POLITECNICO DI MILANO
Facoltà di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica

Analisi della Costruzione Partecipativa di un Wiki con un'Applicazione a Wikipedia

Relatore: Prof. Marco Colombetti
Correlatore: Ing. David Laniado

Tesi di Laurea di:
Riccardo Tasso, matricola 708301

Anno Accademico 2007-2008

A mio zio
Giuseppe

Sommario

La tecnologia dei wiki e il suo più grande esempio, Wikipedia, hanno causato un grandissimo cambiamento nel modo di utilizzare lo strumento del World Wide Web da parte di tutti i suoi navigatori.

Dalla possibilità di inserire i contenuti solo per poche persone si è passati a un ambiente dove chiunque può condividere la propria conoscenza senza limiti se non quelli dettati dall'accordo tra i partecipanti.

Tuttavia all'aumentare dell'importanza di Wikipedia come strumento di diffusione dell'informazione attraverso Internet sono iniziati a sorgere i primi dubbi sulla sua affidabilità come fonte e soprattutto su chi effettivamente ne scriva e ne gestisca i contenuti.

Questo lavoro di tesi vuole proporre una metodologia innovativa di analisi automatizzata della comunità di un generico wiki basata su quattro pilastri fondamentali.

La raccolta di dati dagli enormi archivi del wiki.

L'utilizzo di questi dati per calcolare il contributo di ciascun utente alla costruzione di ogni singola pagina. Per questa fase vengono proposte e analizzate tre differenti metriche.

La selezione degli utenti che maggiormente hanno contribuito al suo sviluppo e hanno quindi potuto influenzarne i contenuti.

La creazione di un modello di Rete Sociale in grado di rappresentare le relazioni tra gli utenti del wiki.

Questa metodologia è stata studiata dal punto di vista teorico, alla luce dell'esperienza maturata dagli altri studi sull'argomento.

Quindi è stata implementata attraverso opportuni strumenti software progettati per essere riutilizzati per un qualsiasi wiki.

Infine è stata applicata a quattro versioni di Wikipedia: quella in italiano del 2005, del 2007 e del 2008; quella in inglese del 2007. Questo ha permesso innanzitutto di confrontare le differenze fra le tre metriche proposte. Quindi di confrontare tra di loro le diverse versioni di Wikipedia. Infine di proporre

un paragone tra le comunità di Wikipedia con alcune di coautori di articoli scientifici e di sviluppatori di comunità open source.

I risultati ottenuti permettono di trarre interessanti considerazioni principalmente su Wikipedia, ma anche sulla tecnologia dei wiki e offrono molteplici spunti per gli sviluppi futuri.

Ringraziamenti

Le prime persone che sento il bisogno di ringraziare sono quelle che mi hanno permesso di giungere a questo importante traguardo.

I miei genitori, Roberto e Marilena, che mi hanno sempre dato tutto ciò che ha contribuito a formarmi fisicamente e spiritualmente.

La nonna Renza, che mi ha tenuto per mano nei primi passi della mia vita, e gli altri nonni che non ho avuto la fortuna di conoscere di persona ma che vivono ancora nei ricordi dei miei genitori.

Mio fratello Federico, che mi completa e mi insegna ogni giorno che le differenze tra di noi non possono che rendere più bello il nostro rapporto.

I miei cugini, Giorgio, Clara, Guglielmo e Giulio, che mi hanno concesso il privilegio di essere un punto di riferimento per loro, così come lo sono stati per me Laura e Massimiliano.

I miei zii, Benito e Gina, Filippo ed Emilia, Giuseppe e Antonella, che mi hanno voluto bene come un figlio e mi hanno insegnato la vita attraverso il loro esempio.

I miei amici, che hanno condiviso con me i momenti più belli della mia vita come quelli di maggiore difficoltà. Nominarli tutti sarebbe impossibile, ma non c'è problema perché nel mio cuore ci state tutti senza riserve.

Chiara e Cristina, che a partire da quest'estate mi hanno dato tutto il loro affetto e hanno illuminato la mia strada coi loro sorrisi raggianti.

Costanza, che ha avuto il coraggio di accettarmi sempre per quello che sono.

Oltre a loro vorrei ringraziare tutte le persone speciali che ho avuto modo di conoscere in questi anni di studio e in particolare in questo momento della mia vita che non scorderò facilmente.

Il *team U.D.S.*, e in particolare Simone, con il quale ho passato i più bei momenti di studio universitario e nonostante ciò mi hanno insegnato che oltre all'università esistono anche i videogiochi, la musica, il *Pyhton* e molte altre cose!

I tesisti dell'*AirLab* che hanno condiviso con me questi ultimi giorni di

preoccupazioni sui dettagli più insignificanti della scrittura di una tesi, come il numero di pagine, la copertina e l'indice delle tabelle.

I dottorandi del *D.E.I.* che mi hanno regalato l'esperienza di chi era già passato per questa impervia strada.

I docenti che ho incontrato nel mio percorso, e in particolare quelli che hanno risposto alle mie domande con entusiasmo e mostrando passione per il loro ruolo.

Il professor de Alfaro, Ian Pye che mi hanno incoraggiato in questo lavoro e i collaboratori del canale *IRC #it-wikipedia*, che mi hanno aiutato tantissimo a comprendere come studiare Wikipedia nel modo corretto.

Davide, che ha sempre trovato il tempo per ascoltarmi e per trasmettermi le sue grandi conoscenze al pari di un fratello maggiore.

David, che ha dato tutto se stesso per permettermi di realizzare la tesi che desideravo e per questo non finirò mai di ringraziarlo.

Vorrei ringraziare ancora una volta Davide e David per l'amicizia che mi hanno donato.

Il professor Colombetti che ben più di una volta mi ha aperto la mente verso nuovi orizzonti e mi ha dato enorme fiducia nelle mie capacità.

Indice

Capitolo 1

Introduzione

Negli ultimi anni una delle tecnologie che ha assunto maggiore importanza nell'ambito del World Wide Web è sicuramente quella dei wiki. La sua peculiarità consiste nell'abbassamento delle barriere d'ingresso alla pubblicazione di contenuti sul Web favorendo la collaborazione di individui accomunati dal medesimo obiettivo o interesse.

Al contrario di un tradizionale sito Web infatti, un wiki assume che il suo visitatore non sia un semplice lettore ma anche un portatore di nuova conoscenza. Un qualunque utente può infatti modificare o inserire informazioni istantaneamente, senza alcuna limitazione e senza avere bisogno di alcuna competenza tecnica avanzata.

Tutto ciò può sembrare di poco conto dal punto di vista tecnico, tuttavia si è visto come attraverso un wiki sia estremamente efficace catturare, organizzare e mantenere aggiornata la conoscenza di gruppi di persone all'aumentare della loro dimensione.

La dimostrazione empirica di questa considerazione è quella data da Wikipedia, una famiglia di progetti finalizzati alla costruzione partecipativa di un'enciclopedia online multilingue. Nel giro di pochi anni infatti è riuscita a raggiungere dimensioni inimmaginabili grazie alle migliaia di collaboratori che volontariamente offrono il loro contributo e a diventare uno dei punti di riferimento del Web.

All'aumentare dell'importanza di Wikipedia nel reperire informazioni per chiunque disponga di un accesso a Internet stanno però emergendo i primi dubbi sulla sua attendibilità. Innanzitutto perché il controllo della validità delle informazioni al suo interno è svolto dagli stessi utenti che ne usufruiscono e non offre garanzie. In secondo luogo perché si teme che gruppi di persone possano voler sfruttare la grande visibilità di Wikipedia per plasmare l'opinione pubblica su determinate tematiche controllandola

segretamente.

Questo è possibile poiché, per quanto possa essere democratica Wikipedia, ad avere l'ultima parola sui suoi contenuti sono sempre quegli individui che possono investire più tempo nel progetto.

Allo stesso modo in un generico wiki non è sempre facile individuare il grado di partecipazione degli utenti e il loro ruolo all'interno della comunità formatasi attorno al progetto.

1.1 Obiettivi e motivazioni

Lo scopo di questo lavoro è quello di mettere a punto delle tecniche per analizzare in modo automatico la comunità che si forma attorno a un wiki.

Uno studio di questo tipo non può essere svolto manualmente perché tipicamente i wiki sono utilizzati in progetti di grandi dimensioni e nei quali non sono definiti a priori una comunità o i ruoli degli individui all'interno di essa. Inoltre le varie sezioni di un wiki vengono portate avanti in parallelo da gruppi di persone con risorse e conoscenze differenti. Di conseguenza un wiki può svilupparsi in modo decisamente poco omogeneo. Ciò rende difficile a un osservatore umano privo di strumenti adeguati per l'analisi la navigazione all'interno di tutte le sue pagine alla ricerca degli interventi degli utenti.

Sarà innanzitutto importante capire quali dati mantenuti dal software possono essere i più interessanti per cogliere le proprietà dei singoli interventi.

Successivamente bisognerà trovare dei modelli adatti a catturare il modo in cui contribuisce un singolo utente all'interno di un wiki o di una sua area.

Infine si renderà necessario estendere questi modelli per catturare le relazioni di collaborazione tra i differenti partecipanti in modo da poter cogliere il loro ruolo all'interno della comunità.

Questi passi dovranno essere implementabili e applicabili allo studio di wiki generici e in particolare a Wikipedia.

In questo modo i risultati ottenuti potranno essere considerati utili a due differenti livelli. Il primo sarà quello di capire come è strutturata la comunità di Wikipedia e come interagiscono tra di loro i suoi autori. Ad un livello più alto invece si potranno trarre conclusioni più generali sui wiki e sul loro funzionamento. Si ritiene che quest'ultimo obiettivo sia di fondamentale importanza per valutare quanto il caso del successo di Wikipedia sia effettivamente dovuto alla tecnologia wiki.

1.2 Contributi originali

Essendo quella dei wiki una tecnologia molto recente, gli studi di tipo quantitativo su di essa sono da considerarsi ancora agli inizi. Restringendo il campo a quelli che hanno come oggetto la comunità e le relazioni che si instaurano tra i suoi membri ci si accorge come siano stati svolti pochissimi lavori in proposito. Ecco perché il primo aspetto interessante di questa tesi è proprio l'oggetto dello studio. Anche il metodo di analisi presenta dei contributi innovativi.

Per la prima volta si è deciso di valutare in modo preciso il contributo di un utente all'interno di ciascuna pagina di un wiki. Sebbene già esistessero degli approcci di questo tipo, essi erano stati applicati solo a livello globale dell'intero wiki. Oltre all'adattamento di uno di questi approcci globali alla singola pagina, viene proposta in questo lavoro una nuova metrica di calcolo del contributo che partendo dalla precedente ne corregge alcuni aspetti causa di imprecisioni.

Viene quindi proposto un algoritmo per determinare, all'interno di ciascuna pagina, l'insieme dei suoi autori più influenti, i quali cioè hanno contribuito in maniera molto più importante degli altri collaboratori e hanno quindi potuto maggiormente guidare la sua evoluzione sia per quanto riguarda la forma che i contenuti.

Successivamente viene proposta una nuova tecnica per modellare le relazioni di questi contributori più influenti in una Rete Sociale. Ricondursi a questo tipo di modello offre il vantaggio di potersi ricondurre a casi notevoli di studio di comunità e ad algoritmi di analisi molto potenti. Essi saranno utilizzati sia per studiare le caratteristiche della comunità nella sua globalità che per trovare i suoi membri più influenti e i loro comportamenti all'interno di essa.

I risultati di questo processo di analisi sono stati quindi applicati a differenti versioni di Wikipedia: tre di esse sono quelle della versione italiana analizzata in tre anni diversi (2005, 2007 e 2008) e una è quella in inglese del 2007. In questo modo è stato possibile confrontarne le differenti comunità sia tra di loro che con quelle degli autori di articoli scientifici e degli sviluppatori di software open source. Un confronto di questo tipo non è mai stato svolto sino ad ora.

1.3 Struttura della tesi

Nel Capitolo 2 si è approfondita la conoscenza di Wikipedia e del software su cui si basano tutti i suoi progetti e la gran parte dei wiki attualmente

sul Web. Sono stati analizzati i dati e gli strumenti messi a disposizione agli utenti per contribuire e per controllare i contributi degli altri membri della comunità. Si è colta l'occasione per introdurre i termini specifici che saranno usati nel corso della trattazione e le norme delle comunità per le versioni di Wikipedia analizzate.

Quindi sono stati analizzati differenti studi fatti negli ultimi anni su Wikipedia. Il primo tipo di studi riguarda quelli con risultati prevalentemente di tipo qualitativo. Poi si è trattato di quelli quantitativi, che possono prendere in considerazione differenti aspetti di Wikipedia. In particolare quelli di maggiore interesse per questo lavoro riguardano quelli sugli utenti di Wikipedia, anche se molti spunti interessanti sono stati tratti da lavori in aree diverse, come ad esempio gli studi sulla misura dei contributi e quelli riguardanti la qualità dei contenuti. In ultimo sono stati affrontati quegli studi che hanno cercato di rappresentare la comunità di Wikipedia come una Rete Sociale e si è avuto modo di vedere come questo tipo di analisi siano tra le meno approfondite.

Il Capitolo 3 affronta il problema di analisi da un punto di vista teorico. In esso vengono definiti e motivati i dati che è necessario estrarre da Wikipedia per questo lavoro. Successivamente vengono affrontate le differenti metriche per calcolare il contributo degli utenti a una pagina di Wikipedia e spiegate quali considerazioni è possibile trarre da ciascuna di esse. Quindi viene descritto il processo di selezione degli autori più importanti per ogni pagina e le informazioni che esso può dare. Infine viene affrontato il problema di modellare una Rete Sociale dei coautori di Wikipedia e di quali caratteristiche della comunità esso può permettere di individuare.

Nel Capitolo 4 vengono spiegate le scelte implementative che hanno portato alla realizzazione dei moduli software in grado di svolgere le analisi affrontate teoricamente nel capitolo precedente.

Nel Capitolo 5 sono definiti i parametri sperimentali che hanno determinato i risultati ottenuti dall'analisi di quattro differenti versioni di Wikipedia: quella italiana (nelle sue tre versioni del 2005, 2007 e 2008) e quella inglese (nella sua versione del 2007). Sono stati poi esposti e interpretati i risultati dei processi di analisi.

Infine nel Capitolo 6 sono state tratte le conclusioni del lavoro svolto e sono stati mostrati gli sviluppi futuri più interessanti alla luce dei risultati ottenuti.

Capitolo 2

Stato dell'arte

2.1 Wikipedia

Wikipedia è un'enciclopedia online, multilingue, a contenuto libero, redatta in modo collaborativo da volontari e sostenuta dalla Wikimedia Foundation, un'organizzazione senza fine di lucro (Wikipedia, 2009a). Essa nasce nel 2001 dalle ceneri di Nupedia, progetto analogo redatto però da esperti di settore, terminato a causa della scarsa competitività con altre enciclopedie. I fondatori del progetto, Jimbo Wales e Larry Sanger, decisero che il metodo tradizionale basato su revisione paritaria era troppo complesso per raggiungere in breve tempo i loro scopi e offrirono la possibilità di scrivere all'interno del loro progetto a tutti gli utenti del Web, utilizzando un tipo software chiamato *wiki*. I software di tipo wiki, il cui maggiore esponente attuale è MediaWiki[1], hanno come caratteristica principale quella di rendere modificabili le loro pagine Web da qualunque suo visitatore[2] con una sintassi semplificata rispetto a quella HTML, in modo da ampliare la possibilità di scrittura di pagine ipertestuali in modo collaborativo. L'originalità dell'idea, unita alla scarsa capacità tecnica richiesta per intervenire in un wiki ed alla crescente popolarità del Web, hanno reso in breve tempo Wikipedia non solo il più grande wiki al mondo, ma anche un punto cardine del cosiddetto passaggio alla seconda generazione del Web improntata proprio sulla costruzione di contenuti in modo collaborativo.

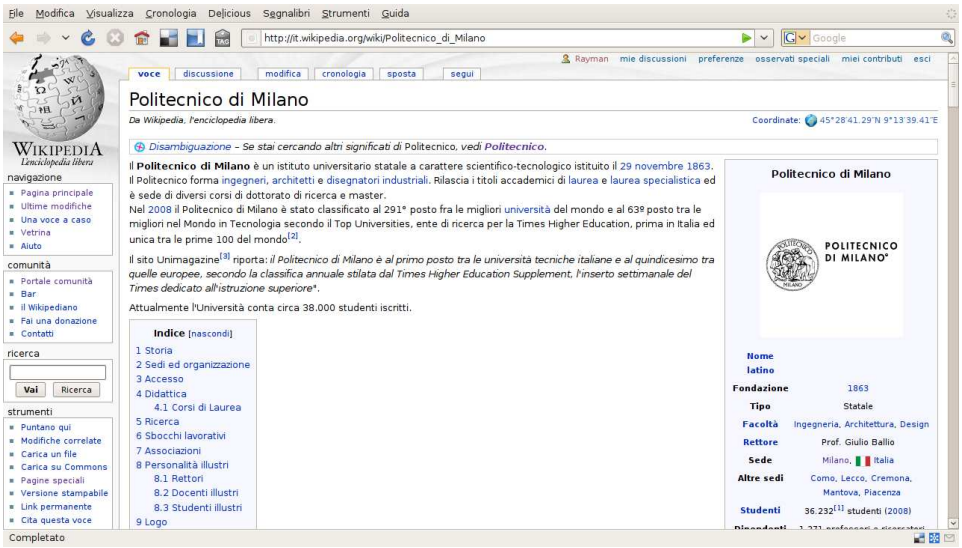
Il software MediaWiki presenta ciascuna pagina in maniera semplice e ordinata ma, al contrario di un qualunque altro sito Web, è subito visibile al visitatore un collegamento alla pagina che permette di modificare il con-

[1] Prodotto proprio per conto di Wikipedia e disponibile per la creazione di altri wiki sul sito `http://www.mediawiki.org/`.

[2] Anche se in alcuni casi è possibile effettuare dei controlli d'accesso degli utenti.

tenuto corrente, come è possibile vedere in figura 2.1. Questa caratteristica

Figura 2.1: Esempio di una pagina della versione italiana di Wikipedia.



del software è importantissima perché permette ai potenziali nuovi partecipanti di cominciare a collaborare immediatamente. Tuttavia l'evoluzione di Wikipedia non sarebbe stata così di successo se non ci fossero state altre funzioni più o meno avanzate per controllare e coordinare il suo sviluppo. Il più grande limite di un progetto wiki aperto a chiunque come Wikipedia è l'ovvia possibilità di essere soggetto a modifiche da parte di utenti malintenzionati. Ed ecco come la prima e inevitabile caratteristica di MediaWiki è quella di salvare ogni versione o *revision* di ciascuna pagina nel suo database, permettendo così di ripristinare senza alcuno sforzo uno stato precedente del sistema considerato migliore (*revert*). Strumento per beneficiare di questa funzionalità è la cronologia di una pagina o *revision history*, una particolare pagina che mostra la lista di tutte le revision ordinate cronologicamente dalla più recente con la possibilità di visualizzarne una qualsiasi oppure di confrontarne due distinte in cerca delle differenze. Per sfruttare meglio questo strumento è stata implementata una funzione, chiamata *diff*, in grado di evidenziare i cambiamenti tra due qualsiasi versioni del medesimo articolo. Molto interessante, ma non è questa la sede più opportuna per approfondire un confronto del genere, l'osservazione che questo modello di scrittura ha molto in comune con quello di sviluppo di software open source (Stvilia et al., 2005). Un ulteriore strumento di controllo è la *Lista degli osservati speciali*

che consente, ad ogni utente registrato, di dichiarare una lista di articoli per i quali si desidera essere informati di ogni cambiamento. Questo strumento è molto usato per rilevare i vandalismi oltre che per essere sempre informati sull'evoluzione di una pagina. Al fine di favorire la cooperazione tra i suoi partecipanti, Wikipedia prevede per ogni articolo una pagina speciale, realizzata sempre tramite la tecnologia wiki, nella quale discutere su questioni inerenti a esso. Questa pagina, detta *Discussion Page*, è in prima istanza usata per pianificare le decisioni sugli sviluppi dell'articolo a cui si riferisce, sia per quelle occasioni in cui non è facile trovare un accordo, sia per suddividere il lavoro di aggiornamento, manutenzione e arricchimento. Si noti anche come questa pagina, seppur non sia il suo scopo principale, sia importantissima per la creazione di una comunità perché è proprio in essa che i nuovi arrivati cominciano a prendere coscienza delle differenti identità che agiscono su un articolo e delle convenzioni e regolamenti dell'enciclopedia.

Poiché, come si è visto nell'esempio delle pagine di discussione, la tecnologia wiki viene utilizzata non solo per la scrittura delle voci dell'enciclopedia, Wikipedia è stata suddivisa in differenti aree chiamate *namespace* (Wikipedia, 2009d). Il namespace più importante è ovviamente quello principale (*Main*) nel quale si trovano tutti gli articoli enciclopedici. Tra gli altri è interessante segnalare quello delle pagine di discussione (*Discussione*), quello di documentazione del progetto (*Wikipedia*), di guida per i nuovi arrivati (*Aiuto*), quello delle pagine con lo scopo di elencare articoli accomunate dallo stesso argomento (*Categoria*) e le pagine personali (*Utente*), create automaticamente per ogni utente al momento della registrazione. Anche le pagine del namespace Utente hanno un importantissimo risvolto nella costruzione di una comunità, poiché in esse ciascun partecipante può esprimere la sua identità sia all'interno che all'esterno di Wikipedia.

Un aspetto forse più importante per i ricercatori della realizzazione tecnica e degli ideali su cui si fonda Wikipedia è proprio quello della comunità di persone che si è formata in modo spontaneo con l'obiettivo comune di costruire un'enciclopedia. Partendo da semplici linee guida e da pochi vincoli tecnologici, essa è stata in grado di costruire delle vere e proprie policy in grado di garantire un processo di scrittura razionale senza degenerare nell'anarchia. La Wikimedia Foundation ha definito cinque linee guida sulle quali si poggiano tutte le regole successivamente introdotte (Wikipedia, 2009c). Si tenga presente che, sebbene le linee guida siano le stesse per tutte le versioni di Wikipedia, le regole prodotte sono differenti in base alla lingua di ciascuna enciclopedia. Questo accade poiché ciascuna versione linguistica di Wikipedia è realizzata su una differente istanza del software MediaWiki e di conseguenza un utente è scoraggiato dal partecipare a differenti versioni

dovendosi registrare a ciascuna di esse in modo distinto. Questo aspetto è responsabile della nascita di comunità distinte operanti sui differenti wiki di ciascuna versione di Wikipedia. Il primo pilastro riguarda l'obiettivo di Wikipedia: la costruzione di un'enciclopedia ed in quanto tale un'informazione può essere inserita solo se rispetta dei criteri di verificabilità [3] e se non si tratta di una ricerca originale [4]. Il secondo punto cardine è l'ambizione del raggiungimento di un punto di vista neutrale o *Neutral Point Of View* (NPOV) che suggerisce di riportare in un articolo tutte le possibili opinioni su un dato argomento senza privilegiarne nessuna. Il terzo punto riguarda invece la libertà del contenuto di Wikipedia: la licenza scelta che implicitamente accetta ogni suo editor è la *Gnu Free Documentation License*[5]; questo aspetto vieta di inserire nell'enciclopedia alcun tipo d'informazione protetta da copyright. Un altro pilastro di Wikipedia sottolinea ancora l'importanza della comunità il cui codice di condotta deve essere improntato sul rispetto di ogni suo partecipante. Infine l'ultima linea guida dichiara che le regole possono cambiare a seconda delle esigenze della comunità al fine di mantenere intatto lo scopo principale dell'enciclopedia. A titolo d'esempio si vuole citare la politica di risoluzione dei conflitti presente nella Wikipedia italiana e descritta come un vero e proprio processo decisionale con tanto di eccezioni (Wikipedia, 2009c). Il punto di partenza prevede ovviamente la strategia di evitare ogni tipo di conflitto. Se questo non fosse possibile il primo passaggio della procedura prevede il dialogo con l'altra parte coinvolta. Se questo non bastasse si rende necessario l'intervento di uno o più mediatori ai quali chiedere un parere sul conflitto. Se le cose non dovessero ancora risolversi la decisione dovrà essere presa tramite votazioni. Il processo di votazione è usato praticamente in ogni caso in cui risulta non immediato prendere una decisione ed è per questo definito in maniera molto precisa.

Il punto di partenza più opportuno per comprendere la comunità di Wikipedia è sempre l'analisi del software MediaWiki (Forte and Bruckman, 2008). Esso prevede una grande distinzione tra *utenti anonimi* e *utenti registrati*. I primi sono identificati dal loro indirizzo IP e hanno il diritto di creare o modificare una qualsiasi pagina purché non rientri nella categoria di quelle protette. Gli utenti registrati invece, come in un comune sistema informativo, sono identificati da un nome e autenticati da una password se-

[3] Che implica la possibilità di verificare le fonti dalle quali è stata presa.

[4] Wikipedia non vuole essere una fonte primaria e cioè un documento storico, bensì una fonte secondaria (analizza, assimila, e/o sintetizza fonti primarie) o terziaria (generalizza ricerche esistenti o fonti secondarie su uno specifico soggetto preso in esame) (Wikipedia, 2009b)

[5] Disponibile nella sua versione italiana sul sito `http://www.softwarelibero.it/gnudoc/fdl.it.html`.

greta. Essi non godono di maggiori diritti sull'editing delle pagine rispetto agli utenti anonimi, fatta eccezione per la modifica delle pagine protette, tuttavia beneficiano di una migliore User Experience e possono partecipare attivamente alla comunità di Wikipedia. Infatti ciascun membro di questa classe di utenti ha a disposizione lo strumento che consente di definire la lista dei suoi osservati speciali e una propria pagina nella quale può presentarsi alla comunità. I contributi di un utente registrato gli sono correttamente attribuiti, mentre non è così per quelli non registrati poiché per ragioni tecniche l'indirizzo IP di un anonimo non garantisce l'identità dell'host. Il fatto che nella cronologia e nelle pagine di discussione un utente registrato venga presentato da una stringa piuttosto che da una serie di numeri, contribuisce sicuramente a rafforzare la sua identità all'interno della comunità. Vi sono quindi dei ruoli tecnici che può ricoprire un utente registrato offerti da MediaWiki. Il più interessante tra questi è quello di *Amministratore*, il quale può proteggere o sproteggere una pagina controversa dagli interventi anonimi, cancellare pagine, rimuovere il diritto di editing ad un utente (*blocco*) e disporre di ulteriori strumenti di revert. Un altro ruolo interessante è quello di *Burocrate* il quale gode della possibilità di investire altri utenti della carica di amministratore o burocrate e di rinominare i nomi di utenti registrati mantenendo però la loro identità in tutta l'enciclopedia. Anche l'assegnazione di questi ruoli avviene per votazione da parte della comunità.

Al di la di questi ruoli imposti dalla tecnologia, che comunque è progettata ad hoc per il caso di Wikipedia, sono poi sorte delle classi formali e non, alle quali ciascun utente dichiara o meno di appartenere. Secondo (Forte and Bruckman, 2008) esse possono essere distinte tra ideologiche, basate sul contenuto e funzionali. Le classi ideologiche e basate sul contenuto raggruppano rispettivamente utenti che condividono le stesse opinioni sulla redazione di un articolo e che decidono di partecipare solo ad alcune aree tematiche di Wikipedia poiché si considerano esperti di quel particolare argomento. Quelle funzionali raggruppano utenti che hanno scelto di contribuire a Wikipedia principalmente svolgendo un particolare compito. Ad esempio a questa classe appartiene il gruppo degli utenti che dedicano la maggior parte del tempo nella lotta contro i vandalismi, quelli che si occupano della correzione di errori ortografici oppure di dare una migliore struttura ad un articolo senza cambiarne il contenuto. Un'ulteriore classe da non dimenticare è quella dei *Bot*, agenti software programmati da esseri umani con scopi differenti. Esistono Bot che svolgono semplici compiti di amministrazione dei contenuti, come ad esempio l'aggiornamento delle pagine riguardanti gli eventi storici accaduti in una certa data. Si pensi a come l'intervento di questi agenti possa contribuire a correggere errori molto noiosi

da individuare e sistemare per un essere umano. Altri ancora semplicemente monitorano alcune pagine in attesa di alcuni eventi o registrando delle statistiche. Infine è stata accertata la presenza di Bot "maligni" programmati con lo scopo di compiere atti vandalici sull'enciclopedia. É presente un regolamento di Wikipedia che impone, tra le altre cose, la dichiarazione dei Bot (Wikipedia, 2009g) tramite la creazione di un account dedicato il cui nome utente contenga la stringa "bot". Ovviamente i creatori di questo genere di Bot non hanno alcun tornaconto nel seguire questo regolamento, hanno anzi tutto l'interesse di camuffare quanto più possibile le loro azioni per non far si che qualcuno li interrompa. Per questo motivo può essere complicato identificare la tipologia di tutti gli agenti operanti su Wikipedia.

In un ambiente così aperto dal punto di vista delle partecipazioni sono sorte spontaneamente anche delle basilari forme di reputazione. Sebbene un certo gruppo di individui condivida la convinzione che l'identità in Wikipedia dovrebbe contare molto poco, è inevitabile guardare agli interventi fatti da utenti anonimi con maggiore sospetto rispetto a quelli degli utenti registrati. Viceversa, sfruttando gli strumenti messi a disposizione da MediaWiki come le pagine di discussione, le cronologia e le pagine personali, ci si può qualitativamente rendere conto di quali siano gli utenti più affidabili e attivi all'interno della comunità. La misura di attività più usata su Wikipedia, a causa della semplicità di calcolo, è sicuramente il conteggio degli interventi o *edit count*. Ogni qualvolta un utente esegue una qualsiasi modifica ad una qualunque pagina, il suo contatore viene incrementato di un'unità. Questo dato non ha valore formale all'interno di Wikipedia, anche perché è considerato troppo poco preciso ed è facilmente alterabile[6]. Tuttavia esso è spesso sfoggiato, assieme alla data di creazione dell'account, come una sorta di titolo nobiliare specialmente dagli utenti di vecchia data che hanno potuto collezionare un alto punteggio nel tempo.

Col tempo è anche emerso dalla comunità un processo per eleggere quegli articoli particolarmente ben riusciti con lo scopo di segnalare ai lettori di quali voci fidarsi maggiormente. Queste pagine particolari sono "messe in Vetrina" (Wikipedia, 2009e) e vengono anche chiamate, nella versione anglofona di Wikipedia, *Featured Articles*. Per raggiungere questo stato, segnalato con una stella dorata in cima alla pagina, un articolo deve essere prima di tutto segnalato da un utente registrato. Partirà da quel momento un processo di valutazione dello stile, della prosa, dell'esaustività e della neutralità, durante il quale gli interessati potranno colmare le lacune evi-

[6]L'edit count è alterabile nel senso che con un minimo sforzo, un qualunque utente può aumentarlo a dismisura anche facendo uso di sistemi automatici quali software.

denziate dalla comunità (Wikipedia, 2009f). Alla fine di questo periodo tutti gli utenti iscritti da almeno trenta giorni e autori di almeno cinquanta contributi, potranno esprimere il loro voto in proposito e, in caso di decisione negativa, dovranno fornirne una valida motivazione. Allo scadere del tempo di voto, una pagina sarà inserita in vetrina solo se avrà ricevuto almeno l'80% dei voti favorevoli su un totale di almeno dieci. In caso di non accettazione dell'articolo la votazione non potrà essere ripetuta prima di tre mesi.

Partendo da un'idea e da un software tutto sommato semplici, Wikipedia è cresciuta sotto diversi punti di vista in modo inimmaginabile in un breve periodo, diventando un punto di riferimento per l'intero World Wide Web in quanto uno dei suoi siti più visitati. Essa conta infatti più di 250 edizioni in lingue differenti la cui maggiore, per numero di articoli, è quella in inglese nella quale si possono trovare circa 2.7 milioni di voci ad oggi. Anche in Italia il progetto ha avuto un notevole successo e la versione italiana di Wikipedia contiene attualmente circa 500 mila articoli. La reazione più immediata è quella di stupore, per un progetto che con una minima coordinazione imposta dall'alto ed esigui finanziamenti è riuscito a raggiungere dei risultati così importanti. Sebbene la qualità dei contenuti sia molto discontinua ed in troppi casi ancora sia facile imbattersi in errori od omissioni, come evidenziato in (Lipczynska, 2005), Wikipedia sta acquistando sempre maggiori consensi dai navigatori del Web che in essa trovano un repository di informazioni tra i più vasti e strutturati di tutta la Rete. Questo processo oggi pare inarrestabile e di conseguenza l'atteggiamento più corretto sembra essere quello di analizzarlo, piuttosto che attaccarlo. Gli svantaggi di Wikipedia rispetto ad un'enciclopedia tradizionale sono ovviamente legati alla sua instabilità. Le informazioni contenute al suo interno possono infatti cambiare in precisione, in attualità e addirittura in correttezza nella frazione di un secondo, come rimanere immutate per anni. Tuttavia questo punto debole potrebbe diventare un punto di forza nel momento in cui si disponesse degli strumenti adatti per valutare la qualità dell'informazione contenuta in Wikipedia. Le tradizionali enciclopedie sono sempre state soggette alla rapida obsolescenza, problema che sta emergendo specialmente in questi anni di rapide innovazioni in svariati campi del sapere umano. Si è avuto modo di vedere, qualitativamente, che alcune voci di Wikipedia sono aggiornate rispetto agli eventi mondiali in tempo reale.

