettagli come ammesso nelle conclusioni. I motivi che hanno ridotto la precisione dei risultati sono dovuti a molti fattori, ad esempio l'impossibilità nel determinare quando un utente si può relazionare con un altro, la difficoltà nel determinare le relazioni vere e quindi più influenti. La determinazione dell'utente influente è stata in parte valutata tenendo conto del profilo, ma le limitazioni del SN nella lettura dei dati ha un costo sia di memoria e di privacy. L'ultimo scoglio della ricerca è stata l'impossibilità di valutare gli effetti delle azioni di marketing sugli influencer, come hanno reagito gli utenti legati ad esso e quindi notare un aumento di interesse portato dalla maggior visibilità.

Molti studi sui comportamenti online degli utenti rivolgono l'attenzione sull'aspetto economico delle pubblicità, nel dettaglio riportiamo il risultato di un esperimento effettuato dall'università del New Jersey in cui è stato chiesto a diversi utenti di ricercare informazioni riguardo a 5 prodotti commerciali, tramite l'utilizzo di forum o delle schede del materiale pubblicitario online. Evidenziamo come l'interazione con utenti del forum che non hanno interesse apparente nella promozione di un prodotto siano più credibili e degni di ascolto rispetto alla descrizione di una pubblicità online, inoltre ciò che ne aumenta l'affidabilità è la possibilità che vengano condivise le esperienze nell'utilizzo del prodotto. Infatti la descrizione di utilizzo di un prodotto fornita dai forumers dovrebbe avvicinarsi alle reali prestazioni del dispositivo rispetto alla descrizione del venditore che probabilmente non ha mai utilizzato l'articolo. I forum inoltre hanno la capacità di creare empatia tra gli utenti, coinvolgendoli in storie e aneddoti che coinvolgono l'oggetto di discussione, mentre non si riscontra la stessa enfasi nel leggere le descrizioni dei rivenditori. Tuttavia la ricerca non vuole ridurre l'importanza di uno o dell'altro metodo di informazione, poiché i dettagli tecnici forniti dalla scheda

informativa del negoziante non può essere sostituita dalle informazioni degli utenti, mentre il forum non è in grado di far cambiare idea o convincere le persone su alcuni concetti, ma entrambi hanno potere diverso di coinvolgimento. L'esperimento coinvolse circa 60 studenti, a cui è stato chiesto di informarsi tramite forum o sito ufficiale riguardo ad uno tra cinque diversi settori: ciclismo, attrezzi per l'esercizio, integratori alimentari, fotografia, o accessori per lo stereo. In maniera casuale veniva assegnato il metodo di ricerca agli studenti o tramite siti di vendita o forum; ogni settimana, per 12 settimane potevano scegliere un forum o un sito diverso per ricercare ulteriori informazioni, annotando le nozioni imparate. Al termine del periodo di ricerca è stato somministrato un questionario riguardo alle 5 categorie che era possibile scegliere all'inizio e in aggiunta altre domande su altri 5 argomenti. I temi principali del questionario hanno misurato i livelli di: probabilità nel acquisto, spesa prevista, conoscenza del prodotto, interesse nella categoria del prodotto, quanto tempo ho pensato alla categoria del prodotto. L'analisi dei risultati utilizzando un MANOVA, ha rilevato diversi livelli di interesse tra topic, ma a parità di argomento una significativa differenza tra l'uso di forum e l'uso di siti commerciali, in favore dei primi per ogni punto analizzato [15].

2.6 Medicina nel web

Uno studio risalente al 2002 del Journal of Medical Internet Research ha stimato che circa 25 milioni di persone in UK hanno accesso a internet e 15 milioni di queste lo usano regolarmente. Il WorldWide vanta più di 500 milioni (dati del 2002) di persone collegate in tutto il mondo. Ognuno di essi ha a disposizione oltre 3 miliardi di documenti web e almeno il 2% dei siti web riguardano la salute. In effetti la ricerca di informazioni mediche è una delle ragioni di accesso alla rete. Una ricerca ha mostrato che dal 50% al 75% degli utenti sono abituati a cercare informazioni mediche con cadenza di 3 volte al mese. Nel dicembre 2001 il sito web di informazione

sulla salute NHS (www.nhsdirect.nhs.uk) ha avuto 5,2 milioni di contatti da 171900 visitatori (fonte JMIR). Questo è stato possibile anche per merito dei medici e chirurghi che hanno sfruttato il web per le consulenze con i pazienti; inoltre la comunicazione via web ha permesso di ridurre le incomprensioni tra medico e paziente sfruttando le esperienze dei singoli utenti che riescono a spiegare i concetti con un linguaggio meno tecnico.
La maggior parte delle persone ricerca soluzioni a problemi specifici tramite motori di ricerca e non usano canali diretti come i siti specializzati; spesso in concomitanza a visite mediche o per raccogliere notizie riguardo a problemi di salute dei propri amici o parenti. La ricerca scientifica che ha come argomento i contenuti medici online serve per garantire la qualità delle informazioni e per capire l'aspettativa dei pazienti circa la cura medica in rete. Sono state definite 3 figure di individui che si servono di internet per le ricerche mediche: gli utenti sani, gli utenti con diagnosi e gli utenti malati cronici. Ognuna di esse ha un approccio diverso nella ricerca delle informazioni: i primi ricercano problemi riguardo la maternità, prevenzione e malattie curabili di lieve entità; gli utenti con diagnosi ricercano approfonditamente notizie riguardo la propria malattia valutando ogni documento ritrovato in rete; i malati cronici e i loro assistenti ricercano notizie sulle nuove medicine, nuovi metodi di trattamento e terapie per il proprio disagio. Powell e Clark, gli autori della relazione, concludono la ricerca evidenziando la crescita esponenziale dei contenuti medici online spostando l'attenzione sui problemi della quantità e qualità delle notizie fruibili agli utenti[16].

Nel 2004 è stato valutato l'effetto dell'interazione peer to peer, cioè lo sfruttamento della comunità virtuale per creare gruppi di supporto elettronico. Lo studio eseguito da bmj.com, il British Medical Journal, cerca evidenze, valutando gli effetti sulla salute e sui riscontri sociali, generati dalle discussioni e gruppi di aiuto online. Sono stati presi in esame gli studi effettuati negli anni ed estratti quelli relativi alle interazioni sociali e la salute, veicolando le informazioni tramite il web. Le ricerche sono state suddivise in

gruppi separati. Le relazioni sui pazienti che non hanno partecipato ai gruppi di discussione o di aiuto sociale facevano parte del gruppo di controllo mentre gli studi dettagliati con dati prima e dopo l'osservazione prendevano parte dell'analisi. Purtroppo i dati analizzati non hanno trovato un evidenza statistica circa gli effetti dell'utilizzo di supporti web oriented. La mancanza di evidenza non è un fattore negativo, le misurazioni possono essere state falsate da alcuni fattori: l'orientamento commerciale di ogni ricerca estratta non poneva l'attenzione sugli effetti sulla salute portata dalla comunità virtuale, ma dal potenziale economico; inoltre, in alcuni casi l'interazione sociale forzata del paziente non ha dato i risultati aspettati. Infine le diversità tra ricerche e la difficoltà nel separare i casi realmente considerati social web oriented ha reso più difficile questo compito. La raccolta di dati della BMJ ha evidenziato come esistano milioni di gruppi di supporto medico online, tra i quali chat room, forum, web blog; alcuni aneddoti indicano che i gruppi di auto aiuto online possano trovare benefici negli utenti, tuttavia bisogna stare sempre in guardia dai gruppi inadatti. Gli studi che fino ad ora hanno riportato effetti positivi tangibili non separano adeguatamente l'apporto dato dall'auto-aiuto rispetto quello dell'intervento medico stesso[17].

Finora abbiamo esaltato le potenzialità del web, della comunità e i benefici che può portare nell'ambito medico, ma come avevamo accennato in precedenza non sempre le informazioni riportate si rivelano corrette[18]. Questo problema è stato trattato dal Health Information Research Unit canadese con un elaborato intitolato: "Classificare le informazioni sanitarie su internet - Navigazione verso la conoscenza o alla confusione". Lo scopo della ricerca è stato quello di individuare dei siti di rating e applicarli a quelli che si occupano di sanità, stabilire la validità dello strumento utilizzato in base ai criteri di scelta e determinare la base per nuove ricerche in questo settore. Tra i campioni da classificare ci sono stati: *MEDLINE*, *CINHAL*, *Library and Information Science Abstract*, motori di ricerca come Lycos, Excite, Yahoo, liste di discussione in internet e diverse pagine web. Gli strumenti utilizzati

per assegnare la qualità dei siti web che trattano l'argomento sanitario sono accessibili direttamente da internet[19][20].

Un recentissimo studio (02/10/2012) del Censis ha rivelato in percentuale dove gli italiani recuperano l'informazione medica, il 55% della popolazione fa riferimento al proprio medico generale ma un largo 32% utilizza internet come mezzo di informazione; a determinare le percentuali del campione è indubbiamente l'istruzione della popolazione intervistata, si rafforza la figura del medico con l'80% nelle persone senza titolo di studio, mentre si riduce al 37% per il campione con titolo universitario. L'alto livello di insoddisfazione circa la relazione tra medico e paziente valutata tra l'85% e il 93% potrebbe indurre il paziente a ricercare le informazioni relative alla propria salute tramite il web. Inoltre la percezione del campione è quella che il medico non si ponga sullo stesso piano del paziente e il 76,3% delle persone intervistate ritiene la decisione del medico non negoziabile, contro il 23,7% che reputa importante il confronto col professionista[21].

Capitolo 3

Metrica basata sui numeri

In questo capitolo viene presentata principalmente l'architettura dello strumento di crawling di un forum, illustrata nella sezione 3.2, e la metrica finale utilizzante i dati estratti con le relative motivazioni sulle scelte, illustrate nella sezione 3.3. Nella sezione 3.1 vengono invece introdotti gli obiettivi principali che ci si è posti nella realizzazione di questo lavoro.

3.1 Obiettivi del progetto

Abbiamo ricavato dagli studi precedenti il successo suscitato dall'introduzione del web e dal suo utilizzo nel campo della sanità. In particolare ha permesso la divulgazione delle conoscenze mediche a tutti i navigatori; con l'evoluzione tecnologica sono nate diverse categorie di strutture web che hanno come argomento la sanità, da gruppi di auto aiuto, a consulenti online. L'aumento delle fonti ha generato un flusso di notizie incontrollato e passivo di errori; per questo motivo vengono effettuati studi relativi al rating delle notizie, tra i quali strumenti presenti nel web in grado di valutare la correttezza delle informazioni e classificazione dei siti, studio di metriche incentrate sul numero di visite e l'analisi della qualità degli autori. L'avvento del web 2.0 ha spostato l'interesse dai classici siti web ai canali sociali, come social network e forum e da questa analisi preliminare cercheremo di appro-

fondire il concetto di metrica, centrando la misurazione sulla struttura web del forum. Nella quasi totalità dei forum su Internet non si è in grado di determinare in maniera rapida ed efficace la sua diffusione e popolarità e persino una persona con competenze elevate si troverebbe in difficoltà nell'analizzare l'immensa quantità di dati che un forum di media popolarità mette a disposizione. La metrica verrà costruita analizzando i numeri disponibili al pubblico ed eventualmente una combinazione di essi per estrarre parametri utili per il calcolo, come il numero di post generati, il numero di thread, la media dei post degli utenti, ecc.

L'operazione di crawling nei forum è stata resa possibile grazie allo sviluppo di software in grado di connettersi ad essi. Dal codice HTML sono stati estratti i topic, i thread e le informazioni degli utenti relativi ad ogni forum nell'arco di sei mesi, salvandoli in un DataBase. Per ogni DB sono stati calcolati i parametri che possono identificare l'attività di un forum. Ovviamente però non tutti i forum sono uguali, ossia non tutti decidono di condividere le stesse informazioni agli utenti e quasi tutti hanno layout differenti dagli altri; quindi non è possibile specificare una lista di quali parametri sono direttamente estraibili e quali invece calcolabili perché questi variano in base al forum che si sta analizzando e quindi può variare anche la loro accuratezza (i dati forniti direttamente dagli amministratori del forum sono ovviamente più accurati che non quelli deducibili da altri dati diretti).
La seconda metrica invece tiene in considerazione solo il contenuto dei post; ogni post viene analizzato da uno strumento in grado di estrarre i soggetti della discussione dividendo la conversazione per categoria, soggetto e brand, identificando la marca di un farmaco e i sintomi o il motivo della discussione. La metrica basata sul contenuto permette all'utente che richiede la valutazione del forum di scegliere per quali contenuti calcolare il punteggio, ad esempio uno stesso forum può avere un punteggio elevato riguardo al brand: Bayer, mentre uno basso per il brand: Pfizer.

In questo modo possiamo valutare i volumi di conversazione di ogni sottogruppo, evidenziando punti deboli o punti di forza di ogni forum. Il punteggio che riguarda la seconda metrica può avere diverse utilità: riconoscere tra i forum presi in considerazione quale ha più informazioni riguardo un farmaco o una malattia, dal punto di vista del marketing è possibile riconoscere il forum migliore in cui pubblicizzare un nuovo prodotto.
Abbiamo preso in considerazione solo quei forum che trattano di farmaceutica e cura medica: Healthboard [22], LowCarber [23], RenewedReflection [24], DiabetesDaily [25], LowCarbFriends [26], ObesityDiscussion [27], ThreeFatChicks [28], U.S. MessageBoard [29], Weightloss [30], WeightlossCenter [31], WomensHealth [32].

Tutto questo lo si è svolto solamente su forum in ambito medico per poter permettere una più efficace analisi dei risultati e poter trarre delle conclusioni che siano il più affidabili e veritiere possibili; tutto questo è trattato e analizzato nel capitolo: "Analisi dei risultati".

3.2 Architettura dello strumento

L'architettura dello strumento, chiamato *Analisi_ Forum*, è mostrata in Figura 3.1. Il programma si divide in in 4 parti principali:

La *prima* parte, denominata in Figura 3.1 semplicemente "Forum", è quella in cui viene scelto un singolo forum che sarà oggetto successivamente del crawling e del calcolo della metrica, sempre in questa fase, oltre alla scelta, via è anche il settaggio di tutti i parametri specifici che sono necessari al Crawler per eseguire il suo compito come:

- caricamento dell'HTML dell'home page
- impostazione della struttura/layout
- svuotamento tabelle necessarie

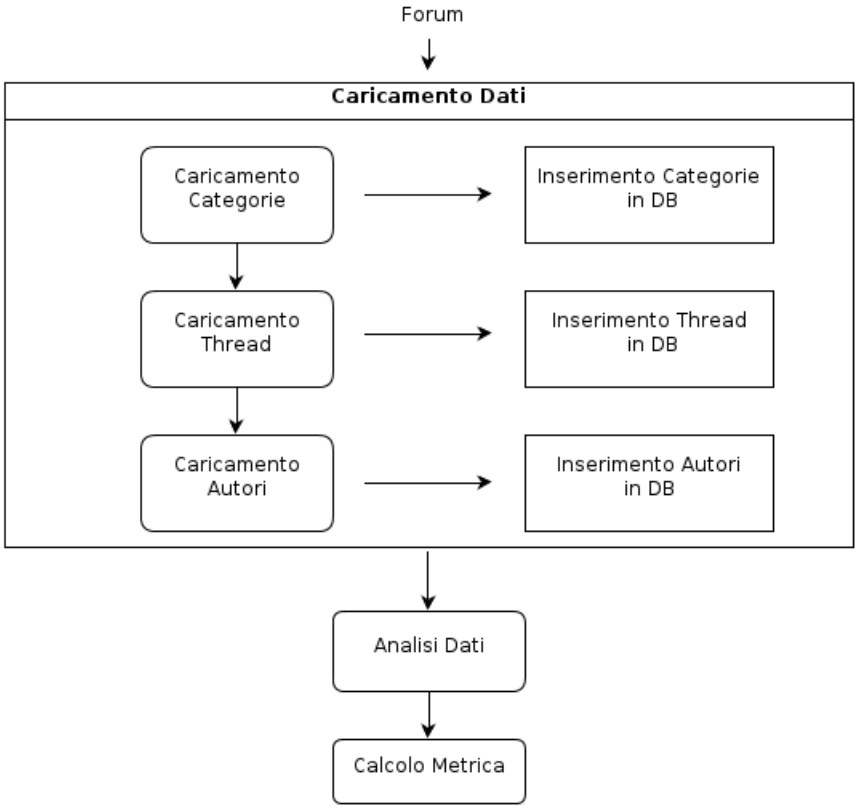

Figura 3.1: Architettura dello strumento Analisi_Forum

La *seconda* parte rappresenta il Crawler vero e proprio e sarà specificato in dettaglio nelle successive sezioni.