Si capisce come Wikipedia sia un interessantissimo caso di studio per le accademie ed in particolare per informatici, gestionali e giornalisti. Non deve quindi stupire come i primi studi su di essa provengano da pubblicazioni da ambiti non strettamente legati alla tecnologia. Tutti gli studi su

Wikipedia possono beneficiare della trasparenza dei processi su cui si basa l'enciclopedia. La difficoltà principale risulta quella di saper gestire e analizzare l'enorme mole di dati prodotta da tutta la comunità. Nonostante i primi studi fatti su Wikipedia fossero di tipo qualitativo, i più interessanti e validi dal punto di vista scientifico sono quelli di tipo quantitativo. Nelle prossime sezioni verranno spiegati e confrontati queste due tipologie di studio in relazione agli obiettivi di questa Tesi.

In ultimo si precisa che, se come punto di riferimento per la descrizione di Wikipedia si è presa in considerazione la sua versione in lingua italiana, ove non diversamente specificato le ricerche presentate nella prossima sezione si rifaranno alla sua versione inglese. Questo interesse per la versione inglese è chiaramente motivato da ragioni di verificabilità, dato che la lingua ufficiale della comunità scientifica è attualmente l'inglese, sia da ragioni di maturità della suddetta versione di Wikipedia rispetto a tutte le altre, perlomeno per quanto riguarda le sue dimensioni.

2.2 Studi qualitativi

Uno dei primissimi studi riguardanti la qualità di Wikipedia è quello di Andrew Lih (Lih, 2004), interessato al fenomeno delle citazioni di articoli dell'enciclopedia da parte di testate giornalistiche[7]. Più che per la comunque interessante tesi dell'articolo, cioè che una pagina Wikipedia citata dalla stampa godrà di un aumento della sua qualità grazie alla maggior affluenza di visitatori, il lavoro è diventato un punto di riferimento degli studi successivi per l'aver individuato per la prima volta due semplici ma precise metriche di qualità. Il *rigore* di un articolo è misurato come il numero totale delle modifiche che esso ha ricevuto. L'assunzione è quella secondo la quale un articolo con un grande numero di modifiche ricevute sarà stato soggetto ad un maggior numero di controlli di qualità. La seconda dimensione individuata da Lih per valutare la qualità di un articolo è la *diversità*, cioè il numero di autori distinti che hanno collaborato alla sua realizzazione. Questa misura è importante per valutare la coralità dell'articolo, assumendo che un numero maggiore di contributori possa influire positivamente sulla sua qualità.

Un articolo che approfondisce e deve essere usato per meglio interpretare il precedente (Lih, 2004) è quello che sottolinea una certa relazione tra ciò che è popolare su Wikipedia e sui motori di ricerca (Spoerri, 2007). Sfruttando dati statistici pubblicamente accessibili del webserver di Wikipedia viene messo in evidenza come ben il 70% del traffico su di essi provenga da

[7] Tra le quali parecchie sono siti Web considerati autorevoli dall'autore.

un motore di ricerca. Unito alla considerazione che l'insieme delle pagine più visitate di Wikipedia ha significative sovrapposizioni con quello delle query maggiormente effettuate sui motori di ricerca, si evince che le dimensioni mostrate da (Lih, 2004) potrebbero essere influenzate, più che dalle citazioni da parte di testate giornalistiche, proprio dai risultati dei motori di ricerca stessi. Questo legame di Wikipedia con i motori di ricerca è evidenziato anche da (Viegas et al., 2007) nel tentativo di motivare l'aumento di vandalismi identificati nell'enciclopedia dal 2005 al 2007. Le pagine di Wikipedia, essendo spesso ai primi posti dei risultati forniti dai motori di ricerca, diventano un bersaglio più interessante per vandalismi o attacchi spam, poiché facilmente raggiungibili all'interno del Web.

Un'interessante articolo riguardante i possibili modelli di valutazione della qualità dell'informazione di una fonte, e in particolare di un articolo Wikipedia, ricorda tuttavia quanto sia difficile e soggettivo questo compito (Stvilia et al., 2005). Il modello considerato più adatto per quest'attività si sviluppa lungo tre dimensioni, ciascuna delle quali prevede una serie di indici che devono essere valutati per ogni articolo: la prima dimensione riguarda la *qualità intrinseca*[8] dell'articolo nella sua globalità; la seconda è la *qualità relazionale/contestuale*[9] delle differenti informazioni all'interno del medesimo articolo; la terza è la *reputazione*[10] dell'informazione in una certa comunità. Il lavoro prende in considerazione dei piccoli insiemi di articoli che verranno studiati secondo i parametri del modello dagli autori in persona. La scelta degli autori è stata quella di prendere alcuni Featured Article, alcuni articoli segnalati dalla comunità come non conformi alla politica di neutralità dell'enciclopedia oppure considerati come non sufficientemente accurati e un ultimo campione di articoli non appartenenti a queste categorie. Il contributo più interessante del lavoro è tuttavia quello che prende in considerazione le pagine di discussione di ciascun articolo e cerca di individuare il processo di dibattito che ha portato le pagine, in particolare quelle che hanno raggiunto la qualifica di Featured, a migliorare secondo i parametri del modello. Si ritiene che una grande quantità di informazioni possa essere ottenuta a partire dalle pagine di discussione, che tuttavia hanno lo svantaggio di non essere strutturate, bensì organizzate come un vero e proprio wiki, e devono quindi essere consultate in modo qualitativo. Questo studio viene utilizzato

[8]La quale comprende gli indici di accuratezza/validità, coesione, complessità, consistenza semantica, consistenza strutturale, attualità, ridondanza, naturalezza, completezza.

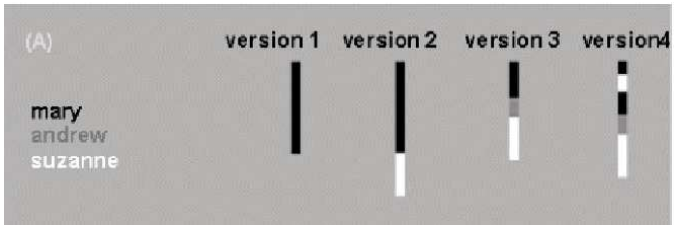
[9]La quale comprende gli indici di accuratezza, complessità, accessibilità, naturalezza, ridondanza, rilevanza, precisione, sicurezza, consistenza semantica, consistenza strutturale, verificabilità, volatilità.

[10]La quale comprende un unico indice di autorità.

anche per trarre delle informazioni, sempre di tipo qualitativo, sulla comunità degli utenti di Wikipedia. Si nota infatti che gli articoli appartenenti alla categoria Featured hanno in comune una ristretta cerchia di individui che condivide delle norme sociali di cooperazione non definite formalmente all'interno della comunità, la quale sembra sviluppare dei meccanismi di reputazione che esulano dalla semplice divisione in categorie di utenti.

Partendo dalla considerazione che molti articoli di Wikipedia risultano effettivamente di qualità, nonostante le sue linee guida di comportamento siano così poco restrittive, si è cercato di capire come non solo le cooperazioni ma anche i conflitti interni tra gli utenti di Wikipedia potessero avere influenza sugli articoli prodotti (Viegas et al., 2004). Vista la difficoltà di analizzare lunghissime revision history, specie per le pagine più importanti, gli autori del lavoro hanno avuto la brillante intuizione di realizzare un software che visualizzasse graficamente l'evoluzione di un articolo. Il software *History Flow*[11] rappresenta l'evoluzione di un articolo tramite un istogramma il cui asse delle ascisse ordina cronologicamente le revision. Ciascuna revision è quindi visualizzata attraverso una linea suddivisa in frammenti proporzionali al numero di parole, ciascuno dei quali colorato in modo che al medesimo autore corrisponda sempre un unico colore. L'autore di un frammento di testo per una revision è la persona che per la prima volta lo ha inserito nella storia della pagina[12]. Si faccia riferimento alla figura 2.2 per un semplice esempio del funzionamento del software. Anche in questo

Figura 2.2: Esempio di un grafico prodotto dal software History Flow



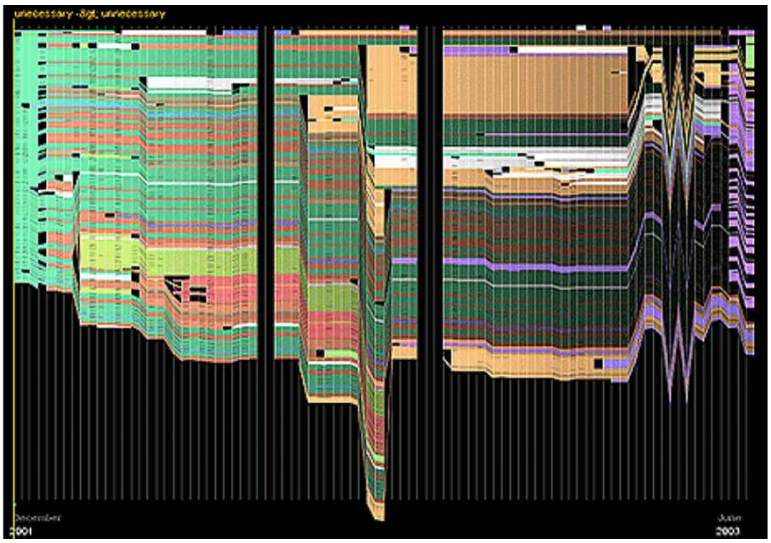
caso, trattandosi di analisi non automatiche poiché svolte dall'occhio umano, la scelta delle pagine in esame è limitata ad un ristretto sottoinsieme di esse. Ciò nonostante gli autori sono stati in grado di individuare alcune dinamiche interessanti. La prima riguarda le cancellazioni di massa, pratica

[11] Disponibile all'indirizzo: `http://www.research.ibm.com/visual/projects/history_flow/` .

[12] Questo dato viene calcolato tramite l'algoritmo di differenze tra files introdotto nell'articolo (Heckel, 1978).

di vandalismo che consiste nella cancellazione del contenuto di una pagina. Si è osservato che questa pratica è abbastanza frequente ma anche che essa viene individuata e corretta da altri utenti in tempi molto brevi. Si veda ad esempio la figura 2.3. Grazie al grafico prodotto dal software History

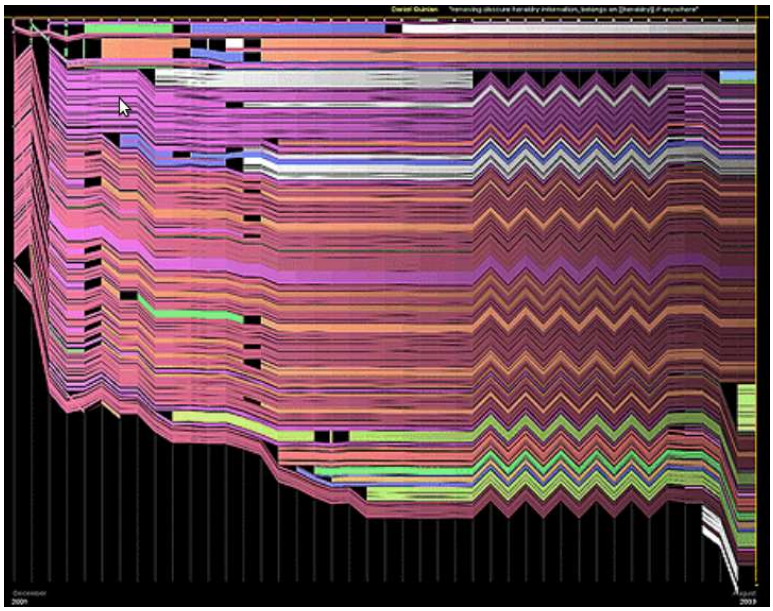
Figura 2.3: Esempio di un grafico prodotto dal software History Flow per la pagina Abortion



Flow sono state osservate delle dinamiche di negoziazione, cioè dei periodi nella storia di una pagina nella quale il contenuto è alternativamente aggiunto da un utente e poi rimosso da un altro[13]. Si veda ad esempio la figura 2.4. Riguardo a questo sistema di trovare un accordo sulla versione finale di una pagina, che comunque è utilizzato prevalentemente in relazione a quegli utenti che non conoscono lo strumento di dibattito delle pagine di discussione, ci si è resi conto del fatto che esso può presentarsi senza particolari differenze sia in pagine che trattano argomenti solitamente ritenuti controversi (come la pagina sull'aborto) che in pagine insospettabili (come la pagina sul Cioccolato). Fatta eccezione degli atti di vandalismo, ci si potrebbe aspettare un'evoluzione degli articoli di Wikipedia di tipo monotona e che tende a stabilizzarsi nel tempo, una volta raggiunta una certa qualità della voce. Numerosi esempi prodotti da History Flow invece dimostrano che spesso, quando una pagina raggiunge una certa dimensione, essa subisce un drastico taglio dopo il quale la sua crescita riprende. Questo fenomeno è da

[13]Meccanismo che può degenerare, se prolungato eccessivamente, in quelle che nel gergo di Wikipedia sono chiamate *edit war*.

Figura 2.4: Esempio di un grafico prodotto dal software History Flow per la pagina Chocolate



imputarsi alla frequente abitudine, all'interno di Wikipedia, di suddividere le voci più grandi in altre più piccole. Se pensiamo ad esempio ad un ipotetico articolo sull'informatica, è normale che all'inizio esso parli in generale di tutte le sue numerose branche. Con l'aumentare dei contenuti la pagina può assumere un aspetto che poco ha della voce enciclopedica e la comunità può decidere di rimuovere, ad esempio la sezione dell'articolo riguardante l'intelligenza artificiale, creando una nuova voce nell'enciclopedia contenente solamente le informazioni su quest'area. Un'altra dinamica osservata qualitativamente riguarda il tasso di sopravvivenza del primo intervento per ogni pagina. I grafici prodotti mostrano infatti che questo intervento è quello che rimane più a lungo nella storia della pagina, forse a causa del fatto che il creatore di un articolo influenza i futuri interventi definendone la tematica principale.

Lo studio dei ruoli non formalizzati degli utenti all'interno della comunità di Wikipedia è intrapreso anche nell'articolo *Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia* (Bryant et al., 2005). Per gli autori anche l'apprendimento delle norme sociali in Wikipedia, come in altre comunità di pratica, è situato ed avviene attraverso la

partecipazione periferica legittimata[14] dalla comunità stessa. I nuovi arrivati, grazie alla possibilità di interagire e osservare direttamente gli esperti del dominio, possono apprendere come meglio adattare i propri sforzi al contesto che li circonda. Col passare del tempo e delle esperienze, il loro grado di partecipazione assumerà forme sempre più centrali per il funzionamento della comunità. Il lavoro utilizza la *teoria delle attività* (Engestrom, 1999), basata su sei dimensioni distinte ma influenzate l'una dall'altra, come struttura per descrivere la trasformazione della partecipazione di un utente nel tempo. La ricerca viene effettuata intervistando un campione di nove volontari più o meno attivi in Wikipedia. Mentre la dimensione dell'*oggetto* rimane l'obiettivo comune di costruire e condividere la conoscenza in un formato enciclopedico, il cambiamento nella dimensione del *soggetto* viene identificato nel passaggio dall'inserimento, da parte dei nuovi arrivati, del solo contenuto di loro interesse apportando perlopiù piccole modifiche, a quello degli utenti esperti, interessati maggiormente alla qualità dell'intera enciclopedia piuttosto che a quella di un singolo articolo, fino al raggiungimento di una situazione nella quale essi saranno più impegnati in quelli che vengono chiamati *meta-task*, e cioè compiti di coordinamento, che non nell'attività stessa di scrittura. Per quanto riguarda la trasformazione nell'*uso degli strumenti*, gli autori individuano un cambiamento nel modo di collaborare nelle esili barriere d'ingresso alla partecipazione in Wikipedia. I nuovi arrivati, con la semplice pressione di un bottone all'interno dell'articolo che stanno consultando, possono ritrovarsi a modificarlo senza dover per forza imparare una sintassi di formattazione del testo. Al tempo stesso sono altrettanto facilmente raggiungibili gli strumenti che permettono ad un utente esperto di controllare da un punto di vista sopraelevato la qualità globale dell'enciclopedia, come le pagine di discussione, la lista delle revision e la notifica delle modifiche per le pagine di maggiore interesse. Infine la *percezione della comunità*, l'*attività di governo* e la *divisione del lavoro* vengono considerate praticamente indistinguibili nel caso di Wikipedia. Il passaggio da utente novizio a utente esperto si ha con l'acquisizione della consapevolezza dell'esistenza di una comunità e con l'apprendimento delle norme che la regolano.

Le tesi di *Becoming Wikipedian* sono condivise anche da (Viegas et al., 2007), tanto da spingere gli autori ad estendere la metodologia applicata in (Viegas et al., 2004) ed il software History Flow. Una primissima osservazione è possibile semplicemente rieseguendo il processo di visualizzazione di una pagina con i dati aggiornati di Wikipedia. La crescita del numero

[14]Introdotta e spiegata in (Lave and Wenger, 1991).

di revision dal 2005 al 2007 è visibilmente più che lineare e questo, a causa della necessità di rappresentare comunque il diagramma nella medesima area di visualizzazione, mette in evidenza schemi di evoluzione leggermente differenti. Ad esempio il tipico andamento periodico delle pagine soggette a edit war è troppo breve per essere notato. Rimangono tuttavia ancora valide le considerazioni sulla rapida identificazione di vandalismi. Una considerazione molto importante di questo lavoro è il tentativo di quantificare il processo di evoluzione di Wikipedia suddividendo lo studio della crescita dei suoi namespace. Sono i namespace non principali, e in particolare quello delle pagine di discussione degli articoli e degli utenti, che contengono le interazioni di coordinamento dell'enciclopedia ed è quindi molto interessante vedere come i risultati in tabella 2.1 evidenzino un tasso di crescita molto ripido proprio per questi namespace riguardanti meta-task. Un'analisi del namespace delle pagine di discussione viene svolta in questo lavoro in modo manuale. Come già spiegato infatti, sebbene le pagine di discussione siano percepite dalla comunità come un forum, la loro implementazione è comunque strutturata secondo la tecnologia del wiki. Chiunque può modificare un intervento altrui e spesso questa pratica è svolta dagli amministratori per liberare la pagina di discussione da interventi riguardanti dispute già risolte. L'analisi su un ristretto numero di pagine tuttavia mostra come esse siano utilizzate non solo per risolvere conflitti all'interno delle pagine, ma anche per pianificare e suddividere il lavoro di scrittura. Un ulteriore ruolo di queste pagine è quello di apprendimento da parte dei nuovi utenti, poiché è in esse che gli esperti rispondono alle loro domande riguardanti le policy di Wikipedia. Esse contengono dunque ancor più informazioni sulla comunità di Wikipedia di quanto gli autori avessero immaginato prima di questa ricerca. Rimane invece misterioso, secondo gli autori, il motivo che spinga così tanti individui a dedicare un così grande impegno ad un progetto che non garantisce nessun beneficio materiale.

In realtà lo studio della comunità di Wikipedia ricorda per molti versi lo studio di una comunità virtuale come affrontato in (Kollock and Smith, 1996). Nonostante la percezione sia quella che i sistemi di comunicazione mediati da sistemi informatici abbiano effetti positivi sulla società, favorendo una maggiore partecipazione, maggiore trasparenza, enfasi sul merito piuttosto che sullo stato e quindi un appiattimento delle gerarchie sociali, la questione critica rimane sempre quella della cooperazione, cioè l'abbattimento della tensione tra razionalità individuale e collettiva. I principi utilizzati da (Kollock and Smith, 1996) per identificare una comunità online in grado di organizzarsi e di autogovernarsi sono quindi applicabili anche a Wikipedia. I *confini della comunità sono definiti chiaramente* dall'obiettivo

Tabella 2.1: Crescita dei namespace di Wikipedia

Namespace	Pagine nel 2003	Pagine nel 2005	Tasso di crescita
Main	170369	1531095	9x
Talk	20067	229999	11x
User	3324	76491	23x
User talk	2564	199264	78x
Wikipedia	1211	81738	68x
Wikipedia talk	441	7267	16x
Image	6970	292451	42x

globale di costruire una conoscenza enciclopedica libera in una determinata lingua; tuttavia le sue dimensioni non sono definite a priori ed in questo senso possono presentarsi dei problemi di coordinamento. Le *regole di governo dei beni collettivi* sono motivate dall'ambiente in cui opera la comunità e possono, per la maggior parte, essere modificate dai membri della comunità. Il *controllo dell'applicazione delle regole e la punizione per la loro infrazione* sono applicati dagli amministratori eletti dalla comunità. Il mezzo informatico in quest'ultimo punto rende molto facile il controllo ma più difficile la punizione e nel caso di Wikipedia tutto ciò si rispecchia nella rapida individuazione dei vandalismi ma al tempo stesso nella difficoltà ad eliminarli in modo permanente.

L'ultimo studio qualitativo su Wikipedia è un'intervista (Riehle, 2006) a tre membri della Wikimedia Foundation: Angela Beesley per la Wikipedia inglese, Elisabeth Bauer per la Wikipedia tedesca e Kizu Naoko per la Wikipedia giapponese. Tutte e tre le intervistate possono citare un percorso di partecipazione molto simile rispetto a quello delineato da *Becoming a Wikipedian* e, proprio perché il loro lavoro si concentra attualmente sui meta-task all'interno di Wikipedia, sono da considerarsi esperte di dominio di tre distinte comunità di Wikipedia. É interessante la dichiarazione, validata da tutte le intervistate, che ciascuna Wikipedia può contare due tipologie di utenti: una prima di persone molto attive e una seconda di editor occasionali; sulla motivazione del primo insieme di individui però, nemmeno loro sanno dare una spiegazione precisa. Sebbene le policy e gli strumenti decisionali di enciclopedie in diversi linguaggi siano distinti, in quanto completamente definite dalla comunità, è possibile delineare alcuni tratti comuni tra molte Wikipedia: un sistema di riconoscimento delle pagine migliori basato sul consenso della maggioranza e promosse ad una maggiore visibilità all'interno del sito, utile incentivo a migliorare l'impegno dei membri della comunità; il conflitto causato dal disaccordo sul contenuto di un articolo; i

problemi di qualità discontinua all'interno dell'enciclopedia. A proposito dei problemi sulla qualità vi è la ferma convinzione che i migliori articoli siano scritti da pochi utenti esperti del domino, piuttosto che da un'intelligenza collettiva. Le speranze per il futuro riguardano il rafforzamento delle idee alla base di Wikipedia: la volontà di mantenere un ambiente di costruzione di conoscenza collettivo con basse barriere d'ingresso dal punto di vista tecnico e basato su una comunità in grado di accogliere ed istruire i nuovi arrivati. L'intervista si conclude con la speranza di avere una sempre migliore comprensione degli utenti che contribuiscono all'enciclopedia.

2.3 Studi quantitativi

2.3.1 Misurare Wikipedia

Per meglio addentrarsi nel consistente e variegato insieme di studi di tipo quantitativo su Wikipedia, si è scelto di farsi guidare da un articolo molto importante per i lavori degli anni a venire[15]: *Measuring Wikipedia* (Voss, 2005). In esso l'autore indirizza le ricerche su Wikipedia in sette direzioni principali. Per la prima tra queste, la *crescita di Wikipedia*, è lo stesso articolo a fornire i primi risultati, mentre per le altre il lavoro può essere più che altro considerato un importantissimo sforzo creativo di cui moltissime ricerche successive hanno beneficiato. Il primo risultato riguarda la considerazione che, dopo un breve periodo di crescita lineare, a partire dall'anno 2002 le dimensioni di Wikipedia sono cresciute in modo esponenziale. Le diverse dimensioni che confermano quest'affermazione sono: la dimensione del database (comprese quindi anche tutte le revisioni passate), il numero totale di parole, numero totale di collegamenti interni, numero di articoli (contenenti almeno un collegamento interno), numero di utenti attivi (autori di almeno cinque contributi in un mese) e numero di utenti molto attivi (autori di almeno cento contributi in un mese). Quindi l'articolo prende in considerazione gli studi sui singoli *Articoli* dell'enciclopedia. L'intuizione più importante tra le considerazioni sugli articoli riguarda principalmente la necessità di distinguere tra namespace differenti, in quanto chiari indicatori di contenuti con scopi differenti. La parte più sostanziosa dell'articolo si concentra tuttavia sullo studio degli *Autori* di Wikipedia. Il paragone di Voss è quello dell'enciclopedia con il mondo delle pubblicazioni scientifiche, la cui più grande differenza è l'enorme quantità di contributori di un articolo di Wikipedia rispetto ad uno scientifico. Tra gli studi considera-

[15]Questo studio è il terzo più citato cercando Wikipedia su `http://scholar.google.com`. Esso ad oggi ha ricevuto ben 110 citazioni.

ti questo è il primo che separa i contributi anonimi da quelli degli utenti registrati. Un dato importante, probabilmente influenzato dalle differenti culture nazionali, riguarda il fatto che il numero dei contributi anonimi è molto variabile tra le edizioni di Wikipedia: l'intervallo parte dal 10% di edit anonimi nella Wikipedia italiana, sino al 40% nella versione in lingua giapponese. Viene inoltre segnalata l'importanza di poter distinguere tra il differente apporto dato da diversi contributi. Per avere valori più precisi sull'attività dei singoli autori è infine necessario considerare unici gli interventi consecutivi da parte di uno stesso utente e tenere conto del comportamento dei Bot di Wikipedia. Per quanto riguarda la produttività degli autori, l'ipotesi verificata da Voss è la validità della legge di Lotka, ulteriore punto di contatto sia con il mondo dei ricercatori che degli sviluppatori di software open source. La legge di Lotka (Lotka, 1926) esprime il concetto che gli autori che maggiormente hanno partecipato al processo di scrittura sono quelli che maggiormente interverranno anche in futuro. Vi è poi interesse da parte dell'autore nelle comunità di utenti. Poiché la tecnologia dei wiki facilita la divisione del lavoro, sarebbe interessante poter distinguere differenti ruoli di utenti in modo automatico. Per questo Voss suggerisce l'utilizzo di tecniche di Social Network Analysis che verranno trattate nella prossima sezione. La quarta direzione di ricerca individuata è quella riguardante lo studio degli *Edit*, i quali potrebbero permettere l'identificazione di particolari pattern d'interazione quali, ad esempio, le edit war. Lo studio dell'entità di un edit potrebbe infine rivelarsi non banale. Quindi una strada differente per misurare Wikipedia viene identificata nello studio della sua *Struttura di collegamenti* ipertestuali che quasi ogni pagina utilizza per rimandare a concetti collegati alla tematica dell'articolo approfonditi all'interno dell'enciclopedia. Gli studi sui quali ci si può appoggiare sono quelli sulla struttura delle reti complesse, per una panoramica approfondita si consulti (Newman, 2003). La particolarità di Wikipedia rispetto a questo settore è la possibilità che un collegamento possa essere rimosso in qualsiasi momento, come accade nel Web. Inoltre i collegamenti di Wikipedia possono anche riferirsi ad articoli non ancora scritti. Collegato in qualche modo a quest'ultima strada, si intravede la possibilità di analizzare i *Contenuti* di Wikipedia, facendo analisi semantiche sulla rete di concetti definita dagli autori dell'enciclopedia tramite i collegamenti interni. Per concludere viene introdotto nel lavoro qualche spunto riguardante il problema della *Qualità*. Seppure venga ribadito che si tratta di un parametro soggettivo e quindi di difficile definizione, si ipotizza un'estensione del noto aforisma di Linus Torvalds (Raymond, 2001) secondo il quale: "Dato un numero sufficiente di

occhi, tutti i bug vengono a galla"[16]. Tuttavia, secondo l'autore, è necessario distinguere tra contenuto enciclopedico, al quale Wikipedia ambisce, e contenuto di una base di conoscenza, categoria alla quale molto probabilmente Wikipedia tutt'ora appartiene.

2.3.2 La crescita di Wikipedia e dei suoi articoli

Gli aspetti della crescita di Wikipedia sono trattati in diversi articoli in relazione a differenti misure. Per questo si è deciso di non raggruppare la loro esposizione in quanto questa direzione della ricerca può essere considerata trasversale alle altre. Se in un primo momento ci si può accontentare di uno studio statico di un fenomeno all'interno di Wikipedia, è molto utile, per confermarne la validità, riconsiderarlo in relazione al tempo. Come questo è utile per molti sistemi dinamici, a maggior ragione lo è per Wikipedia, che può considerarsi ancora in una fase iniziale della sua storia.

Sino ad oggi la maggioranza degli studi quantitativi su Wikipedia si è concentrata sulle voci dell'enciclopedia e quindi sul suo namespace principale. Tuttavia, specialmente per quanto riguarda le ricerche che vogliono studiare le dinamiche sociali di Wikipedia piuttosto che il suo aspetto contenutistico, tutti gli altri namespace possono essere considerati ancor più ricchi di informazioni. L'ostacolo che fino ad oggi ha frenato le analisi combinate sui differenti namespace è probabilmente la disparità di dimensioni tra quello principale e quelli secondari.

2.3.3 Gli utenti di Wikipedia

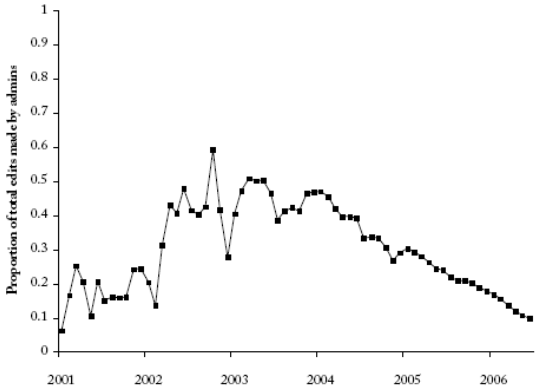
Gli studi sugli utenti di Wikipedia, come sottolineato in (Riehle, 2006), sono ancora pochi attualmente. Il primo d'interesse, *Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie*[17], è da considerarsi tutto sommato recente (Kittur et al., 2007). L'apertura a chiunque degli articoli di Wikipedia suggerisce che essa possa considerarsi scritta da un'*intelligenza collettiva* ma l'esperienza mostra che si può distintamente notare che un certo gruppo di utenti emerge tra gli altri per quantità di impegno speso su di essa. Capire quali siano effettivamente gli utenti che contribuiscono ad un sistema informativo collaborativo è un aspetto importantissimo per la progettazione della sua usabilità, perché può indicare in che direzione concentrare gli sforzi degli sviluppatori. Se effettivamente il sistema deve la sua ricchezza agli utenti esperti, allora sarà il caso di fornire

[16] "Given enough eyeballs, all bugs are shallow".

[17] Potere di pochi contrapposto alla saggezza della massa: Wikipedia e la nascita della borghesia.

strumenti software in loro aiuto. Altrimenti bisognerà soddisfare maggiormente le esigenze di quegli attori che non hanno grande esperienza nell'uso del sistema e necessitano quindi di strumenti che guidino il loro contributo. Le misure usate per compiere uno studio del genere su Wikipedia sono il numero di edit e l'entità degli edit contata in numero di parole. Per calcolare quest'ultima è stata utilizzata la funzione di *diff* fornita dal software MediaWiki ed in particolare è stato contato il numero di parole cambiate da una revision rispetto alla sua precedente. Il primo gruppo di utenti studiato è quello degli amministratori. Come si può vedere in Figura 2.5, il rapporto tra il numero di edit di questo gruppo elitario rispetto al totale sta subendo, all'epoca dello studio, una decrescita lineare opposta al suo andamento iniziale. Due ipotesi per giustificare questo fenomeno vengono quindi smentite

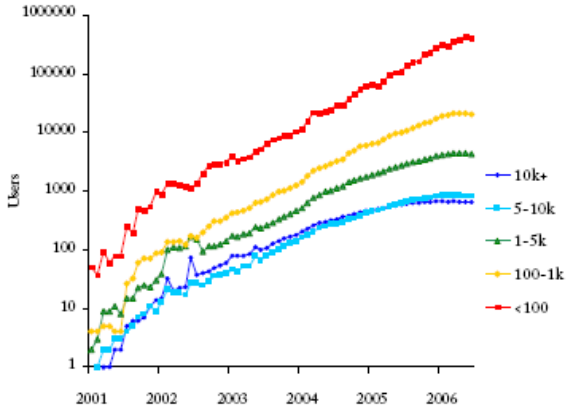
Figura 2.5: Andamento della percentuale del numero di edit fatti dagli amministratori di Wikipedia, secondo (Kittur et al., 2007)



dagli autori. Il valore assoluto di questa metrica non segue il medesimo andamento mostrato in Figura 2.5, anzi il numero di interventi da parte degli amministratori è in crescita. Ciò dimostra che il loro impegno non è diminuito. Anche l'ipotesi secondo la quale l'introduzione dei Bot alleggerisca il compito degli amministratori è smentita tramite il conteggio dei loro edit in relazione al totale, che mostra come questo fenomeno non possa spiegare il declino degli amministratori. In secondo luogo viene svolto uno studio su cinque diverse classi di utenti raggruppate per numero di interventi: meno di cento edit; tra i cento e i mille; tra i mille e i cinquemila; tra i cinquemila e i diecimila; oltre i diecimila. Per ciascuna classe vengono analizzate e confrontate differenti curve. La percentuale di edit della classe sul totale mostra come l'apporto dato dagli utenti con meno di cento edit stia aumen-

tando a discapito di quelli con più di diecimila edit. Il numero assoluto di edit mostra come il fenomeno precedente non sia dovuto ad una diminuzione di attività degli utenti con un numero di edit superiore a diecimila, bensì all'aumentare degli interventi fatti dalla classe di utenti con meno di cento edit. É lo studio del numero medio di edit per utente in un mese (e cioè la misura precedente normalizzata dal numero di utenti contenuti da ciascuna classe) che mostra come in realtà il comportamento degli utenti con pochi edit, mediamente responsabili di circa cinque interventi mensili, sia rimasto costante nel tempo. Si tratta quindi del grandissimo aumento della popolazione assoluta di questi utenti occasionali che influenza in questo modo l'apporto globale di quelli più attivi, come mostrato in Figura 2.6. Questa

Figura 2.6: Aumento della popolazione assoluta per classi di utenza in Wikipedia, secondo (Kittur et al., 2007)



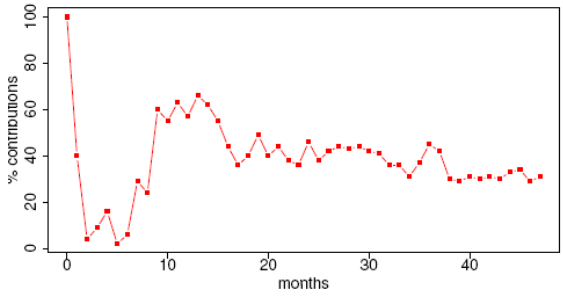
tesi viene confermata anche dall'andamento della popolazione di ciascuna classe relativamente a quella totale. Lo studio viene quindi portato avanti utilizzando, invece che il numero di edit, il numero di parole aggiunte per ogni intervento come misura di contributo. Anche questo approccio mostra le medesime dinamiche e funge quindi come ulteriore conferma della tesi degli autori. Tuttavia si mostra come l'apporto degli utenti più attivi sia più sostanziale, poiché essi non si limitano a piccole modifiche, bensì ad aggiunte di dimensioni mediamente maggiori (e quindi probabilmente più ricche di contenuti).

Si vuole però evidenziare come questo risultato sia da valutare attentamente in relazione alla funzione di *diff* utilizzata. Il rischio è quello, ad esempio, di contare i revert come nuovi contributi quando invece essi sono svolti in brevissimo tempo da chi ha individuato degli interventi tipicamente

di tipo vandalico. Un'ultima tesi dimostrata numericamente dagli autori è quella che gli utenti esperti, al contrario di quelli inesperti, in media aggiungono più parole di quante ne tolgono. Il contributo di questo lavoro è quindi quello di mostrare come studiare le dinamiche di un sistema collaborativo possa avere degli impatti sull'impiego di risorse in talune direzioni. Se all'inizio è importante soddisfare gli utenti più volenterosi e propensi a collaborare, in un secondo momento potrebbe risultare più interessante concentrare i propri sforzi nel facilitare l'intervento di quella moltitudine di persone che contribuisce in piccola misura.

Sebbene pubblicato nello stesso anno, il lavoro di (Ortega and Gonzalez Barahona, 2007) applica le misure dell'articolo appena discusso ad un insieme di dati più recente. Lo studio conferma le tesi del lavoro rivisitato, conferendogli quindi un'ulteriore validità. Il lavoro non si ferma qui ma prosegue estendendo il metodo applicato fino a quel momento. Viene presentato un software, *WikiXRay*[18], in grado di automatizzare per ogni versione locale di Wikipedia l'analisi di (Kittur et al., 2007) e non solo. Il primo risultato dato da questa estensione è che la versione in lingua svedese di Wikipedia non presenta così distintamente un fenomeno di nascita della borghesia. Come si nota in Figura 2.7 infatti, la percentuale degli interventi degli amministratori, a meno di una fase iniziale di assestamento, può considerarsi costante. La motivazione data per spiegare questo fenomeno è

Figura 2.7: Andamento della percentuale del numero di edit fatti dagli amministratori della Wikipedia svedese, secondo (Ortega and Gonzalez Barahona, 2007)



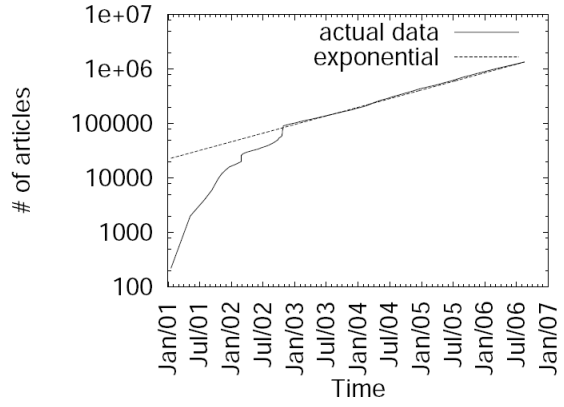
quella della differente politica di selezionamento degli amministratori. Se la carica di amministratore della Wikipedia inglese non ha data di scadenza, quella nella sua versione svedese ha la durata di un anno. Di conseguenza gli utenti che vogliono mantenere il loro ruolo amministrativo sono costretti

[18]Disponibile liberamente sul sito: `http://developer.berlios.de/projects/wikixray/`; e documentato su: `http://meta.wikimedia.org/wiki/WikiXRay`.

ad impegnarsi per dimostrare di meritarlo e quelli che invece non partecipano perdono il loro ruolo. La successiva estensione riguarda la possibilità di modificare gli intervalli di osservazione delle statistiche per raggiungere una maggiore precisione. Infatti la suddivisione delle classi di utenti può variare nel tempo e risulta interessante capire quanto è stabile l'insieme dei contributori più attivi. La selezione delle classi di utenti diventa quindi non più assoluta come prima, ma vengono considerate le classi del primo 5% degli utenti ordinati per il numero di edit. Il risultato ottenuto è che questo piccolo gruppo di utenti molto attivi rimane tale nel tempo e da solo è responsabile del 10% dei contributi totali.

Infine anche il lavoro di (Almeida et al., 2007) trae importanti conclusioni sugli utenti di Wikipedia. L'articolo verifica che la crescita dell'enciclopedia, misurata in numero di nuovi articoli in un intervallo temporale di un mese, è segue una legge power law, come si vede in Figura 2.8. Questo può essere

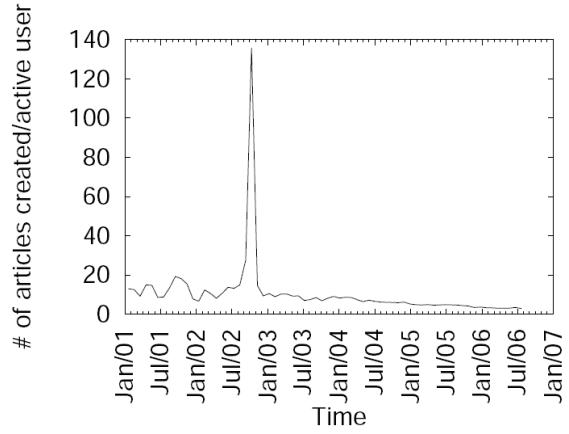
Figura 2.8: La crescita di Wikipedia misurata in numero di nuovi articoli al mese, secondo (Almeida et al., 2007)



spiegato in due modi secondo gli autori: o è aumentata la produttività dei singoli autori oppure il numero di collaboratori. Chiaramente le due ipotesi non sono esclusive e pertanto vengono misurate entrambe. Per misurare la produttività degli autori la scelta ricade sul calcolo del numero medio di articoli creati per autore[19]. Questa misura decresce linearmente, si veda la Figura 2.9, e ciò dimostra che la prima ipotesi non è verificata. Si noti il picco in concomitanza del mese di Ottobre 2002. Esso è spiegabile solo sapendo che durante quel periodo, nella Wikipedia inglese, un Bot si è occupato di creare una pagina per ogni comune degli Stati Uniti. Questo fatto segnala

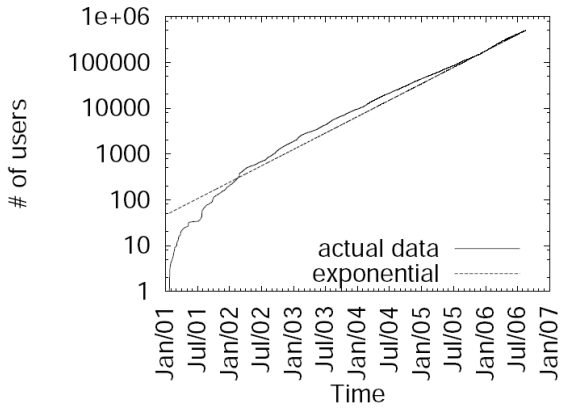
[19]In questo caso vengono considerati come autori solo gli utenti registrati.

Figura 2.9: La crescita di Wikipedia misurata in numero medio di articoli creati per autore, secondo (Almeida et al., 2007)



come l'opera dei Bot possa essere una sorgente di rumore il più delle volte non trascurabile. Per misurare invece l'aumento dei contributori si disegna il grafico del numero distinto di autori per mese su scala logaritmica, Figura 2.10. Si nota che la crescita di questo parametro è esponenziale, il che

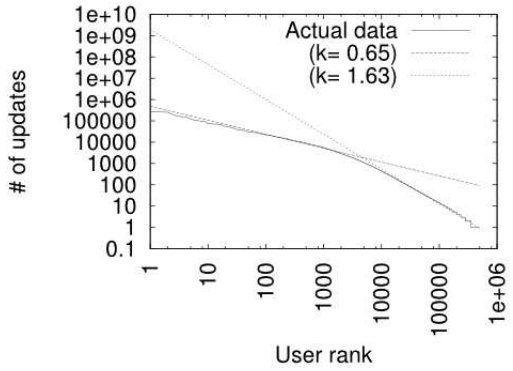
Figura 2.10: La crescita demografica di Wikipedia, secondo (Almeida et al., 2007)



dimostra che l'evoluzione di Wikipedia è dovuta alla rapidissima espansione nel numero dei suoi partecipanti. Gli autori quindi misurano il numero di edit per ogni utente e stilano una classifica che assegna un ranking dal valore più alto al valore più basso. Disegnando il grafico in scala logaritmica su entrambi gli assi di queste quantità, ci si rende conto che l'andamento

è quello di una retta che ad un certo punto cambia di pendenza, come mostrato in Figura 2.11. La conclusione è che ci sono due distinti gruppi di utenti: un piccolo insieme molto produttivo e una grande maggioranza che contribuisce all'enciclopedia in modo significativamente inferiore. Due

Figura 2.11: Classifica dei contributori confrontata col numero di edit (Almeida et al., 2007)



interessanti risultati riguardano i rapporti degli utenti con le pagine. Il primo conferma che il processo di creazione di un articolo è abbastanza esclusivo: infatti solo il 30% degli utenti totali ne ha creato uno, mentre il restante 70% si è limitato a modificarne i contenuti. Il secondo sostiene che gli interessi degli utenti sono abbastanza vari, poiché il numero medio di articoli modificati da ciascun autore è pari a quattro.

2.3.4 La misura dei contributi a Wikipedia

Come si è visto, la misura dell'opera dei contributori di Wikipedia è l'edit. É proprio a partire da questa che è stato proposto uno studio per un sistema di reputazione degli utenti guidato dai contenuti (Adler and de Alfaro, 2007). Un sistema di questo tipo, secondo gli autori, può risultare molto utile in diversi scenari, tra cui la colorazione del testo basata sulla reputazione dell'autore, l'applicazione automatica di restrizioni sulla modifica di articoli particolarmente importanti, l'aiuto nell'identificazione degli interventi vandalici da parte degli amministratori e ancora l'incentivo per gli utenti ad aggiungere contributi di alta qualità. Al contrario di molti sistemi di reputazione[20], questo sistema non è guidato dagli utenti: i voti sono assegnati in base alla semplice idea che quando un autore inserisce nel

[20]Tra i quali, ad esempio, quello del noto sito di aste online Ebay.

sistema una nuova revision, implicitamente esprime un parere su quelle che lo precedono, e di conseguenza sui loro autori. Quanto più la revision corrente manterrà gli apporti di una delle sue precedenti, tanto più le conferirà un voto elevato. In questo modo gli autori dei contributi maggiormente accettati (o modificati di meno) avranno un'alta reputazione nel sistema. La caratteristica di questo metodo di essere guidata dai contenuti porta con se il grande vantaggio di essere assolutamente trasparente agli utilizzatori, ma anche alcuni problemi. Il più importante è sicuramente quello che i contenuti possono essere cambiati per molteplici ragioni, tra le quali non tutte sono indice di scarsa qualità. Il sistema è però in grado di riconoscere quegli edit che sono stati annullati da quelli che invece sono stati successivamente migliorati. In questo modo gli atti vandalici ricevono un rinforzo negativo, mentre i contributi appropriati ma che possono essere migliorati ne ricevono uno parzialmente positivo. La reputazione definita dal lavoro risulta quindi avere un duplice valore. *Prescrittivo* poiché per avere una buona reputazione l'unico modo risulta quello di scrivere contenuti che dureranno nel tempo. *Descrittivo* perché permette di classificare gli utenti in base alla reputazione guadagnata all'interno del sistema.

In questa sezione, più che i dettagli dell'implementazione di questo sistema che verranno affrontati successivamente, è interessante enunciare i risultati ottenuti. Gli autori hanno calcolato la reputazione per tutti gli utenti della Wikipedia francese ed hanno notato una certa correlazione con la durata del contributo nel tempo, sia misurato come numero di modifiche che come quantità di nuovo testo aggiunto. La differenza tra le due misure sta nel fatto che la prima tiene conto anche dello spostamento e della rimozione delle parole tra due versioni. In particolare si nota come il 7.7% del numero di modifiche sia fatto da autori di bassa reputazione, cioè con reputazione inferiore al 20%. Il 32% del testo inserito da queste modifiche è considerabile di breve durata, cioè rimosso dalle successive revisioni per almeno l'80% del suo contenuto. Quindi le modifiche ad opera di autori con bassa reputazione hanno una probabilità di essere di breve durata 4.1 volte maggiore della media. Un discorso molto simile vale anche per il nuovo testo inserito da autori di bassa reputazione, pari all'8.4% del totale. Il 38% di questo testo è da considerarsi di breve durata. Il testo ad opera di autori con bassa reputazione ha quindi una probabilità di essere di breve durata di 4.5 volte superiore alla media. Infine viene cercata la correlazione della reputazione con il numero totale di edit. I risultati ottenuti sono leggermente inferiori rispetto alle precedenti metriche, ma gli autori non prendono in considerazione questa misura poiché facilmente alterabile. Infatti, se un sistema di reputazione basato su edit count venisse effettivamente imple-

mentato, sarebbe banale e anche dannoso per il wiki aumentare a dismisura il proprio punteggio.

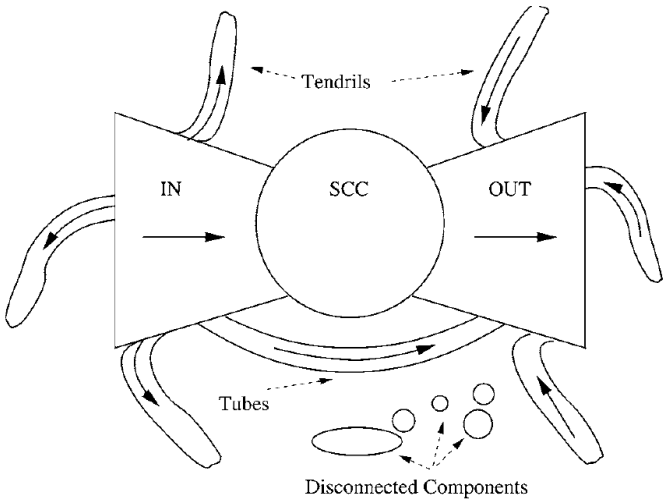
I successivi lavori del medesimo gruppo di ricerca, che ruota attorno al ricercatore Luca de Alfaro si sono focalizzati su due altri interessanti sviluppi del loro primo lavoro nell'ambito dei sistemi collaborativi. Il primo (Adler et al., 2008a) riguarda l'utilizzo delle metriche trovate per valutare una nuova misura di *contributo utile* degli autori di Wikipedia, che si avrà modo di approfondire nel corso di questa trattazione. Il secondo (Chatterjee et al., 2008) riguarda i possibili attacchi a cui il sistema può essere soggetto e le modifiche da apportare agli algoritmi per fronteggiarli.

2.3.5 La struttura dei collegamenti interni

In quanto semplificazione del linguaggio HTML, anche Wikipedia permette la creazione di un grafo orientato di pagine i cui archi sono i collegamenti ipertestuali. Per questo (Capocci et al., 2006) hanno avuto l'idea di studiare Wikipedia con le tecniche tipiche usate per studiare il World Wide Web. Nonostante vi siano delle differenze sostanziali, lo studio di Wikipedia è molto semplificato a causa della sua struttura centralizzata e della disponibilità dei dump del suo database. La più grande differenza con il World Wide Web è nella possibilità, da parte di chiunque, di modificare la struttura dei collegamenti tra le pagine. La prima interessante scoperta degli autori è che Wikipedia condivide con il Web la struttura topologica detta a "cravattino"[21], a causa del suo tipico disegno mostrato in Figura 2.12. Essa prevede una grande componente connessa (SCC) comprendente la maggior parte delle pagine e due ulteriori strutture, una di nodi (IN) che non possono essere raggiunti dalla SCC e un'altra di nodi (OUT) che non possono raggiungere la SCC. Vi sono infine altre componenti marginali come i tubi (Tubes) che collegano l'insieme dei nodi IN con l'insieme dei nodi OUT; i viticci (*tendrils*), sottoinsiemi di nodi che possono solamente raggiungere i nodi OUT oppure essere raggiunti dai nodi IN; le componenti sconnesse (*disconnected components*). Lo studio di questa struttura viene fatta per sei differenti versioni di Wikipedia in lingue differenti che mostrano più o meno simili proporzioni di queste componenti. In secondo luogo lo studio verifica la validità del modello di *preferential attachment* introdotto per la prima volta da (Barabasi and Albert, 1999) e di grande importanza per lo studio delle reti complesse. Secondo questo modello la probabilità che su un nodo incida un nuovo arco è proporzionale al suo grado (cioè il numero di archi

[21] In bibliografia nota come *bow-tie-like structure*.

Figura 2.12: Tipica struttura a "cravattino" delle reti complesse.



incidenti)[22]. Questo risultato è in un certo senso sorprendente. Il modello di preferential attachment è associato a reti nelle quali la diffusione di informazioni non è efficiente e viene assunta un'ipotesi di razionalità limitata. Se per il Web questo è vero perché i link sono aggiunti dai Webmaster solo limitatamente al proprio sito, in Wikipedia non c'è questo limite. Di conseguenza le ipotesi, tutt'ora da verificare, dei ricercatori sono due: gli utenti di Wikipedia non sono ancora in grado di sfruttare le potenzialità tecniche dello strumento dei wiki; oppure il modello è conseguenza dell'organizzazione della conoscenza in generale. Lo studio si conclude con l'individuazione delle sottocomunità del grafo, cioè quei sottografi con una densità superiore alla media. Questo studio è fatto sulla versione portoghese di Wikipedia in quanto, per le sue modeste dimensioni, l'algoritmo di ricerca può essere portato a termine in tempi rapidi. I risultati di quest'indagine mostrano la divisione tematica della conoscenza dell'Uomo. Le prime tre comunità trovate, in ordine di densità, risultano essere: quella dei film e attori di film brasiliani (circa 100 nodi); quella dei termini medici (circa 300 nodi); quella dei termini geografici (circa 200 nodi).

Dopo aver analizzato le proprietà generali della struttura dei collegamenti interni di Wikipedia, (Buriol et al., 2006) si occupano dell'evoluzione del suo grafo nel tempo. Una prima parte di questo lavoro si concentra sulla

[22]Si noti che nel caso di un grafo orientato questa proprietà va verificata nelle due dimensioni degli archi entranti e uscenti.

crescita di alcuni parametri numerici di Wikipedia. Il numero di articoli, di interventi, di visitatori e di autori distinti sono stati soggetti, sino al momento dello studio, ad una crescita esponenziale. Per quanto riguarda gli articoli, non solo è aumentato il numero ma anche la dimensione del contenuto. La crescita degli articoli sembra dipendere dalla loro età: infatti viene mostrato come i nuovi articoli siano mediamente più corti rispetto a quelli esistenti da più tempo ma anche come il loro tasso di crescita sia caratterizzato dalla medesima costante. L'evoluzione delle dimensioni di un articolo pare essere quindi una caratteristica immutata nella storia di Wikipedia. Lo stesso può essere detto per la dimensione media di un intervento, misurata dagli autori in un intervallo compreso tra i 300 e i 500 bytes, l'equivalente all'incirca di un parametro di testo. La distribuzione degli interventi per pagina ha invece un andamento power-law. Circa il 53% delle pagine ha ricevuto più di 10 edit ma solo il 5% ne ha ricevuti più di 100. Questa caratteristica è spiegata dal fatto che la popolarità di una pagina può essere correlata con il suo numero di visite Essendo anche ragionevole pensare che una pagina molto visitata riceva più edit, si può dedurre che la distribuzione degli interventi segua quella dell'importanza di una pagina. C'è anche da prendere in considerazione una ragione più tecnica: le liste di osservazione notificano gli utenti interessati ad una pagina di ogni loro modifica. In questo modo le pagine che ricevono più edit, ne riceveranno molti altri. Il numero di revert per numero di interventi è invece in debole crescita e attualmente si attesta al 6% del totale. Questo dato segnala un aumento dei vandalismi ma non pare essere correlato con il numero di interventi anonimi che invece è rimasto costante e compreso tra il 20% ed il 30% del totale. A questo proposito è da segnalare come ben il 70% dei ripristini sia effettuato entro un'ora dall'intervento rimosso. Gli autori notano quindi che diverse tipologie di pagine presentano diversi profili di aggiornamento. Alcune ricevono numeri variabili di aggiornamenti in risposta ad eventi esterni, come ad esempio le pagine riguardanti la politica. Altre invece ricevono aggiornamenti in modo più uniformemente distribuito nel tempo, come ad esempio gli articoli riguardanti la matematica o la biologia. Suddividono quindi le pagine esistenti da almeno tre anni in quattro differenti profili di aggiornamento: quelle aggiornate costantemente nel tempo e quelle che sono state maggiormente aggiornate rispettivamente negli ultimi nove, sei o tre mesi. Osservando la lista degli articoli di ciascun gruppo, sembrano non esserci relazioni tra i loro argomenti. Inoltre, se un evento esterno può aver scatenato un grande numero di edit, non è verificata statisticamente la possibilità che esso abbia fatto scaturire un aumento degli edit totali. La seconda parte dell'articolo presenta invece il maggiore contributo del lavoro: lo studio della struttura

dei collegamenti di Wikipedia nel tempo. La prima interessante scoperta è che il grafo di Wikipedia così studiato è una rete ad invarianza di scala[23], cioè la cui distribuzione dell'in-degree dei nodi[24] segue ha un andamento di tipo power law. Lo stesso tipo di distribuzione è osservata anche per il grafo del Web, secondo altri studi. La distribuzione dell'out-degree è invece di tipo log-normale. Entrambi questi fenomeni sono segni di maturità della rete. Gli autori notano inoltre che il la rete di Wikipedia sta diventando più densa. Infatti la media del numero di collegamenti per articolo è passato da sette a sedici nel giro di due anni e mezzo. Successivamente viene calcolato che la quantità di testo, misurata in bytes, per collegamento è aumentata. Questo dimostra che in effetti la rete sta diventando più densa non soltanto a causa dell'aumento dei suoi contenuti. Quindi viene analizzata la distribuzione del PageRank delle pagine. La distribuzione del PageRank[25] rappresenta le proprietà mesoscopiche di un grafo, in contrapposizione alla distribuzione del grado che invece rappresenta le sue proprietà microscopiche. Anche in questo il grafo di Wikipedia è molto simile al caso del World Wide Web, cioè la distribuzione del PageRank ha un andamento power law con esponente circa pari a -2. La differenza tra Wikipedia e il Web è però che tipicamente la correlazione tra PageRank e in-degree di un nodo è molto bassa per il secondo. In Wikipedia accade l'opposto, il che è un segno di maturità della rete, poiché le sue misure microscopiche tendono a quelle mesoscopiche. La correlazione nel caso di Wikipedia è aumentata nel corso del tempo, sino a raggiungere un valore quasi dell'80%. Questo significa, secondo gli autori, che la bassa correlazione dell'in-degree con il PageRank osservata per il World Wide Web potrebbe essere un fenomeno transitorio, dovuto alla sua scarsa maturità. Tra le altre tipiche misure microscopiche di una rete complessa, il coefficiente di clustering[26] e la percentuale di archi reciproci[27], Wikipedia sembra essersi stabilizzata dopo un breve periodo di assestamento. L'analisi dell'evoluzione della struttura macroscopica della rete viene fatta riferendosi al precedente studio di (Capocci et al., 2006) e alla Figura 2.12. La prima osservazione sull'evoluzione di questa struttura

[23]Nella letteratura inglese nota come scale free network.

[24]Si definisce in-degree di un nodo il totale dei suoi archi entranti. Analogamente viene definito l'out-degree di un nodo come il numero degli archi uscenti da esso.

[25]Per maggiori dettagli sulla misura del PageRank e sul suo utilizzo per studiare i grafi del Web si veda, ad esempio, (Pandurangan et al., 2002).

[26]Il coefficiente di clustering è definito come la percentuale di triple di nodi transitivi sul totale delle triple di nodi della rete. Esso viene utilizzato tipicamente per misurare la *coesione* tra i nodi di un grafo.

[27]Un arco reciproco è tracciato solamente se tra due nodi ci sono due archi di verso opposto.

è che la componente connessa sta diventando più grande in termini relativi al numero di pagine. Attualmente essa contiene il 66% circa delle pagine. Questo è consistente con le misure più recenti del Web. La percentuale dei membri instabili della componente connessa, e cioè quei nodi collegati ad essa solo tramite un collegamento entrante ed uno uscente, è però molto più bassa in Wikipedia rispetto al Web. Questo indica una connessione più compatta di Wikipedia ed è spiegata dagli autori tramite l'osservazione che la struttura enciclopedica incoraggia la connessione tra argomenti collegati nei contenuti.

2.3.6 I contenuti di Wikipedia

Già la precedente sezione lasciava intravedere che la struttura dei collegamenti aveva a che fare con i contenuti di Wikipedia. In questo senso un lavoro di collegamento tra le due sezioni è quello di (Bellomi and Bonato, 2005). Gli autori, esaminando la struttura a grafo dei collegamenti interni di Wikipedia, applicano ad essa due algoritmi molto famosi per il ranking delle pagine Web: PageRank (Page et al., 1998) e HITS (Kleinberg, 1999). Per quanto riguarda l'algoritmo HITS, è bene precisare che il lavoro considera solo la metrica di autorità. Le pagine considerate più rilevanti riguardano principalmente argomenti di geografia ed eventi storici. In secondo piano si scorgono nomi di personaggi famosi e parole d'uso comune. In realtà è l'algoritmo di PageRank che rivela le maggiori sorprese. Infatti la maggior parte delle pagine considerate maggiormente rilevanti riguarda argomenti a sfondo religioso. Lo studio sembra a questo punto richiedere una maggiore segmentazione delle pagine. Per questo motivo vengono analizzate solamente le pagine riguardanti la categoria geografica di Wikipedia. In questo caso le osservazioni riguardano il fatto che PageRank sembra trascendere dagli eventi politici recenti privilegiando una più ampia prospettiva storica culturale. PageRank sembra anche offrire una visione meno occidentale delle pagine più rilevanti[28]. Poi lo studio si sposta sugli eventi storici. Anche in questo caso HITS sembra privilegiare le pagine che trattano argomenti relativi alla cultura americana. Lo studio dei personaggi famosi evidenzia come HITS prediliga i leader politici attuali, mentre PageRank i personaggi legati alla religione, alla filosofia e alla società. Interessante notare l'assenza di personaggi femminili dalla cima di entrambe le classifiche. Lo studio dei nomi comuni non fa che confermare la "preferenza" dell'algoritmo PageRank per argomenti di tipo religioso.

Un'altra applicazione della conoscenza della struttura dei collegamenti

[28] Stiamo comunque parlando di uno studio fatto sulla versione inglese di Wikipedia.

viene sfruttata da (Adafre and de Rijke, 2005) per un obiettivo molto concreto. La piena potenzialità di Wikipedia è sfruttata utilizzando in modo opportuno i collegamenti ipertestuali. Questo tuttavia richiede un lavoro di valutazione su quali parole di un articolo necessitano di un approfondimento. Questo lavoro è svolto dagli autori dell'articolo e talvolta ci sono proprio degli amministratori che si occupano di leggere un articolo e di indentificare i collegamenti mancanti. L'approccio per l'identificazione proposto dagli autori dell'articolo è del tutto automatico, il che aiuterebbe sia gli autori che gli amministratori di Wikipedia nel loro lavoro permettendogli di concentrarsi sui contenuti. L'algoritmo proposto si divide in due passi principali. Il primo di questi consiste nel clustering delle pagine simili per argomento trattato. Esso è svolto secondo l'assunzione che pagine simili sono citate dalle stesse pagine. Per questo si rappresenta ogni pagina come l'insieme di parole che sono le voci dell'enciclopedia che attualmente citano la pagina, detto la *full representation* della pagina. Tutte questi insiemi vengono indicizzati dal motore di ricerca Lucene[29]. Si utilizza quindi come query ogni full representation per trovare la classifica delle pagine più simili per ciascuna di esse. Di questa si prendono le prime N pagine[30] e si riesegue l'indicizzazione in un nuovo motore Lucene. Il processo di ricerca delle pagine simili a partire dalla query formata dalla rappresentazione compatta di una pagina viene nuovamente eseguito ma questa volta vengono scelte come simili le pagine con un punteggio di similarità superiore ad una certa soglia[31] α. Il secondo passo consiste nell'inserire i collegamenti ad una pagina conoscendo le sue simili. L'assunzione in questo caso è quella per cui pagine simili dovrebbero avere una struttura simile di collegamenti. Quindi per ogni pagina simile a quella data si verifica se le due hanno parole in comune e si replicano gli eventuali collegamenti della seconda nella prima. Il problema di questo sistema è la valutazione, dato che non è disponibile una lista dei link mancanti. La valutazione non può essere fatta che in modo qualitativo dagli autori per un sottoinsieme di cento collegamenti trovati in tutta Wikipedia. Le conclusioni sono che il sistema è ancora perfettibile ma che il risultato ottenuto è incoraggiante, poiché il 68% dei collegamenti trovati è stato giudicato rilevante.

Il lavoro di (Holloway et al., 2005) ambisce alla generazione di una mappa di Wikipedia basata sulla co-occorrenza delle categorie negli articoli. La versione di Wikipedia inglese ai tempi dello studio, comprendeva 78977 cate-

[29]Lucene è un motore di ricerca testuale scritto in Java. Per il codice e maggiori informazioni sulla sua implementazione si consulti: `http://lucene.apache.org/`.

[30]Nell'esperimento $N = 100$.