La *terza* parte, quella denominata "Analisi Dati", costituisce il core del calcolo della metrica ed la più teorica di tutto il programma, in questa parte si vanno ad analizzare completamente le tabelle fornire dalla fase precedente, quella del crawling, e si vanno ad estrarre tutti i parametri/dati che saranno necessari alla fase successiva per il calcolo vero e proprio della metrica. Sempre in questa fase vengono solamente estratti e forniti i dati diretti e semplici, alcuni sono addirittura estratti già nella fase di crawling perché direttamente disponibili in lettura all'interno dell'HTML delle pagine, gli al-

tri invece sono calcolati a partire dai dati contenuti nelle tabelle poiché non presenti direttamente o necessitano di qualche calcolo ulteriore.

Nella *quarta* ed ultima parte, quella di "Calcolo Metrica", vengono innanzitutto calcolati i dati complessi, costituiti da una combinazione di più dati semplici che costituiscono le singole parti facenti parte della metrica finale. Queste parti verranno poi normalizzate e unite per definire il valore finale di importanza del forum.

3.2.1 Crawler

Vengono qui descritte in dettaglio le varie parti che costituiscono il core del Crawler.

3.2.1.1 Caricamento Categorie

Questa parte rappresenta la prima fase delle tre che costituiscono il Crawler del forum, quella in cui vi è anche una primo caricamento dei dati che saranno poi forniti alle fasi successive.

Prima di tutto vi è da specificare a cosa corrispondono le "Categorie", queste non sono altro che tutte le varie tipologie di sottosezioni in cui è caratterizzabile il topic del Forum in analisi. Per esempio nel nostro caso, ambito medico, un Forum potrebbe definire Categorie come: Traumi, malattie ereditarie, malattie congenite, rotture ecc; oppure un Forum che tratta di diete potrebbe voler distinguere tra: vegetariane, a base di insalata, a basso contenuto proteico ecc. Si può dire che qualsiasi sia la metodologia di categorizzazione scelta dell'amministratore di un Forum, noi a livello di analisi definiamo come "Categorie" tutti i link che si trovano nella prima pagina e che portano a un elenco di Thread contenenti post di utenti.

Come si può notare dalla Figura 3.2 ci si può trovare di fronte a due scenari distinti quando si è in questa fase, la possibilità che vi siano o meno delle Sottocategorie. Dal lato sia logico che applicativo non vi è molta differenza

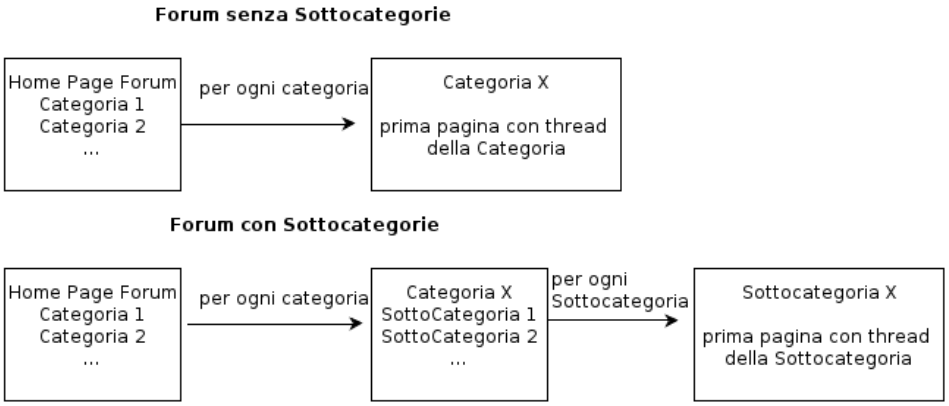

Figura 3.2: Fase Caricamento Categorie

tra un caso e l'altro, il programma è in grado di riconoscere in automatico in quale situazione si trova e comportarsi di conseguenza. In entrambi i casi comunque quando il Crawler incontra una pagina contenente Categorie va a caricare l'HTML del link della singola Categoria, sarà poi nella fase successiva, quella di Caricamento Thread, che capirà in quali delle due situazioni si trovi e determinerà se dovrà rieseguire la parte di Caricamento Categorie poiché avrà trovato delle Sottocategorie o meno. In questa fase viene sempre eseguito lo stesso algoritmo indipendentemente dal Layout del Forum e vi è anche il caricamento del primo dato semplice utile, quando reso disponibile dagli amministratori, ossia in numero di membri del forum. Quando non disponibile verrà mostrato successivamente come questo viene calcolato.

3.2.1.2 Caricamento Thread

Questa è senz'altro la fase più complicata di tutto il progetto di crawler, è la parte in cui vi sono il maggior numero di diversità anche tra Forum che utilizzano lo stesso tipo di Layout. In questa sezione si va ad applicare la stessa procedura per tutte le Categorie analizzate precedentemente, ossia si vanno ad analizzare tutte le pagine HTML delle singole Categorie. Come illustrato in Figura 3.3, il primo passo è determinare in quale caso ci si trova

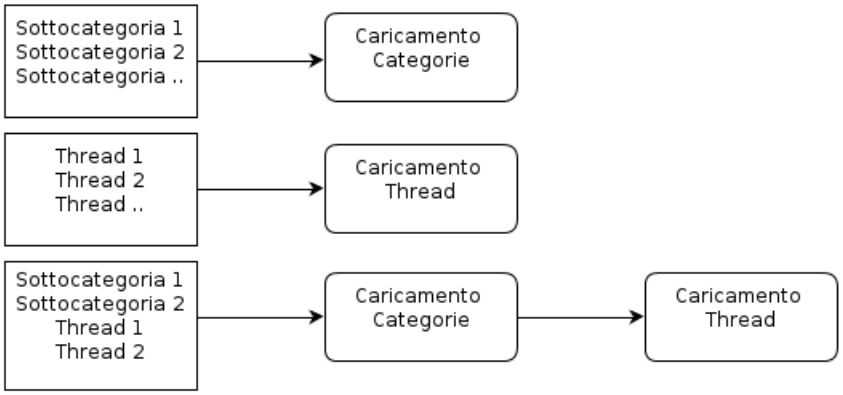

Figura 3.3: Fase Caricamento Thread

tra i 3 possibili:

- Pagina di sole Sottocategorie
- Pagina di soli Thread
- Pagina con Sottocategorie e Thread

Nel primo e nel terzo caso non vi è altro che un richiamo della fase di "Caricamento Categorie", ma la maggior parte delle volte ci si trova a dover analizzare una pagina contenente solo, o al più, Thread. Per ogni Categoria si vanno ad analizzare solamente i thread che hanno un ultimo post di un utente più recente di un periodo ti tempo dato come input al programma, questi dati saranno poi salvati perché da fornire come ingresso alla fase successiva, quella di "Caricamento Autori". In questa fase possono essere estratti dei dati semplici come il numero di thread totali e il numero di post totali di un Forum.

3.2.1.3 Caricamento Autori

In questa terza e ultima fase vengono trattati e aggiungi al database tutti gli autori che hanno scritto post in qualche Thread del Forum, vengono memorizzati anche dati tipo quanto tempo è passato tra il primo e il secondo post

di un thread(nel caso vi sia almeno una risposta) e la data di iscrizione di ogni autore, dato che si rivelerà fondamentale successivamente per determinare e variare il numero di utenti attivi. Un dato invece che è indispensabile estrarre in questa fase è il numero di post che ogni utente ha scritto da quando è iscritto al forum, in molti casi questo è direttamente disponibile come informazione esattamente come lo è la sua data di iscrizione, in altri casi invece questo dato va calcolato. Quando ci si ritrova in questa situazione non si può fare altro che incrementare di 1 i post scritti da quell'utente ogni volta che se ne incontra uno, tuttavia come è logico accorgersi, questo sistema è meno preciso rispetto al precedente, poiché potrebbero esserci post degli utenti in thread che non stiamo analizzando perché l'ultimo aggiornamento è precedente al periodo in esame. Tuttavia rappresenta una buona, e unica possibile, approssimazione, al limite si potrebbe far crawlare tutto il Forum da quando questo è stato creato e si otterrebbero gli stessi dati come se fossero stati forniti dall'amministratore.

3.2.2 Analisi Dati e Calcolo Metrica

Terminata la parte di crawling del forum specifico si passa a quella di analisi dei dati. In questa fase ci si trova come input un database completo con le tabelle Thread e Autori piene di dati e da cui è possibile estrarre tutte le informazioni necessarie anche al calcolo dei dati più complessi. Vediamo ora singolarmente quali sono e cosa rappresentano tutti i vari dati che sono estraibili, più o meno direttamente, dal Forum:

- **Numero Utenti Iscritti:** questo dato rappresenta uno delle più importanti, se non la più importante, informazioni che può servire a caratterizzare l'importanza di un Forum. Nella maggior parte dei casi è un dato che gli amministratori non hanno problemi nel rivelare e si trova direttamente nella Home Page, in altri invece è un dato non disponibile immediatamente e quindi va calcolato andando a contare quanti autori si incontrano nell'analisi del Forum. Ovviamente in

questo secondo caso non sarà così accurato come quello fornito poiché vi possono essere molte complicazioni come l'impossibilità di crawlare tutta la storia del Forum da quando è nato oppure perché tanti iscritti potrebbero non aver mai scritto in qualche Thread ma si limitano a leggere ed è quindi impossibile scoprire della loro esistenza ecc. In ogni caso questo è sicuramente un dato che da solo rappresenta già buona parte dell'importanza di un Forum, non potrà mai essere definito importante un Forum se non visto da una grande quantità di persone, ossia che non conti un buon numero di iscritti.

- **Numero Utenti Attivi:** anche questo dato sta a rappresentare una delle motivazioni indispensabili per la categorizzazione di un Forum. Tuttavia questo dato non è ovviamente già fornito direttamente dagli Amministratori ma va calcolato in base ai dati memorizzati dal Crawler e la maniera più corretta e logica per fare questo è senz'altro quella di andare a verificare la quantità di attività generata dai singoli iscritti all'interno di tutti i Thread del Forum. Oltre alla grande importanza, questo elemento ha anche un'alta soggettività che dipende tantissimo non solo dal Forum in analisi ma soprattutto da come la persona che sta calcolando la metrica vuole che questo dati sia rappresentato. Poiché all'interno dei dati forniti dal Crawler si ha anche il tempo di iscrizione di ogni singolo utente, questo dato è sempre calcolato come:

$$\frac{n^\circ \text{ post autore}}{mesi} > soglia$$

dove "n°post autore" rappresenta il numero di post che ogni autore ha scritto da quando iscritto al forum, "mesi" rappresenta il numero di mesi passati da quando l'autore si è iscritto, "soglia" rappresenta l'elemento che l'utente che sta calcolando la metrica vuole usare come discriminante per valutare se un dato utente è da considerarsi attivo. Questo elemento, insieme all'arco temporale di ultimo aggiornamento

in cui andare a caricare i Thread, rappresenta l'elemento più soggettivo e potenzialmente più indicativo nella formula di calcolo della metrica. Nel nostro caso abbiamo considerato come attivi gli utenti che dal momento dell'iscrizione al momento dell'analisi avessero una media almeno di 4 post/mese, questo dato è puramente indicativo e può essere facilmente modificato perché impostato come input dato al Crawler.

- **Numero Post:** questo dato è uno di quelli che si può sempre estrarre quasi direttamente dalle informazioni fornite da un Forum, in tutti infatti vi è sempre la dicitura di quanti post ci sono all'interno di un Thread, la somma di questi numero rappresenta il numero totale di post del Forum. Rappresenta principalmente la quantità di attività che si è generata all'interno del Forum e ovviamente, più attività vi è, più è probabile si generi in futuro e quindi sarebbe ottimale come sito per la diffusione di idee o rivalutazione di un marchio.

- **Numero Post Medio:** dato molto semplice da calcolare:

$$\frac{n^{\circ}\text{post totali}}{n^{\circ}\text{ Thread}}$$

Questo dato sembra banale e dotato di poco significato ma in realtà ha una informazione intrinseca che potrebbe ritenersi importante se non addirittura fondamentale per una persona che sta valutando un Forum. In questo elaborato sono stati trattati solo Forum in ambito medico poiché così era più semplice fare un raffronto sull'importanza di questi paramenti e sul valori finali della metrica, tutte discussioni fatte all'interno del capitolo:"Analisi dei risultati", ma se si sposta l'attenzione anche su altri tipi di Forum come per esempio quelli di risoluzione di problemi, o di riparazioni o simili, ecco che questo dato diventa molto interessante. In un Forum di risoluzione di problemi per esempio

questo elemento potrebbe rappresentare la qualità e competenza delle persona iscritte e attive a questo Forum e quindi in grado di risolvere anche problemi complessi in pochi passaggi, cosa che invoglierebbe moltissimo una persona a postare qui la sua problematica aumentando di conseguenza l'importanza generale del Forum. Questo dato potrebbe anche essere abbinato ad un altro elemento che non abbiamo considerato poiché in ambito medico poco rilevante: il numero di Thread chiusi (facilmente implementabile all'interno del Crawler). In un forum di risoluzione di problematiche un alto numero di thread chiusi potrebbe rappresentare un alta percentuale di problematiche risolte e quindi alta probabilità o di trovare già una soluzione al problema che si ha, o di poterla trovare nel giro di poco tempo. Ora dovrebbe essere chiaro come un basso numero di post per thread unito ad un alto numero di Thread chiusi potrebbe rappresentare una fonte di informazione e importanza incredibilmente grande per una persona che sta analizzando un Forum.

- **Distanza tra i primi 2 post:** rappresenta la media del tempo che intercorre tra il primo e il secondo post di ogni Thread del Forum, ovviamente solamente tra quelli che hanno una risposta:

$$\frac{\sum tempi}{n^\circ \text{ Thread}}$$

Questo dato rappresenta meno informazione rispetto ai precedenti tuttavia non è assolutamente da scartare perché potrebbe essere discriminante tra un buon ed un cattivo Forum in determinati contesti, tra questi potrebbe anche rientrare quello medico. Una persona che sta analizzando e calcolando la metrica per un determinato Forum potrebbe considerare come importante non solamente la correttezza delle risposte ad un determinato problema ma anche la velocità con cui l'utenza è in grado di generare una risposta plausibile. Questo tipo di ragionamenti può essere fatto in qualsiasi ambito: dalle problemati-

ca informatiche fino ad arrivare alle soluzioni per problemi allergici in campo medico. Si potrebbe pensare di affiancarlo anche alla rapidità con cui un Thread passa da essere creato a quando viene chiuso perché "risolto" o quanti Thread non hanno una risposta o altri tipi di dati ma sono tutti discorsi e ragionamenti che verranno trattati successivamente.