[31]Nell'esperimento $\alpha = 0.3$.

gorie distinte organizzate in maniera semi-gerarchica, ovvero con possibilità di incontrare dei cicli o di avere alcune categorie sconnesse dalla componente principale (per un totale di 1069). Ciascun articolo può appartenere a più categorie, compresa la possibilità di appartenere o non appartenere a due categorie in relazione gerarchica tra di loro. Lo studio assume che due categorie sono simili se sono usate nello stesso articolo. Più precisamente viene utilizzata la *Cosine Similarity* per trovare un valore di similarità per ogni coppia di categorie. Coi risultati ottenuti viene quindi tracciata e disegnata una rete composta da 56609 nodi rappresentanti le categorie e 2190700 archi pesati. Purtroppo non vengono fatti studi ulteriori su questa rete, che viene solo analizzata graficamente. Lo stesso grafo viene però colorato in modo da mettere in evidenza l'ultimo aggiornamento di ciascuna categoria. Questo esperimento permette di osservare come la maggior parte di esse siano mantenute al passo con l'evoluzione dell'enciclopedia. Successivamente il grafo viene colorato in base all'autore di ogni categoria. Un utente viene considerato autore di una categoria se egli è l'ultimo ad averla modificata. Il quadro che ne emerge è quello di alcuni utenti, alcuni dei quali Bot, che si occupano di popolare le categorie di loro interesse.

2.3.7 La qualità di Wikipedia

Il problema della qualità degli articoli di Wikipedia è sempre stato molto sentito. Gli scarsissimi strumenti per verificare la validità delle informazioni presenti al suo interno la rendono, nonostante le ambizioni dei padri del progetto e degli utenti più volenterosi, più che un'enciclopedia un grande repository di informazioni più o meno aggiornate, il più delle volte poco precise o lacunose. Il processo di selezione degli articoli in vetrina è in qualche modo un tentativo di risolvere questo problema, ma non è certo privo di difetti. Prima di tutto dipende in modo critico dagli utenti che partecipano a questo processo di selezione. Solitamente si tratta di utenti di fiducia, ma nessuno può garantire che non ci sia qualche interesse nascosto o che la disattenzione di qualche amministratore porti a situazioni rischiose. Per questo i tentativi di trovare un sistema automatico di valutazione degli articoli non sono pochi.

Uno dei primi studi in questa direzione è quello di (McGuinness et al., 2006). Dopo aver descritto in linea di principio come sia possibile suddividere un articolo in frammenti caratterizzati dal proprio autore, vengono discussi sia un modo per annotare questi frammenti con informazioni di fiducia, sia come integrare queste informazioni nell'attuale interfaccia utente di Wikipedia. In particolare è interessante la soluzione trovata per quest'ul-

timo sistema, cioè la presenza di un nuovo strumento, a disposizione di un utente di Wikipedia, in grado di colorare il testo di un frammento di testo in base alla sua attendibilità. Quindi l'articolo propone due possibili metriche molto semplici per valutare automaticamente la qualità di una pagina. Esse tengono conto soltanto della struttura dei collegamenti tra pagine in quanto gli autori considerano ancora prematuro includere la reputazione degli utenti Wikipedia nel loro modello. Il primo approccio prende il nome di *Link-Ratio*. Esso si basa sull'idea che ogni qualvolta un utente decide di collegare una parola in un testo ad un articolo, implicitamente gli sta assegnando un certo grado di fiducia. Di conseguenza il grado di qualità di una pagina è calcolato come il numero di occorrenze del suo titolo che la raggiungono tramite un collegamento ipertestuale in rapporto al totale in tutta l'enciclopedia. Ad esempio, se su cento occorrenze della parola *Internet* ben ottanta recassero un collegamento alla rispettiva voce su Wikipedia, essa risulterebbe per il sistema molto migliore di un'altra per la quale ci sono solo due occorrenze in tutta l'enciclopedia ma di queste soltanto una la cita. Quest'approssimazione non è così scontata, se si pensa che in Wikipedia i collegamenti possono puntare anche a pagine non ancora scritte. Inoltre parole di uso comune, come ad esempio *Amore*, vengono collegate raramente al rispettivo articolo, a prescindere dalla sua qualità, poiché spesso non è richiesto approfondirne il significato. Ciò nonostante gli autori verificano che gli articoli in vetrina hanno un Link-Ratio medio leggermente più alto rispetto al totale (34% contro 26%). In realtà questa statistica risulta sbilanciata dal fatto che gli articoli in vetrina sono solo lo 0.1% del totale. Inoltre si evidenza anche come il Link-Ratio di parole comuni sia molto basso, in confronto ad esempio a quello di persone famose. La conclusione è che un approccio del genere assume di significato solo se usato per confrontare articoli della stessa specificità. L'altro approccio usato dagli autori è quello di usare il noto algoritmo di PageRank sfruttando la struttura di collegamenti di Wikipedia. Secondo gli autori, non essendoci correlazione tra PageRank e Link-Ratio, i due criteri si prestano bene ad essere combinati per ottenere un valore di affidabilità ben più preciso di quanto possibile usandone uno solo.

Un modello di fiducia più complesso, con i medesimi obiettivi del lavoro precedente, è stato successivamente proposto da (Zeng et al., 2006). Esso è formalizzato tramite una rete dinamica di bayesiana in cui le variabili che influenzano la qualità di una revisione sono: la qualità della revisione precedente; il grado di fiducia negli autori della revisione precedente; la quantità di testo inserito o cancellato rispetto alla revisione precedente. Per quanto riguarda la valutazione dei risultati, vengono selezionati dagli autori un totale di 868 articoli appartenenti alla categoria delle voci geografiche. Di

essi 50 sono featured e altrettanti sono di tipo *clean-up*, cioè considerati dalla comunità maggiormente bisognosi di lavori di correzione e integrazione. La fiducia degli autori è modellata con una distribuzione beta distorta su differenti valori di fiducia in base all'appartenenza dell'utente ad una delle seguenti categorie: amministratore, registrato, anonimo, bloccato. L'algoritmo per valutare le differenze tra due revision, contata alla granularità della singola parola, è invece quello di più lunga sottosequenza in comune. Dagli esperimenti fatti con questo modello sull'insieme dei dati si evince che il valore medio di fiducia per gli articoli in vetrina è più alto di quello degli articoli non classificati, a loro volta superiori a quelli clean-up. Sempre su questo modello gli autori costruiscono un classificatore addestrato con i 100 articoli appartenenti alle classi featured e clean-up. Su un insieme di test di 200 nuovi articoli la percentuale di articoli di buona qualità sul totale è pari all'82%, mentre per quelli di cattiva qualità è pari all'84%.

Un interessante estensione del lavoro di (Adler and de Alfaro, 2007) sul calcolo automatico della reputazione degli autori di Wikipedia è quello presentato in (Adler et al., 2008b). Più che nel trovare un unico valore di qualità per l'intero articolo, l'interessante idea del lavoro consiste nell'assegnare un valore di fiducia in ciascuna delle sue parole. Questo è motivato dal fatto che, essendo il tipico articolo di Wikipedia scritto da molte persone, non è detto che sia possibile considerarlo come un'unica entità informativa. Basandosi sull'algoritmo di calcolo della reputazione di un autore, anch'esso completamente automatico come già spiegato, vengono proposti un modello di base e due modelli corretti. Il primo modello calcola il valore di fiducia per ogni parola di una revision a partire da quelli della revision precedente, dalla reputazione del suo autore e dalla distanza dalla precedente versione. Il punto di partenza del sistema è la lista delle parole inserite, cancellate e spostate tra le due revision. Per quanto riguarda le nuove parole, esse prendono il valore di fiducia del loro autore. Le parole appartenenti a frammenti di testo spostati invece assumono un grado di fiducia variabile: quelle all'esterno del blocco, cioè all'inizio oppure alla fine, ricevono una fiducia molto simile a quella del nuovo testo; man mano che ci si avvicina al centro del blocco questo effetto decresce esponenzialmente in modo da far prevalere la fiducia passata. Ovviamente alle parole cancellate non deve essere assegnato un nuovo valore di fiducia. Quello ottenuto in questo modo non è ancora il valore finale delle parole nella nuova revision. Nel caso in cui l'autore della modifica abbia una reputazione più alta della fiducia di qualche parola, essa riceve un incremento proporzionale alla distanza tra le due[32]. Questo mo-

[32]Questa differenza è sicuramente positiva perché abbiamo detto che la reputazione

della il fatto che se un autore di alta reputazione mantiene del testo (anche spostandolo eventualmente), automaticamente gli conferisce una certa autorità data dalla sua approvazione. Per quanto ricco, questo modello deve ancora essere dettagliato per far fronte ad alcune particolarità di Wikipedia. In prima istanza bisogna tener conto delle parole reinserite dopo essere state cancellate in passato. Perciò si tiene traccia della reputazione di ogni *dead chunk*, ovvero un frammento di testo cancellato, decrementandola di una quantità proporzionale alla reputazione dell'autore della cancellazione. Nel momento in cui questo frammento dovesse essere reinserito, sarà quella la sua reputazione. Inoltre l'assunzione per cui un autore con alta reputazione incrementa la fiducia nel testo che modifica è vera solo in parte. In particolare la sua attenzione sarà più alta per quei frammenti di testo vicini alle modifiche che sta facendo. Per modellare anche questo aspetto si utilizza un nuovo bonus a partire dalla reputazione dell'autore della revisione, del tutto simile a quello descritto precedentemente, solo per quei paragrafi considerati vicini al suo intervento. Il sistema viene quindi studiato dal punto di vista dei possibili attacchi in grado di manometterlo. Uno dei principali risulta essere quello di atti vandalici che, riorganizzando più volte il testo di un frammento con un account dalla bassa reputazione, mirano a far perdere l'informazione di fiducia all'interno del testo. Un successivo tipo di attacco, chiamato *tampering*, consiste nel fare molti piccoli edit successivi in grado ciascuno di aumentare la fiducia del testo globale sino alla reputazione dell'autore malintenzionato. Entrambi questi attacchi vengono risolti grazie all'individuazione di queste dinamiche. La fase di valutazione dei risultati è molto importante in questo lavoro, poiché serve a tarare le costanti delle formule di aggiornamento della fiducia delle parole. L'importante risultato, ottenuto su un dataset di 1000 voci con almeno 200 revision, è quello secondo il quale il 60% delle parole identificate con bassa fiducia, viene cancellato nella revisione successiva. La valutazione viene effettuata anche senza l'utilizzo di un sistema di reputazione degli utenti che fornisce, come prevedibile, una precisione minore. Il sistema, ai tempi dell'articolo funzionante solo in modalità batch, è ora implementato come plugin[33] di MediaWiki e quindi nella sua versione online.

Sempre nella direzione di classificare un articolo di Wikipedia in modo automatico si muove il lavoro di (Rassbach et al., 2007). L'idea è quella di addestrare un classificatore a massima entropia con informazioni su articoli la cui qualità è accertata da esperti di settore. Ottenere questi dati non sarebbe semplice se non fosse che esiste un progetto, all'interno della

dell'autore è più alta della fiducia in quella parola.

[33]Disponibile con licenza BSD sul sito `http://trust.cse.ucsc.edu/`.

versione in inglese di Wikipedia, volto proprio a categorizzare gli articoli nelle seguenti categorie, in ordine di qualità decrescente: *Featured*, *A*, *Good*, *B*, *Start* e *Stub*. Di circa 600 mila articoli esaminati dalla comunità di Wikipedia ben il 71% del totale rientra nella categoria Stub, ovvero articoli molto corti e incompleti. Poiché lo studio si vuole focalizzare principalmente sulla bontà del testo come indicatore di qualità di una pagina, le feature scelte per addestrare il classificatore sono tutte estrapolate dall'ultima revisione di ciascun articolo. É per questo motivo che gli autori hanno deciso di escludere dalla fase di addestramento gli Stub, in quanto troppo poveri di caratteristiche da cogliere. Dei rimanenti il Le feature, in totale 50, sono state scelte in quattro categorie principali. Gli indicatori di dimensione riguardano misure quali il conteggio delle parole o il numero di paragrafi. L'uso delle *part-of-speech* conta l'occorrenza delle differenti parti del discorso nel parsing sintattico delle frasi. Le feature web-specifiche misurano quanto sono state utilizzate le possibilità ipertestuali all'interno dell'articolo. Infine le metriche di leggibilità sono indicatori standard[34] utilizzati per la complessità e la comprensibilità di un testo in prosa. I risultati sull'insieme di test sono misurate in base all'accuratezza normalizzata per il predittore di ciascuna classe. Il 50% delle classi è identificata correttamente. Tuttavia se, invece che utilizzare cinque categorie, ne vengono utilizzate soltanto tre (*Great* che include gli articoli Featured ed A; *Good*; *Bad* che contiene gli articoli B e Start) le performance del classificatore salgono al 69%. Lo studio si conclude suggerendo ulteriori tecniche che potranno essere esplorate in futuro per migliorare i risultati della categorizzazione. L'aggiunta di nuove features, come la misura della coesione e della coerenza del testo. L'analisi dell'evoluzione di un articolo basandosi sulla sua cronologia. L'utilizzo della conoscenza della sintassi di Wikipedia e delle pratiche di strutturazione degli articoli. Lo sfruttamento dell'analisi delle immagini all'interno di una pagina. Lo studio della struttura dei link di Wikipedia anche considerando le pagine che dall'esterno di essa citano i suoi articoli (quindi una versione del PageRank non solo locale a Wikipedia come suggerito in (McGuinness et al., 2006). Infine viene espressa la curiosità di provare nuovi modelli di classificatori, intravedendo in quelli basati su Support Vector Machines una buona possibilità di miglioramento.

La diffusione delle tecniche basate su classificatori ha visto un ennesimo tentativo di identificare le pagine di qualità in Wikipedia nei lavori di (Blumenstock, 2008a,b). Il corpus degli articoli di Wikipedia viene ridotto togliendo gli articoli più brevi di 50 parole e tutti quegli articoli non consi-

[34] Tra i quali: Kincaid, Coleman-Liau, Flesch, Fog, Lix, SMOG e Wiener Sachtextforme.

derabili delle vere e proprie voci enciclopediche (come le immagini e le liste di altri articoli). Tra questi vengono campionati casualmente 9513 articoli e selezionati tutti e 1554 gli articoli featured. Anche in questo caso le feature estratte riguardano solo la versione attuale degli articoli e la cronologia non viene presa in considerazione. Le features sono raggruppate in quattro categorie principali per un totale di più di 100. Le dimensioni superficiali sono, ad esempio, il numero di parole, il numero di caratteri e il numero di frasi. Le caratteristiche strutturali contengono, tra le altre, misure del numero di link interni o esterni, il conteggio delle immagini e delle citazioni. Gli ultimi due gruppi di features sono quelli delle metriche di leggibilità e dei tag delle part-of-speech. Le prove di classificazione sono state eseguite con differenti classificatori, dai più semplici ai più complessi, usando due terzi degli articoli per la fase di training e la parte rimanente per quella di testing, facendo in modo di avere la stessa percentuale di featured articles in entrambi gli insiemi. Il miglior risultato ottenuto è risultato essere quello dato da un modello basato su multi-layer perceptron, con un'accuratezza del 97.99%. In realtà la scoperta più interessante riguarda il fatto che anche solo con un semplice predittore che classifica un qualsiasi articolo con un numero di parole superiore a 2 mila come featured, la precisione ottenuta risulta del 96.31%. La conclusione è dunque quella che gli articoli featured non sono soltanto gli articoli migliori di Wikipedia, ma anche i più lunghi. Questo è spiegato dall'autore dall'ipotesi che il processo collaborativo di Wikipedia forza, in qualche modo, gli articoli più lunghi ad essere di qualità.

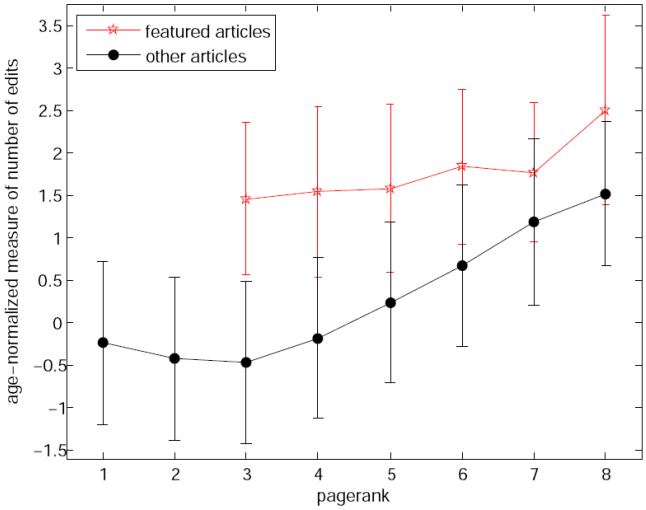
Un'ulteriore proposta di modelli per la valutazione della qualità di un articolo di Wikipedia è presentata in (Hu et al., 2007). Il lavoro comincia formalizzando tre modelli di crescente complessità. Il modello *Basic* si basa sull'assunzione che un articolo di qualità sarà scritto da autori di qualità. Dato che i due parametri si rinforzano a vicenda, possono essere calcolati con algoritmi iterativi sino alla convergenza. Il modello *PeerReview*, estensione del precedente, si basa invece sull'assunzione che ogni autore conferisce una certa fiducia a quelle parole di un articolo che non modifica nel corso del suo contributo. Autori di qualità non solo scriveranno, come nel modello Basic, ma revisioneranno anche articoli di qualità. Il limite di quest'ultimo modello salta subito all'occhio: non sempre un autore ha tempo di leggere e di esaminare tutto il contenuto di un articolo. Di conseguenza l'autorità conferita ad un testo da parte di un autore non può essere considerata uguale per tutta la sua interezza. Il modello *ProbReview* cerca di modellare quest'osservazione attraverso una funzione di probabilità che non assegna l'autorità in modo uniforme a tutto l'articolo modificato da un autore. Questa funzione deve essere decrescente e monotona in funzione della distanza

tra il testo approvato e il testo inserito. Infine viene proposto un modello *Naïve*, utile per il confronto con gli altri modelli, che semplicemente valuta la qualità di un articolo con il numero delle sue parole. Come nel caso del lavoro di (Rassbach et al., 2007), viene utilizzato l'utile insieme degli articoli della versione inglese di Wikipedia, classificati come Featured, A, Good, B, Starting e Stub. L'insieme di prova, con solo 230 articoli, è da considerarsi di piccole dimensioni. Gli autori hanno deciso di rimuovere i Bot dal conteggio della reputazione, in quanto essi non possono essere considerati autori di nuovo contenuto. Inoltre hanno collassato tutte le revisioni adiacenti dello stesso autore e rimosso le stop-words, cioè le parole di uso più comune come ad esempio gli articoli, nel calcolo della differenze tra due versioni. Per quanto riguarda gli utenti anonimi, essi vengono considerati distinti in base all'indirizzo IP dichiarato ai server di Wikipedia. La valutazione dei classificatori implementati partendo dai modelli descritti mostra risultati degni di nota. Se entrambi i modelli PeerReview e ProbReview, quest'ultimo realizzato utilizzando tre differenti funzioni di probabilità, mostrano performance superiori a quelle del modello Naïve, così non è per il modello Basic. Ciò spinge gli autori alla conclusione che può essere interessante integrare i modelli ideati con un fattore in grado di tener conto della lunghezza del testo di un articolo. Quindi i test sono rieseguiti su questo modello ibrido e tra i due modelli di partenza, PeerReview e ProbReview, solo il primo beneficia lievemente di questa integrazione. Nonostante ciò il modello ProbReview rimane quello di maggior successo.

L'ultimo lavoro che si reputa interessante citare sull'argomento della valutazione della qualità degli articoli di Wikipedia è quello di (Wilkinson and Huberman, 2007). In prima istanza lo studio cerca di valutare l'impatto del numero di edit sulla qualità di una voce. Per questo viene utilizzato come parametro di qualità la classificazione dei Featured Articles. Nel conteggio del numero di edit delle pagine Featured però, vengono rimossi quelli appartenenti alle due settimane di maggiore attività. Questo perché una pagina Featured riceve maggiore visibilità sia durante la settimana nella quale viene votata la sua classificazione, sia appena dopo la sua "promozione", periodo durante il quale essa viene esposta nella vetrina delle pagine migliori nella prima pagina di Wikipedia. La brillante osservazione degli autori è però quella secondo la quale il numero di edit ad una pagina dipende fortemente dalla sua popolarità. Una pagina di alta qualità riguardante un argomento sconosciuto ai più, potrebbe infatti ricevere un minor numero di edit rispetto ad un'altra dedicata ad un argomento molto popolare ma non per questo di qualità superiore. L'idea è quindi quella di utilizzare il notissimo algoritmo di PageRank per valutare la popolarità di una pagina. Per ciascun articolo

gli autori calcolano il numero medio di edit normalizzato in base all'età di ciascuno, in modo da non avvantaggiare gli articoli meno recenti che hanno avuto modo di ricevere un maggior numero di edit nel corso della loro storia. Infine le pagine vengono suddivise in nove classi distinte in base al loro PageRank e di ciascuna viene calcolato il numero medio di edit, come mostrato in Figura 2.13. Il medesimo studio è condotto con la misura di

Figura 2.13: Media e deviazione standard della misura del numero di edit delle pagine di Wikipedia normalizzato in base all'età e raggruppate per PageRank.



diversità, ovvero il numero distinto di utenti che hanno collaborato ad un articolo. I risultati mostrano una certa evidenza empirica della tesi che sia il numero di edit che la diversità hanno effettivamente influenza positiva sulla qualità di un articolo. In ultimo questo approccio è applicato anche ad una semplice misura di cooperazione all'interno di una pagina. Attraverso lo studio delle pagine di discussione gli autori considerano il numero di edit come una misura di interazione tra gli utenti della pagina. Il risultato di questo studio mostra, in modo ancor più evidente rispetto ai precedenti due esperimenti, come anche questo valore sia correlato alla qualità di una pagina. La conclusione degli autori è dunque quella secondo la quale sarebbe molto interessante approfondire più complesse misure di collaborazione all'interno di una pagina.

2.4 Studi sulle reti complesse e reti sociali

Nonostante siano già stati elencati dei lavori in grado di analizzare la struttura dei collegamenti all'interno di Wikipedia, nessuno di essi può essere considerata un'analisi sociometrica. Questo genere di analisi ha molto in comune con quello delle reti complesse. La sua caratteristica distintiva è però quella di avere come oggetto dello studio la rete di rapporti interpersonali all'interno di una popolazione. I primi studi degni di nota sulle reti sociali (*social network*) risalgono agli anni '60, ma solo grazie all'aumento delle capacità computazionali degli elaboratori odierni hanno potuto essere ampliati a casi reali (Scott, 2000, Wasserman and Faust, 1994). Gli studi più recenti in quest'area di ricerca molto attiva attualmente riguardano le reti di autori della comunità scientifica, a titolo di esempio si possono citare gli articoli (**?**Newman, 2001a,b, Cotta and Merelo, 2006, 2007b,a), e degli sviluppatori di software open source, si veda (Madey et al., 2002, Gao et al., 2003, von Krogh and Spaeth, 2007, Barbagallo et al., 2008). Alla luce del precedente parallelo tra Wikipedia, comunità scientifica e ancor di più con la comunità degli sviluppatori di software open source, potrebbe sembrare ovvio applicare studi di tipo sociometrico a Wikipedia. Nonostante ciò, attualmente un solo studio di questo genere è presente tra le pubblicazioni scientifiche.

Gli studi di (Korfiatis, 2006, Korfiatis et al., 2006) sono da considerarsi più che altro preliminari e prendono in considerazioni differenti possibilità di costruzione di una Rete Sociale a partire dalle informazioni di Wikipedia. La prima proposta riguarda la costruzione di una rete bipartita, dotata cioè di due tipologie di nodi. La prima classe di nodi è quella già nota degli articoli dell'enciclopedia, relazionati tra loro tramite collegamenti ipertestuali. La seconda classe di nodi è invece quella degli utenti, le cui relazioni vengono considerate per la prima volta proprio in questo articolo. La principale relazione sociale considerata è quella che associa un utente a quello che ha scritto una revision prima di lui nel medesimo articolo. Essa non è simmetrica e può essere pesata in funzione delle differenze tra le due revisioni. Tuttavia per semplicità il peso degli archi non viene considerato nello studio. Ad un livello d'astrazione più alto viene presa in considerazione la possibilità di mettere in relazione due autori nel caso in cui essi abbiano scritto in due articoli appartenenti alla medesima categoria. Questo perché il lavoro cerca di dimostrare che tanto più un autore scrive in una data sezione di Wikipedia, tanto più egli potrà essere considerato esperto di quell'area di conoscenza. Questa rete però non viene studiata su un caso reale. Viene quindi studiata una rete del primo tipo, composta da dieci pagine appartenenti alla catego-

ria degli articoli sulla filosofia e accomunati da un pari numero di edit. Per una rete di questo tipo viene quindi definita una misura di *centralità degli autori*, calcolata come il rapporto tra numero di archi connessi ad esso e il massimo numero di archi che potrebbe avere. Si noti che il massimo numero di relazioni che un autore può avere è dato dal numero totale di nodi meno uno. Questo tipo di centralità è considerata una proprietà negativa per un utente, dato che un arco entrante in un nodo rappresenta una modifica al suo contributo. Viene quindi definita la *centralizzazione di un articolo* come la distanza media tra la centralità massima di un autore e quella di tutti gli altri nello stesso articolo. L'interpretazione di questa metrica è che al diminuire del suo valore aumenta il grado di distribuzione del processo sociale nella costruzione dell'articolo. I risultati dello studio sono decisamente limitati dallo scarso numero di pagine considerate e per questo le considerazioni risultano essere solamente di tipo qualitativo.

Capitolo 3

Il processo di analisi di Wikipedia

3.1 Requisiti

Nel precedente capitolo si è visto come gli studi su Wikipedia possano prendere moltissime direzioni. Questo è possibile grazie all'enorme quantità di dati resi disponibili dagli archivi[1] di WikiMedia Foundation che, per dare ancora maggior trasparenza al suo servizio, pubblica a intervalli di tempo regolari i registri (log) di tutta la cronologia per ciascuna versione dell'enciclopedia. Il primo problema da affrontare risulta quindi quello della gestione e dell'estrazione delle informazioni d'interesse da questi log che possono raggiungere dimensioni considerevoli[2]. Questo fattore ha un certo impatto sul processo di analisi che deve dunque avere dei requisiti quanto più possibile definiti prima del suo inizio.

Innanzitutto è necessario chiarire l'**oggetto dell'analisi**. In questo studio si è scelto di analizzare la comunità di utenti di Wikipedia. Il motivo di questa scelta si ritrova nella semplice considerazione che il più grande fattore in grado di influenzare le sorti di un wiki è la comunità. La principale differenza di questa recente tecnologia con una qualsiasi risorsa informativa nel Web è infatti il continuo e simultaneo lavoro di revisione a cui il contenuto è sottoposto. Nel caso di Wikipedia si tratta di migliaia di persone accomunate solamente dalla voglia di contribuire al progetto, anche in maniera molto differente tra di loro.

Da un lato lo studio è fondamentale in quanto è in grado di estrarre

[1] Accessibili tramite Web all'url `http://download.wikipedia.org` .

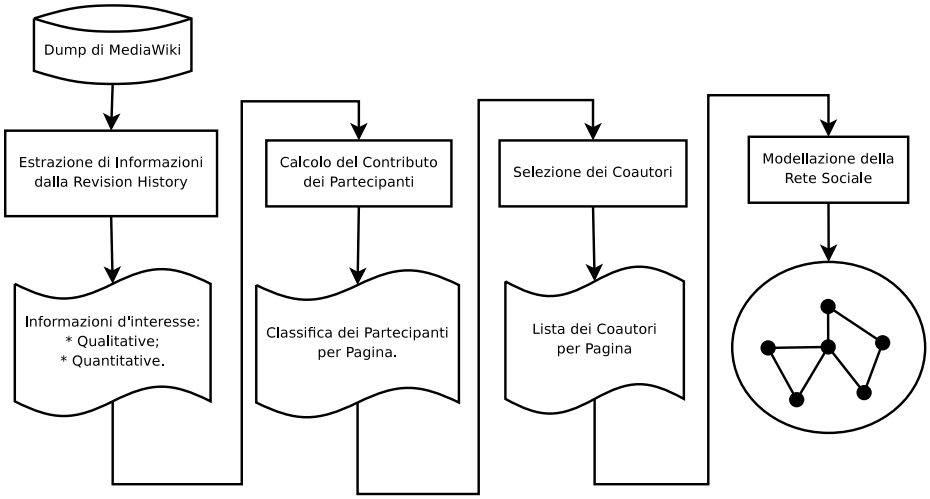
[2] A titolo d'esempio si pensi che il file di log della versione italiana di Wikipedia ad oggi occupa uno spazio su disco, in formato xml non compresso, di 222.3GB.

caratteristiche proprie di Wikipedia. Essa è infatti da considerarsi uno dei principali strumenti di diffusione dell'informazione nel Web. Si tratta di uno dei suoi siti più visitati e spesso i risultati dei principali motori di ricerca indirizzano ad esso.

Da un altro lato questo studio vuole considerarsi di **valenza più generale**. La tecnologia wiki sta assumendo sempre più importanza anche per progetti che non riguardano la scrittura di un'enciclopedia. La sua semplicità sembra essere adatta in quei casi in cui risulta utile catturare la conoscenza di un gruppo di persone accomunate dal medesimo obiettivo, ad esempio la conoscenza implicita di un'azienda. Si ritiene dunque che questo studio possa essere utilizzato per l'analisi della comunità di un qualunque wiki al fine di comprendere meglio le dinamiche tra i suoi attori. Questo è possibile grazie al fatto che la piattaforma di creazione di un wiki più diffusa è proprio quella di MediaWiki, usata da Wikipedia stessa. Lo schema della base di dati di MediaWiki è ben documentato ed esistono svariati strumenti software per analizzare i dati strutturati con esso.

Il processo di analisi ha un ulteriore requisito di **modularità**, dettato invece da questioni più pratiche. Come si avrà modo di approfondire nelle prossime sezioni, il processo di estrazione è sensibile a diversi parametri. Per avere l'opportunità di modificarli senza dover necessariamente rieseguire gran parte degli stessi calcoli è stato quindi adottato l'approccio di progettare diversi moduli software, ciascuno in grado di elaborare dei dati in ingresso producendo un risultato utile alla fase successiva. Il processo di analisi può essere dunque scomposto in quattro sottoprocessi intermedi, come schematizzato in Figura 3.1. Il primo si occupa, prendendo come input un log di Wikipedia, di estrarre le informazioni rilevanti dalla revision history riguardanti le singole revisioni. Il secondo, partendo dai risultati appena prodotti, calcola il contributo di ciascun utente per ogni pagina, secondo diverse metriche che verranno approfondite nella sezione relativa. Il terzo seleziona quegli utenti che hanno maggiormente contribuito allo sviluppo di ciascuna pagina. In ultimo, il quarto sottoprocesso modella questi dati in una Rete Sociale di utenti in grado di evidenziare le dinamiche interpersonali nella comunità di Wikipedia. Un ulteriore vantaggio di questo sistema riguarda il fatto che i risultati intermedi potranno essere a loro volta interpretati al fine di estrarre da essi nuova conoscenza su Wikipedia. Verranno ora approfondite le motivazioni e l'approccio seguito nella realizzazione di ciascuno dei sottoprocessi.

Figura 3.1: Schema dei processi in cui è scomposta l'analisi di Wikipedia.



3.2 Estrazione di informazioni dalla cronologia di un wiki

Il primo processo è esplicitamente finalizzato al successivo di calcolo del contributo dei partecipanti di un wiki. Questo significa che i dati da esso calcolati sono probabilmente i meno interessanti da analizzare indipendentemente. Tuttavia questo processo è fondamentale in quanto base su cui poggiano tutti gli studi successivi.

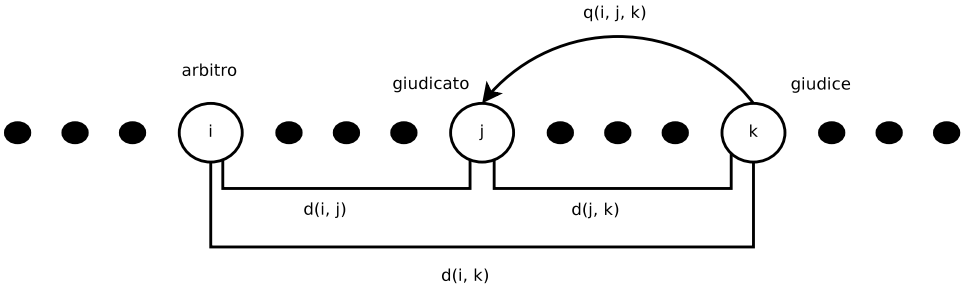
Per meglio comprendere il processo e i dati da esso calcolati è necessario introdurre il concetto di *qualità di una revisione* come proposto da (Adler and de Alfaro, 2007). Si è già avuto modo di vedere la complessità del problema di valutazione automatica di un articolo di Wikipedia. Questo approccio non ambisce in prima battuta a risolvere il suddetto problema, bensì definisce una metrica per stimare l'opinione di un autore di una certa revision v_k, chiamato *giudice*, sulla qualità di una revision v_j di un altro autore, *giudicato*, basandosi su una sua revisione precedente v_i come punto di riferimento. Supponendo di saper calcolare tramite una funzione $d(v, v\prime)$ le differenze tra due versioni dello stesso testo, si potrebbe interpretare questo valore come la *fatica* che un autore deve compiere per trasformare la versione più vecchia in quella più recente. Partendo da questo presupposto l'idea è quella di valutare quanto la versione v_j ha aiutato l'autore della versione v_k nel compito di rendere la versione v_i uguale alla sua. In un certo senso è come se l'obiettivo dell'autore di v_k fosse quello di giungere alla sua versione

ed egli valutasse l'autore di v_j in funzione di quanto questo ha reso v_i più vicina al suo scopo. Di conseguenza l'aiuto dato da v_j all'autore di v_k risulta, in termini assoluti, pari a $d(v_i, v_k) - d(v_j, v_k)$. In termini relativi alla fatica svolta dall'autore di v_j, la formula che esprime il parere di qualità dell'autore di v_k nei confronti di v_j diventa:

$$q(v_i, v_j, v_k) = \frac{d(v_i, v_k) - d(v_j, v_k)}{d(v_i, v_j)}$$

Per una rappresentazione grafica delle quantità utilizzate dalla formula si faccia riferimento alla Figura 3.2. Si presti attenzione al fatto che non ci so-

Figura 3.2: Voto della qualità della revisione j, da parte di k usando come arbitro i.



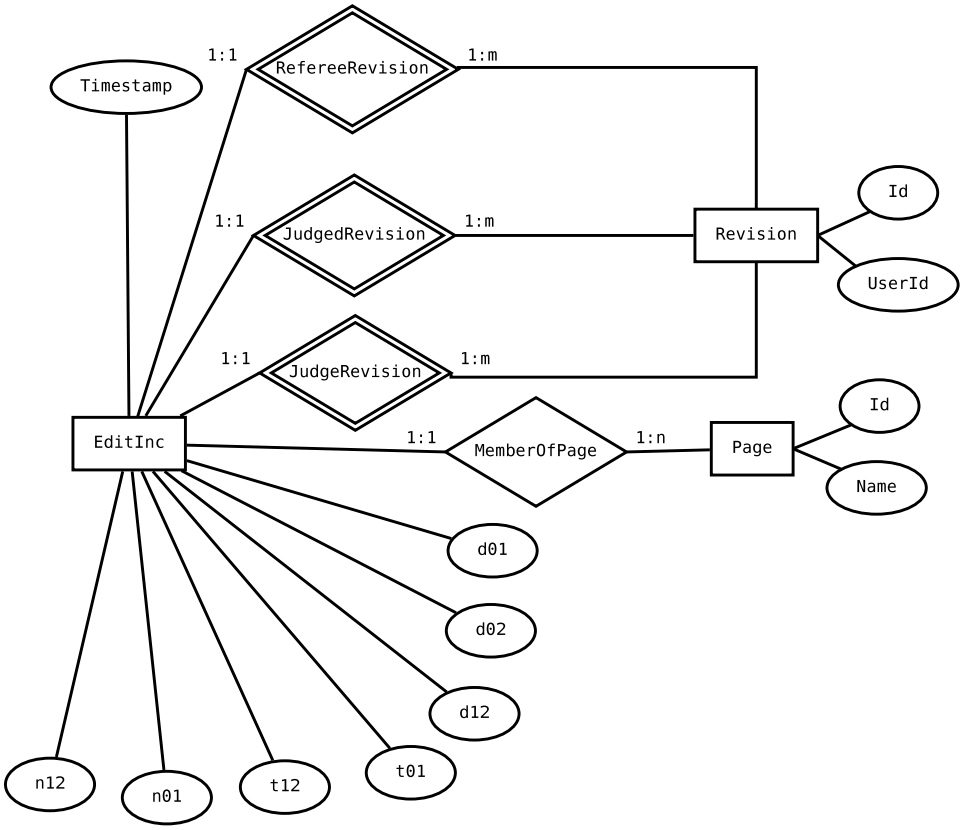
no vincoli sulla similarità tra revisioni. Una data revisione può essere molto più simile a un'altra a essa molto precedente, al limite uguale nel caso di revert, piuttosto che a una molto vicina. Se questa, utilizzando come punto di riferimento la revisione lontana, esprime un parere su quella più vicina, il numeratore della formula di qualità sarà negativo. Si capisce quindi come il valore di $q(v_i, v_j, v_k)$ possa assumere valori all'interno dell'intervallo reale $[-1; +1]$. Nel caso in cui $q(v_i, v_j, v_k) = +1$, v_k avrà mantenuto completamente il contenuto inserito da v_j. Nel caso opposto, cioè quello in cui $q(v_i, v_j, v_k) = -1$, la fatica dell'autore di v_j sarà considerata vana, poiché v_k ha annullato tutti i suoi cambiamenti fatti rispetto a v_i. I casi intermedi esprimeranno sfumature di questi concetti. Un aspetto molto importante di questo calcolo di qualità è che, fissata una revisione, esso è in grado di catturare i pareri di diverse persone (seppure impliciti) basandosi su differenti punti di riferimento. In un ambiente aperto come quello di Wikipedia è indispensabile non accettare mai l'opinione di un'unica voce. Si pensi ad esempio a un conto di qualità in cui vengono considerate solamente le triple del tipo: (v_{i-1}, v_i, v_{i+1}). Se la versione v_{i+1} fosse di tipo vandalico e quindi cancellasse completamente v_i, gli assegnerebbe implicitamente un

voto negativo. Questo perché la differenza tra le due versioni v_i e v_k risulterebbe minore di quella tra v_j e v_k e quindi il numeratore della formula di qualità sarebbe negativo. Se ora si considerano anche le successive versioni $v_{i+2}, v_{i+3}, \cdots$ si può supporre che una qualche versione successiva ripristinerà v_i e da quel momento in avanti essa riprenderà ad avere voti positivi. Tanto più la versione v_i verrà mantenuta negli edit successivi, tanto più essa riceverà voti positivi. In realtà è interessante anche conservare il parere espresso da v_{i+1} in quanto, ammettendo che non sia facile stabilire in modo univoco se si tratta di un vandalo, rappresenterebbe comunque la voce di un'opinione contraria alla comunità. Si vuole ora mettere in luce anche un limite di questa metrica. La prima revision, non avendo precedenti versioni, non può ricevere voti. Allo stesso modo la sua successiva ha una sola revisione che può essere usata come punto di riferimento e quindi non gode dei vantaggi dati da un sistema che usa come punto di forza la pluralità di voti. Così vale per le revision immediatamente successive alla prima che, in linea di principio, possono ricevere meno voti rispetto alle versioni successive a regime. Facendo la considerazione simmetrica, anche l'ultima versione non può ricevere voti, banalmente perché per essa non può esistere alcun giudice. La penultima versione riceverà voti solamente dal suo unico successore e così via. Si può quindi concludere che, se a regime il sistema può considerarsi equilibrato, per pagine con poche versioni esso può portare più facilmente a degli errori di valutazione.

I dati calcolati nel processo, per la maggior parte progettati per essere utilizzati nel calcolo della qualità di una revisione, sono riconducibili a due entità principali che prendono il nome di *EditInc* e di *EditLife*. Esse verranno descritte attraverso uno schema di tipo Entità-Relazione per meglio comprenderne la struttura.

L'entità di tipo EditInc, come mostrata in figura 3.3, è identificata da una tripla di revisioni che rappresentano quella arbitro (*Referee Revision*), quella giudicata (*Judged Revision*) e quella giudice (*Judge Revision*). Per ciascuna di esse i dati a disposizione sono l'*identificativo della revisione* e l'*identificativo dell'utente* suo autore, entrambi univoci e assegnati da MediaWiki. Ciascuna revisione della tripla appartiene al medesimo articolo dell'enciclopedia, il quale viene indicato dalla relazione *Member of Page*. Le triple di revisioni, tra tutte le possibili all'interno dell'insieme di quelle di Wikipedia, vengono scelte secondo il seguente criterio. Chiamato $\mathbb{R} \equiv \{v_0, v_1, \cdots, v_M\}$ l'insieme di tutte le M revision di una data pagina, ogni EditInc esprime il rapporto fra tre di esse (v_i, v_j, v_k) tali per cui $0 < i < j < k < m$. Si noti come, rispetto alla precedente spiegazione generale, sia stata introdotta una finestra basata sul numero di revisioni.

Figura 3.3: Entità EditInc.



La revision v_j verrà valutata solo dalle sue m successive, le quali potranno usare come riferimento una revisione antecedente non più di m posizioni. Questa finestra è utile, oltre che per alleggerire il processo di calcolo, anche per limitare il numero di giudizi che può ricevere una revision. Infatti all'aumentare della distanza tra due revision il voto dato perde di significato poiché è normale, nell'evoluzione della pagina di un wiki, che un testo venga stravolto col passare del tempo anche se al momento della sua creazione esso è stato considerato di qualità. Un ulteriore vincolo nella selezione delle triple è che l'autore di v_k dev'essere differente dall'autore di v_j, poiché è scorretto permettere ad un utente di valutare la propria revision con una nuova. In questo caso l'insieme dei giudici non è ridotto di cardinalità, ma la finestra viene allargata in modo tale da cercare di ottenere sempre il massimo numero di voti possibile. In particolare le informazioni estratte dal log riguardano la *distanza di edit*, la *distanza di numero di versioni* e la *distanza temporale* fra le tre revisioni.

La distanza di edit $d(v, v')$ tra due versioni è calcolata attraverso un algoritmo greedy in grado di stimare una lista di elementi che descrivono come v' può essere ottenuta a partire da v. Da questa lista viene calcolata una quantità di parole inserite I_{TOT}, una di parole cancellate D_{TOT} e una di parole spostate di posizione M_{TOT}. Il calcolo della distanza di edit tra due revisioni è quindi dato dalla formula:

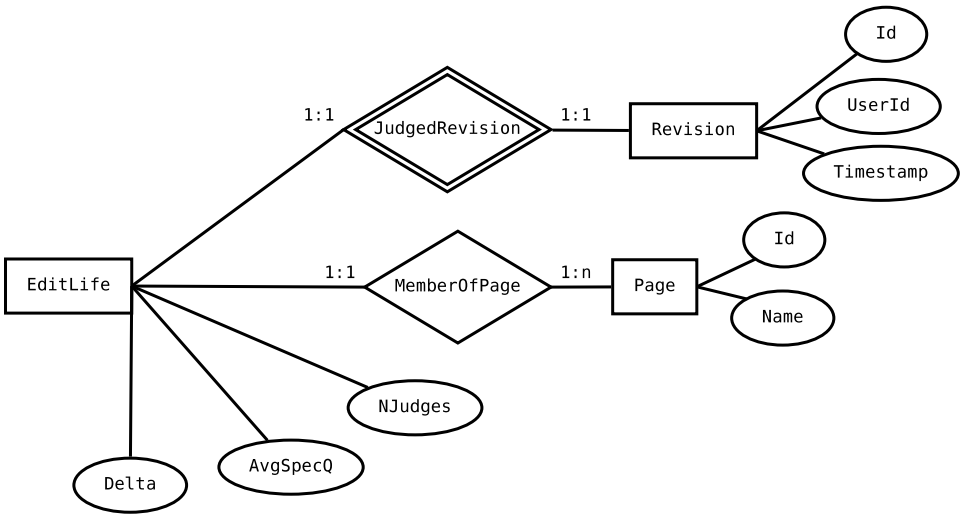
$$d(v, v\prime) = I_{TOT} + D_{TOT} + M_{TOT} - \frac{1}{2} min(I_{TOT}, D_{TOT})$$

Mentre il numero di parole spostate contribuisce pienamente alla distanza tra due revisioni, così non è per le parole inserite e cancellate. Ciò è motivato dal fatto che spesso le parole cancellate hanno un rapporto con quelle inserite. Si consideri ad esempio una revision v_{i-1}, alla quale un autore a_i aggiunge un paragrafo ottenendo così la versione v_i. Si ipotizzi inoltre che il numero di parole del nuovo paragrafo sia pari a k. Se a questo punto un nuovo autore a_{i+1} modificasse completamente il paragrafo inserito da a_i, mantenendo però lo stesso numero di parole, la somma delle modifiche tra v_i e v_{i+1} risulterebbe pari a $2k$, mentre quella tra v_{i-1} e v_{i+1} sarebbe k. Con la definizione di $d(v, v\prime)$ definita precedentemente, la distanza tra v_i e v_{i+1} viene scontata in modo tale che una sostituzione sia considerata equivalente solamente a metà del cambiamento. In questo caso $d(v_i, v_{i+1}) = \frac{k}{2}$ mentre $d(v_{i-1}, v_{i+1}) = k$. Si ritiene che questo comportamento sia in grado di catturare in modo migliore la comune percezione del contributo di un autore nella maggior parte dei casi. Tuttavia, trattandosi di un euristica, non è possibile garantire che valga la diseguaglianza triangolare tale

per cui: $d(v_{i-1}, v_i) + d(v_i, v_{i+1}) \leq d(v_{i-1}, v_{i+1})$. É comunque stato dimostrato da (Adler and de Alfaro, 2007) che essa vale per oltre il 99% delle triple[3]. Data un'entità di tipo EditInc identificata da una tripla di revisioni (v_i, v_j, v_k), i dati *d01*, *d02* e *d12* contengono rispettivamente i valori calcolati per $d(v_i, v_j)$, $d(v_i, v_k)$ e $d(v_j, v_k)$. Vi sono infine degli attributi di supporto come il *Timestamp*, che indica l'istante di tempo in cui è stata inserita la revisione v_k, e come *t01* e *t12* che misurano la differenza temporale rispettivamente tra la revision v_i e v_j e tra v_j e v_k. Questi valori sono calcolabili in modo esatto e non con euristiche come nel caso delle distanze di edit, dunque ci si può aspettare che l'istante di creazione della revision v_i, ad esempio, sia calcolabile come: $Timestamp_i = Timestamp - t01 - t12$. Allo stesso modo gli attributi *n01* e *n12* misurano il numero di versioni che intercorrono rispettivamente tra v_i e v_j e tra v_j e v_k, che si ricordano essere variabili entro una finestra di dimensioni fissate dal parametro m.

Un'entità di tipo EditLife, schematizzata nella Figura 3.4, è ottenuta semplicemente aggregando le informazioni di tutte le entità EditInc con la medesima revisione giudicata (*JudgedRevision*). Tra esse però vengono scel-

Figura 3.4: Entità EditLife.



te quelle che valutano una revision in base solo alla sua precedente. Questo vincolo alleggerisce il processo di calcolo e non mostra particolari perdite di precisione. C'è quindi una e una sola EditLife per ogni JudgedRevision, a sua volta composta da identificativo di revisione (*Id*), identificativo del

[3]I dati sono stati ottenuti sulla versione in lingua italiana di Wikipedia del 2005.

suo autore (*UserId*) e istante di inserimento nell'enciclopedia (*Timestamp*). Ogni EditLife è relativa a una e una sola pagina la quale ha, in generale, almeno una revision. L'attributo *NJudges* esprime il numero di revision per le quali è stato possibile calcolare un voto implicito. Il valor medio dei voti di qualità è salvato nell'attributo *AvgSpecQ* ed è quindi calcolato, per ogni revision v_j e su tutte le triple (v_{j-1}, v_j, v_k) tali per cui $0 < j-1 < j < k < m$, nel seguente modo:

$$AvgSpecQ(v_j) = \frac{\sum_j q(v_{j-1}, v_j, v_k)}{NJudges}$$

L'attributo NJudges è di particolare interesse. Sebbene a regime possa essere considerato costante e pari ad $m + 1$, bisogna ricordare che le ultime m versioni riceveranno meno voti rispetto alle altre in quanto dotate di un minor numero di revisioni successive.

I dati appena descritti possono essere estratti dal log di Wikipedia dal software realizzato dai membri del progetto WikiTrust, come descritto negli articoli (Adler et al., 2008a, Adler and de Alfaro, 2007). Esso è disponibile liberamente e ampiamente documentato[4].

3.3 Calcolo del contributo dei partecipanti

Il problema del calcolo del contributo di un utente di Wikipedia è stato poco affrontato sino ad oggi. La maggiore difficoltà consiste nel trovare un modello che ben interpreti il comportamento dei singoli attori in gioco. Questo è complicato dal fatto che su Wikipedia agiscono contemporaneamente attori molto attivi e occasionali, non necessariamente umani e potenzialmente anonimi. Un'altra importante osservazione sugli utenti di Wikipedia è che non è sempre facile distinguere le loro intenzioni. Non solo per quanto riguarda gli atti di vandalismo, si pensi anche alle differenti opinioni che possono scaturire da argomenti controversi. Poiché l'oggetto di questo studio è proprio la comunità di Wikipedia è importantissimo esplorare diverse metriche in grado di cogliere l'entità, e in qualche modo l'utilità, dei comportamenti dei suoi partecipanti. Si è ritenuto opportuno concentrarsi sul calcolo dei contributi di ciascun utente per ogni singola pagina, poiché essa è da considerarsi l'elemento atomico della conoscenza contenuta nell'enciclopedia. Ciò nonostante potrà essere utile soffermarsi anche su misure di tipo globale, in grado di dare una visione d'insieme dello stato dell'enciclopedia.

[4]Si faccia riferimento al sito ufficiale del progetto: `http://trust.cse.ucsc.edu/` .

3.3.1 Metriche

Il conteggio, la quantità e la qualità degli interventi

Il *conteggio degli interventi* (edit count) è una delle metriche più usate per la stima del contributo di un utente. L'assunzione è che tanti più interventi avrà fatto un utente all'interno di una pagina, di una categoria o dell'intera Wikipedia, tanto più egli avrà contribuito al suo sviluppo. Partendo dall'osservazione del fatto che spesso più edit consecutivi vengono utilizzati per continuare un lavoro di inserimento dei contenuti interrotto, si è deciso di collassarli in un unico intervento equivalente alla somma di quelli adiacenti dello stesso autore. Anche in questo caso però, come già spiegato, questa misura può essere considerata molto poco precisa perché non può dire nulla nè sulla quantità nè sulla qualità degli interventi. Il numero di edit in se è molto più utile per misurare l'attività di un dato utente in un wiki, piuttosto che il suo contributo. Esso è interessante da studiare anche in questo lavoro, oltre che per la facilità di conteggio, soprattutto per l'attenzione che esso riceve da parte dei membri della comunità di Wikipedia.

Si può già iniziare a capire come la considerazione dell'*aspetto quantitativo* sia il prossimo passo importante nel calcolo dei contributi, perché aiuta a discriminare chi effettua solo piccole correzioni ai contenuti da chi invece fa interventi più sostanziali. Il motivo per cui questo aspetto non è stato ancora sposato dalla comunità di Wikipedia è da ricercarsi nella difficoltà di definire in modo preciso il concetto di quantità di contributo. Esso si potrebbe semplicemente identificare con le parole inserite da una revisione rispetto alla sua precedente. Eppure in questo modo si perderebbe la stima del lavoro di tutte quegli utenti che svolgono altri compiti di pari valore. Ad esempio un lavoro di riorganizzazione della struttura dei paragrafi di un articolo può cambiare pochissime parole ma allo stesso tempo migliorarne notevolmente la fruibilità. Allo stesso modo un lavoro di snellimento di un articolo delle sue parti considerate ridondanti non aumenta la sua dimensione, anzi la riduce. Ecco perché come misura quantitativa si ritiene molto valida quella di differenza di edit tra due versioni, come spiegata nella sezione 3.2. Essa è in grado di tener conto degli inserimenti, delle cancellazioni, delle sostituzioni e degli spostamenti di contenuto, tutti potenziali indicatori del contributo di una revisione rispetto ad un'altra. In particolare l'aspetto quantitativo di un contributo potrà essere valutato rifacendosi alle revisioni passate.

L'*aspetto qualitativo* entra in gioco proprio a supporto dell'aspetto quantitativo. É scorretto considerare solo la dimensione di un contributo, poiché esso potrebbe essere di scarso valore. Il caso di un atto vandalico è ancora

una volta un buon esempio, ma lo stesso vale anche per un inserimento non volutamente di scarsa qualità o approssimativo. Un buon modo automatico per valutare la qualità di un contributo è quello spiegato nella sezione 3.2. Esso fornisce un insieme di opinioni da parte di utenti distinti sulla bontà di una revisione. É quindi basato sulla stima di quanto di una revisione verrà mantenuta nelle sue successive. Anche considerare solamente la qualità come unico parametro per valutare un contributo è da considerarsi scorretto, poiché da sola essa privilegia chi fa piccole modifiche sempre accettate dalla comunità. Un tipico esempio è quello di un correttore ortografico, compito che potrebbe essere tranquillamente svolto da un Bot, che pur non svolgendo modifiche di tipo concettuale può ritrovarsi nella situazione di averne fatte molte di ottima qualità. Ecco perché la soluzione più interessante risulta quella che combina le due metriche al fine di ottenere un indicatore di contributo più completo. Nella sezione successiva si prenderanno in esame due possibili metriche di contributo di un utente, entrambe basate su differenti modi di interpretare la quantità e la qualità dell'apporto di una revisione. Entrambe le metriche di partenza sono relative alla singola revisione, ma risulta banale estenderle alla valutazione del contributo di un autore, essendo sempre possibile attribuire a una revisione il suo autore. L'eccezione a questa regola è rappresentata dagli utenti anonimi. Sebbene essi, per ragioni di sicurezza, siano comunque registrati nel database di MediaWiki con il loro indirizzo IP, non è così semplice individuare in esso l'identità. Per molti motivi che non è il caso di approfondire in questa sede, l'indirizzo IP di un host può non essere associato univocamente a un unico individuo. Quindi non ha particolarmente senso calcolare il contributo di un utente anonimo se non come valore aggregato. In questo caso si potrà valutare l'impatto di tutti gli utenti anonimi in una singola pagina oppure nell'intera Wikipedia.

La longevità di un intervento

La *longevità di un intervento* (Edit Longevity) è stata introdotta per la prima volta nell'articolo (Adler et al., 2008a) con l'intenzione di calcolare il contributo degli utenti in tutta Wikipedia. Il problema di valutare la quantità di contributo di una revisione è risolto nel modo più immediato possibile. Viene definita una funzione della revisione, chiamata *contributo dell'intervento*, come la differenza di edit tra essa e la sua precedente. Più formalmente:

$$d(r_i) = d(r_{i-1}, r_i)$$

Uno dei casi in cui questa misura, sempre da considerarsi un'euristica, evidenzia i suoi limiti è quello rappresentato dal seguente scenario. Si immagini

che un vandalo cancelli completamente il contenuto di una pagina e un utente si accorga di questo danno. Grazie al software MediaWiki egli potrà, con una semplice operazione di revert, annullare il contributo del vandalo. In questo caso la differenza del suo intervento con la revisione precedente sarà tanto più alta quanto è di grandi dimensioni il testo che egli ha ripristinato. Non si vuole certo sminuire l'importanza di un intervento considerato importante per la qualità del wiki, tuttavia è bene tenere conto di come il presentarsi di episodi di questo tipo potrà influenzare la misura. Per quanto riguarda invece la valutazione della qualità di un contributo, la scelta ricade sull'aggregare tutte le opinioni espresse dagli autori delle m revisioni successive a quella data basandosi solo sul contributo precedente (cioè lo stesso usato per calcolare il contributo dell'intervento). In questo caso si considera la *qualità media dell'intervento* di un contributo la quantità espressa dall'attributo $AvgSpecQ$ nell'entità $EditLife$. L'Edit Longevity è quindi semplicemente ricavata dalla sommatoria per tutte le revisioni di un autore del prodotto tra il contributo e la qualità media dell'intervento. Detto $\mathbb{A}$ l'insieme degli utenti di Wikipedia, $\mathbb{P}$ l'insieme delle pagine ed E la funzione che dati un autore e una pagina restituisce l'insieme delle revision di quell'autore in quella pagina, si può calcolare la longevità dell'edit per ogni autore nel seguente modo:

$$\forall a \in \mathbb{A} : EditLongevity(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} AvgSpecQ(r) \cdot d(r)$$

Nonostante questa misura sia stata introdotta per misurare il contributo degli utenti in tutta Wikipedia, la sua estensione locale, e cioè alla singola pagina, è banale. In realtà si può notare una differenza tra versione locale e globale della misura. Si ricordi infatti che due revision per ciascuna pagina non possono ricevere un voto di qualità. Esse sono la prima revisione, la quale non avendo versioni precedenti con cui essere confrontata non può ricevere voti, e l'ultima, che non ha versioni successive che esprimano voti su di lei. Se a livello globale questo effetto può considerarsi lieve, all'interno della singola pagina, specialmente se essa è dotata di poche revisioni, possono essere persi due contributi fondamentali: il primo e l'ultimo.

Il vantaggio di questa metrica è che essa è molto semplice da calcolare a partire dai dati della prima fase, in particolare da quelli dell'entità EditLife. Essa privilegia le grandi modifiche che durano nel tempo e può essere misurata in numero di parole, essendo derivata dal prodotto di un numero di parole per una quantità adimensionale.

La longevità di un intervento valutata rispetto alla sua versione più simile

Verrà a questo punto descritta una nuova metrica di valutazione del contributo di un autore chiamata *longevità di un intervento valutata rispetto alla sua versione più simile*. Essa parte dalla medesima idea su cui si poggia la misura di longevità di un intervento, descritta nel paragrafo precedente, ma la estende in tre principali direzioni.

La prima prova a limitare uno dei problemi evidenziati, per il quale l'utente che compie un revert beneficia di un incremento di contributo proporzionale alle dimensioni del testo ripristinato, pur non avendone effettivamente il pieno merito. La considerazione iniziale per individuare una situazione di questo genere partendo dai dati è quella per cui se un utente effettua un ripristino, ci sarà una versione a lui precedente con distanza di edit nulla. In realtà la situazione può essere ulteriormente complicata nel momento in cui il responsabile di un revert non si limiti a quest'azione, ma modifichi a sua volta il contenuto della versione ripristinata, anche in piccola parte. L'evento di questo scenario potrebbe forse sembrare troppo raro perché ce ne si preoccupi, visto che l'azione di revert è il più delle volte svolta in modo atomico tramite l'interfaccia di MediaWiki. In realtà, per i motivi spiegati precedentemente, più interventi consecutivi del medesimo autore vengono collassati in un unico intervento e perciò si è ritenuto importante tenerne conto.

Si vuole quindi mettere in evidenza la differenza di edit con una particolare versione precedente a quella considerata, chiamata la *versione più simile*. In uno scenario ideale, dove cioè tutti gli utenti sono d'accordo e aggiungono sempre migliorie alla versione attuale di una pagina, la versione più simile per ciascuna revision è sempre la precedente. Nella realtà, dove ci sono sia casi di vandalismo che di divergenze di opinione, la versione più simile a una data revsion può essere un'altra tra le precedenti. Oltre a utilizzare questa tecnica per l'individuazione dei revert, si ritiene opportuno riformulare la misura di longevità di un intervento verso una nuova direzione.

Data una revisione, la sua versione più simile tra le precedenti, rappresenta il punto di partenza del contributo del suo autore. In generale egli avrà coscienza di tutte le revisioni successive e sarà da ritenersi responsabile della loro modifica. Si può chiamare questo gruppo come l'*insieme delle versioni rilevanti* per valutare una revision. Per quanto riguarda le differenze con le versioni precedenti rispetto a quella più simile, si reputa invece opportuno considerare che l'autore della revision non abbia percezione di

esse e pertanto non venga giudicato rispetto a loro.

Ecco quindi due differenze della nuova metrica rispetto alla *longevità di un intervento*. La prima riguarda il calcolo del contributo di una revision, non più calcolato come la distanza di edit con la versione precedente, bensì come la minima distanza di edit rispetto ad una tra quelle precedenti. Anche in questo caso si è scelto di utilizzare una finestra di m edit entro la quale limitare la ricerca della versione più simile. Questo, oltre che per i soliti motivi di performance, risulta sensato perché trattandosi comunque di un'euristica, si reputa troppo poco probabile che un autore prenda spunto da una versione molto lontana dalla sua in numero di edit, nonostante questa possa avere una distanza di edit con la sua minima rispetto alle altre. La seconda differenza, derivante direttamente dalla prima, è che con l'assunzione di considerare l'autore di una revision come responsabile della modifica di tutte le revision comprese tra la sua e quella a lui più simile tra le precedenti, assume un valore ben preciso, ai fini della valutazione della revision, il confronto con queste versioni intermedie. Quindi i pareri degli autori delle revision successive su una data versione prenderanno come punto di riferimento (o arbitro) non solo la sua precedente, ma anche tutte quelle considerate rilevanti. Detta v_j la versione analizzata e v_n la sua versione più simile nella finestra considerata, si definisce quindi una distanza da essa come:

$$d_M(v_j) = min\left[d(v_j, v_{j-1}), \cdots, \boldsymbol{d(v_j, v_n)}, \cdots, d(v_j, v_{j-m})\right] = d(v_j, v_n)$$

L'insieme delle valutazioni rilevanti per v_j sarà invece definito come:

$$Q_{v_j} \equiv \{q(v_i, v_j, v_k) | i \geq n \wedge k < j + m\}$$

La prima osservazione da fare a questo proposito è che, rispetto al precedente criterio, ciascuna revision si ritroverà più voti dalla medesima revisione giudice. In particolare è interessante considerare due versioni giudici agli estremi della finestra. La versione successiva a v_i potrebbe, in linea di principio, assegnare m voti, basandosi sulle m revision precedenti a v_i. In realtà questi voti saranno limitati dalla distanza in numero di edit tra v_i e la versione più simile v_n. L'ultima revision prima del termine della finestra, v_{i+m}, esprimerà invece sempre e comunque un unico parere basandosi solo sulla versione v_{i-1}. Si noti come, nel caso particolare in cui la versione più simile a v_i sia la sua precedente, l'insieme dei voti ricevuti per revision nelle due metriche di valutazione del contributo coincidano. Se si adottasse, come nel caso della longevità di un intervento, la semplice media dei voti assegnati per stimare la qualità di una revisione, essa sarebbe sbilanciata a favore di

quelle revisioni che hanno potuto esprimere più pareri, che sicuramente sono le più vicine in numero di edit alla versione valutata. Questo è proprio in contrasto con l'obiettivo della metrica, poiché il voto di una versione distante sembrerebbe avere maggiore credito in quanto il suo autore ha potuto meglio osservare l'evoluzione della pagina. D'altro canto sono le versioni più vicine ad intervenire maggiormente sulla revision in analisi. Si preferisce quindi dare pari peso ai voti espressi da tutte le versioni successive.

Prima di spiegare come è stata risolta questa problematica, si rifletta sulla seguente considerazione. Se le versioni giudici sono tutte scritte da autori distinti, esse rappresentano tanti pareri indipendenti. Se però, come può normalmente accadere in un qualsiasi wiki, tra le m revisioni giudici ce n'è più di una dello stesso autore, è presumibile che questo avrà mantenuto la sua idea sullo sviluppo della pagina e che quindi il sui voti saranno non del tutto indipendenti. Per questo si è deciso di ottenere un parere medio di ciascun autore riguardo alla singola revisione e di stimare la sua qualità come la media dei pareri di ciascun autore. Questo approccio ha il vantaggio di considerare un voto non più come a se stante, bensì come parte di una valutazione individuale. L'opinione di un autore a su una singola versione v_j sarà quindi formalizzato come:

$$q(a, v_j) = average\left[q(v_i, v_j, v_k) \in Q_{v_j} | author(v_k) = a\right]$$

Dove *author* è la funzione che restituisce l'autore di una versione data e ovviamente *average* quella che restituisce la media campionaria dei valori appartenenti a un certo insieme.