- **Numero di visite:** Questa è una informazione di importanza particolarmente strategica per un Forum e potenzialmente di grande interesse per molti utenti che andranno a calcolare la metrica. Potrebbe essere definita come il "bacino d'utenza" del Forum, ossia la quantità di persone che è in grado di raggiungere con la sua informazione ed il suo grado di diffusione all'interno del Web, ed in molti ambiti non solo informatici questo fattore rappresenta la chiave che fa la differenza tra il successo ed il fallimento. Si potrebbe far l'esempio di una qualsiasi industria interessata a far rivalutare la propria immagine, un Forum con anche poche persone attive ma che ne parlano bene e che è letto da veramente molte persone sarebbe l'ideale per loro e questo potrebbe essere l'unico parametro che per loro avrebbe significato, o quello che ne possiede la maggior parte. Sfortunatamente nel nostro caso questo è un dato che non siamo stati in grado di estrarre, il sito fornisce solamente il numero di visitatori presenti in quel momento e non quelli medi durante un arco temporale. Per recuperare questa informazione sarebbe stato necessario monitorare questo dato per molto tempo e alla fine fare la media di tutti i valori letti, tuttavia il Crawler funziona in modo "single-shot", ossia estrae tutte le informazioni in sequenza a partire dall'istante in cui è stato lanciato e quando ha estratto tutto termina l'esecuzione, il dato che si otterrebbe sarebbe quindi solo un rilievo istantaneo e non attendibile e rilevante ai fini statistici poiché troppo dipendente dall'orario in cui è stato lanciato (una rilevazione durante le ore diurne avrebbe infatti un valore logicamente molto mag-

giore rispetto a un rilievo fatto durante le ore notturne).

- **Numero di discussioni chiuse:** questo parametro a una prima analisi superficiale potrebbe sembrare rilevante ai fini di ottenere una valutazione finale valida e completa, potrebbe rappresentare infatti il numero di problemi risolti e, calcolando la distanza temporale tra il primo e l'ultimo intervento, la rapidità con cui questo avviene. Tuttavia ad un'analisi più accurata si scopre che le ragioni per cui un Thread risulta chiuso sono svariate e, per la maggior parte poco interessanti: si potrebbero avere infatti topic chiusi dagli amministratori perché creati nella categoria sbagliata, oppure da sempre chiusi perché contenenti regole di comportamento per gli iscritti al forum e che quindi non necessitano di ulteriori interventi da parte degli utenti, oppure ancora chiusi in automatico perché non hanno un aggiornamento da troppo tempo (e quindi il problema oggetto della discussione non è stato risolto) . Dovrebbe essere chiaro quindi che, anche se l'estrazione di questo valore numerico risulterebbe fattibile, non rappresenterebbe un effettivo incremento di informazione sull'importanza del Forum.

- **Media giornaliera di post:** rappresenta quanta attività si genera giornalmente in tutto il Forum, dato ricavabile con discreta facilità ma che non aggiunge informazione a quanto già ricavato con il numero di post medio che rappresenta già un indicatore sufficiente di questo fattore.

Terminata la parte di analisi e calcolo delle singole variabili, più o meno complesse, si passa alla fase di assemblaggio della metrica vera e propria. L'obiettivo è quello di ottenere un valore compreso in un range a cui avremo preventivamente attribuito una scala di importanza. Sono stati considerati

sufficientemente importanti ai fini di essere inseriti nel calcolo della metrica i seguenti parametri:

1. Numero di iscritti
2. Numero di utenti attivi
3. Distanza tra i primi 2 post
4. Numero di post medio
5. Numero di post

Questi valori, una volta estratti, risultano evidentemente molto diversi tra loro, basti pensare al numero di iscritti che potrebbe raggiungere anche le centinaia di migliaia ed alla distanza tra i primi due post che invece tende ad acquisire importanza al diminuire del suo valore assoluto. Si rende necessario quindi effettuare un'operazione di normalizzazione delle singole variabili, cosa che tuttavia non fattibile se si va ad estrarre i valori solamente da un Forum poiché non vi è una scala di valutazione con soglie assolutamente riconosciuta, si rende necessario quindi ottenere valori da più campioni possibili così da avere un calcolo, e quindi un valore finale anche di metrica, più preciso. Dopo avere estratto tutti i valori per un numero considerevole di Forum, nel nostro caso ne abbiamo analizzati 11, si può procedere ad effettuare le normalizzazioni dei singoli valori; abbiamo utilizzato due formule diverse[33], entrambe però che ci restituiscono valori compresi tra 0 e 1, in cui 0 rappresenta poca importanza mentre 1 rappresenta molta importanza.

Le prima formula è:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}}$$

dove x rappresenta il valore in considerazione, x_{min} rappresenta il valore minimo di quel tipo tra tutti quelli in analisi e x_{max} il suo opposto cioè il valore massimo. É stata utilizzata solamente per i parametri 1, 2 e 5 perché

sono i dati che al crescere del loro valore assoluto aumentano di importanza.

La seconda formula è:

$$y = \frac{x_{max} - x}{x_{max} - x_{min}}$$

dove le variabili della formula rappresentano gli stessi dati della precedente. In questo caso è stata utilizzata per i parametri 3 e 4 poiché questi aumentano di importanza al diminuire del loro valore assoluto.

Una volta ottenuti tutti i dati normalizzati si è dovuto pensare ad un modo per inserirli in un'unica formula matematica in modo da ottenere dei valori finali che rappresentassero la reale importanza del forum ma soprattutto che tra un valore e l'altro ci fosse una distanza che rappresentasse in maniera il più realistica possibile la differenza di importanza esistente tra 2 Forum. La formula finale elaborata è la seguente:

- **La somma:**

$$tot = \alpha 1 + \beta 2 + \gamma 3 + \delta 4 + \phi 5$$

  Rappresenta la soluzione più semplice possibile ma anche la più rappresentativa, i singoli valori compresi tra 0 e 1 vengono pesati in maniera differente in base alle preferenze dell'utente che sta calcolando la metrica. Sarà sufficiente dare un peso minore alla variabile (α, β, γ, δ o ϕ) associata al valore che si pensa abbia meno importanza rispetto agli altri e la valutazione finale sarà adeguata di conseguenza. L'introduzione dei pesi è stata pensata per rendere questa metrica il più generale possibile; essendo i dati estraibili da un qualsiasi Forum Web sempre gli stessi, si voleva fare in modo che una stessa formula andasse bene per ogni utente e per ogni Forum, con il sistema dei pesi questo è possibile poiché è sufficiente impostare come importante quello che per un altro utente invece risulta del tutto irrilevante. Nel nostro caso abbiamo optato per lasciare tutti i valori di pari importanza e quindi

le nostre valutazioni saranno sempre comprese tra 0 e 5.

Non sono state prese in considerazione ulteriori combinazioni perché non è stata trovata una motivazione ragionevole che ulteriori metriche possano aggiungere significatività a quella appena descritta. L'unico accorgimento che si potrebbe prendere in considerazione è di, nel caso si volesse mantenere il valore finale di valutazione compreso tra 0 e 1, prendere il valore ottenuto precedentemente e dividerlo per 5; in questo caso si avrà la media dei parametri mantenendo le stesse differenze di importanza tra i Forum ma contenute in un range più ristretto. Nel Capitolo 5 vengono presentati i risultati ottenuti sui Forum presi in considerazione.

3.3 Analisi statistica dei dati

I dati raccolti con il programma "Analisi_Forum" vengono passati al software statistico SPSS per analizzare il comportamento delle variabili e definire la distribuzione statistica del campione composto da 11 elementi. Il software statistico ci permette di determinare la bontà del campione e sottoporre i dati ad approfondite analisi, riguardo alla dipendenza tra le variabili. Dopo aver determinato il comportamento delle distribuzioni statistiche siamo in grado di scegliere quale coefficiente di correlazione determinare, tra Spearman, Pearson e Kendall. I dati a disposizione su cui avvengono le analisi statistiche sono di tipo numerico e si utilizza il loro valore grezzo senza normalizzazione, per non perdere informazione utile. Dopo questa fase di studio è possibile prendere decisioni riguardo le dipendenze tra parametri e agire nella modifica della metrica.
Se ci fossero dei valori fortemente correlati e tutti presenti nella metrica si avrebbe un valore che dipende troppo da un singolo fattore, andando a dare meno rilevanza agli altri che non sono con questo, e magari anche tra loro, correlati. Per completezza va detto che il campione potrebbe essere trop-

po piccolo per poter avere una scelta completamente definitiva e potrebbero esserci delle variazioni sul grado delle dipendenze.

Capitolo 4

Metrica basata sui contenuti

In questo capitolo viene trattata una estensione della metrica basata sui numeri presentata precedente lavorando sul contenuto testuale dei commenti degli utenti. Si descrivono i passi effettuati nella scelta delle parole chiave e dei parametri da modificare, per la composizione della metrica basata sul contenuto. Le metriche, ottenute utilizzando le parole chiave, sono state sottoposte ad uno studio approfondito per determinare che tipo di variazione si osserva rispetto ai risultati ottenuti dall'applicazione delle altre metriche. Si esamina inoltre, la possibilità di valutare un metodo più preciso di esaminare i contenuti dei post degli utenti sfruttando un applicativo, chiamato "SentEngine", in grado di estrarre la conoscenza dai messaggi e di essere più precisi circa l'argomento o marchio aziendale discusso dagli utenti.
Nella sezione 4.1 vengono trattate le motivazioni che hanno portato alla considerazione di una metrica che misuri l'importanza di un forum rispetto ad un argomento, prendendo in considerazione parole chiave o il contenuto semantico dei messaggi.
Nella sezione 4.2 si descrive la metrica con il filtro su parole chiave mentre nella sezione 4.3 viene descritto lo strumento che permette di estrarre i dati di tipo semantico di cui necessitiamo, infine nella sezione 4.4 viene descritto il metodo per la valutazione dei forum con l'utilizzo dei valori semantici e come è possibile raggiungere una metrica globale che tenga conto dell'aspetto

dimensionale ma anche di quello relativo ai contenuti.

4.1 Motivazioni

La prima metrica che è stata definita, descritta nel capitolo precedente, assegna già un punteggio che da l'idea dell'importanza di un Forum sul Web, ci sono però dei fattori importanti che i soli dati numerici calcolati o raccolti non sono in grado di descrivere. É noto dagli informatici e internauti che i Forum sono una notevole fonte di informazione e che l'importanza e il valore di un Forum dipende principalmente, se non soltanto, dal grado di conoscenza in materia posseduto dagli utenti iscritti che rispondono ai messaggi nei topic dedicati e dalla loro numerosità. Per esempio si potrebbe individuare un Forum dotato di tantissimi utenti iscritti ma poco acculturati sull'argomento in questione che generano quindi molta informazione inutile, questa situazione è sicuramente impossibile da individuare se ci affidassimo solamente al valore della metrica generata fino a questo momento poiché quest'ultima va a determinare principalmente la quantità di traffico che si viene a generare, e quindi il valore ottenuto sarebbe poco veritiero se il nostro scopo fosse misurare la qualità effettiva del Forum. D'altra parte si potrebbe scegliere un forum con altre peculiarità, facciamo l'esempio di uno che tratta di medicina generale, in cui ci sia un piccolo gruppo di dottori che a rotazione rispondono 24 ore su 24 alle domande degli utenti, in questo caso il parametro che enumera le persone attive potrebbero risultare piccolo, poiché le persone che visitano il sito presenterebbero la propria situazione e una volta avuta risposta non si ripresenterebbero per un po' di tempo.

Un'altra interpretazione della situazione precedente può essere spiegata come, la nascita di un sito emergente in cui il bacino d'utenza non risulta ancora molto ampio, sarebbe facile intuire la qualità del Forum medico con una metrica qualitativa e classificarlo tra i migliori, mentre a livello di traffico dati non ancora classificabile tramite la metrica definita sui numeri. Risulta

quindi evidente il fatto che sia necessario fare un ampliamento alla metrica fin qui calcolata nel caso si reputino importanti altre informazioni oltre a quelle relative al movimento di informazione generato. Come si può misurare la qualità degli utenti e di ciò che viene prodotto all'interno di un Forum? Si devono analizzare i post.

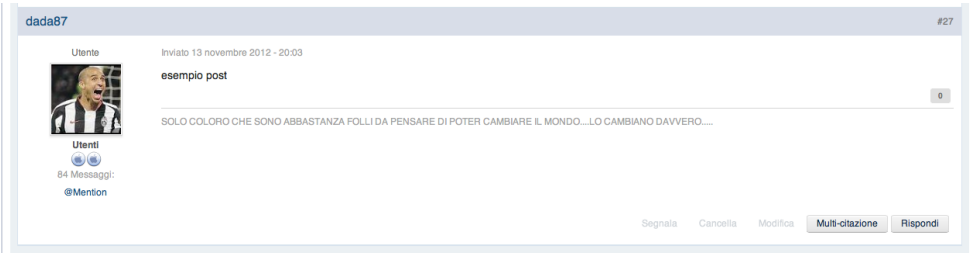

Figura 4.1: Esempio di post

I post sono dei singoli interventi che gli utenti scrivono all'interno del Forum, e più precisamente all'interno della categoria del Forum a cui il messaggio dovrebbe far riferimento e/o trattare come rappresentato in Figura 4.1. Questo è l'unico modo che hanno gli utenti di interagire tra loro all'interno di un Forum (non si considerano i messaggi privati che possono essere equiparati a semplici e-mail) e di aggiungere ulteriore informazione a quella già presente.
Bisogna sottolineare che non tutti i post siano portatori assoluti di nuova informazione, infatti in numerosi casi i post sono comunque indispensabili per caratterizzare la vita del sito Web. Alcuni scenari d'esempio:

- un utente che effettua un nuovo intervento per fare in modo che il Thread da lui creato ritorni in cima alla lista di visualizzazione (nella maggior parte dei Forum i Thread all'interno di ogni categoria sono organizzati in modo da essere in ordine da quello aggiornato più di recente a quello meno). L'utilità di riportare in evidenza il Thread potrebbe essere quello per aumentare la visibilità di un prodotto in vendita.

- potrebbero essere semplicemente risposte per intrattenere una relazione sociale tra persone e non aventi alcun tipo di informazione riferita all'oggetto del Thread.
- ci potrebbero essere post generati da computer in forum molto visitati creati allo scopo di pubblicizzare qualche prodotto.
- ecc.

Tanti messaggi di questo tipo, specialmente quelli pubblicitari o contenenti parole offensive, vengono eliminati prontamente dai Moderatori del Forum, questi sono persone che si assumono il compito di controllare e tenere pulito il contenuto di tutti i Thread di alcune categorie e di evitare che le discussioni degenerino in contenuti inappropriati o si vada fuori tema, *offtopic*. Tuttavia si verificherà sempre la presenza di questi messaggi di scarso valore informativo ma questi verranno riconosciuti e scartati. Vengono affrontati due metodi per aggiungere alla metrica basata sui numeri informazioni relative ai contenuti dei post, una è il risultato della selezione dei post in cui sono presenti parole chiave e l'altra è una selezione più precisa sulla semantica sfruttando uno strumento di terze parti, chiamato SentEngine.

4.2 Adattamento metrica con filtro testuale

Il primo metodo per ampliare le metriche calcolate è considerare la presenza all'interno dei post del Forum una determinata parola e valutare i parametri in base questo fattore. Ad esempio il punteggio della metrica così calcolato potrebbe essere utile per una azienda che deve mettere sul mercato un nuovo prodotto farmaceutico, a scegliere il forum adatto per introdurre immagini e video pubblicitari. La classificazione dei forum con questo metodo di giudizio serve per farsi un'idea dell'opinione della popolazione riguardo al proprio marchio, individuare le discussioni che trattano di argomenti specifici, rilevare il livello di conoscenza delle persone riguardo a temi specifici ecc.. A questo punto, occorre solo scegliere le parole chiave, poi per l'azienda è facile

scegliere il forum che ha ottenuto il punteggio più alto in cui pubblicizzare un prodotto o semplicemente intervenire nelle discussioni per aggiustare le informazioni relativa al proprio marchio.
La valutazione della metrica con il filtro sul testo dei messaggi però non può essere considerata come indice di qualità dei messaggi, non è stato possibile assegnare un giudizio sul contenuto. Ipotizziamo di esprimere il punteggio di un Forum ricercando il nome di un prodotto, il confronto tra forum non è più valutato solo in base all'attività degli utenti, ma anche vincolato alla presenza all'interno dei post della parola chiave. I test sono stati condotti utilizzando solo alcune parole chiave relative ai problemi comuni e non troppo specifici come "diet", "diabete", "cholesterol" per poter confrontare risultati su argomenti di ampio interesse e prevedendo numeri adeguati. La metrica viene calcolata utilizzando il metodo della somma pesata come in precedenza, i parametri che vengono utilizzati sono estratti dal programma in base alle parole chiave mentre i pesi hanno valore 1. Il concetto di filtrazione prevede la ricerca della parola chiave all'interno di tutti i messaggi di ogni forum, valutando i thread non più vecchi di sei mesi, in questo modo si ottiene il numero esatto di post che contengono il termine. Da questa informazione è possibile risalire, interrogando il DB, all'elenco dei thread che contengono i post ed estrarre le informazioni per i termini mancanti. Gli utenti attivi non si determinano come descritto precedentemente, cioè conteggiando tutti gli utenti con una media di scrittura di messaggi mensile superiore ad un valore assegnato, ma vengono presi tutti gli autori presenti nel thread dove è stato trovato almeno un post con la parola chiave. In questo modo si considera l'insieme degli utenti che si riferisce al tema della parola chiave. Il termine che definisce la distanza media in giorni tra due post viene calcolato sui thread che contengono almeno un post con la parola chiave. Nel caso di forum di tipo farmaceutico e medicinale ci si aspetta di trovare una struttura fatta da domande e risposte, la successione di messaggi dello stesso autore può essere considerata come messaggio/richiesta unica, infatti la differenza temporale avviene tra il primo post del thread e il primo post successivo

con autore diverso evitando così l'eventualità di calcolare la distanza tra due post con lo stesso autore. Il termine che definisce il numero medio di post per ogni thread è calcolato come in precedenza, prendendo il numero totale di post presenti in ogni thread in cui compare almeno un post con la parola chiave e dividerlo per il numero dei thread.

I numeri delle frequenze vengono normalizzati seguendo la stessa tecnica descritta nel capitolo precedente, applicando la metrica della somma pesata utilizzando i parametri filtrati si ottiene la formula:

$$tot = \alpha 1 + \beta 2.bis + \gamma 3.bis + \delta 4.bis + \phi 5.bis$$

In cui i parametri equivalgono a :

1. Numero di utenti iscritti
2. bis: Numero di utenti che hanno scritto nei thread in cui compare un post col testo filtrato.
3. bis: Distanza tra i primi 2 post nei thread in cui compare almeno un post col testo filtrato.
4. bis: Numero di post medio per thread, tra quei thread in cui è presente almeno un post col testo filtrato.
5. bis: Numero di post che sono presenti nei thread in cui ricorre la parola chiave.

É stato possibile condurre i test per la realizzazione della metrica sui forum su cui si è potuto salvare i dati dei post, ovvero: WeightLossCenter, LowCarbFriends, DiabetesDaily, ObesityDiscussion, WeightLoss, WomensHealth, USMessageBoard, ThreeFatChicks.

4.3 Architettura dello strumento "SentiEngine"

I dati utilizzati per questa seconda parte sono stati estratti tramite lo strumento *SentiEngine* mostrato in Figura 4.2.

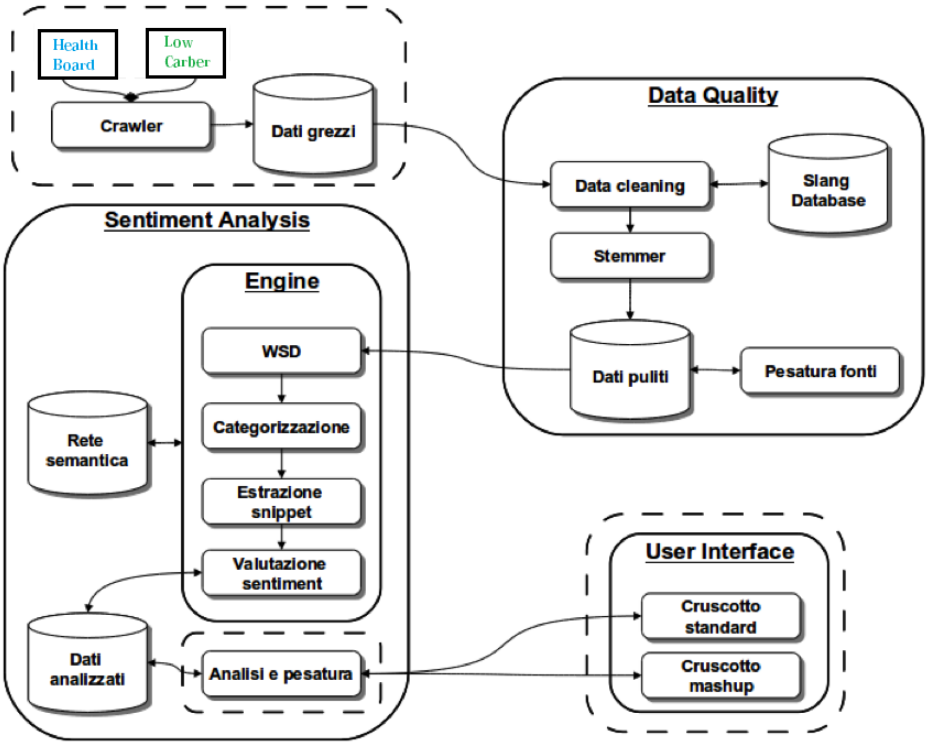

Figura 4.2: Architettura dello strumento di Web reputation

Ci è stato fornito un database formato da 4 tabelle per ciascun Forum: *post, autori, thread, snippet.*
La tabella *post* contiene tutti gli interventi effettuati dagli utenti nel Forum, ogni riga è caratterizzata anche dall'id del thread in cui l'intervento si trova, dell'autore che l'ha creato e dalla sua data di pubblicazione.

La tabella *autori* contiene tutti gli utenti che hanno effettuato almeno un intervento nel Forum, viene estratto anche il nome e gli viene associato un

identificativo per poterci fare riferimento nelle altre tabelle.

La tabella *thread* contiene tutti i dati necessari a caratterizzare una discussione specifica, ossia autore che l'ha creato,data di creazione, titolo e URL.

La tabella *snippet* racchiude le informazione relative al contenuto dei messaggi contestualizzando le parole, i campi che determinano la tabella sono: un id univoco per ogni analisi del post, un campo categoria che racchiude il significato generico del post individuato in un elenco di problemi legati al campo medico, un campo brand indica il marchio aziendale se il contenuto del post richiama un prodotto farmaceutico presente nella base di conoscenza dello strumento SentiEngine, un campo subject definisce l'argomento del post, esempio: "ASPIRIN", "HEADACHE", "THERAPY", la data del post, il testo, e altri indici che permettono l'estrazione delle informazioni legate alle altre tabelle.

4.4 Metrica sui contenuti

La possibilità di definire una metrica basata sui contenuti semantici è stata condizionata dalla analisi del DataBase fornito dal "SentiEngine", che raccoglie i dati relativi ai forum di LowCarber e tre categorie di HealthBoard. Il campione non è sufficientemente ampio per verificare l'efficacia della misurazione ma è stato possibile osservare il confronto tra due forum diametralmente opposti in quanto a dimensioni. L'idea è quella di sfruttare il lavoro dello strumento SentEngine determinando tutti i messaggi relativi ad argomenti di nostro interesse, evitando le ambiguità provocate dalla sola ricerca testuale della parola chiave, per creare una metrica semantica precisa. In questa metrica non si deve tenere conto delle dimensioni del forum, anche se la frequenza degli argomenti è fortemente influenzata dal numero di post, dagli utenti attivi che ci scrivono, ecc. É possibile verificare se LowCarber è più adatto a trattare alcune argomentazioni rispetto all'al-

tro forum di riferimento HealthBoard evidenziando il numero di ricorrenze estratte dallo strumento. Nella descrizione del "SentEngine" è stato elencato l'insieme dei parametri che vengono presi in considerazione per la valutazione semantica, in particolare i campi del DataBase soggetto, categoria e brand rappresentano il perno dello studio effettuato.

Capitolo 5

Analisi dei Risultati

Questo capitolo risulta essere di fondamentale importanza perché raccoglie le osservazioni di ogni fase dell'analisi dei forum evidenziando i punti di forza di ogni metrica utilizzata. Per calcolare il punteggio della metrica si estraggono i valori dei parametri tramite il software di crawling, ma a causa della differenza di unità di misura e dimensione occorre applicare preventivamente una normalizzazione dei dati.
Viene descritta l'analisi statistica per determinare la presenza di variabili correlate e in base all'esito di essa sono state generate diverse metriche osservando risultati differenti. La metrica che assegna il punteggio di importanza dei forum in base alle parole chiave viene ampiamente esaminata per definire la sua validità e l'utilità. Infine vengono discussi i risultati dell'analisi semantica.

5.1 Analisi dei dati

Il software di crawling ha scansionato tutti i forum salvando per ognuno le informazioni in un DB, il passo successivo è stato calcolare i parametri e visualizzarli in Tabella 5.1. Da una prima lettura dei dati si notano i forum più popolati osservando il numero di iscritti, HealthBoard conduce la classifica con circa 900mila iscritti, seguono a lunga distanza ThreeFatChicks e

	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post [giorni]	media di post per thread	n° di post
1	diabetesdaily	85580	210	36	4,08142	11930
2	healthboard	928360	3900	27	3,94504	165029
3	lowcarber	145474	240	29	23,9294	38646
4	lowcarbfriends	66469	3116	0	65,4213	184161
5	obesitydiscussion	1969	1199	2	5,71815	81678
6	renewedreflections	27429	510	3	10,8761	51955
7	threefatchicks	155724	14523	0	17,6199	236177
8	usmessageboard	30617	10176	0	31	4173860
9	weightloss	5789	60	29	1,98033	1812
10	weightlosscenter	2400	215	5	5,35379	4101
11	womenshealth	21996	5597	1	9,69658	254506

Tabella 5.1: Tabella dei parametri

LowCarber con circa 150mila utenti registrati mentre gli altri forum seguono con numeri di molto inferiori. Gli utenti attivi dimostrano la vivacità di un forum e rispetto al numero di iscritti determina il numero di utenti che hanno scritto con continuità negli ultimi 6 mesi. Valutando solo questo parametro ThreeFatChicks è il forum che conta il maggior numero di utenti attivi con 14000 seguito da USMessageBoard con 10000 utenti mentre HealthBoard conta "solo" 3900 utenti attivi. La differenza tra il numero di utenti iscritti e utenti attivi di HealthBoard può essere spiegata dal tipo di utente che frequenta il sito HB, si iscrive per ottenere informazioni e richiedere aiuto, ottiene una risposta soddisfacente e non si presenta più. In questo modo il numero di iscritti incrementa nel tempo, ma la bassa permanenza si scopre nel numero degli utenti attivi.

Il tipo di forum che viene trattato è prevalentemente costruito sull'aiuto peer to peer, cioè domande e risposte fornite da persone comuni; dal punto di vista dell'utente è preferibile individuare un forum con un tempo di risposta vicino ai bisogni dell'utente e ottenere le informazioni nel più breve tempo possibile; quello che ci si aspetta dai forum con molti utenti attivi è che il valore di attesa sia più basso rispetto ai forum con pochi utenti attivi. Dalla stessa tabella 5.1 osserviamo che i forum USMessageBoard, ThreeFatChicks, LowCarbFriends e WomensHealth hanno tempi di risposta tra 0-1 giorni, ciò significa che in media occorre attendere meno di 24 ore per otte-

nere una risposta da parte di un altro utente. Il risultato di questi forum poteva essere previsto data la loro dimensione mentre ciò che stupisce è che gli utenti di HealthBoard attendono circa 27 giorni per ottenere la prima risposta all'apertura di un thread, paragonandosi a forum come LowCarber e WeightLoss con tempi medi attorni i 29 giorni, ma utenti attivi rispettivamente 240 per il primo e 60 per il secondo. In positivo si notano i forum di RenewedReflections e WeightLossCenter che modesti nel numero di utenti sia iscritti che attivi hanno tempi di attesa dai 3 a i 5 giorni in media. Il risultato ottenuto per questo parametro è comunque da valutare in base alle dimensioni in gioco, riferendosi ad Healthboard e l'ipotesi che gli utenti scrivano maggiormente per chiedere piuttosto che risolvere i problemi di altri, si prevede che giornalmente ci sia uno sbilanciamento tra domande e risposte.

L'alta generazione di thread rende più lunga l'attesa di una risposta mentre è logico pensare che i forum con numero di utenti basso hanno una produzione di thread facilmente gestibile, riuscendo a soddisfare gli utenti in minor tempo. Il termine che indica quanti messaggi in media sono presenti in un thread è utile per provare a definire la chiarezza di un sito di discussione, se il numero si aggira tra i 5 e i 10 post in media sappiamo che la discussione potrebbe essere racchiusa in una pagina. Supponendo il termine adatto a rappresentare la realtà si può affermare che la comodità di leggere i commenti su una sola pagina sia un vantaggio, qualora si cerca una risposta. D'altra parte valutando solo in maniera numerica questo aspetto non si può speculare nulla né sui tempi né sull'effettivo recupero delle informazioni.

I dati estratti dal programma evidenziano LowCarbFriends come un forum in cui i thread vengono utilizzati come canali di informazione generica, osservando attivamente la composizione del sito è stata provata la presenza di pochi thread in cui sono presenti molte pagine di post e non vengono generate nuove discussioni per domande specifiche. D'altra parte i forum WeightLoss ed HealthBoard con valori rispettivamente di 1,9 e 3,9 sono dovuti ai thread

ancora senza risposta che incidono sul calcolo della media. Infine il termine dei post conteggia tutti i post scritti negli ultimi 6 mesi mettendo in evidenza i forum che nel periodo di osservazioni dimostrano maggior attività, USMessageBoard, ThreeFatChicks e WomensHealth.

		dati normalizzati				
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post	media di post per thread	n° di post
1	diabetesdaily	0,090254547	0,010371292	0	0,966881181	0,002425188
2	healthboard	1	0,265505082	0,25	0,969030896	0,039121554
3	lowcarber	0,154907593	0,012445551	0,194444444	0,654023733	0,008828757
4	lowcarbfriends	0,069625029	0,211297794	1	0	0,043707311
5	obesitydiscussion	0	0,078752679	0,944444444	0,941081922	0,019143116
6	renewedreflections	0,027482996	0,031113877	0,916666667	0,859778783	0,012018797
7	threefatchicks	0,165972036	1	1	0,753478391	0,056175049
8	usmessageboard	0,030924307	0,69943995	1	0,542572095	1
9	weightloss	0,004123529	0	0,194444444	1	0
10	weightlosscenter	0,000465246	0,010717002	0,861111111	0,946825214	0,000548651
11	womenshealth	0,021618302	0,382838968	0,972222222	0,878371185	0,060568335

Tabella 5.2: Tabella dei parametri normalizzati

La normalizzazione dei dati è un passo obbligato per continuare nello studio della metrica e utilizzare i parametri in modo da assegnare la stessa importanza. Per merito della normalizzazione si ottengono valori compresi l'intervallo 0 e 1, proporzionando le differenze numeriche.