Un'ulteriore scelta fatta nella stima del contributo di un utente è stata quella di non penalizzare l'autore di un intervento di qualità negativa. Lo scopo di questa misura non è tanto quello di individuare chi non compie mai errori, bensì chi maggiormente influisce in modo permanente sullo sviluppo di una data pagina. Perciò la definizione formale della *longevità di un intervento valutato rispetto alla sua versione più simile* (ELS) è la seguente:

$$\forall a \in \mathbb{A} : ELS(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} average\left[q(a, r)\right] \cdot d_M(r)$$

L'ultima direzione verso la quale si è reputato interessante muoversi riguarda la precisione del calcolo dei contributi. Si è detto di come la misura di Edit Longevity non includa nel conto della reputazione l'ultimo e il primo intervento di ciascuna pagina. Per quanto riguarda l'ultimo edit, si è scelto di stimarne la qualità solo nel caso in cui il suo autore sia già intervenuto all'interno della pagina. In particolare la qualità dell'ultimo edit sarà pari alla qualità media di tutti gli interventi fatti dal medesimo autore nella stessa

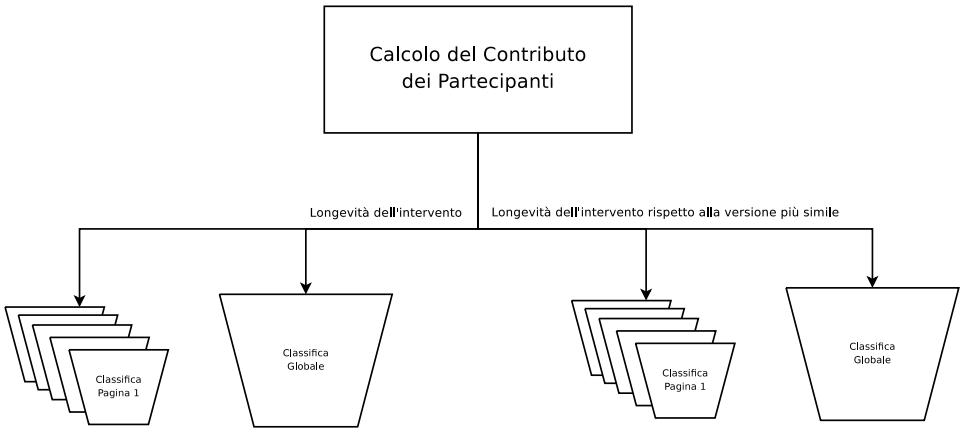
pagina. In questo modo si cerca di limitare la perdita dell'ultimo intervento assumendo che il suo autore si comporterà allo stesso modo di come si è comportato lungo l'intera pagina. Rimane tuttavia il limite di non tener conto del primo intervento.

Anche questa metrica può essere misurata in numero di parole, come quella a cui si ispira.

3.3.2 Considerazioni globali

I processi per calcolare i valori delle tre metriche descritte verrà svolto da opportuni moduli software, la cui implementazione è spiegata nel Capitolo 4, in grado di produrre una classifica sia degli utenti per pagina che degli utenti in tutta Wikipedia, come mostrato in Figura 3.5. Il tipo di classifica

Figura 3.5: Differenti risultati ottenibili dal processo di calcolo del contributo dei partecipanti.



più interessante per questo studio è il primo tipo, quello locale, perché consentirà di scoprire eventuali relazioni tra gli utenti in base al loro contributo all'interno della singola pagina. Tuttavia sarà senza dubbio interessante soffermarsi anche sul secondo tipo di classifica, cioè quella globale.

Tanto per cominciare sarà interessante scoprire quali saranno gli utenti più importanti per le tre metriche basate su assunzioni diverse. Esse mettono in luce tre tipologie di utenti. In particolare si è visto come il conteggio degli interventi possa essere considerato una buona misura di attività di un utente. La longevità di un intervento e la sua variante valutata rispetto alla versione più simile hanno invece lo scopo di individuare gli utenti che maggiormente contribuiscono in modo utile all'enciclopedia. Sarà interessante valutare in

che misura le due metriche differiscono. Per ciascuna metrica inoltre sarà interessante osservare se ci sono delle determinate classi di utenti che contribuiscono in modo simile. In ultimo sarà estremamente importante valutare il contributo degli utenti anonimi. Questo studio è forse il più importante di questo processo. Gli utenti anonimi infatti possono essere considerati solo a livello globale, poiché nei loro interventi non è possibile identificare singoli individui. Le tre metriche possono essere calcolate anche per gli interventi anonimi a patto di considerare gli autori come un unico utente e quindi producendo un risultato aggregato. Essi saranno esclusi da ogni considerazione sugli individui della comunità e quindi risulterà estremamente importante capire l'impatto che essi hanno su Wikipedia. Se i contributi anonimi risultassero essere la maggioranza, lo studio della comunità perderebbe di interesse. Non tanto perché gli utenti anonimi non possano essere considerati parte della comunità, quanto perché essi sono per definizione non intenzionati a far emergere la propria individualità all'interno del progetto, per concentrarsi sui contenuti. Oltre a ciò si può considerare la possibilità di intervenire in modo anonimo una delle più interessanti caratteristiche distintive di Wikipedia. Nè il mondo open source nè tantomeno la comunità scientifica accettano infatti interventi di questo tipo.

3.4 Selezione dei coautori

Lo scopo di questo processo è quello di elaborare le informazioni prodotte dal precedente passaggio di misura del contributo degli utenti di Wikipedia per ciascuna delle sue pagine. Questi dati possono dire molto sulle singole pagine e in particolare si ritiene che essi permettano di misurare quanto è stata partecipativa la scrittura di ciascun articolo. Una delle prime misure utilizzate per questo scopo è quella di *diversità di una pagina*, introdotta da (Lih, 2004). Mentre questa misura si limita a contare il numero distinto di utenti per ciascuna pagina, con le misure introdotte nella sezione precedente si è in grado di valutare quanto ciascuno dei partecipanti ha influito su di esse. Si ricorda che le metriche di Edit Longevity e di longevità dell'intervento valutata rispetto alla versione più simile valutano non solo i parametri quantitativi di un intervento, ma anche quelli qualitativi. Si prendano, a titolo di esempio, due pagine con numero di edit e di diversità paragonabili. Nella prima un solo utente totalizza un alto punteggio di contributo, mentre tutti gli altri risulteranno contributori in maniera molto minore al suo confronto. Nella seconda invece tutti gli utenti totalizzano punteggi di contributo molto simili. La prima dovrà essere considerata una pagina poco collaborativa poiché, a prescindere dal numero dei partecipanti alla sua stesura, uno di

essi ha dominato su tutti gli altri. Inevitabilmente il suo modo di vedere gli argomenti affrontati dall'articolo sarà prevalso. La seconda pagina invece sarà effettivamente frutto di una collaborazione più ampia, nella quale ciascun partecipante ha espresso la sua voce in capitolo senza essere sovrastato da quella degli altri. Per studiare in modo più rigoroso questo fenomeno si è ritenuto interessante selezionare, tra tutti gli autori di ciascuna pagina, proprio quelli che possono essere considerati i suoi contributori principali o *Top User*. Nonostante la semplicità di intervento in Wikipedia possa permettere a chiunque di contribuire a un articolo, non è affatto vero che tutti i suoi contributori hanno la stessa responsabilità sul suo contenuto. Saranno i Top User gli utenti ai quali questo lavoro attribuirà la responsabilità di ciascuna pagina. Essi infatti rappresentano quegli individui che hanno al tempo stesso partecipato tanto e in modo tale che la maggior parte delle modifiche fatte fosse mantenuta dagli utenti successivi.

3.4.1 Metodi per la selezione

Per individuare l'insieme dei Top User (TU), si è deciso di basarsi sulle due metriche di valutazione del contributo più precise, tra le tre presentate: la longevità dell'intevento e la sua variante valutata rispetto alla versione più simile. Si ritiene infatti che il conteggio degli interventi privilegi eccessivamente chi si occupa di fare piccole correzioni che, seppur importanti per l'aspetto qualitativo di un articolo, sono da considerarsi poco significative dal punto di vista dei contenuti.

L'input dell'algoritmo di selezione dei Top User è la lista dei contributori di ciascuna pagina, ordinati in modo decrescente rispetto al loro contributo, come calcolato dal processo precedente. L'obiettivo è quello di scegliere dalla lista quegli utenti che possono essere considerati i maggiori responsabili del contenuto dell'articolo. Basandosi su un solo indicatore è ovvio che un membro dei Top User debba avere un punteggio più alto rispetto a chi non appartiene a questa categoria e cioè:

$$\forall t \in TU, \forall u \notin TU : c(t) > c(u)$$

dove $c(u)$ è la funzione che restituisce il valore di reputazione dell'utente u. Per questo viene preso un utente alla volta secondo l'ordinamento della lista e viene popolato così l'insieme dei Top User il quale dovrà rispettare determinati requisiti.

Il primo requisito è da considerarsi sul singolo utente. Esso impedisce a un individuo di appartenere all'insieme dei Top User nel caso in cui non abbia totalizzato un valore di contributo superiore a una certa quantità detta

W ed è formalizzato nel modo seguente:

$$\forall u \in TU : c(u) > W$$

Questo criterio è motivato dal fatto che si vogliono escludere dal conteggio le pagine di dimensioni troppo ridotte. Poiché il ruolo di Top User assumerà un significato di responsabilità, non si vuole essere costretti a includere un utente solo perché nessun altro ha partecipato più di lui. Non si potranno quindi identificare con questo metodo le caratteristiche di collaborazione per pagine troppo piccole[5]. Normalmente questo accade per quelle pagine non ancora complete o agli inizi. Questo criterio ha ancora più senso se si pensa alla classifica prodotta dalla metrica di longevità dell'intervento, che può assegnare a taluni utenti dei valori negativi. Chiaramente questa proprietà è da verificarsi a priori dell'inserimento di un utente nell'insieme dei Top User. Inoltre dopo aver verificato che un utente non può essere incluso nell'insieme dei Top User, tutti gli altri successivi nella classifica sicuramente non soddisferanno la condizione e non verranno considerati.

Il secondo requisito riguarda invece una proprietà globale dell'insieme dei Top User. Si vuole considerare l'insieme di cardinalità minima la cui somma dei contributi dei partecipanti supera una certa soglia percentuale $T \in [0; 1]$ se confrontato con il totale dei contributi all'interno di una pagina. Quindi alla formalizzazione si aggiungono due vincoli ulteriori:

$$min(|TU|)$$

$$\frac{\sum_{u \in TU} c(u)}{UserContribution} > T$$

Dove U è l'insieme degli utenti di una data pagina e

$$UserContribution = \sum_{u \in U} c(u)$$

É ancora una volta il caso della longevità di un intervento a richiedere delle precisazioni. Essendoci la possibilità di avere dei contributi negativi per alcuni utenti si è scelto di considerare il sottoinsieme degli utenti aventi contributi positivi:

$$U \supseteq U^+ \equiv \{u | c(u) > 0\}$$

Quindi il vincolo principale diventa:

$$\frac{\sum_{u \in TU} c(u)}{UserContribution^+} > T$$

[5]La quantificazione di questo aspetto dipende ovviamente dal parametro W.

Dove:

$$UserContribution^{+} = \sum_{u \in U^{+}} c(u)$$

Anche in questo caso però si può sfruttare l'ordinamento della lista dei contributi. Infatti per soddisfare questo vincolo basterà prendere il prossimo utente in classifica e verificare, a posteriori del suo inserimento nell'insieme TU, se la percentuale di reputazione totalizzata dai membri dell'insieme supera la soglia T. In questo caso si può interrompere la ricerca dei Top User, poiché nessun altro utente potrà essere inserito nell'insieme senza violare i vincoli. Altrimenti si può passare alla valutazione dell'utente successivo.

L'ultima questione da affrontare rimane la gestione degli utenti anonimi in questo processo di selezione. Si è avuto modo di vedere come essi non possano essere contati tra i singoli utenti e, di conseguenza, non possono essere considerati dei Top User. Eppure in non pochi casi il loro contributo collettivo all'interno delle singole pagine è ben visibile. Nel caso in cui sia positivo, il contributo collettivo degli utenti anonimi è dunque da includersi nel conteggio dei contributi totali di una pagina. Quindi nuovamente:

$$\frac{\sum_{u \in TU} c(u)}{TotalContribution^{+}} > T$$

Dove:

$$TotalContribution^{+} = UserContribution^{+} + AnonymousContribution^{+}$$

Si rifletta sul fatto che all'aumentare del contributo anonimo percentuale diminuisce quello degli utenti registrati. Di conseguenza si può giungere a una situazione limite, nella quale il contributo anonimo percentuale è pari a $1 - T$, in cui tutti gli utenti registrati possono essere inclusi nell'insieme dei Top User. Per evitare che la soglia assuma un significato eccessivamente differente in funzione del contributo anonimo, si è deciso dunque di scontarla di un fattore a esso proporzionale. L'assunzione è quella secondo la quale, non potendo stabilire se tra gli utenti anonimi qualcuno avrebbe soddisfatto i vincoli per rientrare nell'insieme dei Top User, il loro contributo sia equamente suddiviso tra le due partizioni degli utenti. La nuova soglia T_A per la selezione dei Top User risulta quindi essere la seguente:

$$T_A = T - T \cdot \frac{AnonymousContribution^{+}}{TotalContribution^{+}}$$

3.4.2 Considerazioni locali

A questo punto del processo ci si ritrova dunque con una lista di utenti considerati importanti per ciascuna pagina, i Top User. Dalle fasi precedenti

inoltre si conoscono anche altri dati d'interesse sulle singole pagine: quelli riguardanti i contributi totali e quelli riguardanti gli utenti anonimi. I primi due dati molto semplici, tradizionalmente utilizzati per misurare la qualità di una pagina, sono il numero totale di interventi (*EditCount*) e il numero distinto di utenti intervenuti (*Diversity*).

Per quanto riguarda quest'ultima metrica si è scelto di rappresentare gli utenti anonimi come un unico utente autore di tutti gli edit svolti da utenti anonimi. La motivazione di questa scelta è da ricercarsi prima di tutto nell'impossibilità tecnica di identificarli. In realtà tutti questi utenti diversi hanno un fattor comune: il disinteresse per una reputazione all'interno di Wikipedia. Se questo disinteresse potrebbe apparire come una caratteristica in grado di evidenziare un comportamento dannoso nei confronti dell'enciclopedia, si vuole far notare come questo non sia scontato. In realtà nemmeno gli utenti registrati ricevono una ricompensa materiale per il loro operato. La loro motivazione è dunque da ricercare in altri tipi di soddisfazione. Anche un utente anonimo può risultare gratificato del suo operato senza che gli altri sappiano di lui.

Un altro dato d'interesse per la pagina è il contributo globale positivo ($TotalContribution^{+}$), indice di quanto sforzo è stato impiegato nella scrittura della pagina da parte dei suoi autori.

Quindi si dispone delle informazioni relative a un importante gruppo di autori della pagina: quello dei Top User. A partire dalla lista di questi utenti, possono essere calcolati degli indici che si ritiene siano d'interesse per misurare il grado di collaborazione all'interno di una pagina. Tra di essi il più immediato è la cardinalità di questo insieme (*NumberOfTopUser*), che esprime il valore assoluto dei partecipanti più importanti della pagina.

Il dato derivato da quest'ultimo è la percentuale di Top User (*TopUserPercentage*), in grado di esprimere il rapporto tra la quantità dei Top User e il numero totale degli utenti di una pagina, quindi calcolato come:

$$TopUserPercentage = \frac{NumberOfTopUser}{Diversity}$$

Infine, sempre a partire dall'insieme dei Top User, si possono ricavare altre due interessanti quantità derivate. Esse sono la percentuale di interventi totalizzati dai Top User sul totale (*TopUserEditsPercentage*) e la percentuale di contributi totalizzati da essi sul totale (*TopUserContributionPercentage*). La prima, definita la funzione $e(u)$ che restituisce il numero di interventi fatti da un utente u in una data pagina, è calcolabile come:

$$TopUserEditsPercentage = \frac{\sum_{u \in TU} e(u)}{EditCount}$$

Lo scopo di questo indice è, ancora una volta, quello di permettere di confrontare le metriche di contributo con il conteggio degli edit. La seconda invece è, come spiegato nel paragrafo 3.4.1, calcolabile con la seguente formula:

$$TopUserContributionPercentage = \frac{\sum_{u \in TU} c(u)}{TotalContribution^{+}}$$

Questo indice rappresenta la forza del gruppo dei Top User all'interno della pagina. Infatti, tanto più alto sarà il suo valore, tanto più gli interventi considerati utili dalla comunità saranno quelli dei Top User.

Le ultime informazioni d'interesse che sono state selezionate per lo studio di una singola pagina sono quelle relative agli utenti anonimi che hanno operato al suo interno. Poiché gli utenti anonimi possono essere visti come un gruppo di utenti al pari dei Top User, ovviamente con significato differente, i medesimi indici appena descritti per essi possono essere riutilizzati. L'unica differenza è che con il gruppo di utenti anonimi non si può scendere in dettaglio sino al singolo individuo. Sono quindi consultabili i soli indici di percentuale di interventi totalizzati da utenti anonimi sul totale (*AnonymousEditsPercentage*) e la percentuale di contributi totalizzati da essi sul totale (*AnonymousContributionPercentage*) definiti in modo analogo agli ultimi due indici descritti. In realtà quest'ultimo valore percentuale verrà considerato nullo nel caso in cui il contributo anonimo assoluto risultasse negativo.

Questo genere di dati è stato utilizzato in questo lavoro solo per monitorare il processo di selezione dei coautori. Tuttavia si ritiene che essi possano dire molto sulle proprietà di una singola pagina.

3.4.3 Considerazioni globali

A partire dai dati sulle singole pagine risulterà quindi interessante aggregare i valori trovati per trarre delle conclusioni più generali. Il primo passo da fare sarà però quello di rimuovere dallo studio le pagine per le quali non è stato possibile trovare dei Top User. Questo perché, come spiegato precedentemente, l'assenza di utenti appartenenti a questa categoria è dovuta all'insufficienza d'informazioni sulle pagine in questione. I dati ricavabili dopo quest'operazione sono svariati. Può essere interessante calcolare le statistiche sul numero di Top User medio per pagina o gli stessi valori aggregati per gli anonimi.

Queste statistiche possono inoltre essere estratte per differenti insiemi di pagine, per capire se esse possano avere caratteristiche in comune. Ad

esempio si proverà a raggruppare le pagine Featured, riconosciute cioè dalla comunità come le migliori di Wikipedia, e per esse sarà possibile vedere quanti Top User in media sono stati rilevati dall'algoritmo di selezione.

Considerazioni simili possono essere inoltre fatte sui dati riguardanti gli utenti anonimi. La loro presenza in gruppi di pagine di un certo tipo potrà permettere di valutarne l'impatto non solo da un punto di vista quantitativo ma anche qualitativo.

3.5 Costruzione di una Rete Sociale

Riprendendo la strada principale di questo lavoro, si vuole a questo punto descrivere il quarto e ultimo processo di elaborazione dei dati che porterà alla costruzione di una Rete Sociale (*Social Network*) degli utenti di Wikipedia o, si ricorda, di un qualunque altro progetto basato sulla tecnologia wiki di cui si dispongano i dump del database. Prima di questo però è importante capire come mai proprio una Rete Sociale è lo strumento modellistico migliore per gli scopi di questo lavoro.

Un aspetto fondamentale di un wiki è, come già detto, l'approccio collaborativo alla scrittura del contenuto. É all'aumentare delle dimensioni del progetto che questa caratteristica assume particolare importanza. Infatti con questo tipo di tecnologia ciascun partecipante può concentrarsi sui singoli aspetti dei quali si ritiene più competente. In questo modo viene applicata una suddivisione volontaria del lavoro per la quale il progetto può crescere in modo del tutto non uniforme nelle sue singole componenti. Nelle parti con meno partecipanti il processo di scrittura di un articolo sarà più fluido poiché sarà più facile trovare un accordo. Viceversa in quelle aree con tanti partecipanti ci sarà una maggiore probabilità di scontro tra opinioni avverse. Tenendo conto del fatto che spesso le pagine sulle quali tante persone dedicano il loro tempo sono proprio quelle di maggiore interesse anche per chi non partecipa, in linea di principio questa maggiore tendenza a far dialogare portatori di opinioni differenti dovrebbe garantire una maggior neutralità. Tutto ciò funzionerebbe bene in un mondo chiuso, dove cioè ogni singolo autore può essere identificato e non ha dunque vantaggio nell'ostacolare il progetto. Esso sarebbe tuttavia limitato da un ristretto numero di partecipanti.

In un mondo aperto come quello del Web, dove sostanzialmente è molto difficile identificare gli autori di atti vandalici e di conseguenza prendere provvedimenti contro di essi, si rende quindi necessaria l'istituzione di particolari ruoli di controllo, gli amministratori, in grado di alterare il normale processo di scrittura collaborativa. Essi possono infatti impedire a un indivi-

duo o a gruppi di utenti di collaborare. Oltre a ciò il grado di partecipazione non è limitato, né inferiormente né superiormente, cioè ognuno può dedicare tutto il tempo che vuole al progetto. Questo in un certo senso avvantaggia quegli utenti della comunità che, per motivi personali, dispongono di una maggior quantità di risorse da investire nel progetto. Tutti questi aspetti possono influire notevolmente sulla scrittura degli articoli e quindi sul contenuto e l'andamento di Wikipedia. Paradossalmente tutti questi processi di collaborazione e di conflitto sono nascosti dall'interfaccia del wiki, che in un certo senso inganna il lettore facendogli credere che quello che sta leggendo è scritto in modo coerente e uniforme. Si ritiene che sia questa la vera insidia di Wikipedia. Da un lato la grande mole di informazioni contenute in un unico repository organizzato in modo ordinato, predisposto alla ricerca e alla navigazione tramite collegamenti ipertestuali rendono molto attraente per chiunque la sua consultazione. Dall'altro lato il mascheramento dei processi di scrittura rende pressoché impossibile per un lettore capire da quale mano provengano i contenuti che sta leggendo. A un livello di osservazione più elevato questo si traduce nella difficoltà di catturare informazioni sulla reale base di utenti che opera sui contenuti di Wikipedia. Quindi su quanti siano effettivamente gli autori più coinvolti e in grado di influenzare l'andamento di uno dei siti più visitati del Web. Per questo si ritiene indispensabile sfruttare i rapporti di tipo sociale tra di essi, poiché solo in questo modo si possono mettere alla luce caratteristiche quali gli individui più centrali nel progetto, la possibilità di propagarsi delle informazioni di coordinazione e tutte le misure tradizionalmente usate negli studi sociometrici.

Il problema si sposta quindi sulla selezione dei dati relazionali da considerare nella costruzione della Rete Sociale. Esso è di cardinale importanza poiché chiaramente la scelta delle relazioni tra gli utenti influirà in maniera determinante sui risultati di ogni studio della rete. Il compito è particolarmente impegnativo in generale poiché le relazioni di tipo sociale sono difficili da definire in maniera formale e soprattutto possono essere soggettive. Nel caso specifico di Wikipedia si è visto addirittura come un utente possa operare all'insaputa degli altri membri della comunità. Nel lavoro di (Korfiatis, 2006, Korfiatis et al., 2006) viene considerata una relazione tra due individui nel caso in cui essi abbiano scritto almeno un articolo nella stessa categoria di Wikipedia. Alternativamente gli autori propongono di mettere in una relazione asimmetrica l'autore di una revision con quello della sua successiva.

Si ritiene che entrambe le ipotesi catturino relazioni troppo poco significative per lo scopo di questo lavoro. Infatti la prima, considerando che le categorie di Wikipedia possono comprendere migliaia di articoli, potrebbe mettere in relazione due utenti che non hanno nemmeno partecipato allo

stesso articolo e quindi con pochissime caratteristiche in comune. La seconda potrebbe invece dipendere da fattori casuali, come ad esempio il fatto che due autori collaborano nel medesimo periodo temporale alla stessa pagina. É in questa situazione che potrebbero essere tracciate relazioni tra utenti i quali effettivamente ignorano l'esistenza uno dell'altro o la cui partecipazione è solo un evento occasionale.

Si è scelto dunque di ricondursi al tipo di relazione individuata dagli studi bibliometrici o sulle reti di sviluppatori di progetti open source, che ben più di una volta sono stati confrontati con Wikipedia in questa sede. La relazione di cui si parla è quella di *coautore* e associa tipicamente due autori del medesimo articolo o software. L'aspetto importante però è quello che non verranno considerati tutti gli editor di una voce di Wikipedia come i suoi coautori, perché altrimenti si ricadrebbe nei medesimi problemi individuati per lo studio di (Korfiatis, 2006, Korfiatis et al., 2006). Verranno considerati come *coautori di una pagina di Wikipedia* tutti i membri dell'insieme dei Top User per essa. Essi infatti rappresentano quegli utenti che maggiormente hanno contribuito allo sviluppo della pagina ed è da considerarsi molto difficile che non abbiano, prima o poi, avuto coscienza dell'operato degli altri individui appartenenti alla loro categoria. Due utenti connessi in una Rete Sociale di questo tipo hanno con buona probabilità collaborato e condividono le stesse norme di comportamento e di gestione di un articolo. Non necessariamente essi avranno la medesima opinione, poiché Wikipedia ambisce a una scrittura neutrale degli articoli, nè le stesse competenze, poiché la tecnologia dei wiki facilita la suddivisione del lavoro ed è quindi possibile che i due utenti in questione si siano compensati a vicenda. Una rete di questo tipo, che potrebbe essere chiamata *rete dei collaboratori più importanti*, avrà un insieme di nodi costituito da un sottoinsieme di quello degli utenti di Wikipedia. In particolare sarà creato un nodo per ogni utente che è nell'insieme dei Top User di una pagina assieme ad almeno un altro utente. Per quanto riguarda quegli utenti che da soli sono considerati Top User di una pagina si è scelto di non includerli nello studio poiché di essi non si può dire se effettivamente abbiano collaborato con qualcuno. Questo tuttavia non è da escludersi, poiché potrebbe trattarsi di utenti che eseguono compiti per conto di altri, magari amministratori, che coordinano dalla pagina di discussione i loro interventi. Attualmente però non ci sono mezzi per cogliere questo tipo di relazioni nascoste. La rete così formata non potrà dire nulla sui collaboratori occasionali, ma sarà adatta a studiare le personalità che maggiormente influenzano Wikipedia. Per le considerazioni già approfondite nei precedenti paragrafi risulterà chiaro che nella rete non saranno rappresentati gli utenti anonimi, non perché essi siano necessaria-

mente di scarso impatto sul wiki, ma perché non c'è modo di identificarli. Si descriveranno ora più in dettaglio gli studi che si vogliono fare su questa rete e l'interpretazione delle misure che verranno effettuate.

3.5.1 Studio della rete a livello macroscopico

Lo studio macroscopico della Rete Sociale costruita come appena descritto consente di farsi un'idea generale della comunità degli utenti. Questo tipo di analisi è utile per fare confronti tra differenti reti e, in questo caso, si è molto interessati al confronto con le reti di coautori di articoli scientifici. Per questo si prenderanno come riferimento le misure indicate negli studi precedenti proprio su questo tipo di reti (Newman, 2001a, Cotta and Merelo, 2006).

La prima misura d'interesse, il *numero di articoli*, non esprime una proprietà della rete. Essa più che altro è utile per comprendere le dimensioni dell'universo considerato. È presumibile che un ampio numero di articoli implichi un alto numero di autori e un maggior numero di relazioni tra di essi. Si ricordi però che non tutti gli articoli del wiki sono considerati significativi ai fini della costruzione della rete. In particolare quelli con meno di due Top User non contengono alcun'informazione riguardante le relazioni tra utenti. Sarà dunque importante considerare quanti *articoli con almeno un autore* e quanti *articoli con più di un autore* sono stati calcolati per rendersi conto delle dimensioni reali dell'insieme che ha dato origine alla rete.

Il *numero di autori* è invece una proprietà della rete, in particolare espressa dal numero di nodi. Per come si costruirà la rete non tutti gli utenti di Wikipedia figureranno al suo interno. Come in una rete di coautori chi non pubblica articoli non verrà contato nella rete, anche nel nostro caso chi non è mai tra gli utenti più importanti di una pagina oppure chi è sempre considerato l'unico autore in tutte le pagine in cui esprime dei contributi notevoli, non sarà presente. D'altra parte, sia che i suoi contributi siano poco rilevanti sia che egli lavori sempre da solo, non si può dire nulla sulle sue interazioni all'interno della comunità. Al contrario del caso delle comunità scientifiche però, per un wiki generalmente si dispone della lista di tutti gli utenti registrati e di conseguenza si potrà confrontare il numero degli utenti considerati coautori con quello totale. Inoltre al confronto si potrà anche aggiungere il *numero di autori*, includendo cioè nel conteggio quegli utenti che, pur essendo stati Top User in qualche articolo, non risultano aver mai collaborato con altri. Questi dati consentiranno di valutare le reali dimensioni del gruppo dei collaboratori più importanti di Wikipedia.

Il numero di *autori per articolo* e il numero di *articoli per autore* mette-

ranno in evidenza altre due proprietà della comunità che non possono essere estratte dalla rete solamente. Sarà interessante studiarne l'andamento per rendersi conto effettivamente di quanto possa essere collaborativa la scrittura in Wikipedia.

Il *numero di collaboratori* per autore è invece una misura molto usata negli studi sociometrici. Esso corrisponde al grado di un nodo, cioè il numero di archi ai quali esso è connesso. Di questa misura sarà interessante valutare sia il valor medio sia la distribuzione, che nel caso risultasse avere un andamento di tipo power law, potrebbe mostrare la conferma della legge di preferential attachment come descritta da (Barabasi et al., 2002). Il significato della legge è il seguente: autori che sono stati più a contatto con altri raggiungono una visibilità maggiore all'interno della comunità e dunque hanno la possibilità di accrescere ulteriormente il numero delle loro collaborazioni.

Il passo successivo è lo studio delle *componenti connesse*, cioè quegli insiemi di nodi all'interno dei quali ciascun individuo può raggiungere, attravreso un percorso di archi, un qualsiasi altro individuo dello stesso insieme. In una rete dove i collegamenti rappresentano la diffusione di informazioni, nel caso della rete di Wikipedia le norme della comunità, sarà interessante vedere come e quanto essa è frammentata. Il *numero di componenti connesse* e la *dimensione della componente più grande*, anche relativo al numero di nodi, saranno importanti per valutare questa caratteristica. Lo studio delle componenti connesse è anche il primo passo per valutare l'esistenza di sottocomunità all'interno dell'universo in analisi.

Il *coefficiente di clustering* è un altra misura di coesione della rete. Esso è definito come la percentuale di triple di nodi per i quali vale la proprietà transitiva sul totale delle possibili triple e cioè:

$$C = \frac{3 \times \text{numero di triangoli nel grafo}}{\text{numero di triple di nodi connesse}}$$

Sebbene si debba tenere conto del fatto che il valore calcolato è influenzato dal numero di articoli con almeno tre coautori, un alto valore del coefficiente di clustering può indicare il fatto che aver collaborato con un certo autore può facilitare la collaborazione con un altro suo collaboratore. Quindi ancora una volta si tratta di verificare il transito delle informazioni all'interno della comunità. Inoltre si pensi al fatto che una struttura di rete ad albero, in quanto priva di cicli, avrà coefficiente di clustering nullo. Al contrario un grafo completamente connesso totalizzerà il massimo punteggio per questa metrica, pari a uno. Questo significa che le reti con una struttura fortemente

gerarchica avranno coefficiente di clustering basso, altro dato d'interesse per Wikipedia.

In ultimo si prenderanno in considerazione le misure del *grado di separazione*. La prima misura è la media della minima distanza tra due nodi qualsiasi della rete. Essa, insieme al *diametro* della rete (la lunghezza del più lungo cammino minimo tra tutte le coppie di vertici) da un'idea di quanto siano vicini tra di loro gli utenti di Wikipedia. Una bassa distanza in questo caso significa un maggior grado di partecipazione, da parte degli utenti, a tutti gli argomenti dell'enciclopedia. Si precisa che queste misure andranno calcolate escludendo i nodi appartenenti a differenti componenti, considerati altrimenti separati da una distanza infinita che annullerebbe il senso dello studio.

3.5.2 Studio delle sociometric star

Le misure di centralità dei nodi di una rete permettono di identificare le cosiddette *sociometric star*, individui di spicco all'interno della comunità secondo diversi criteri. Nel caso di Wikipedia si ha già una misura di personalità importanti, che è quella data dal calcolo del contributo totale come approfondito nella sezione 3.3. Tuttavia questa metrica non prende in considerazione le relazioni tra gli utenti all'interno della comunità. Sarà quindi molto interessante confrontarla con le misure di centralità descritte in questo paragrafo. Per approfondimenti sulle misure di centralità si vedano (Wasserman and Faust, 1994, Scott, 2000, Liu, 2007).

La prima e più semplice misura da studiarsi è la *Degree Centrality*. Essa assegna a ogni individuo nella rete un valore di centralità in base al suo grado $d(n)$, cioè il numero di collegamenti con gli altri individui. Spesso si è soliti normalizzare questa misura sul numero massimo di connessioni che può avere un qualunque nodo della rete, calcolato come il numero di nodi N escluso se stesso:

$$C_D(n) = \frac{d(n)}{N-1}$$

L'assunzione di questa misura di centralità è quella secondo la quale un utente è tanto più centrale quanto più è coinvolto in legami sociali. Essa è da considerarsi una misura locale poiché considera solo i nodi nel vicinato di ciascun individuo, cioè quelli a lui direttamente connessi. Tra l'altro questi nodi sono considerati in modo del tutto indifferente tra di loro. Nel caso della Rete Sociale di Wikipedia, gli individui più centrali secondo questo criterio sono quelli che hanno partecipato in modo più significativo con differenti persone. In un certo senso questa è una misura di quanto essi sono stati in grado di relazionarsi, in modo positivo e cioè senza generare conflitti, con

i differenti membri della comunità. Individui centrali per questa metrica saranno quindi da considerarsi esperti delle norme vigenti all'interno della comunità e saranno particolarmente dotati di qualità come convincimento e diplomazia.