La metrica di somma pesata viene osservata ponendo tutti i pesi uguali a 1, lasciando ad uno studio futuro l'onere di evidenziare le differenze e peculiarità della scelta di ogni peso, in questa situazione viene valutato l'impatto di ogni indice in maniera equa e l'assegnamento dei punteggi di ogni forum risulta composta dal contributo di tutti i parametri. Il massimo punteggio si verifica nel caso forum in fase di normalizzazione risulta avere tutti gli indici uguali a uno ottenendo il punteggio di 5, questo scenario si può verificare se il confronto avviene con un campione di forum piccolo o sproporzionato, ad esempio tra ThreeFatChicks e LowCarber, il primo assumerebbe punteggio

	nome_forum	sum
1	diabetesdaily	1,069932208
2	healthboard	2,523657532
3	lowcarber	1,024650078
4	lowcarbfriends	1,324630134
5	obesitydiscussion	1,983422162
6	renewedreflections	1,84706112
7	threefatchicks	2,975625475
8	usmessageboard	3,272936353
9	weightloss	1,198567973
10	weightlosscenter	1,819667225
11	womenshealth	2,315619012

Tabella 5.3: Tabella punteggi della metrica di somma pesata

massimo mentre il secondo 0. Dalla tabella 5.3 della metrica "sum" possiamo notare fin da subito che i migliori forum sono nell'ordine USmessageBoard, ThreeFatChicks e HealthBoard, il punteggio non è determinato esclusivamente dalle dimensioni del forum ma è composto dal termine che definisce la velocità di risposta, e dal numero non troppo elevato di post per Thread. Maggior interesse in questa metrica è stato suscitato osservando il forum con dimensioni più contenute sul numero di utenti, ObesityDiscussion, non si trovi all'ultima posizione nella classifica per merito degli utenti attivi, prontezza di risposta e dimensione dei thread contenuta, come evidenzia il grafico.

La metrica tende ad aiutare il forum con numero di iscritti maggiore, HealthBoard, che nonostante gli indici di attività siano inferiori rispetto agli altri forum si mantiene in terza posizione, questa agevolazione però non è sufficiente. Si conferma la supremazia dei forum di USMessageBoard e ThreeFatChicks con punteggi rispettivamente di 3,2 e 2,9; hanno entrambi indici di attività e numero di post superiori ad HealthBoard, mentre cedono sul contributo del numero di iscritti. Un'osservazione in negativo la si può dare del forum Low Carber, la presenza elevata del numero di iscritti non

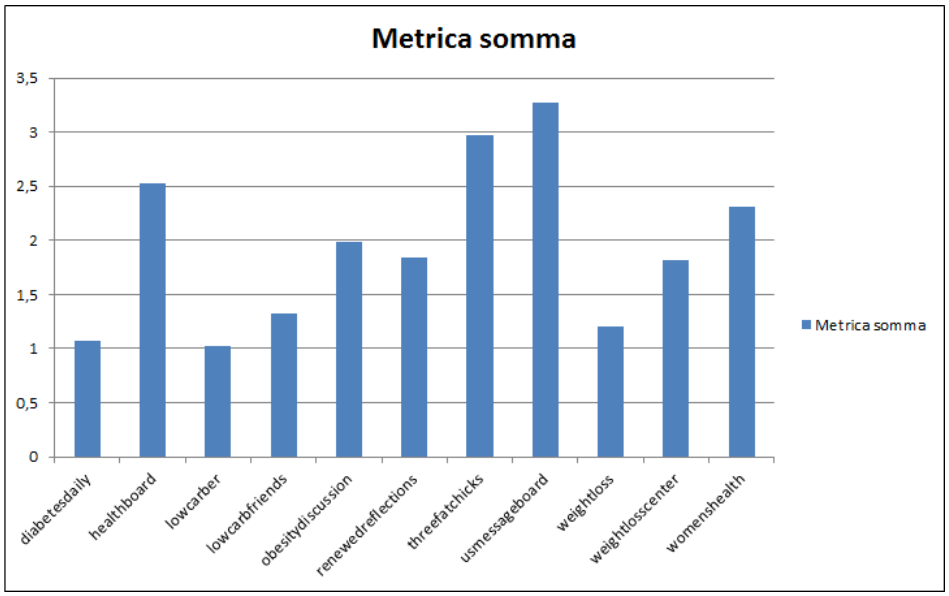

Figura 5.1: Istogramma punteggi della metrica di somma pesata

si conferma nei dati dell'attività del forum, come possibile causa di questa situazione l'abbandono della visita del sito da parte degli utenti, nel periodo di osservazione il forum non risulta particolarmente buono per nuovi utenti o per aprire un thread a meno di attendere un mese prima di ottenere la prima risposta.

La metrica risulta essere veritiera secondo il punto di vista logico dell'osservazione dei forum, assegnando punteggi elevati per quei forum con attività degli utenti elevata e considerando il numero di iscritti come potenzialmente in grado di intervenire, anche nella scelta tra i forum intermedi la metrica riesce a dividere piuttosto bene i punteggi, per RenewedReflection il maggior numero di utenti iscritti permette di prevalere su WeightlossCenter che primeggia leggermente nel resto dei parametri, entrambi ottengono un punteggio attorno a 1,8. É necessario riflettere se la metrica così utilizzata sia priva di elementi ripetuti, cioè se i parametri siano tutti necessari per rilevare l'aspetto numerico dei forum o se è possibile ridursi a conteggiare solo alcuni di essi. A sostegno di questa osservazione valutiamo l'indagine statistica

della correlazione nella sezione successiva.

5.2 Risultati della correlazione statistica

É stata presentata nella sezione precedente la metrica basata sui numeri, ma non è stata valutata la possibilità che i parametri utilizzati siano dipendenti l'uno dal altro. Per verificare questa ipotesi è stata condotta un'analisi statistica per calcolare la correlazione tra i dati e in caso positivo raccogliere informazioni con metriche semplificate. La correlazione tra due variabili si calcola utilizzando una correlazione bi-variata che è una metodologia simmetrica in cui si considerano due variabili X e Y sullo stesso piano causale. Il coefficiente di correlazione lineare di Bravais Pearson misura il tipo e l'intensità della relazione lineare tra due variabili X e Y, indicando il livello di significatività dei risultati e nel caso dare le giuste osservazioni nonostante il numero del campione limitato.
Purtroppo non sempre è possibile utilizzare il coefficiente di Pearson per valutare la correlazione tra due variabili. I casi in cui, una o entrambe le variabili di interesse sono ordinali, una o entrambe le variabili non sono distribuite normalmente e la relazione che si intende misurare non è di tipo lineare. In questi casi è possibile utilizzare il coefficiente ρ di Spearman o il coefficiente τ di Kendall [34], questi coefficienti infatti non richiedono particolari assunzioni sulla distribuzione delle variabili studiate, si usa parlare di "correlazione non parametrica". Il coefficiente τ di Kendall è particolarmente utilizzato con riferimento a dati qualitativi ordinali, oppure a dati quantitativi ma con un numero ridotto di modalità distinte; mentre il coefficiente ρ di Spearman quando riferito a caratteri quantitativi, è particolarmente robusto rispetto a valori estremi (outliers), in quanto basato sui ranghi e non sulle modalità effettive; questo fatto lo rende più significativo, in certi casi, di altri indici di associazione per caratteri quantitativi[35].

In Tabella 5.4 e Tabella 5.5 si evidenzia la forte correlazione statistica tra

Correlazioni			n_iscritti	n_utenti_attivi	distanza_2_post	media_post_thread	n_post
Rho di Spearman	n_iscritti	Coefficiente di correlazione	1,000	,355	,046	,209	,282
		Sig. (2-code)		,285	,893	,537	,401
		N	11	11	11	11	11
	n_utenti_attivi	Coefficiente di correlazione	,355	1,000	-,823**	,536	,955**
		Sig. (2-code)	,285		,002	,089	,000
		N	11	11	11	11	11
	distanza_2_post	Coefficiente di correlazione	,046	-,823**	1,000	-,690*	-,823**
		Sig. (2-code)	,893	,002		,019	,002
		N	11	11	11	11	11
	media_post_thread	Coefficiente di correlazione	,209	,536	-,690*	1,000	,609*
		Sig. (2-code)	,537	,089	,019	.	,047
		N	11	11	11	11	11
	n_post	Coefficiente di correlazione	,282	,955**	-,823**	,609*	1,000
		Sig. (2-code)	,401	,000	,002	,047	.
		N	11	11	11	11	11

**. La correlazione è significativa al livello 0,01 (2-code).

*. La correlazione è significativa al livello 0,05 (2-code).

Tabella 5.4: Tabella di correlazione di Spearman

il numero di post totali e il numero di utenti attivi con un coefficiente di correlazione di Spearman di 0,955 mentre per Kendall 0,855 ed un livello di significatività che conferma il risultato. Una correlazione inversa tra il numero di post totale e la distanza in giorni tra due messaggi con coefficiente di Spearman -0,823 e di Kendall -0,661 con livello di significatività nei limiti per considerare affidabile l'ipotesi di dipendenza. Infine anche tra utenti attivi e la distanza tra due post si dimostra dipendenza, al crescere degli utenti attivi si riducono i tempi di attesa tra domanda e risposta, con indici di correlazione rispettivamente di -0,823 di Spearman e -0,699 di Kendall con significatività minore di 0,03 a confermare il concetto. La dipendenza tra questi fattori era prevedibile dal fatto che misurano indirettamente l'attività del forum utilizzando i post come discriminante. Dalla tabella inoltre è possibile notare la presenza di altre correlazioni, per esempio tra il numero di post totali e il numero di post medio per thread e tra la distanza tra post e il numero medio di post per thread, tuttavia il campione non è abbastanza numeroso per poter confermare la dipendenza tra le variabili come evidenzia il livello di significatività troppo alto. A causa del numero limitato di forum a nostra disposizione non si può speculare ulteriormente sui parametri in

Correlazioni			n_iscritti	n_utenti_attivi	distanza_2_post	media_post_thread	n_post
Tau_b di Kendall	n_iscritti	Coefficiente di correlazione	1,000	,236	-,019	,164	,164
		Sig. (2-code)	.	,312	,937	,484	,484
		N	11	11	11	11	11
	n_utenti_attivi	Coefficiente di correlazione	,236	1,000	-,699**	,418	,855**
		Sig. (2-code)	,312	.	,003	,073	,000
		N	11	11	11	11	11
	distanza_2_post	Coefficiente di correlazione	-,019	-,699**	1,000	-,585*	-,661**
		Sig. (2-code)	,937	,003	.	,014	,006
		N	11	11	11	11	11
	media_post_thread	Coefficiente di correlazione	,164	,418	-,585*	1,000	,418
		Sig. (2-code)	,484	,073	,014	.	,073
		N	11	11	11	11	11
	n_post	Coefficiente di correlazione	,164	,855**	-,661**	,418	1,000
		Sig. (2-code)	,484	,000	,006	,073	.
		N	11	11	11	11	11

**. La correlazione è significativa al livello 0,01 (2-code).

*. La correlazione è significativa al livello 0,05 (2-code).

Tabella 5.5: Tabella di correlazione di Kendal

tabella.

Il passo successivo è quello di applicare le osservazioni emerse nello studio di correlazione alla metrica della somma pesata per classificare i forum. L'idea è quella di calcolare il punteggio di un forum utilizzando come parametri i due termini indipendenti e di aggiungere a turno uno o più parametri correlati, fino ad ottenere sei diverse interpretazioni.

5.3 Risultati delle metriche post correlazione

La correlazione statistica ha evidenziato dipendenza tra i parametri che identificano gli utenti attivi, la distanza tra due post e il numero totale di post, è possibile affrontare la modifica della metrica vista precedentemente utilizzando le informazioni restituite dall'analisi statistica.
É utile presentare le diverse evoluzioni della metrica con alcune premesse: i pesi delle variabili sono fissati a 1, ma è possibile da parte degli utenti personalizzare la scelta di assegnamento, tuttavia non sarà discussa l'inci-

denza sulle variazione dei pesi; il risultato ottenuto dalle sei variazioni della metrica determinano il punteggio del forum in modo diverso, ma acquisiscono lo stesso significato; si identificheranno le differenze rispetto ai risultati ottenuti dalla metrica di riferimento "somma pesata" vista in precedenza; le metriche vengono calcolate utilizzando i parametri normalizzati già calcolati sul campione degli 11 forum; i termini del numero di iscritti e media di post per thread sono sempre utilizzati.

Post correlazione	A: senza U_attivi	B: senza dist2p	C: senza post	D: senza ua e d2p	E: senza u_attivi e post	F: senza d2_p e post_i
diabetesdaily	1,059560916	1,069932208	1,06750702	1,059560916	1,057135728	1,067507
healthboard	2,25815245	2,273657532	2,484535978	2,00815245	2,219030896	2,234536
lowcarber	1,012204528	0,830205634	1,015821321	0,817760083	1,00337577	0,821377
lowcarbfriends	1,11333234	0,324630134	1,280922823	0,11333234	1,069625029	0,280923
obesitydiscussion	1,904669483	1,038977718	1,964279046	0,960225039	1,885526367	1,019835
renewedreflections	1,815947243	0,930394453	1,835042323	0,899280577	1,803928446	0,918376
threefatchicks	1,975625475	1,975625475	2,919450427	0,975625475	1,919450427	1,919450
usmessageboard	2,573496402	2,272936353	2,272936353	1,573496402	1,573496402	1,272936
weightloss	1,198567973	1,004123529	1,198567973	1,004123529	1,198567973	1,004124
weightlosscenter	1,808950223	0,958556114	1,819118573	0,947839112	1,808401571	0,958007
womenshealth	1,932780044	1,34339679	2,255050677	0,960557821	1,872211709	1,282828

Tabella 5.6: Punteggio forum post correlazione

Facendo riferimento alla tabella 5.6 si possono descrivere le metriche ricavate, la prima che è stata definita come A riporta i risultati della metrica basata sulla somma pesata con 4 parametri: numero di utenti iscritti, distanza media in giorni tra due post, numero medio di post per thread, numero totale dei post scritti in 6 mesi.

La metrica B definisce i punteggi dei forum con 4 parametri: numero di utenti iscritti, numero di utenti attivi, numero medio di post per thread, numero totale dei post scritti in 6 mesi.

La metrica C definisce i punteggi dei forum con 4 parametri: numero di utenti iscritti, numero di utenti attivi, distanza media in giorni tra due post, numero medio di post per thread.

La metrica D definisce i punteggi dei forum con 3 parametri: numero di

utenti iscritti, numero medio di post per thread, numero totale dei post scritti in 6 mesi.

La metrica E definisce i punteggi dei forum con 3 parametri: numero di utenti iscritti, distanza media in giorni tra due post, numero medio di post per thread.

La metrica F definisce i punteggi dei forum con 3 parametri: numero di utenti iscritti, numero di utenti attivi, numero medio di post per thread.
Il confronto delle metriche su base numerica è piuttosto gravoso e per tale motivo sfruttiamo un grafico a istogramma 5.2 per notare le differenze in maniera visiva.