La seconda misura d'interesse è la *Closeness Centrality*. L'assunzione questa volta è quella secondo la quale un nodo è centrale se può facilmente interagire con tutti gli altri individui. Per calcolarla si confronta la distanza minima che un nodo potrebbe avere da tutti gli altri nodi (ancora una volta $N-1$) con la distanza minima che egli ha effettivamente da tutti gli altri nodi a lui connessi:

$$C_C(n) = \frac{N-1}{\sum_{m=1}^{N} d(n,m)}$$

Dove $d(n,m)$ e la misura del più breve percorso tra due nodi n e m. Si noti che anche in questo conto, come per le misure globali del grado di separazione medio, la distanza di un nodo con se stesso è nulla, ma bisogna prestare particolare attenzione alle distanze tra nodi non connessi, poiché sarebbero da considerare infinite. Per approfondire la trattazione di queste casistiche eccezionali si rimanda ad esempio a (Wasserman and Faust, 1994, Scott, 2000, Liu, 2007). Gli utenti della Rete Sociale di Wikipedia centrali per questa metrica sono quelli che hanno maggiore possibilità di diffondere le norme di comportamento al suo interno. Infatti se un utente adotterà delle particolari convenzioni nella scrittura di un articolo, come ad esempio la tendenza a citare molte fonti nel suo discorso, i suoi coautori le recepiranno come valide, poiché rispettate da un autore importante nella pagina che anche loro curano. Un utente molto vicino a tutti gli altri potrà influenzare molto rapidamente un gran numero di utenti importanti di Wikipedia.

Un'altra misura molto importante nella diffusione delle informazioni all'interno della rete è la *Betweenness Centrality*. Essa individua come centrali quegli autori per i quali passano un grande numero di percorsi più brevi (*shortest path*) tra i nodi della rete. Definiti il numero di shortest path tra due nodi i e j, si indicherà il loro numero con p_{ij} con $p_{ij}(n)$ quelli passanti per un certo nodo n. Quindi la misura di centralità, normalizzata sulla massima centralità possibile per n (pari cioè alla metà del prodotto tra $N-1$ e $N-2$), sarà:

$$C_B(n) = \frac{2\sum_{i<j} \frac{p_{ij}(n)}{p_{ij}}}{(N-1))(N-2)}$$

Gli utenti della Rete Sociale centrali per questa metrica sono quelli che hanno maggiore controllo sulla diffusione delle informazioni al suo interno. Senza

di essi infatti la distanza media tra i nodi della rete non può che aumentare, con conseguenti rallentamenti nella diffusione delle norme.

L'ultima misura per individuare le sociometric star presa in considerazione da questo studio è quella di *Eigenvector Centrality*. Essa è considerata una misura di centralità di tipo globale, in contrapposizione ad esempio alla degree centrality, poiché la reputazione di un nodo non è valutata solamente a partire dal numero dei suoi vicini, ma anche in base alla loro importanza. Dunque un nodo con tanti vicini considerati poco importanti può benissimo essere considerato meno centrale di un altro con pochi vicini molto centrali nella rete. La versione più generale di questo tipo di centralità può essere applicata a grafi con archi pesati, quale può essere considerata la Rete Sociale degli utenti di un wiki nel caso in cui si consideri come valore di un arco il numero di volte in cui i due individui da esso connessi sono stati coautori della stessa pagina. L'idea è quella per cui la centralità di un nodo sarà maggiormente determinata da quella dei suoi vicini con arco di peso maggiore. Non si vuole approfondire in questa sede il processo di calcolo di questa misura per la quale esistono differenti algoritmi. Per maggiori dettagli sulla Eigenvector Centrality si consulti (Newman, 2003). In generale questa misura può essere considerata un calcolo più preciso della degree centrality, basato su proprietà globali della rete piuttosto che locali. Sarà interessante confrontarla proprio con la degree centrality poiché una rete in cui le proprietà locali coincidono con quelle globali è da considerarsi stabile.

Capitolo 4

Scelte implementative

In questo capitolo verranno discusse le scelte implementative che hanno portato allo sviluppo dei diversi sottoprocessi di analisi di un wiki come descritti nel capitolo precedente. Tutti i software sono stati scritti nel linguaggio di programmazione Java il cui maggiore vantaggio in questo progetto risulta essere la portabilità. Infatti i software che stanno per essere descritti sono stati scritti e testati su dati di dimensioni ridotte su una macchina con architettura a 32bit, mentre i veri processi di analisi sono stati eseguiti su una macchina con architettura a 64bit dotata di maggiori risorse.

4.1 Processo di estrazione di informazioni dalla cronologia di un wiki

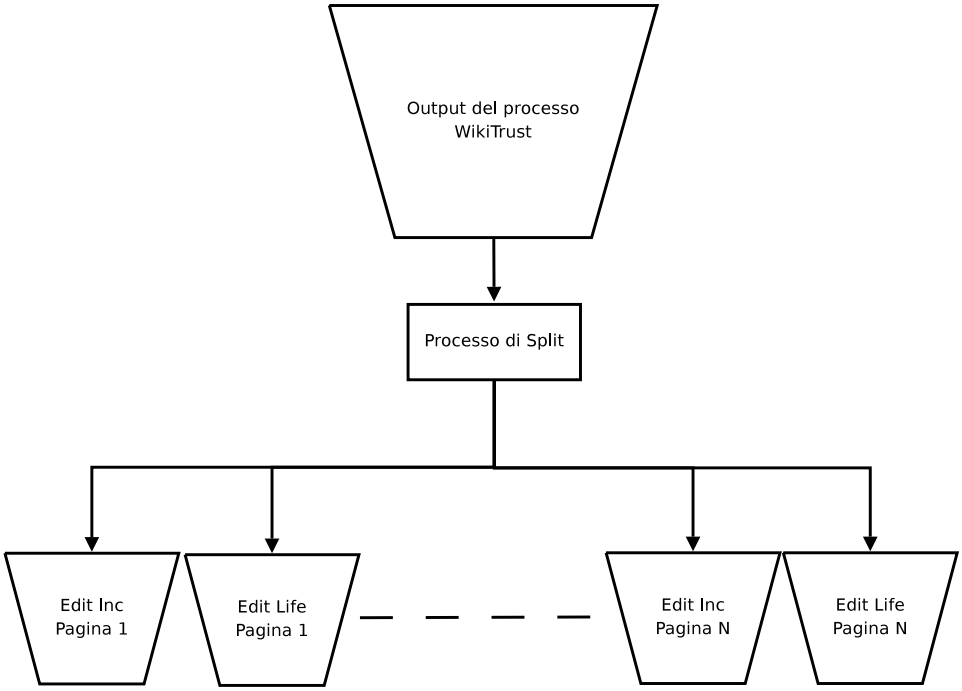
Come già spiegato il primo sottoprocesso è già stato implementato dai membri del progetto WikiTrust[1] e di conseguenza non è stato necessario occuparsi di questo aspetto dell'analisi.

Tuttavia è stato necessario adattare il file output del processo del software WikiTrust alle esigenze di quest'analisi. Il file prodotto è di tipo testuale e racchiude al suo interno, ordinate cronologicamente, tutte le entità di tipo Edit Inc e di tipo Edit Life, queste ultime derivate dalle prime. Il motivo di questo tipo di output è che il progetto originale prevede il calcolo di una reputazione e di un contributo degli autori all'interno dell'intero wiki studiato. In questo caso invece l'analisi è incentrata su ciascuna pagina del wiki considerata nella sua atomicità e quindi, per non dover gestire parallelamente il grandissimo numero di pagine dei wiki che si vogliono analizzare, si è optato per una suddivisione del file originale in una moltitudine di file più

[1] Si faccia riferimento al sito ufficiale del progetto: `http://trust.cse.ucsc.edu/` .

piccoli. In particolare è stato progettato e implementato un applicativo in grado di riconoscere, all'interno di ogni singola linea del file originale, sia la pagina di appartenenza che il tipo di entità rappresentata. Questo processo, chiamato di *Split* e descritto graficamente in Figura 4.1, riceve in ingresso il file di output del processo realizzato per il progetto WikiTrust. Quindi

Figura 4.1: Diagrama del processo di Split del file output del processo realizzato per il progetto WikiTrust



crea due file per ogni pagina del wiki: il primo contenente tutte le linee rappresentanti entità di tipo Edit Inc e il secondo tutte quelle rappresentanti le entità Edit Life. I file del primo tipo verranno utilizzati per calcolare il contributo di ciascuna pagina secondo la metrica di longevità dell'intervento valutata rispetto alla sua versione più simile. Essi potrebbero benissimo essere utilizzati anche per il calcolo della metrica di longevità dell'intervento, ma in questo caso torna più utile il secondo tipo di file, che già contiene le informazioni per il calcolo di questa metrica, ottenute aggregando le entità di tipo Edit Inc come descritto nel capitolo precedente. Questo aspetto renderà il calcolo dei valori della metrica di longevità dell'intervento più leggeri rispetto a quelli della sua variante, poiché già in questa fase sono stati elaborati i dati per essa.

In questo modo, dopo il sottoprocesso di estrazione, sarà possibile calcolare i valori di una delle due metriche per una qualsiasi pagina presa singolarmente. Alternativamente sarebbe stato necessario scorrere tutto il file originale mantenendo in memoria i dati calcolati per ciascuna metrica su ogni singola pagina, oppure scorrerlo più volte, e in ognuna di esse si sarebbe dovuto cercare quelle linee riguardanti una certa pagina per una certa metrica. È chiaro come entrambe queste scelte alternative siano poco praticabili in un caso di studio reale: la prima a causa delle elevatissime risorse richieste, la seconda per via dell'inefficienza di dover scorrere più volte un file per il quale la maggior parte delle informazioni risultano inutili.

Sono stati inoltre estratti altri dati dai wiki analizzati. Alcuni dei sottoprocessi successivi richiederanno l'elenco degli utenti amministratori e dei Bot operanti all'interno del wiki. Per recuperare queste informazioni da Wikipedia si è provveduto a cercare le liste di queste tipologie di utenti e a ripulirle da informazioni inutili tramite espressioni regolari. Per questi compiti non sono stati scritti processi appositi, poiché la localizzazione e la formattazione di queste liste non è universale. Wiki diversi possono dichiarare o non dichiarare i loro amministratori e Bot e soprattutto il formato di queste liste, scritte da utenti sempre attraverso la tecnologia wiki, non è standard. Se per l'individuazione degli amministratori questa tecnica si è rivelata priva di errori, non è sempre facile individuare un Bot. Questo perché le liste dei Bot sono aggiornate dalle comunità con minore rigore, ma soprattutto perché non tutti gli utenti Bot sono dichiarati come tali. Una semplice euristica utilizzata più volte in letteratura è quella di considerare Bot tutti quegli utenti che hanno per nome una stringa che termina con i tre caratteri "bot". Si è deciso tuttavia di non implementare quest'euristica per evitare di incorrere in falsi positivi.

Un altro tipo d'informazione che verrà utilizzata successivamente è quella relativa agli elenchi di Featured Article. Anche per questi la maggior parte delle versioni di Wikipedia mantengono una lista che è stata recuperata manualmente e ripulita tramite espressioni regolari.

Ovviamente queste liste sono aggiornate alla data a cui risale il log analizzato.

4.2 Processi di calcolo del contributo dei partecipanti

I processi di calcolo del contributo dei partecipanti sono due, come già accennato nella sezione precedente. Il primo è quello per calcolare il contributo

di un utente sia all'interno di una singola pagina che a livello globale con la metrica di longevità dell'intervento. Il secondo è analogo con la differenza di calcolare la metrica di longevità dell'intervento valutata rispetto alla sua versione più simile. Entrambi calcolano anche altri dati d'interesse per ciascuna pagina, come il numero di interventi totali, il numero di utenti distinti (diversità) e la reputazione totale positiva.

Oltre a questi dati aggregati, entrambi contano il numero di interventi per ciascun utente all'interno di una pagina. Infine ciascuno calcola il contributo, diverso al variare della metrica utilizzata, e il numero di interventi totali al livello del wiki. Il conteggio degli interventi, sia per la singola pagina che globale, calcolato dai due processi ovviamente coincide. Tuttavia si è reputato utile implementare questo calcolo in entrambi poiché la metrica di conteggio degli interventi è il punto di riferimento col quale confrontare ogni altra. In questo modo non si è vincolati a calcolare entrambe le metriche per ogni wiki, ma si può decidere quella più adatta in ogni occasione. Inoltre il conteggio degli interventi è banale da calcolare con i dati in ingresso e non rende più complessa l'esecuzione dei processi.

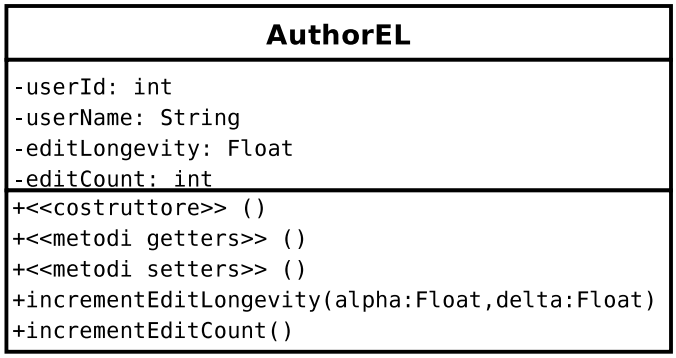
Per tutte le classifiche globali si è inoltre utilizzato un semplice metodo per rendere un po' più rapida l'individuazione delle differenze tra esse. Verranno considerati i primi 100 utenti classificati e per ognuno di essi si valuterà, grazie alle liste ricavate dal wiki, se si tratta di un Bot o di un amministratore. Quindi sarà possibile contare quanti utenti sui primi 100 appartengono a una determinata categoria.

L'ultimo aspetto da chiarire risulta la gestione degli utenti anonimi. Essi sono identificati, all'interno del file di output del processo WikiTrust, con identificativo utente pari a 0. Risulta quindi banale trattarli come un singolo utente.

4.2.1 Processo di calcolo della longevità di un intervento

Questo processo risulta essere molto semplice grazie al fatto che esso può utilizzare solamente le entità Edit Life che sono ottenute da elaborazioni di quelle di tipo Edit Inc. Le classi implementate modellano le due entità in gioco. La classe *Edit Life* è molto semplice e la sua unica particolarità è quella di ricevere come parametro del suo costruttore la linea testuale che riguarda un'entità per renderla accessibile tramite gli ovvi metodi getters. L'entità relativa all'autore è invece modellata tramite la classe *AuthorEL*, mostrata in Figura 4.2. Anche essa è molto semplice e incapsula gli attributi di interesse per ciascun autore. Solo il metodo di *incrementEditLongevity* richiede una breve spiegazione. Esso riceve in ingresso due parametri: *alpha*

Figura 4.2: Diagramma della classe AuthorEL



corrisponde alla qualità di una revision, indicata con *AvgSpecQ* nell'entità Edit Life; *delta* rappresenta la quantità di modifiche apportate da una revision rispetto alla sua precedente. Il metodo incrementa l'attributo *editLife* dell'autore di una quantità pari al prodotto tra *alpha* e *delta*.

La semplice implementazione del processo è descritta in pseudocodice nell'Algoritmo 1. Per ottenere una maggiore chiarezza si è deciso di omettere le istruzioni riguardanti il calcolo delle statistiche relative a ciascuna pagina.

4.2.2 Processo di calcolo della longevità di un intervento valutata rispetto alla sua versione più simile

Questo processo è da considerarsi meno semplice del precedente a causa del fatto che esso utilizza le informazioni contenute nelle entità Edit Inc che devono essere a loro volta elaborate per ottenere il valore di contributo finale che verrà attribuito all'autore di ciascuna revision. Innanzitutto sono stati incapsulati gli attributi dell'entità Edit Inc in un oggetto la cui unica particolarità è quella di accettare come parametro del costruttore la stringa di testo contenente tutte le informazioni necessarie a definire l'entità. In secondo luogo è stato necessario implementare un oggetto in grado di contenere tutte le informazioni d'interesse per una revision. Questo compito, svolto per il processo precedente dall'entità Edit Life, è la vera grande differenza tra i due processi dal punto di vista dell'implementazione.

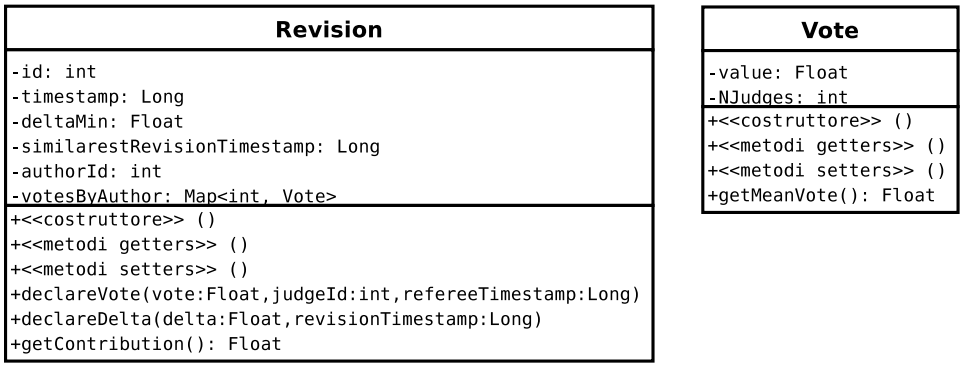
La classe *Revision*, come mostrata in Figura 4.3, contiene le informazioni necessarie per identificare una versione all'interno di una pagina, cioè il suo identificatore univoco *id* e il suo *timestamp*.

Inoltre vi sono due variabili, *deltaMin* e *similarestRevisionTimestamp*,

Algorithm 1: Algoritmo di calcolo della longevità di un intervento

```
1  begin
2      globalAuthorSet ← ∅
3      foreach (p ∈ setOfPages) do
4          localAuthorSet ← ∅
5          foreach (line ∈ p) do
6              el ← new EditLife(line)
7              if (el.getAuthor() ∉ globalAuthorSet) then
8                  recupera le informazioni sull'autore da el
9                  inserisci l'autore nell'insieme globalAuthorSet
10             if (el.getAuthor() ∉ localAuthorSet) then
11                 recupera le informazioni sull'autore da el
12                 inserisci l'autore nell'insieme localAuthorSet
13             alpha ← el.getAvgSpecQ()
14             delta ← el.getDelta()
15             a ← globalAuthorSet.element(el.getJudgedAuthor())
16             a.incrementEditLongevity(alpha, delta)
17             a.incrementEditCount()
18             a ← localAuthorSet.element(el.getJudgedAuthor())
19             a.incrementEditLongevity(alpha, delta)
20             a.incrementEditCount()
21         write(p, localAuthorSet, ordinato per contributo decrescente)
22     write(globalAuthorSet, ordinato per contributo decrescente)
23 end
```

Figura 4.3: Diagramma delle classi Revision e Vote



che verranno utilizzate per memorizzare rispettivamente la differenza, in numero di parole, tra la versione attuale e quella che risulterà essere la sua più simile e il timestamp di quest'ultima. La variabile *authorId* memorizza al suo interno l'identificativo dell'autore della revision. Infine la mappa *votesByAuthor* tiene traccia di tutti i voti ricevuti da una revision raggruppati per autore.

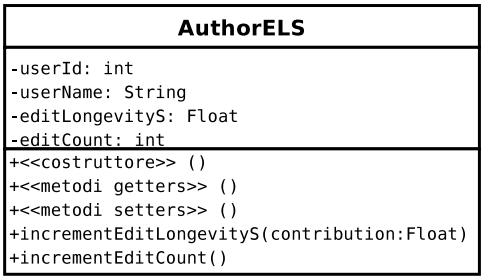
La classe *Vote*, sempre rappresentata in Figura 4.3, colleziona nella variabile *value* la somma di tutti i voti di uno stesso tipo e in *NJudges* il numero dei voti rappresentati e restituisce il valore del voto medio con il metodo *getMeanVote*.

I metodi importanti per la classe *Revision* sono tre. Il metodo *declareVote* è usato dal chiamante per dichiarare un voto ricevuto da una certa revisione. Non è importante solo il valore di questo voto, ma anche il suo autore e il timestamp della revision usata come giudice. Si ricorda infatti che per la metrica che si vuole calcolare i voti saranno raggruppati per utente e saranno accettati solo se il giudice sarà una revision successiva a quella più simile per quella attuale. Il metodo *declareDelta* viene utilizzato nel modo seguente. Quando esso riceve un valore che rappresenta la differenza in numero di parole tra una revision e quella corrente, confronta questo valore con l'attuale *deltaMin*. Se il valore passato risulta essere inferiore a quello di *deltaMin* significa che c'è una revision più simile a quella in oggetto, e di conseguenza sarà necessario aggiornare questa variabile e quella di *similarestRevisionTimestamp*. Si presti attenzione al fatto che prima di dichiarare i voti è necessario trovare la versione più simile, altrimenti non si può decidere se essi vadano o meno accettati. Infine il metodo *getContribution* aggrega i voti di tutti gli utenti e li restituisce mediati in modo coerente alle definizioni date nel Capitolo 3.

Un'ultima classe implementata per questo processo è quella chiamata *AuthorELS*, che modella un autore in modo molto simile a quella di *AuthorEL*, presentata precedentemente. Come si può vedere in Figura 4.4, le due entità sono veramente molto simili e pertanto non si ritiene interessante approfondirne la descrizione. L'unica novità di questa classe è il metodo *incrementEditLongevityS*, che incrementa l'attributo *editLongevityS* del valore passato come parametro.

Anche l'implementazione di questo processo è molto simile a quella relativa alla metrica di *longevità dell'intervento*, come si può notare osservando lo pseudocodice dell'Algoritmo 2. La sua differenza principale con l'Algoritmo 1 risiede nel calcolo di un insieme di Revision per ciascuna pagina che si attua in due scansioni del suo file di log. La prima cerca e valuta le informazioni sulla versione più simile per ciascuna Revision, mentre la secon-

Figura 4.4: Diagramma della classe AuthorELS



da ne calcola il valore di qualità. Dopo questa fase preliminare l'algoritmo prosegue in modo analogo a quello già descritto.

4.3 Processo di selezione dei coautori

Una volta prodotte le classifiche degli utenti per ogni pagina, il processo di selezione dei coautori non presenta particolari difficoltà. Esso è schematizzato con un diagramma di flusso in Figura 4.5. Il processo deve popolare l'insieme dei Top User per ogni pagina, verificando di volta in volta i vincoli teorici spiegati nel Capitolo 3. L'algoritmo più efficiente risulta essere quello che considera il prossimo utente con contributo maggiore e ne valuta l'inserimento nell'insieme dei Top User. I controlli da effettuare per rispettare i vincoli sono due.

Il primo dev'essere fatto a priori dell'inserimento e risulta rispettato se il contributo dell'utente corrente supera la soglia minima di parole W.

Il secondo controllo va fatto dopo l'inserimento e consiste nel verificare che il contributo totale dell'insieme dei Top User attuale abbia o meno superato la soglia percentuale decisa all'inizio dell'algoritmo. Questa soglia dipende dal contributo anonimo pagina per pagina e dunque, prima di iniziare la selezione, è necessario recuperare questo dato, problema che si risolve semplicemente andando a cercare nella lista degli utenti della pagina il contributo A dell'utente anonimo che rappresenta il totale degli utenti non registrati.

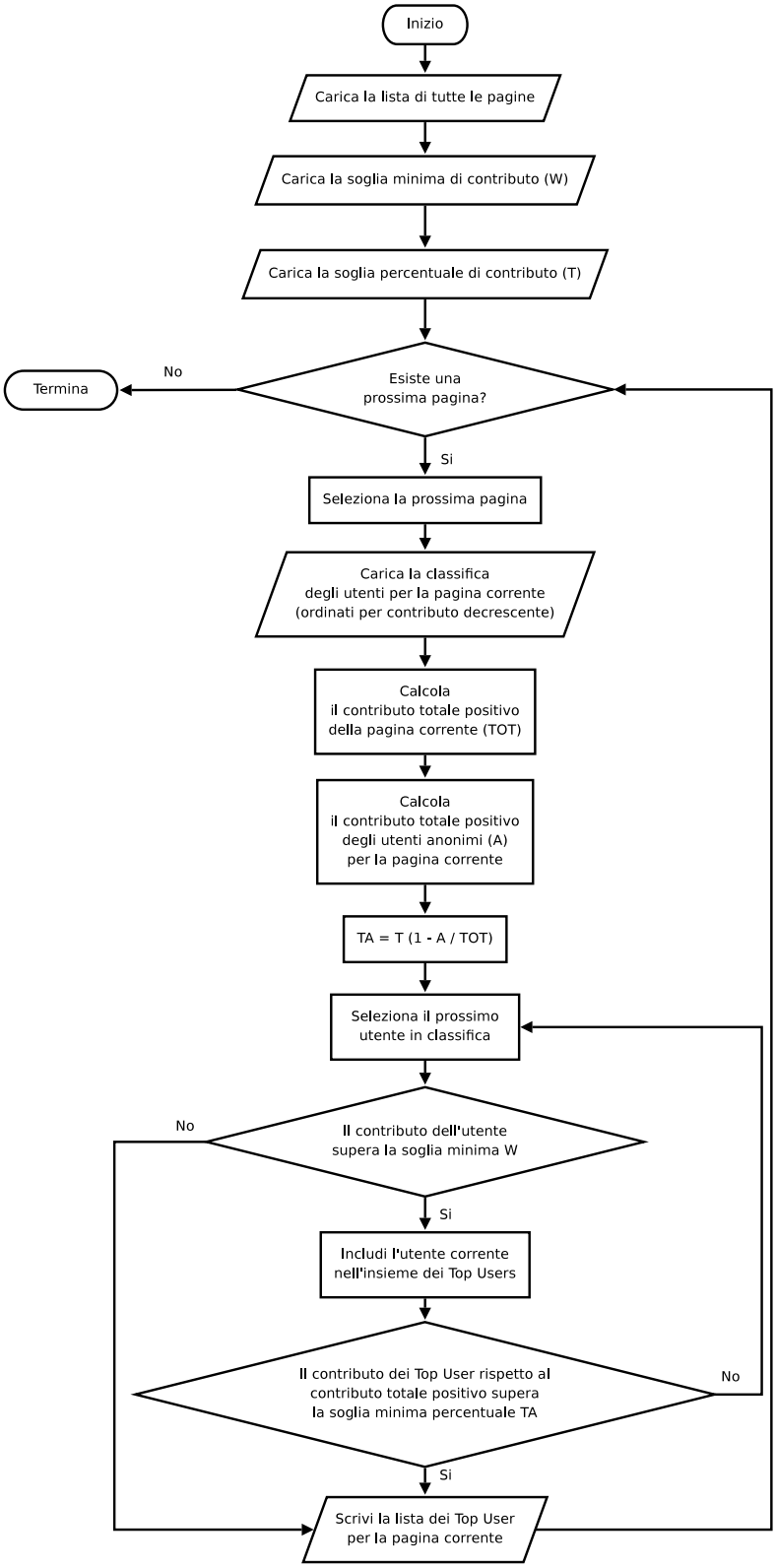
Si noti che è sufficiente che venga violato uno solo di questi vincoli per decidere che gli utenti successivi non potranno essere dei Top User. Prima di passare alla pagina seguente è infine necessario salvare l'elenco dei Top User.

Nella descrizione del processo si è volutamente omessa la parte riguar-

Algorithm 2: Algoritmo di calcolo della longevità di un intervento valutata rispetto alla sua versione più simile

```
1  begin
2      globalAuthorSet ← ∅
3      foreach (p ∈ setOfPages) do
4          revisionSet ← ∅
5          foreach (line ∈ p) do
6              calcola la versione più simile per ciascuna revision e popola
               l'insieme revisionSet
7          foreach (line ∈ p) do
8              calcola la qualità per ciascuna revision
9          localAuthorSet ← ∅
10         foreach (r ∈ revisionSet) do
11             if (r.getAuthor() ∉ globalAuthorSet) then
12                 recupera le informazioni sull'autore da r
13                 inserisci l'autore nell'insieme globalAuthorSet
14             if (r.getAuthor() ∉ localAuthorSet) then
15                 recupera le informazioni sull'autore da r
16                 inserisci l'autore nell'insieme localAuthorSet
17             incrementELS ← r.getContribution()
18             a ← globalAuthorSet.element(r.getJudgedAuthor())
19             a.incrementEditLongevityS(incrementELS)
20             a.incrementEditCount()
21             a ← localAuthorSet.element(r.getJudgedAuthor())
22             a.incrementEditLongevityS(incrementELS)
23             a.incrementEditCount()
24         write(p, localAuthorSet, ordinato per contributo decrescente)
25     write(globalAuthorSet, ordinato per contributo decrescente)
26 end
```

Figura 4.5: Diagramma di flusso del processo di selezione dei coautori



dante il calcolo delle statistiche, banalmente ricavate a partire da variabili intermedie. Come spiegato nel Capitolo 3 queste statistiche verranno analizzate sia a livello globale che raggruppandole per tipologia di pagina. In particoalre si è fatto uso della lista dei Featured Article di un wiki per separare le statistiche relative a questo tipo di pagina dalle altre.

Le analisi statistiche sono state effettuate con il pacchetto software[2] *R*, in grado di importare facilmente dati in differenti formati, di calcolare statistiche per essi, quali correlazione, valori minimi, massimi, media e mediana e di rappresentarli graficamente. Il software si è dimostrato molto efficiente e ottimamente documentato.

4.4 Processo di costruzione di una Social Network

Per realizzare la Rete Sociale di un wiki si è utilizzata la libreria *Java Universal Network/Graph* (JUNG), software[3] che fornisce un comodo strumento di alto livello per la modellazione di dati rappresentabili come un grafo. La versione utilizzata è la 2.0 beta1 che ha tra le sue caratteristiche l'utilizzo dei tipi generici, che consentono di realizzare una rete utilizzando dei qualsiasi oggetti Java come nodi o archi.

Il processo implementa l'Algoritmo 3 che rappresenta la banale realizzazione dei concetti teorici espressi nel Capitolo 3. Per ogni pagina si recupera la lista dei suoi Top User, considerati i suoi coautori, e li si inserisce come nodi nel grafo. Tutti i coautori di una pagina sono collegati da un arco pesato. Il peso di un arco tra due nodi viene incrementato di uno ogni qualvolta due utenti già direttamente connessi nella rete si trovano nuovamente a essere coautori della medesima pagina.

Il processo di costruzione della rete calcola anche alcune statistiche, non inserite nell'Algoritmo 3 per evitare di rendere poco leggibile lo pseudocodice. Esse tuttavia sono molto semplici da calcolare e sono: il *numero di pagine esaminate*, il *numero di articoli con almeno un Top User*, il *numero di articoli con almeno due Top User*, il *numero totale di Top User distinti*, il *numero di autori nella rete*, il *numero medio di articoli per autore*, il *numero medio di autori per articolo* e il *numero medio di autori per articolo considerando solo gli articoli con almeno due autori*. Si noti che tutte queste statistiche non sono proprietà della rete ma possono essere calcolate solo in questo processo di costruzione.

[2]Disponibile liberamente al sito `http://www.r-project.org/` .

[3]Disponibile con licenza open source al sito `http://jung.sourceforge.net/`

Algorithm 3: Algoritmo di costruzione della Social Network dei coautori di un wiki

```
1  begin
2      Graph g ← nuovo grafo, non diretto, pesato
3      foreach (p ∈ setOfPages) do
4          List topUsers ← getTopUsers(p)
5          for (i = 0; i < size(topUsers); i++) do
6              for (j = i + 1; j < size(topUsers); j++) do
7                  if (g contiene arco(i, j)) then
8                      incrementa il peso dell'arco(i, j) di uno
9                  else
10                     if (g non contiene nodo(i)) then
11                         aggiungi a g, nodo(i)
12                     if (g non contiene nodo(j)) then
13                         aggiungi a g, nodo(j)
14                     crea l'arco(i, j) con peso uno
15 end
```

Infine il grafo viene salvato su disco in formato Pajek[4]. Questo formato è di tipo testuale ed è stato scelto per la sua semplicità e la sua diffusione nell'ambito degli studi sulle Reti Sociali. Molti altri pacchetti software sono in grado di importarlo e si è ritenuta quest'opportunità decisamente importante.

Sebbene JUNG abbia un metodo per esportare un grafo nel formato Pajek, si è scelto di implementarne uno adatto alle esigenze del lavoro. Innanzitutto questo nuovo metodo gestisce correttamente l'esportazione di grafi pesati. In secondo luogo esso verifica sulle liste degli Amministratori e dei Bot estratti da Wikipedia la tipologia di utente rappresentato da ciascun nodo. Nel caso in cui esso sia un amministratore o un Bot, il metodo associa al nodo un colore rispettivamente *verde* o *blu*. Se esso potrebbe essere un Bot, poiché non appartiene alla lista dei Bot ma il suo nome utente termina con la stringa "bot", il colore assegnato è il *marrone*. Altrimenti il colore di un utente non appartenente a nessuna categoria particolare è *rosso*. È chiaro come queste informazioni tornino utili principalmente per la visualizzazione

[4]Per ulteriori informazioni su questo formato si consulti il sito Web `http://pajek.imfm.si/doku.php` .

della rete, ma esse possono essere sfruttate anche come label di un nodo per valutazioni di tipo quantitativo.