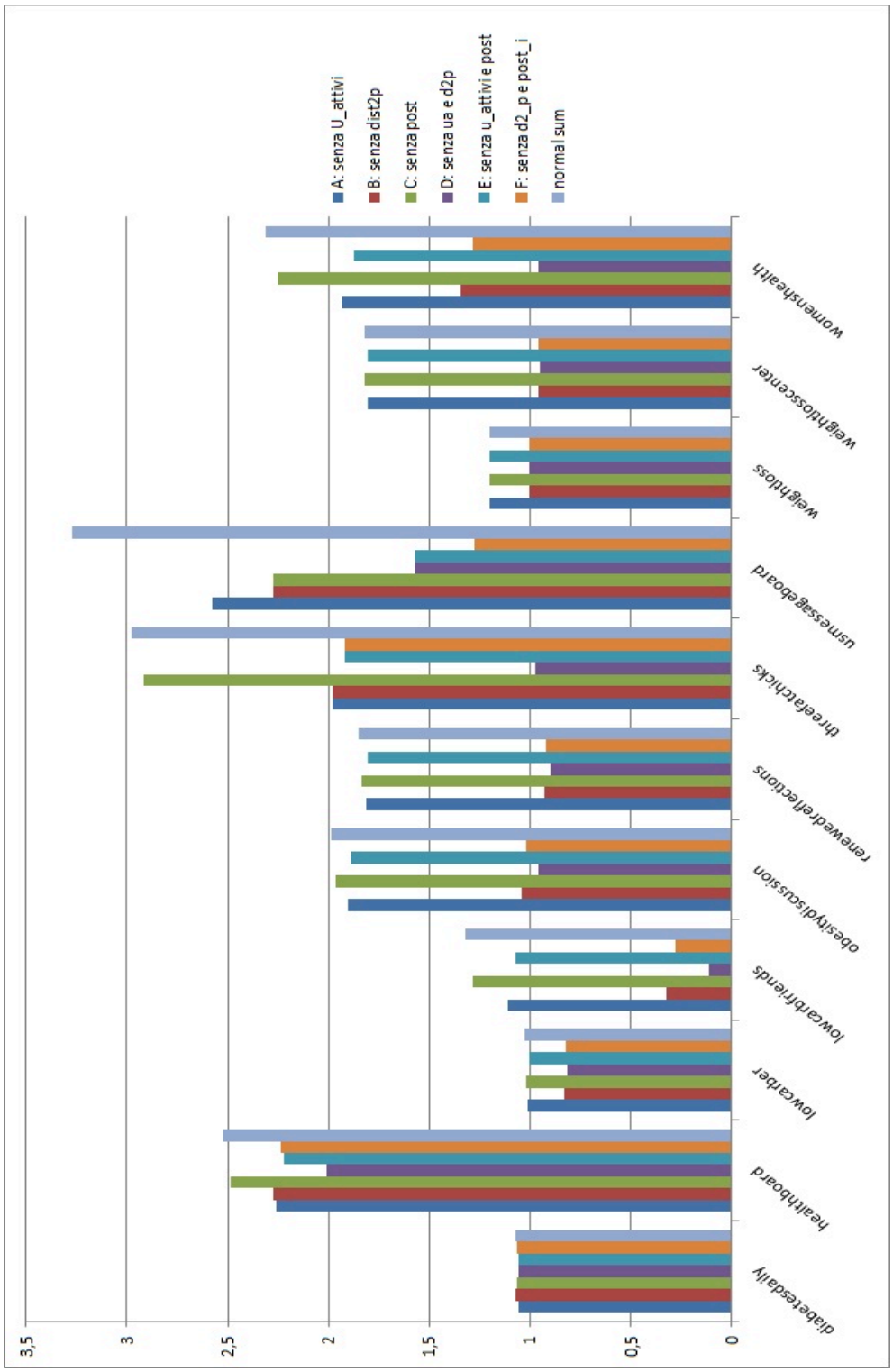

Figura 5.2: Istogramma di Correlazione

Come è stato possibile intuire dalla spiegazione delle metriche, ognuna è stata ottenuta da ogni possibile combinazione con i parametri tra loro dipendenti. Ogni metrica applicata ai forum viene confrontata una ad una con la metrica di riferimento denominata nel grafico "normal sum", che è la stessa descritta nel capitolo 5.1. Nell'istogramma è possibile notare 3 forum in cui le diverse metriche non portano cambiamenti sostanziali, DiabetesDaily, WeightLoss e Low Carber non sembrano rispondere alle diversità del calcolo del punteggio, questa situazione è confermata osservando i valori numerici dei parametri che non rientrano nella metrica. Per DiabetesDaily il contributo degli utenti attivi è di 0,01, la distanza tra 2 post è 0 mentre il contributo dei post totali è 0,002; con queste oscillazioni è impercettibile nel grafico rilevare differenze rispetto la metrica di riferimento. Lo stesso ragionamento si estende a LowCarber e WeightLoss in cui i termini incidono rispettivamente 0,012 - 0,19 - 0,008 per il primo e 0 - 0,19 - 0 per il secondo. Si ricorda che la normalizzazione dei parametri dipende dal campione e il numero di post totali è fortemente influenzato da USMessageboard col numero di 4173860 post in 6 mesi e per riportare la proporzione tra 0 e 1, il termine degli altri forum è poco incisivo ai fini del punteggio.

Le metriche A, B e C sono composte da 4 parametri e rispetto alla metrica *sum* condividono in gran parte la stessa classificazione dei forum. In A i primi tre forum in ordine di importanza risultano senza sorprese USMessageBoard, HealthBoard e ThreeFatChicks, quello che va evidenziato è che senza utenti attivi, si toglie importanza alla presenza degli utenti che scrivono e diventa preponderante il numero relativo agli utenti iscritti, questo si riscontra per HealthBoard che guadagna una posizione, mentre rimane invariato il resto della classifica. Per la metrica B il parametro scartato, la distanza in giorni tra due post, penalizza i forum LowCarbFriends, ThreeFatChicks, USMessageBoard e WomensHealth il cui indice è a punteggio pieno, mentre influiva in maniera negativa per HealthBoard in cui i giorni tra due messaggi è in media di 27 giorni. In questo scenario balza così al primo

posto il forum di HealthBoard seguito da quello di USMessageBoard quasi a pari punti con un punteggio di valore 2,273 contro 2,272 e ThreeFatChicks a 1,97, un altro risultato significativo relativo agli altri forum riguarda DiabetesDaily che passa dalla decima posizione in ordine di importanza al quinto posto.

La metrica C valuta i forum senza il contributo del numero di post, in questo modo l'attività del forum ha come parametri la distanza in giorni dei post e gli utenti attivi. La motivazione che può spingere nell'utilizzo di questa metrica è quello di ritenere la produzione di post inutile ai fini della qualità delle risposte; cioè è possibile ritenere ottimali due forum con egual numero di utenti attivi e tempi di risposta, ma che esauriscono l'argomento con un numero di messaggi diverso. É possibile considerare buoni, in questo modo, i forum in cui gli utenti del thread dopo aver discusso dell'argomento chiudono il thread, ma anche quelli in cui il thread si trasforma in una discussione sociale, senza informazioni mediche utili. Tale metrica penalizza solo il forum USMessageBoard, che si piazza al terzo posto preceduto da HealthBoard e ThreeFatChicks al primo posto, mentre non si rileva alcun cambiamento rilevante tra gli altri forum.

In D la metrica utilizza solo 3 parametri riducendo ulteriormente significato al punteggio, l'unico parametro che indica l'attività del forum è riposta nel numero di post scritti. Chiunque ritenesse superflua l'attesa di oltre 20 giorni di una risposta al proprio messaggio e della possibilità di ricevere la replica da un numero esiguo di utenti può utilizzare questa metrica. Da questo calcolo i tre forum con punteggio più alto risultano: HealthBoard, USMessageBoard e DiabetesDaily, quest'ultimo è sempre stato penalizzato dal valore assunto dagli indici scartati quando gli altri forum godevano di un notevole incremento, occupando il 10° posto con la metrica di riferimento.

La metrica E utilizza la distanza tra due post come indice di attività per

il calcolo del punteggio, da questo scenario HealthBoard riceve il miglior punteggio con 2,21 seguito da ThreeFatChicks e da ObesityDiscussion. L'ordine dei forum inizia a diversificarsi rispetto le altre metriche imputando la causa all'incidenza che hanno i 3 parametri; in particolare il numero di iscritti permette ad HealthBoard di primeggiare, mentre ObesityDiscussion è favorita grazie gli altri 2 parametri. Per i forum di tipo medico, di auto aiuto questo tipo di metrica non è ideale, basti pensare che USMessageBoard risulta al 7 posto tra i forum in esame nonostante abbia una comunità molto attiva, i tempi di risposta sono brevi e produce una quantità di post imparagonabile rispetto a tutti gli altri siti di discussione e meriterebbe un punteggio superiore rispetto ad esempio, DiabetesDaily.

L'ultima metrica F può contare sui contributi degli utenti attivi, gli utenti iscritti e la media dei post per thread. Si presta per classificare i forum in cui non è importante il tempo di attesa di risposta né dal numero di post generati. Un esempio di scenario in cui è valido esemplificare la metrica di riferimento è quello di forum tecnici, in cui sono presenti utenti specializzati e in pochi messaggi è sufficiente soddisfare gli utenti e il tempo di risposta non è importante, inoltre è possibile limitare le differenze dovute al numero di post che vengono scritti da forum di dimensioni diverse. Nella nostra configurazione la metrica premia HealthBoard, ThreeFatChicks e WomensHealth seguito da USMessageBoard.
Ogni metrica trattata può essere utilizzata per altri tipi di forum e a seconda delle necessità dell'utente è possibile modificare i pesi a piacere per limitare il contributo di alcuni termini particolarmente ingombranti. La riduzione dei termini da 5 a 3 ha dato più importanza ai parametri rimasti e all'assegnamento di opportuni pesi, questi stratagemmi non sono esaustivi per definire una metrica univoca per classificare i forum, bisogna sviluppare uno strumento che ne determini la qualità dei contenuti. Un tentativo è stato affrontato nella sezione seguente.

5.4 Risultati della metrica numerica con filtro testuale

In questa sezione si vogliono confrontare i forum sulla base del contenuto dei messaggi scambiati tra gli utenti, per soddisfare questa richiesta è stato ampliato il programma estendendo le funzioni del calcolo della metrica all'estrazione di testo. Si lavora sulla base dei 5 parametri già definiti nel capitolo 4.2, valutando le loro occorrenze sulla base della parola chiave. Per questo test sono state utilizzate alcune parole che riteniamo possano ricorrere frequentemente nei forum in esame: "diet", "diabete" e "cholesterol", in futuro è possibile eseguire il programma "Analisi_Forum" per il ricalcolo delle informazioni utilizzando parole chiave diverse. Si suddividono i risultati e si riportano in tabella: 5.7
Il campione ora è di 10 elementi, per avere i dati di RenewedReflection occorrerà sviluppare il software per l'autenticazione automatica al forum, poichè solo con tali permessi è possibile salvare i post.

			diet			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post [giorni]	media di post per thread	n° di post
1	diabetesdaily	85580	240	10	78	612
2	healthboard	928360	232	5	22	801
3	lowcarber	145474	161	17	9	444
4	lowcarbfriends	66469	3407	155	0	113289
5	obesitydiscussion	1969	281	59	3	10156
7	threefatchicks	155724	4429	28	0	87623
8	usmessageboard	30617	812	221	0	70325
9	weightloss	5789	70	3	149	125
10	weightlosscenter	2400	148	6	200	591
11	womenshealth	21996	302	6	1	1167
			diabete			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post [giorni]	media di post per thread	n° di post
1	diabetesdaily	85580	9	5	3	20
2	healthboard	928360	180	9	21	600
3	lowcarber	145474	102	19	1	313
4	lowcarbfriends	66469	1481	353	0	46963
5	obesitydiscussion	1969	69	91	0	1747
7	threefatchicks	155724	1668	75	0	20225
8	usmessageboard	30617	535	139	0	10777
9	weightloss	5789	15	7	8	15
10	weightlosscenter	2400	5	5	5	5
11	womenshealth	21996	71	13	0	354
			cholesterol			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post [giorni]	media di post per thread	n° di post
1	diabetesdaily	85580	14	5	13	16
2	healthboard	928360	82	6	2	304
3	lowcarber	145474	85	12	0	211
4	lowcarbfriends	66469	1358	406	0	53653
5	obesitydiscussion	1969	68	112	0	1351
7	threefatchicks	155724	1225	84	0	14189
8	usmessageboard	30617	368	256	0	6922
9	weightloss	5789	3	1	144	3
10	weightlosscenter	2400	18	5	7	25
11	womenshealth	21996	29	5	0	57

Tabella 5.7: Dati con applicazione del filtro

Analizzando la tabella si vede il numero degli utenti iscritti ai siti di discussione che è un parametro necessario per il calcolo della metrica e non assume diverso valore se calcolato con il filtro testuale, poiché definisce il bacino d'utenza di un forum, il suo valore equivale all'ultima rilevazione datata 30/10/2012. Diversamente dal numero di iscritti gli altri termini sono

raccolti per parole chiave: “diet”,”diabete”, “cholesterol”.

I punteggi dei forum con i 5 parametri appena descritti sono poco influenzati dal termine del numero di iscritti facendo prevalere il contributo dei termini che si riferiscono all’argomento ricercato. La prima notevole differenza con le informazioni estratte dalla metrica basata sui numeri risiede nel termine degli utenti attivi, in precedenza sono stati considerati attivi solo se la media di scrittura dei messaggi valutata nei 6 mesi fosse superiore di 4 [post/mese], in questo scenario invece sono attivi tutti quelli che hanno scritto almeno un post nello stesso thread dove è stato rilevato un messaggio dove viene citata la parola chiave. Le due interpretazioni permettono di determinare, sotto determinate ipotesi, quali forum producono contenuti grazie agli utenti occasionali e quali per mezzo dei frequentatori abituali. Le tabelle su cui fare il confronto dirette sono 5.1 e per ogni parola chiave da 5.7, cominciando dalle informazioni relative a “diet” si analizza il numero di utenti attivi.

Il forum DiabetesDaily vede coinvolte 240 persone nella produzione di 612 post in cui alcuni di questi è contenuta la parola “diet”, rispetto alla prima metrica c’è un aumento di 30 utenti che possiamo considerare occasionali. Sono occasionali poiché hanno media di scrittura inferiore ai 4 post/mese e per questo motivo non sono stati rilevati nella raccolta dati in precedenza; riguardo la lunghezza media delle discussione possiamo rilevare un buon valore, con 10 post per thread è infatti possibile visualizzare una discussione in una sola pagina, ciò che non è piacevole per questo forum riguarda il tempo medio di attesa della prima risposta a quota 78 giorni. In HealthBoard le dimensioni della comunità favoriscono la produzione di post con 800 messaggi, ma riguardo gli utenti coinvolti nelle discussioni, rispetto ai 3900 che frequentano il forum assiduamente solo 300 autori circa hanno scritto riguardo all’argomento “diet”. Questo risultato può specchiarsi per diverse situazioni, come ipotizzato in precedenza gli utenti si iscrivono per ottenere informazioni nel breve periodo, abbandonando la frequenza del forum una

volta ottenute, riguardo all'attività dei 3900 utenti è possibile che la grandezza del forum prevedendo molte categorie di discussione ha permesso la loro canalizzazione in base ai propri interessi, contando 310 utenti interessati all'argomento "diet". La media di post per thread riferita all'argomento si mantiene bassa, anche questo scenario può essere spiegato dal tipo di utenti che frequenta HealthBoard, favorendo un alto tasso di nascita di nuovi thread sul argomento non ancora ampiamente trattati.

Riguardo i tempi di attesa tra due post, HealthBoard conferma i dati rilevati dalla precedente analisi. Nel sito di LowCarber il termine "diet" coinvolge 161 autori che in media attendono 9 giorni per leggere la prima risposta sull'argomento e dei 444 post scritti nei thread la media di messaggi per discussione è di 17, scoprendo indirettamente che si è richiamato il termine "diet" in 26 thread. LowCarbFriends è certamente un forum molto attivo, dei suoi 65000 utenti iscritti circa 3400 risultano aver dato contributo nella generazione dei post relativi a "diet", la quale risulta essere presente nel 60% dei messaggi raccolti negli ultimi sei mesi. Per il termine in analisi non bisogna attendere nemmeno 24 ore per ottenere una risposta all'apertura del thread, ma per trovare un informazione può essere necessario leggere ben 115 post. ObesityDiscussion raccoglie molti post riferiti alla parola chiave, poteva essere chiaro già leggendo il nome del sito che l'argomento in questione è trattato abbondantemente; in questo sito di discussione, nonostante l'altissimo numero di messaggi vige il motto "pochi ma buoni", infatti degli oltre 10000 messaggi ci sono solo 280 autori diversi che si rivelano essere molto attivi e veloci nel rispondere con un attesa di 3 giorni tra domanda e risposta. L'elevato numero di messaggi si rispecchia anche nella media di post per thread a quota 59 [post/thread], peggiorando in termine di chiarezza e semplicità di raccogliere le informazioni rispetto la media maturata valutando l'intero forum. Un altro forum di ampie dimensioni è ThreeFatChicks, la parola "diet" coinvolge 4400 utenti diversi producendo una quantità di 87000 messaggi; se i dati medi riferiti ai tempi di attesa degli utenti alla prima risposta e alla media

di post per thread sono confermati anche nella realtà, si può affermare la buona strutturazione del sito, che ha convogliato il gran numero di messaggi in thread ben divisi, con 28 post per thread, e tempi di attesa inferiori le 24 ore. USMessageBoard con circa 70000 post e tempi di attesa inferiori le 24 ore non è stata in grado di mantenere semplicità e chiarezza all'interno delle discussioni passando il problema agli 800 utenti che hanno scritto riguardo a "diet", che dovranno leggere in media 221 messaggi per thread. WeightLoss nonostante il nome non raggiunge numeri di partecipazione elevati riguardo al termine, la spiegazione di questa situazione si può dare dalla lettura dei dati, dei 125 post coinvolti è presente una diversità di 70 utenti che devono attendere 5 mesi circa prima di avere una risposta e solo 3 messaggi per thread; tutti i valori indicano che i thread che coinvolgono la parola "diet" non sono adeguatamente trattati e nemmeno la quantità di messaggi prodotti dalla discussione vale l'attesa di una risposta. Lo stesso giudizio si può estendere anche al forum WeightLossCenter con tempi di attesa più elevati, ben 200 giorni, nonostante una frequentazione più alta con 148 utenti e 591 messaggi a trattare l'argomento. WomensHealth riguardo alla parola chiave "diet" può essere paragonato ad HealthBoard, con più di 300 utenti diversi che hanno prodotto oltre 1100 messaggi. Rispetto ad HealthBoard si riscontrano tempi di attesa dall'apertura di un thread di un giorno e la struttura della discussione ha in media 6 post per thread.

Riguardo a "diet", "diabete", "cholesterol" si vuole determinare a parità di dimensione di forum una maggiore predisposizione di uno rispetto ad un altro riguardo ad argomenti diversi. Osserviamo ora i dati relativi al termine di ricerca "diabete" lasciando al termine dei confronti la riflessione circa la determinazione dei migliori forum in base all'argomento.
É incredibile osservare come il forum DiabetesDaily sia povero di contenuti quando si tratta proprio di diabete, solo 20 post sono coinvolti negli ultimi sei mesi prodotti dai 9 utenti diversi che hanno scritto. La scarsità di dati rende poco significativo commentare gli altri parametri del forum anche se ai fini

del punteggio incidono molto positivamente, grazie al tempo di risposta sui 3 giorni e il numero di post per thread di 5. Per HealthBoard si confermano gli stessi ordini di grandezza riscontrati dalle precedenti misurazioni, in cui i tempi medi di attesa di una risposta sono di 21 giorni e la chiarezza delle discussioni è in media di 9 post per thread. Riguardo il coinvolgimento degli utenti riscontriamo un numero di 195 diversi autori su 600 messaggi scritti. LowCarber si avvicina molto ai dati di HealthBoard in quanto a produzione di informazioni in cui compare la parola "diabete", nonostante una disparità notevole nel potenziale dato dagli utenti iscritti. Dei 313 post si rilevano 102 autori diversi, dalla nascita di un thread si attende circa un giorno per ottenere una risposta e tutte le informazioni delle discussioni sono raggiungibili da una lettura media di 19 post. LowCarbFriends dimostra ancora di essere un forum molto attivo anche a proposito dell'argomento "diabete" registrando la più alta produzione di post a quasi 47000 messaggi, scritti da 1481 utenti diversi. Non si può dire nulla sul contenuto dei messaggi, per questo prendiamo il solo significato numerico che assegna l'idea del traffico generato dagli utenti. Da LowCarbFriends gli utenti non attendono nemmeno un giorno per ottenere la prima risposta, ma devono ricercare le informazioni tra 350 post della stessa discussione.
Il gran numero di post può essere giustificato dall'ipotesi che il sito di discussioni ha poche discussioni molto dibattute e qualcuna mediamente sviluppata, infatti navigando all'interno del forum si possono rilevare diverse discussioni con numero di post che supera le migliaia e altre discussioni invece poco sviluppate, a incidere sulla media. ObesityDiscussion con i suoi 69 utenti ha prodotto ben 1747 post riguardo all'argomento diabete e rispetto i dati riferiti al forum in generale i tempi di attesa media rispetto la prima risposta sono inferiori al giorno, mentre è stata persa la chiarezza dovuta alla suddivisione dei post in diverse discussioni, con una media di 91 messaggi per thread. ThreeFatChicks mantiene elevati i numeri relativi sia alla produzione dei post sia alla diversità degli autori che interagiscono confrontato con gli altri forum, ma rispetto al termine "diet" la flessione è notevole, una

riduzione di post e autori di 1/4 rispetto al rilevamento precedente. Si rileva inoltre la stessa prontezza nei tempi di risposta e una presenza di post per thread maggiore, a quota 75 messaggi per discussione. USMessageBoard con una produzione di circa 10000 messaggi rivolti al termine "diabete" ha coinvolto un numero di utenti diversi pari a 535, in grado di rispettare i tempi di risposta entro un giorno, mentre per ricavare le informazioni contenute nelle discussioni gli utenti devono leggere quasi 140 post per thread. WeightLoss e WeighLossCenter possono essere valutati nello stesso modo perché generano un traffico di pochi post, e non sono importanti relativamente al termine di ricerca. Entrambi hanno un numero di post che eguagliano il numero di utenti coinvolti, ma le grandezze in gioco rimangono sulle dieci unità, decretando la quasi totale estraneità all'argomento discusso. WomensHealth, in questo scenario può essere interamente confrontato con i risultati di LowCarber con cui condivide la maggior parte dei risultati, soltanto nel numero di utenti attivi WomensHealth è inferiore con 71 persone rispetto le 103 dell'altro forum. La superiorità evidente di LowCarber nei confronti di WomensHealth riguardo la parola "diabete" è un indizio per la conferma dell'ipotesi che i forum possono essere rivalutati in base ai contenuti e nel caso essere più adatti in alcuni argomenti. Purtroppo non si ha la certezza della qualità dei messaggi, se i contenuti sono relativi effettivamente all'argomento oppure è solo una citazione senza contestualizzazione, si possono solo assegnare i dati relativi al traffico di dati e utenti.

Si descrivono infine i dati raccolti dalla ricerca nei messaggi della parola "cholesterol", cominciando dal sito DiabetesDaily sono stati conteggiati solo 16 post relativi al termine chiave, coinvolgendo nella discussione 14 autori diversi, l'attesa per ottenere la prima risposta è stata misurata a 13 giorni. HealthBoard rispetto i precedenti argomenti non coinvolge un gran numero di utenti, solo 82 autori a cui fanno riferimento i 304 post conteggiati dai thread in cui si è parlato di "cholesterol", tuttavia rispetto ai tempi medi di risposta all'apertura dei thread rilevati in ogni situazione, HealthBoard

ha fatto registrare un miglioramento attestando a 2 giorni l'attesa, svelando una comunità più sensibile all'argomento del colesterolo. LowCarber conta 85 utenti impegnati a portare avanti le discussioni sul colesterolo e disponibili a rispondere immediatamente all'apertura di un nuovo thread. Nei sei mesi di osservazioni i thread si articolano con una media di 12 post, per un totale di 211 messaggi. LowCarbFriends è senza dubbio il sito di discussioni che ha generato il più ampio traffico relativo al "cholesterol", con un attività dei suoi utenti registrata a 1358 persone coinvolte e ben 53653 post correlati al termine. Il valore elevato dei post è da imputarsi al metodo di calcolo del parametro legato alla struttura del forum, possiamo osservare direttamente visitando il sito la presenza di tanti forum in cui le discussioni sono formate da centinaia di post e la presenza di un occorrenza del termine "cholesterol" condiziona il conteggio dell'intero numero di messaggi del thread.

Come già espresso in precedenza il metodo che conteggia il numero di post non è in grado di discriminare i post il cui significato è legato direttamente a "cholesterol" oppure è un messaggio riferito ad altri post, l'unico vincolo è la presenza del post all'interno dello stesso thread in cui compare effettivamente un occorrenza del termine ricercato. Si può definire tale parametro come un limite superiore del numero di post, in grado di conteggiare nella migliore delle ipotesi il numero massimo di messaggi che si riferiscono ad una parola chiave, d'altra parte definiamo limite inferiore il numero di post definito dal conteggio dei soli post che contengono esattamente il termine di ricerca, senza contare le relative risposte ottenute nello stesso thread. LowCarbFriends conta 406 post in media per ogni thread che contiene il temine "cholesterol" e il numero elevato di utenti attivi permette risposte immediate per ogni nuova discussione. Il filtraggio con la nuova parola chiave applicato ad ObesityDiscussion riporta risultati simili al filtraggio con la parola "diabete", 68 utenti molto attivi che han prodotto 1351 messaggi, suddivisi in discussioni con in media 112 post e il tempo medio di risposta dal primo messaggio inferiori a 24 ore. Un altro forum che ha ottenuto buoni risul-

tati confermando l'importanza che gli viene assegnata dalla prima metrica è ThreeFatChicks, con 1225 utenti attivi; i thread coinvolti dall'argomento contano un totale di 14189 messaggi con una suddivisione media di 84 post per thread e distanza tra i primi due post in linea con i tempi definiti dal forum senza parola filtro. USMessageBoard ha già dimostrato di non avere una suddivisione ottimale dei post ma nemmeno identificando i soli thread riferiti alla parola "cholesterol" riesce a convogliare i messaggi in più discussioni, con una media 256 messaggi per thread, questo parametro può essere però considerato positivamente solo se indica l'approfondimento degli argomenti, ma senza uno strumento che ne determina il contenuto non è possibile confermare l'ipotesi. In USMessageBoard il numero di utenti coinvolti è 368 mentre i post relativi alla parola chiave sono 6922. WeighLoss negli ultimi sei mesi ha registrato l'apertura di 3 thread e 3 utenti coinvolti, l'attesa di una risposta è durata circa 140 giorni. WeightLossCenter non è da ritenersi un buon forum relativo alla parola "cholesterol", nonostante l'attesa di 5 giorni per ottenere una risposta, nel periodo di osservazione sono stati selezionati solo 25 post da 18 utenti diversi. Infine WomensHealth nonostante un buon punteggio riguardo la metrica basata sui numeri non sviluppa molti contenuti riguardo all'argomento del colesterolo, con un coinvolgimento di 29 utenti risultando inferiore ad altri forum che LowCarber.

A questa analisi bisogna aggiungere il confronto diretto tra forum in base alle parole chiave e determinare una classificazione orientata ai contenuti, per farlo si sfrutta di nuovo la normalizzazione dei dati riportata dalla tabella 5.8.

			diet			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post	media di post per thread	n° di post
1	diabetesdaily	0,090254547	0,038999771	0,967889908	0,61	0,004303489
2	healthboard	1	0,037164487	0,990825688	0,89	0,005973631
3	lowcarber	0,154907593	0,020876348	0,935779817	0,955	0,002818918
4	lowcarbfriends	0,069625029	0,765542556	0,302752294	1	1
5	obesitydiscussion	0	0,048405598	0,743119266	0,985	0,088641264
7	threefatchicks	0,165972036	1	0,885321101	1	0,773196423
8	usmessageboard	0,030924307	0,170222528	0	1	0,620338624
9	weightloss	0,004123529	0	1	0,255	0
10	weightlosscenter	0,000465246	0,017894012	0,986238532	0	0,004117917
11	womenshealth	0,021618302	0,053223216	0,986238532	0,995	0,009207875
			diabete			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post	media di post per thread	n° di post
1	diabetesdaily	0,090254547	0,002405292	1	0,857142857	0,000319434
2	healthboard	1	0,105231509	0,988505747	0	0,012670897
3	lowcarber	0,154907593	0,058328322	0,959770115	0,952380952	0,006559053
4	lowcarbfriends	0,069625029	0,887552616	0	1	1
5	obesitydiscussion	0	0,038484666	0,752873563	1	0,03709698
7	threefatchicks	0,165972036	1	0,798850575	1	0,430597555
8	usmessageboard	0,030924307	0,318701143	0,614942529	1	0,229396482
9	weightloss	0,004123529	0,006013229	0,994252874	0,619047619	0,000212956
10	weightlosscenter	0,000465246	0	1	0,761904762	0
11	womenshealth	0,021618302	0,039687312	0,977011494	1	0,007432173
			cholesterol			
	nome_forum	n° di iscritti	n° utenti attivi	Distanza tra 2 post	media di post per thread	n° di post
1	diabetesdaily	0,090254547	0,008118081	0,990123457	0,909722222	0,000242311
2	healthboard	1	0,058302583	0,987654321	0,986111111	0,005610438
3	lowcarber	0,154907593	0,060516605	0,972839506	1	0,00387698
4	lowcarbfriends	0,069625029	1	0	1	1
5	obesitydiscussion	0	0,04797048	0,725925926	1	0,025125815
7	threefatchicks	0,165972036	0,901845018	0,795061728	1	0,264417521
8	usmessageboard	0,030924307	0,269372694	0,37037037	1	0,128965517
9	weightloss	0,004123529	0	1	0	0
10	weightlosscenter	0,000465246	0,011070111	0,990123457	0,951388889	0,000410065
11	womenshealth	0,021618302	0,019188192	0,990123457	1	0,001006524

Tabella 5.8: Normalizzazione dei dati con applicazione del filtro

Successivamente si applica la formula della metrica utilizzando i parametri estratti utilizzando le parole chiave prescelte, da cui otteniamo i punteggi di ogni forum 5.9:

metriche sui post con text	diet	diabete	cholesterol	normal
diabetesdaily	1,711447715	1,95012213	1,998460618	1,069932208
healthboard	2,923963807	2,106408154	3,037678453	2,523657532
lowcarber	2,069382675	2,131946035	2,192140685	1,024650078
lowcarbfriends	3,137919878	2,957177644	3,069625029	1,324630134
obesitydiscussion	1,865166128	1,82845521	1,799022221	1,983422162
renewedreflections	0	0	0	1,84706112
threefatchicks	3,824489559	3,395420166	3,127296303	2,975625475
usmessageboard	1,821485459	2,193964461	1,799632889	3,272936353
weightloss	1,259123529	1,623650207	1,004123529	1,198567973
weightlosscenter	1,008715708	1,762370008	1,953457768	1,819667225
womenshealth	2,065287925	2,045749281	2,031936474	2,315619012

Tabella 5.9: Risultati metrica basata sui contenuti

Il confronto dei forum avviene tramite il grafico ad istogramma per determinare l'ordine di importanza in maniera diretta in figura 5.3 .

I risultati della metrica basata sui numeri aveva riportato una classifica dei forum ponendo tra i primi posti USMessageBoard, ThreeFatChicks ed HealthBoard per il loro alto contenuto di utenti e produzione di post, la situazione è cambiata con l'applicazione del filtro con le parole chiave, si può da subito assegnare il primato al forum ThreeFatChicks riguardo a tutte e tre le parole chiave utilizzate. Per quanto riguarda "diet" si rileva un notevole passo in avanti per il forum LowCarbFriends che diventa il secondo forum in ordine di importanza con un punteggio di metrica superiore a 3. HealthBoard mantiene un punteggio elevato garantendosi il terzo posto tra i forum in esame, seguito da LowCarber e WomensHealth. LowCarber, ricordiamo ha sempre occupato l'ultimo posto con la metrica basata sui numeri mentre filtrando i messaggi con l'argomento "diet" è risultato superiore alla metà dei forum in esame, confermando l'ipotesi che valutando i contenuti è possibile riclassificare l'ordine di importanza dei forum fino a preferire un forum meno popolato.

Applicando la metrica con la parola filtro "diabete", il secondo forum in

ordine di importanza rimane LowCarbFriends, seguito da USMessageBoard e LowCarber. Ciò che risalta maggiormente da questa analisi è la riduzione delle differenze tra forum di diverse dimensioni, permettendo di far prevalere i siti con maggior contenuto specifico. Tra i forum con punteggio minore WeightLoss, ObesityDiscussion e WeightLossCenter si può sottolineare questa caratteristica, con la metrica basata sui numeri infatti la definizione del miglior forum sarebbe spettata a ObesityDiscussion, mentre applicando la metrica con filtro "diabete" la differenza di punteggio si assottiglia rendendo equivalente l'importanza rivestita.

Per quanto riguarda la parola "cholesterol" possiamo mettere sullo stesso piano i risultati dei punteggi di ThreeFatChicks, HealthBoard e LowCarbFriends che riescono a raggiungere un valore simile usufruendo un contributo diverso da parte dei vari parametri. La comunità di LowCarber è da preferirsi rispetto al resto dei forum, comprovando la capacità a generare flussi di informazioni orientati ai contenuti, nonostante numericamente non è stata in grado di confrontarsi con forum più ampi, come ObesityDiscussion o USMessageBoard.

Dall'esperienza descritta possiamo affermare che un forum può ottenere diversi risultati se si applica la metrica basata sui numeri oppure quella basata sul contenuto dei messaggi. Le prove sono evidenti nel campione analizzato, in cui il forum di LowCarber dal punto di vista numerico non ottiene un punteggio rilevante mentre applicando la metrica sui contenuti "diet", "diabete" e "cholesterol" si distingue tra i più importanti; la controprova è nel forum di USMessageBoard, molto attivo a livello di post e utenti, ma adottando la metrica sul contenuto non ha grande interesse nei termini ricercati. L'utilizzo di entrambe le metriche fornisce gli strumenti necessari per determinare l'importanza del forum non solo a livello di produzione di informazioni, ma anche rispetto a determinati argomenti rendendo fondamentale l'analisi dei contenuti.

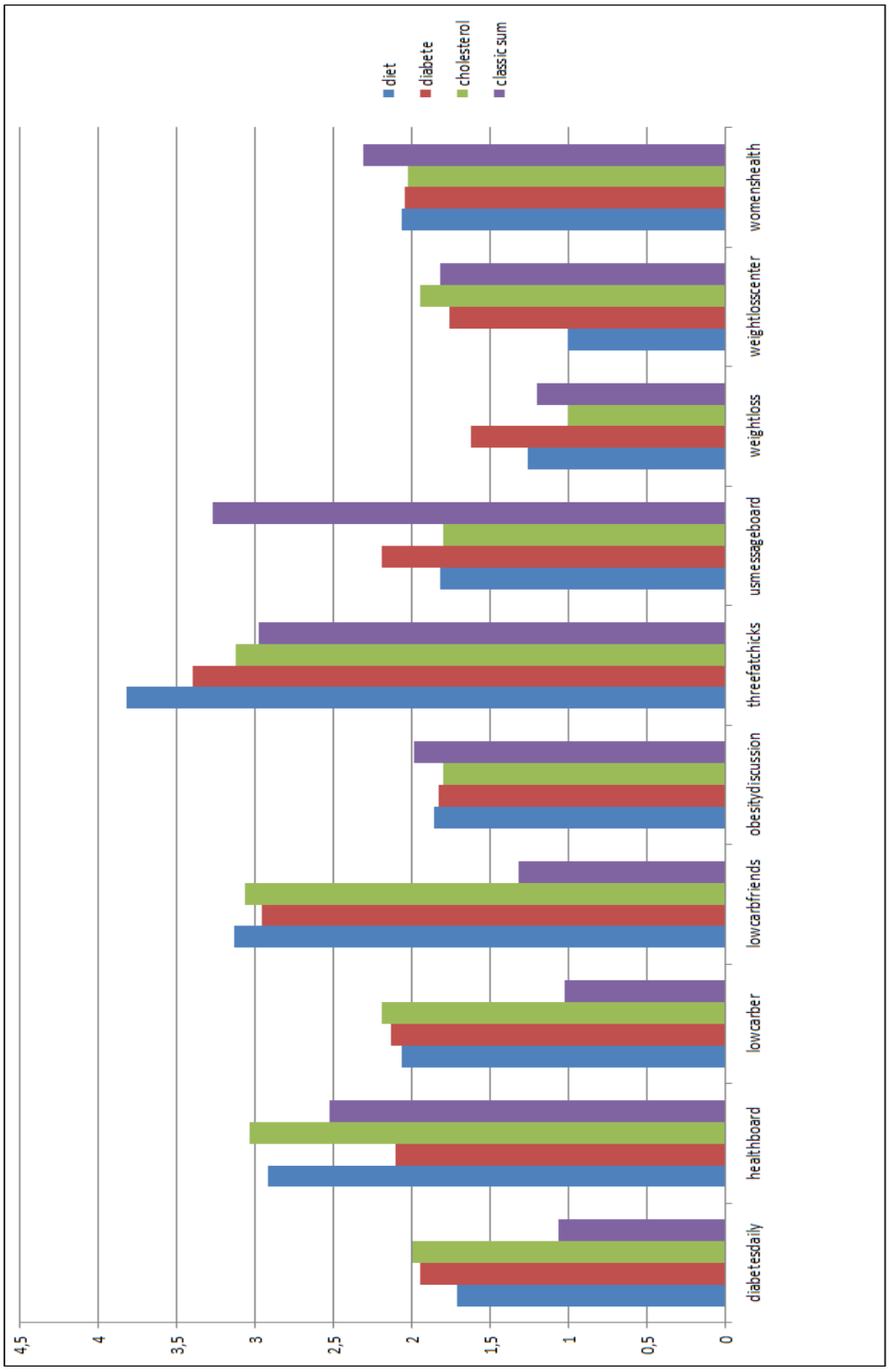

Figura 5.3: Grafico relativo alla metrica dei contenuti

5.5 Risultato della metrica semantica

Il DataBase che contiene le informazioni dei Forum analizzati dallo strumento SentiEngine è stato interrogato per ottenere un elenco degli argomenti su cui orientare la metrica. L'estrazione è stata eseguita per ottenere un confronto tra i forum sulla base degli stessi soggetti di ricerca, per determinare quello con maggior influenza nell'argomento. Il risultato di questa azione è stato quello di ottenere una quantità estremamente elevata di voci divise in "category", "brand", "subject" su cui poter effettuare il confronto dei siti e selezionare quelle più significative per proseguire nell'individuazione di una metrica adatta. La metrica non può essere determinata in base alla scelta delle voci estratte poiché riguardano molto specificatamente prodotti farmaceutici e non si otterrebbe una misurazione generale dei forum, si aggiunge anche il fattore limitante del numero del campione che rende ogni osservazione priva di significato statistico. L'idea che è stata alla base della ricerca è quella di assegnare il punteggio di influenza semantico al forum in base al numero di occorrenze che si registrano per le parole relative alla category, al brand e al subject e di comporre la metrica pesando il risultato dei tre parametri. La grande specializzazione dello strumento SentEngine ha reso difficoltosa l'operazione di confronto tra forum poiché l'osservazione di un prodotto in particolare riferita a una categoria di disturbi e argomenti sanitari non è abbastanza esauriente ed esaustiva riguardo all'argomento stesso, perché non si tiene conto di altri prodotti che concorrono nella trattazione del disturbo. A causa di questi problemi non è stata prodotta alcuna metrica riferita ai contenuti semantici, lasciando però lo spunto per chi volesse proseguire nella sua determinazione suggerendo un ampliamento del campione e lo sviluppo di un modello di dominio farmaceutico orientato all'ambito customer.

Capitolo 6

Conclusioni

A partire dalle ricerche già effettuate sulla ricerca delle fonti, la loro importanza e conseguente evoluzione tramite internet hanno portato allo studio mirato delle comunità virtuali, i forum, concentrando le osservazioni attorno ad un ampio argomento che è la farmaceutica e la cura del paziente. Questa specificazione ha permesso di raggiungere l'obiettivo prefissato e cioè determinare un metodo per classificare i forum in ordine di influenza e portata visiva da parte degli utenti online, creando alle aziende interessate delle opportunità di tipo pubblicitario e divulgativo.
Una prima fase di analisi qualitativa delle fonti ha permesso di determinare 5 parametri fondamentali in grado di descrivere l'attività degli utenti e il traffico di informazioni di ogni forum solo basandosi sui contenuti numerici ricavabili dal sito, tra cui il numero degli utenti iscritti, il numero di utenti attivi, il numero di post totali, la distanza in giorni tra l'apertura di una discussione e la prima risposta, la media di post per ogni discussione. Per ritenere il punteggio "attuale", la scansione dei forum ha preso in considerazione gli ultimi sei mesi, approfittando della caratteristica dei forum che mantengono in memoria le discussioni e disponibili per la consultazione in ogni momento.
Qualora non sia possibile determinare alcuni dei 5 termini sono state proposte varianti della metrica principale definendo superflui alcuni termini,

poiché considerati ridondanti nello studio di correlazione statistica, tra di essi: il numero di utenti attivi, il numero di post totali e la distanza tra apertura thread e prima risposta.

Il punteggio dei forum può assumere diverso valore informativo se vengono modificati i pesi della metrica, inducendo a privilegiare il contributo di determinati parametri. Nel caso di pesi equivalenti e metrica definita solo numericamente sono stati classificati molto influenti i seguenti forum: U.S. MessageBoard, ThreeFatChicks, HealthBoard e WomensHealth. Utilizzando le metriche dopo lo studio statistico non ci sono state riclassificazioni eccessive, mantenendo l'ordine di importanza con poche sorprese. Oltre l'analisi numerica dei forum è stata determinata la metrica che rileva il traffico numerico in base a parole chiave, cercando di avvicinarsi ad una classificazione orientata al significato. Come ipotizzato i dati ricavati tramite parole filtro hanno permesso di rivalutare alcuni forum, assegnando loro un punteggio superiore rispetto a quello ottenuto in precedenza. In questo scenario i forum più influenti relativi alla parola chiave "diet" sono stati: ThreeFatChicks, LowCarbFriends, HealthBoard; per la parola chiave "diabete": ThreeFatChicks, LowCarbFriends, U.S. MessageBoard; mentre per la parola filtro "cholesterol": ThreeFatChicks, LowCarbFriends ed HealthBoard exequo, LowCarber. Un ulteriore passo è stato quello di determinare precisamente il traffico relativo ad un argomento sfruttando uno strumento in grado di determinare semanticamente ogni messaggio contenuto nei forum, ma a causa della mancanza di materiale sufficiente dovuta all'analisi dello strumento di soli 2 forum non è stato possibile ipotizzare un metodo generale per poi accertare la sua verità.

6.1 Problemi e soluzioni

Non tutti i forum condividono al pubblico le stesse informazioni, a volte è sufficiente iscriversi al sito per ottenerle in maniera diretta. Nel caso in esa-

me solo i forum di Diabetesdaily e Womenshealth non hanno reso pubblico il numero degli utenti iscritti, l'informazione necessaria è stata approssimata conteggiando tutti gli utenti intervenuti nell'ultimo anno.

6.2 Sviluppi Futuri

In futuro si può pensare all'implementazione all'interno del programma dell'autenticazione automatica ai forum per ricavare, quando possibile, i dati aggiuntivi degli utenti iscritti come:media messaggi, data di iscrizione, ecc. informazioni disponibili solo dalla comunità.
Della ricerca fin qui rappresentata manca la metrica semantica, è possibile continuare a partire dalle considerazioni fatte nel capitolo 5.5 e avvalersi di ulteriori campioni per giustificare i risultati conseguiti.
Si può sfruttare questa ricerca per estendere lo studio della metrica anche ad altri tipi di forum (viaggi, telefonia, automobili, ecc.), ripercorrendo i passi descritti, dall'individuazione dei parametri che identificano l'influenza del forum, lo studio statistico per semplificare la metrica e infine una metrica semantica.

Bibliografia

[1] Julie B. Morrison, Peter Pirolli, and Stuart K. Card. A taxonomic analysis of what world wide web activities significantly impact people's decision and actions. *Interactive Posters*, 2001.

[2] Giuseppina Lombardo, Barbara Caci, and Maurizio Cardaci. Dalla credibilità offline alla web-credibility: dimensioni psicologiche del costrutto. *Psychofenia*, 10(16), 2007.

[3] Università commerciale Luigi Bocconi. L'affidabilità delle fonti. *Laboratorio di economia e gestione delle istituzioni e delle iniziative artistiche e culturali.*

[4] Shaohan Cai and Minjoon Jun. Internet users' perceptions of online service quality: a comparison of online buyers and information searchers. *Managing Service Quality*, 13(6), 2003.

[5] Fiona Fui-Hoon Nah. A study on tolerable waiting time: how long are web users willing to wait?

[6] Soo Young Rieh. Judgment of information quality and cognitive authority in the web. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 2002.

[7] Janet E. Alexander and Marsha Ann Tate. Web wisdom: How to evaluate and crate information quality on the web. 1999.

[8] Naomi Mandel and Eric J. Johson. When web pages influence choice: Effects of visual primes on experts and novices. *JOURNAL OF CONSUMER RESEARCH,*, 29, 2002.

[9] Stuart J. Barnes and Richard Vidgen. Measuring web site quality improvements: a case study of the forum on strategic management knowledge exchange. *Measuring Web site quality improvements: a case study of the forum on strategic management knowledge exchange*, 2003.

[10] Raymond Morin. 9 indicators of a social web influencer. *http://www.intelegia.com/en/2012/03/09/9-indicators-of-a-social-web-influencer/.*

[11] T. Balasubramanian and R.Umarani. Clustering as a data mining technique in health hazards of high levels of fluoride in potable water. *(IJACSA) International Journal of Advanced Computer Science and Applications,*, 3(2), 2012.

[12] Markus Weimer, Iryna Gurevych, and Max Muhlhauser. Automatically assessing the post quality in online discussions on software. *Proceedings of the ACL 2007 Demo and Poster Sessions,*, pages 125–128, June 2007.

[13] Rasmus Sundberg, Anders Eriksson, Johan Bini, and Pierre Nugues. Visualizing sentiment analysis on a user forum.

[14] Michael Trusov, Anand V. Bodapati, and Randolph E. Bucklin. Determining influential users in internet social networks. April 20, 2009.

[15] Barbara Bickart and Robert M. Schindler. ï¿Œinternet forums as influential ï¿Œinternet forums as influenti sources of consumer information. *Journal of Interactive Marketing*, 15(3), 2001.

[16] Jonh Powell and Aileen Clarke. The www of the world wide web: Who, what, and why? *Journal of Medical Internet Research*, February 2002.

[17] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, 328, 2004.

[18] Gunther Eysenbach, Jonh Powell, Olivier Kuss, and Eun-Ryoung Sa. Empirical studies assessing the quality of health information for consumers on the world wide web. *Jama*, 287(20), May 2002.

[19] Alejandro R. Jadad and Anna Gagliardi. Rating health information on the internet: Navigation to knowledge or to babel? *Jama*, 279(8), February 1998.

[20] Gretchen K. Berland, Marc N. Elliot, Leo S. Morales, Jeffrey I. Algazy, Richard L. Kravitz, Michael S. Broder, David E. Kanouse, Jorge A. Munoz, Juan-Antonio Puyol, Marielena Lara, Katherine E. Watkins, Hannah Yang, and Elizabeth A. McGlynn. Health information on the internet: Accessibility, quality, and readability in english and spanish. *Jama*, 285(20), May 2001.

[21] Censis. Quale futuro per il rapporto medico paziente nella nuova sanità? *Forum per la ricerca Biomedica*, 2012.

[22] http://www.healthboards.com.

[23] http://forum.lowcarber.org.

[24] http://www.renewedreflections.com/forums/.

[25] http://www.diabetesdaily.com.

[26] http://www.lowcarbfriends.com.

[27] http://www.obesitydiscussion.com/forums.

[28] http://www.3fatchicks.com/forum/.

[29] http://www.usmessageboard.com.

[30] http://www.weightlossforums.org.

[31] http://www.weight-loss-center.net.

[32] http://www.womens-health.com/boards/forum.php.

[33] Written and Frank. *Data Mining*. 2005.

[34] Sidney Siegel. Nonparametric statistics. *The Pennsylvania State University*.

[35] Università della Bocconi di Milano. Richiami metodologici. 2012.