Le misure sulla rete sono invece state fatte con la libreria[5] del pacchetto statistico *R* chiamata *Igraph* nella sua versione 0.5. Questa libreria dispone innanzitutto di un parser di reti in formato Pajek, nonché di numerose funzioni molto ben documentate in grado di calcolare proprietà macroscopiche di una rete e i valori di centralità dei suoi nodi. Per le classifiche prodotte da questi studi si è deciso di applicare la stessa tecnica descritta per il processo di calcolo del contributo degli utenti. Per i primi 100 classificati si calcolerà, sempre tramite le liste degli amministratori e dei Bot, quanti appartengono a ciascuna categoria. Infine la libreria *Igraph* dispone di differenti algoritmi per la visualizzazione di reti complesse. Sebbene alcune di queste funzioni fossero disponibili direttamente in JUNG, risultati sperimentali hanno dimostrato una maggior efficienza della libreria *Igraph*.

[5]Disponibile liberamente al sito `http://igraph.sourceforge.net/` .

Capitolo 5

Risultati sperimentali

L'obiettivo di questo capitolo è quello di mostrare e interpretare i risultati dell'esecuzione dei quattro processi di analisi di Wikipedia descritti nel Capitolo 3. Per questo la struttura dei due capitoli è analoga ma, piuttosto che concentrarsi sulla descrizione del metodo, questo fornirà e motiverà i valori utilizzati per i parametri di ciascun processo e ne descriverà gli output prodotti. Si ricorda che, data la generalità di ciascun processo, gli esperimenti potranno essere riprodotti su un qualsiasi wiki basato sulla piattaforma MediaWiki, con la possibilità dunque di studiare differenti comunità e quindi confrontarle con quelle analizzate in questo lavoro. Facendo variare i parametri dei singoli processi si potrà invece cercare di ottenere nuovi risultati ai quali assegnare interpretazioni differenti.

5.1 Estrazione di informazioni dalla cronologia di Wikipedia

La prima decisione da prendere per analizzare un wiki con il processo descritto nel Capitolo 3 è proprio quello di scegliere l'oggetto dello studio.

Essendo questa prima fase di analisi abbastanza onerosa dal punto di vista computazionale si è approfittato della disponibilità degli autori del progetto WikiTrust, i quali hanno fornito i risultati intermedi (*log*) prodotti dal loro software su alcune versioni di Wikipedia.

Il primo log riguarda la versione in lingua italiana di Wikipedia relativa all'11 Dicembre 2005. Un primo studio verrà quindi svolto su questa versione con lo scopo di confrontare le due metriche di longevità di un intervento e della sua variante appena introdotta. Questa versione è opportuna per le sue ridotte dimensioni che rendono i successivi processi di analisi, da eseguirsi una volta per ciascuna metrica, più rapido. Essa conta un totale di 93 mila

voci, conteggiate tra tutte le pagine del namespace principale escludendo le pagine di redirect e tutte quelle pagine con meno di due interventi consecutivi da parte di autori distinti. Per queste ultime le metriche di qualità non possono essere applicate poiché c'è bisogno di almeno tre versioni di autori distinti per valutare la qualità di una singola revision. La perdita delle statistiche su queste pagine non è da considerarsi grave, poiché esse sono di scarso interesse per la valutazione delle relazioni tra gli utenti di Wikipedia.

Sempre con questo obiettivo è stato scelto di analizzare un secondo log, relativo alla versione in lingua italiana di Wikipedia del 17 Marzo 2008. Di essa però sono state considerate solamente un numero di pagine pari al 25% del totale, estratte in maniera uniforme. Il dataset è tutto sommato di grandi dimensioni, contando circa 127 mila voci, e per questo si ritiene che le valutazioni effettuate avranno comunque validità generale per quanto riguarda il confronto tra le due metriche.

Tra i log a disposizione ne sono stati scelti inoltre due in grado di favorire un secondo tipo di studio. Esso verrà effettuato per confrontare le caratteristiche di due versioni di Wikipedia molto differenti tra loro: quella in lingua italiana e quella in lingua inglese.

Gli utenti della prima sono per la maggior parte concentrati nel territorio italiano, dove essa è parlata principalmente da circa 60 milioni di persone madrelingua. La lingua inglese invece è parlata, in gran parte anche da persone non madrelingua, in tutto il mondo ed è considerata la lingua franca della comunità scientifica ed economica. Si stima che le persone di madrelingua inglese siano circa 309 milioni e che questa lingua sia parlata in totale da 1.5 miliardi di persone in tutto il mondo (Gordon, 2005). Questo influenza parecchio le dinamiche della comunità delle rispettive enciclopedie e in particolare la versione inglese di Wikipedia può beneficiare di molti più contributi distribuiti più o meno uniformemente in tutto il mondo. Si noti che non è importante solo il numero dei potenziali contributori, che sicuramente influisce sulle dimensioni dell'enciclopedia, ma anche la grande ricchezza di opinioni portata da persone con culture così eterogenee tra loro. Oltre a ciò le differenze tra le due versioni di Wikipedia riguardano anche il fatto che quella inglese è nata un anno prima di quella italiana e per questo la sua comunità può considerarsi, per questi primi anni di vita, più matura.

Dal punto di vista tecnico la versione di Wikipedia in lingua italiana analizzata risale al 22 Maggio 2007 e conta poco più di 301 mila voci. La versione analizzata per la lingua inglese invece risale al 6 Febbraio 2007 e, con solo un anno di vita in più rispetto a quella in italiano, conta poco meno di 2 milioni di pagine. La scelta di studiare due versioni così distanti è motivata dal fatto che in questo modo si potranno osservare similarità e

differenze tra di esse.

I log di queste due versioni tuttavia non sono così ricchi di informazioni come quelli delle versione italiane del 2005 e del 2008. In particolare sono privi delle informazioni relative alle distanze temporali e in numero di versioni tra le entità EditInc che sono essenziali per calcolare la metrica di longevità di un intervento valutata rispetto alla sua versione più simile. Per questo il confronto tra queste due versioni di Wikipedia utilizzerà la più semplice metrica di longevità di un intervento. I risultati saranno comunque interessanti, a patto di tener conto dei limiti evidenziati nel Capitolo 3, per il fatto che le due versioni saranno confrontate rispetto alla stessa metrica.

Un parametro di questo processo è la dimensione m della finestra entro la quale cercare i giudizi di qualità per ciascuna revision. Da un lato la finestra non deve essere troppo piccola poiché non sempre le prime versioni successive ad una data riescono ad esprimere un buon giudizio su di essa. Si pensi, a titolo di esempio, a una versione contenente informazioni sbagliate. Gli interventi immediatamente successivi possono non correggere le inesattezze e quindi non assegnare un valore di qualità negativo a essa. Si potrebbe pensare che maggiori sono le dimensioni della finesta, migliore sarà la precisione dei risultati ottenuti. Da un altro punto di vista però la finestra non può essere di dimensioni troppo elevate poiché è naturale che una pagina cambi nel tempo, anche se i contenuti sono esatti al momento del suo inserimento. Non si vuole dunque penalizzare l'autore di una delle prime revision solo perché dopo molti interventi il suo contenuto non è stato mantenuto.

Il valore scelto per le dimensioni della finestra è pari a 10 e si ritiene che esso sia adatto a valutare in modo corretto la qualità di una revision secondo i criteri appena spiegati. Da studi precedenti è infatti noto che normalmente gli atti di vandalismo più notevoli vengono individuati e corretti dagli amministratori di Wikipedia in tempi molto brevi (Viegas et al., 2004, 2007).

Questo valore della finestra è anche quello che verrà utilizzato per cercare, tra le revision precedenti, quella più simile a quella data. La motivazione di questa scelta è analoga a quella per la scelta della finestra temporale entro la quale una revision può ricevere giudizi.

5.2 Calcolo del contributo dei partecipanti

Questo sottoprocesso si occupa di analizzare ogni pagina del wiki oggetto dello studio e di calcolare per essa la classifica dei contributori secondo le

differenti metriche descritte nel Capitolo 3. Chiaramente esso dipende fortemente dalla metrica di contributo considerata che può quindi considerarsi l'unica sua variabile. Come già detto si è scelto di calcolare, per ogni pagina, la classifica dei suoi autori secondo le due metriche di longevità di un intervento e di longevità di un intervento rispetto alla sua versione più simile. Questi dati saranno l'input delle fasi successive e verranno dunque analizzati in modo automatico dal sottoprocesso di selezione dei coautori.

In questa sezione si preferisce analizzare e confrontare le classifiche globali prodotte dalle due metriche di valutazione del contributo di un utente, con lo scopo di cogliere le differenze tra esse. Inoltre, solo in questo tipo di studio, verrà calcolata anche la classifica globale del conteggio degli interventi (*edit count*), che verrà usata come punto di riferimento per il confronto tra le due metriche. Le classifiche delle prime venti posizioni per le differenti metriche calcolate sono consultabili nell'allegato A.

5.2.1 Il conteggio degli interventi

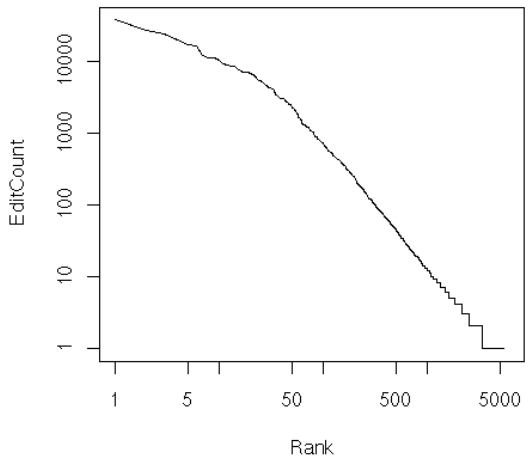
Wikipedia Italiana, versione 2005

Il conteggio degli interventi per la versione in lingua italiana di Wikipedia del 2005, come si può in parte verificare nella Tabella A.1 in allegato, mostra chiaramente come ai primi posti ci siano degli attori non umani. I Bot, che all'epoca risultano essere 18, ricoprono ben 14 dei primi 100 posti di questa classifica, sintomo il fatto che essi compiono un grandissimo numero di interventi di non precisata qualità o quantità. Anche gli amministratori, su un totale di 37, ricoprono ben 26 posti della classifica. È interessante comunque notare come i primi due posti siano assegnati a due Bot, *Gacbot* e *Zerobot*, il primo dei quali distacca il primo utente umano *Snowdog* di ben 11 mila interventi. Altra interessante considerazione è quella sull'utente *Twice25*, che pur non ricoprendo alcuna carica di amministrazione all'interno di Wikipedia riesce a essere sesto nella classifica degli interventi.

Per quanto riguarda l'influenza degli utenti anonimi, secondo questa metrica essa può essere quantificata pari al 14.96% del totale, dunque molto ridotta.

Si visualizza ora, in Figura 5.1, l'andamento della classifica del numero di interventi mostrando entrambi gli assi in scala logaritmica. In questo modo si vede che questa curva segue l'andamento di due Zipf's Law con pendenza differente. Come già osservato da (Almeida et al., 2007) per la versione inglese di Wikipedia, questo andamento segnala la presenza di due distinti gruppi di utenti all'interno della versione di Wikipedia analizzata. Un primo gruppo, che termina al variare della pendenza della curva, di utenti molto

Figura 5.1: Andamento logaritmico del numero di interventi di un utente in funzione della sua posizione in classifica (Wikipedia in italiano, 2005).



attivi dal punto di vista degli interventi e un secondo gruppo più moderato. La dimensione del primo gruppo, come si può notare in Figura 5.1, è circa di 40 utenti.

Wikipedia Italiana, versione 2008

Il numero di Bot agenti sulla versione italiana di Wikipedia è aumentato rispetto al 2005. Si è passati da 18 a 163 Bot dichiarati e quindi ci si aspetta che essi vengano messi in risalto dalla metrica del conteggio degli interventi. Come al solito in allegato si può trovare la Tabella A.18 relativa ai primi 20 posti della classifica. Infatti ben 40 dei primi 100 posti della classifica sono ricoperti da Bot. Inoltre i primi 11 posti sono tutti assegnati ad agenti software, eccezion fatta per l'utente *FlaBot*. Anche intuitivamente ci si può rendere conto che in realtà si tratta di un errato riconoscimento. Solo andando a vedere la pagina relativa all'utente tuttavia si è certi del fatto che si tratta di un Bot non incluso, per qualche motivo, nella lista di quelli dichiarati. Il successivo e reale utente umano nella classifica è l'amministratore *Snowdog*, che già nel 2005 era il primo editor escludendo i Bot. Nei primi cento posti in classifica si trovano 26 amministratori su un totale di 94. Tra i semplici utenti registrati si trova invece, in sedicesima posizione, *Gac* che

notiamo essere l'autore e manovratore di uno dei primi Bot in classifica, *Gacbot*. Se si considera che un Bot esegue generalmente i compiti assegnatigli dal suo autore o gestore ci si rende conto di come l'apporto dell'utente *Gac* sia da considerarsi molto pesante sull'intero andamento della Wikipedia italiana sino al 2008.

L'influenza degli utenti anonimi è stimata, secondo questa metrica, pari al 23.14% del totale.

Confronto tra due versioni di Wikipedia Italiana: dal 2005 al 2008

Un'osservazione interessante riguarda l'aumento degli interventi da parte degli utenti anonimi sul totale. Questo può spiegarsi con l'aumento di popolarità in Italia di Wikipedia, che per ovvi motivi, ha probabilmente attirato una maggior quantità di visitatori occasionali provenienti dagli ambienti più vari e quindi più che altro interessati a intervenire sui contenuti piuttosto che a diventare parte integrante della comunità.

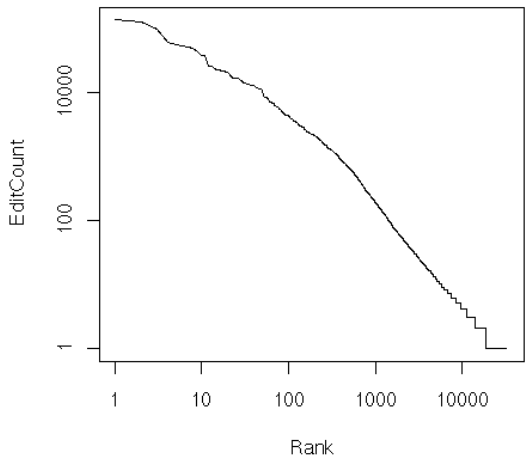
Wikipedia Italiana, versione 2007

Dal punto di vista del conteggio degli interventi la versione italiana di Wikipedia del 2007 non si discosta molto da quella del 2008 appena esaminata. I primi dieci posti della classifica sono occupati da Bot, come si può vedere nella Tabella A.12, sebbene per esserne sicuri si debba controllare sulle rispettive pagine utente di *FlaBot* ed *Eskimbot* che lo siano effettivamente. Il motivo del perché alcuni Bot non risultino inclusi nelle rispettive liste può essere il seguente. Analizzando la descrizione del loro funzionamento nelle pagine utente, ci si accorge che entrambi si occupano di collegamenti *interwiki*, cioè collegano pagine di differenti versioni linguistiche di Wikipedia che hanno lo stesso significato. Probabilmente i gestori di questi Bot non sono veri e propri utenti della Wikipedia in lingua italiana e di conseguenza non hanno dichiarato i loro software come tali in tutte le comunità all'interno delle quali essi si trovano a operare. Anche i rapporti tra tipologie di utenti tra i primi 100 editor in classifica si confermano essere simili a quelli del 2008. Vi sono 27 Bot su un totale di 77 dichiarati e 33 amministratori su un totale di 85 (38%). Il primo utente umano risulta essere sempre l'amministratore *Snowdog*, mentre il primo utente non amministratore è *Twice25* che si posiziona in ventesima posizione.

L'influenza degli utenti anonimi riserva invece qualche sorpresa. Essa risulta infatti essere pari al 13.23%, molto simile a quella del 2005.

Anche in questo caso l'andamento della classifica, su scala bilogaritmica, mostra un andamento lineare con due differenti pendenze. Il gruppo di utenti

Figura 5.2: Andamento logaritmico del numero di interventi di un utente in funzione della sua posizione in classifica (Wikipedia in italiano, 2007)



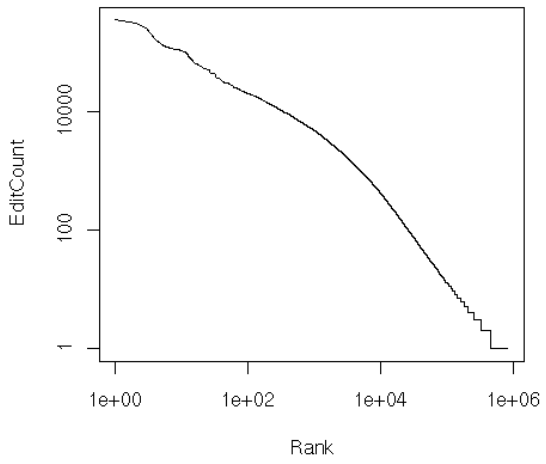
molto attivi è cresciuto sino a raggiungere circa 800 utenti, come mostrato in Figura 5.2.

Wikipedia Inglese, versione 2007

Anche per quanto riguarda la versione inglese di Wikipedia, il conteggio degli interventi vede nelle prime posizioni della classifica, come si può osservare in Tabella A.29, parecchi Bot. Nelle sue prime 15 posizioni, eccezion fatta per l'utente amministratore *Rich Farmbrough* in sesta posizione, vi sono solamente utenti non umani. Nella classifica delle prime 100 posizioni si trovano 26 Bot su un totale di 158 dichiarati. Gli amministratori nelle prime 100 posizioni sono invece 52 su un totale di 1148. Come già detto il primo amministratore in classifica è *Rich Farmbrough*, il quale è anche gestore di *SmackBot*, secondo utente con maggior numero di interventi. Anche in questo caso, come per quello dell'utente *Gac* nella Wikipedia italiana, ci si rende conto di come un utente possa operare in diversi modi non solo attraverso il proprio account, ma anche attraverso dei Bot. Il primo utente non amministratore e non Bot è invece *Bobblewik*.

La percentuale di interventi anonimi in questa versione di Wikipedia è del 24.36%, un valore non indifferente.

Figura 5.3: Andamento logaritmico del numero di interventi di un utente in funzione della sua posizione in classifica (Wikipedia in inglese, 2007)



L'analisi della curva in Figura 5.3 mostra ancora una volta un andamento di tipo rettilineo con cambio di pendenza, confermando i risultati di (Almeida et al., 2007) secondo i quali il gruppo di utenti molto attivo nella Wikipedia inglese del 2007 è composto da circa 5000 persone.

Confronto tra differenti versioni di Wikipedia: italiana e inglese

Al di la delle dimensioni, il confronto tra le versioni in lingua italiana con quella in lingua inglese di Wikipedia secondo la metrica del conteggio degli interventi mostra poche differenze.

5.2.2 La longevità di un intervento

Wikipedia Italiana, versione 2005

Già nella versione italiana di Wikipedia del 2005 si può notare, facendo riferimento alla Tabella A.2, come la presenza dei Bot ai primi posti della classifica sia decisamente inferiore rispetto alla metrica di conteggio degli interventi. Tra i primi 100 ve ne sono infatti solo 4, su un totale di 18. Inoltre il primo Bot nella classifica, *Luki-Bot*, è solamente in undicesima posizione. Questo si spiega principalmente in due modi. Innanzitutto i Bot compiono

tanti interventi, ma tra questi molti sono di piccola entità, come le correzioni ortografiche. In secondo luogo la metrica di longevità di un intervento tiene conto della qualità di un intervento. Essa è tanto più alta quanto nelle successive m revision, in questo studio pari a 10, viene conservato l'intervento. Il decadimento della reputazione dei Bot si spiega dunque col fatto che essi, pur facendo tante modifiche, spesso ricevono delle penalità poiché esse non sono perfette, ma vanno a loro volta sistemate da qualcuno. Ecco una nuova interpretazione del fatto che spesso gli utenti che gestiscono un Bot hanno un alto numero di interventi. Molti di questi potrebbero proprio essere dedicati alla correzione del lavoro del proprio software.

Il numero di amministratori tra i primi 100 posti della classifica, pari a 25 su un totale di 37, rimane molto simile a quello nella classifica degli utenti con maggior numero di interventi. Al primo posto si trova l'amministratore *Snowdog* mentre al secondo posto l'utente registrato *Twice25*.

L'apporto dato dagli utenti anonimi, sul totale degli interventi positivi, è invece pari al 5.31%. Questo dato è molto interessante perché significa che, pur essendo pochi e pur considerando le penalità per i cattivi interventi, gli interventi degli utenti anonimi hanno comunque un bilancio positivo all'interno di questa versione di Wikipedia

In ultimo si vuole considerare la correlazione tra i punteggi assegnati da questa classifica e quella del conteggio degli interventi. Essa è pari a 0.62, un valore tutto sommato alto.

Wikipedia Italiana, versione 2008

Anche in questa versione di Wikipedia si nota, riferendosi alla Tabella A.19, che i Bot hanno perso posizioni nella classifica dei maggiori contributori. Tra i primi 100 posti in classifica se ne trovano solamente 5 sui 163 dichiarati, dunque una bassissima percentuale. Gli amministratori sono invece 41 su 94. Il primo in classifica è l'amministratore *Snowdog*, mentre il primo utente non dotato di privilegi è *M7*.

Una considerazione importante da fare per questa versione di Wikipedia riguarda gli utenti anonimi. In questo caso il contributo totale dato da utenti non registrati risulta essere negativo e pari a -2.29 milioni. Considerando che il miglior contributore secondo questa metrica totalizza un punteggio di 1.12 milioni ci si rende conto come una grandissima parte degli interventi anonimi non vengano accettati. Questo non è necessariamente da considerarsi come una prova del fatto che in generale essi compiano dei danni su Wikipedia. É possibile piuttosto che la loro scarsa conoscenza delle norme della comunità li faccia intervenire in modo poco opportuno e che essi vengano dunque spesso

corretti da un punto di vista stilistico piuttosto che contenutistico. Questo aspetto tuttavia è molto difficile da approfondire con i dati a disposizione.

La correlazione dei punteggi assegnati da questa metrica con quelli del conteggio degli interventi è pari a 0.36, che indica una bassa dipendenza statistica tra le due metriche.

Confronto tra due versioni di Wikipedia Italiana: dal 2005 al 2008

In questo confronto temporale è interessante notare come il punteggio degli utenti anonimi sia passato da un valore positivo a uno molto negativo. Questo significa che la maggior parte degli interventi anonimi ha ricevuto voti negativi e cioè sono rimasti poco all'interno di Wikipedia nella loro forma originale. In realtà questo dato dice poco sulla popolazione degli utenti anonimi. Infatti per come è pensata la metrica di longevità dell'intervento è sufficiente un vandalismo di cancellazione di massa, caratterizzato cioè da una grande modifica per la pagina valutata con qualità negativa, per annullare il contributo positivo dato da molti piccoli interventi anonimi.

È anche interessante notare come, dal 2005 al 2008, la correlazione tra i punteggi di questa metrica con quelli assegnati tramite il conteggio degli interventi si sia abbassata notevolmente.

Wikipedia Italiana, versione 2007

Sebbene il primo contributore per la classifica della longevità d'intervento sia il Bot *Gacbot*, su un totale di 77 solo 6 di essi sono tra i primi 100. Molti dei primi contributori sono amministratori, ben 38 sugli 85 totali, dei quali il primo è *Snowdog* in seconda posizione. Il primo utente non amministratore è invece *.snoopy.* in sesta posizione. I primi 20 posti della classifica possono essere trovati in Tabella A.13

Anche in questo caso gli interventi anonimi totalizzano un punteggio negativo, pari a -2.74 milioni, valore confrontabile con il primo utente in classifica, che totalizza un punteggio di 3.75 milioni di contributo positivo.

La correlazione con il conteggio degli interventi in questo caso è pari a 0.52.

Wikipedia Inglese, versione 2007

Anche nella Wikipedia inglese, nonostante il primo classificato sia il Bot *AntiVandalBot*, la presenza degli utenti umani prevale tra i primi 100 posti della classifica, come si può in parte osservare in Tabella A.30. Solo 5 Bot su un totale di 158 dichiarati risultano in classifica, mentre gli amministratori

sono 73 su 1148. La presenza di un Bot nei primi posti della classifica si può spiegare principalmente con il fatto che esso, nel caso in cui si occupi di riconoscimento di vandalismi come nel caso di *AntiVandalBot*, quando ripristina una vecchia versione di un articolo riceve un credito pari alla sua dimensione anche se in realtà egli non ne è il vero autore.

Il primo amministratore in classifica risulta essere *Curps* in terza posizione, mentre il primo utente non dotato di privilegi è *Naxon* in settima posizione. L'utente *Rich Farmbrough*, sesto per numero di interventi, in questa classifica si ritrova in posizione 279. È lecito supporre dunque che, nonostante egli faccia molti interventi, questi siano di poco conto. Infatti guardando la sua pagina utente è proprio riportato il rammarico da parte dell'utente di riuscire a intervenire solo come correttore piuttosto che come scrittore di nuovi articoli[1].

Il contributo totale degli utenti anonimi è ancora una volta molto negativo, pari a -1.09 miliardi di punteggio, ben un ordine di grandezza superiore rispetto al primo classificato e quindi decisamente influente.

La correlazione con il conteggio degli interventi è invece molto bassa, pari a 0.26.

Confronto tra differenti versioni di Wikipedia: italiana e inglese

A prescindere dalle dimensioni le versioni di Wikipedia in inglese e in italiano risultano tutto sommato abbastanza simili per questo tipo di metrica.

La principale differenza risulta essere l'enorme contributo negativo da parte degli anonimi rispetto al primo in classifica nella Wikipedia in lingua inglese rispetto a quella in italiano.

Anche la correlazione con il conteggio degli interventi è significativamente differente, a dimostrazione del fatto che all'aumentare degli interventi la longevità dell'intervento si discosta sempre di più dalla prima metrica.

5.2.3 Longevità di un intervento rispetto alla sua versione più simile

Confronto tra due metriche proposte: longevità di un intervento e longevità di un intervento valutata rispetto alla sua versione più simile

Nella versione italiana di Wikipedia del 2005 le differenze tra le metriche di longevità di un intervento e di longevità di un intervento valutata rispetto

[1] "I would like to be writing great articles, but actually I'm an inveterate fixer of anything that I see that looks wrong - mainly links and copyedits"

alla sua versione più simile non sono molto accentuate. La correlazione tra i punteggi è infatti pari a 0.98. Ne consegue che anche la correlazione tra longevità di un intervento valutata rispetto alla sua versione più simile e il semplice conteggio degli interventi sarà molto vicina a quella dell'altra metrica. Infatti essa vale 0.60.

Anche i primi utenti in classifica, a cominciare dal primo *Snowdog* e dal secondo *Twice25*, mantengono posizioni se non identiche molto vicine tra loro. I primi 20 posti in classifica possono essere osservati nelle Tabelle A.2 e A.3. Le stesse considerazioni sono valide per i rapporti dei gruppi di utenti tra i primi 100 posti in classifica: sono esattamente identici a quelli della classifica considerata con la metrica di longevità di un intervento.

La grande differenza della metrica attualmente studiata rispetto a quella della longevità di un intervento è che essa assegna solo punteggi positivi. In questo modo risulta più sensato valutare il contributo complessivo degli utenti anonimi, poiché esso prende in considerazione quegli interventi di qualità. Questo si riflette nel calcolo del valore aggregato del contributo positivo degli anonimi che risulta essere pari al 18.44% del totale.

La versione del 2008 invece mostra maggiori differenze tra le due metriche proposte. Innanzi tutto la correlazione tra i valori delle due metriche è più bassa, pari a 0.79. Essa quindi è correlata in modo differente anche con il conteggio degli interventi e in particolare con un valore di 0.44.

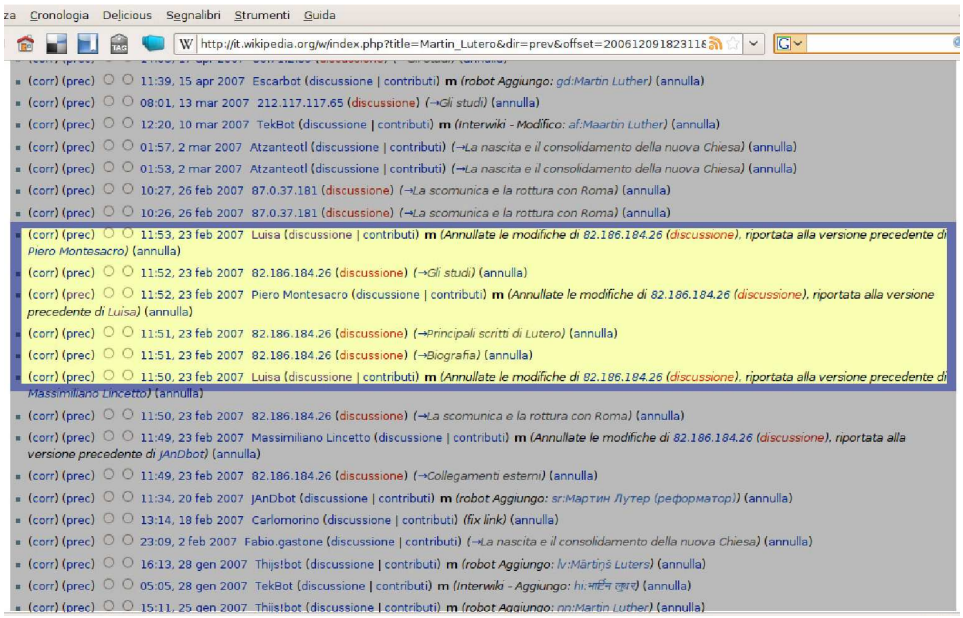
Si notano immediatamente le differenze con i primi utenti in classifica, anche osservando le Tabelle A.19 e A.20. Sebbene il primo sia sempre l'amministratore *Snowdog*, il secondo classificato per la vecchia metrica *M7* è ora in posizione 79. Al secondo posto si trova invece l'utente non dotato di privilegi *Murray*. È noto che questa metrica non dia punti a chi attua dei revert. Di conseguenza si può facilmente capire come probabilmente *M7* si sia occupato principalmente di questo genere d'interventi su Wikipedia. Sempre a causa di questo aspetto si nota come nella classifica dei primi 100 contributori ci siano molti meno aministratori rispetto a quella precedente: solo 21 invece che 41. Evidentemente molti di essi svolgono il compito di revert da vandalismi. Il numero di Bot tra i primi 100 posti della classifica aumenta invece da 5 a 7, un cambiamento tutto sommato poco rilevante. Questo è dovuto al fatto che probabilmente i Bot sono usati più per modificare o aggiungere contenuto che non per compiere dei revert.

Il contributo degli utenti anonimi calcolati con questa metrica risulta essere il 24% del totale, molto simile al medesimo rapporto calcolato sul conteggio degli interventi. È molto interessante notare che, oltre al numero di interventi anonimi, dal 2005 al 2008 è aumentato anche il contributo positivo anonimo. Se confrontati con i valori della metrica di longevità del-

l'intervento, che tiene invece conto anche delle penalità date dagli interventi di qualità negativa, ci si rende conto di come dal 2005 al 2008 sia aumentata, per la Wikipedia italiana, anche l'entità dei contributi anonimi negativi.

Prima di passare alla sezione seguente si vuole mostrare un semplice esempio tangibile delle differenze tra le due metriche. Nella pagina *Martin Lutero*, analizzata nella sua versione italiana del 2008, sono stati compiuti ripetutamente dei vandalismi da parte di un utente anonimo. Gli utenti *Luisa* e *Piero Montesacro* se ne sono accorti e hanno ripristinato le versioni della pagina non vandalizzate, come si può vedere dalla Figura 5.4.

Figura 5.4: Schermata della cronologia per la pagina Martin Lutero per la versione italiana di Wikipedia del 2008



La pagina ha 115 utenti distinti ma, come si può vedere nella Tabella 5.1, i due utenti che hanno ripristinato i vandalismi sono entrambi nelle prime 10 posizioni secondo la metrica di longevità dell'intervento. Così non è per la metrica di longevità dell'intervento valutata rispetto alla sua versione più simile, come si vede chiaramente in Tabella 5.2. Si vede quindi come per la prima metrica questo genere di intervento venga valutato molto.

Un'altra cosa interessante da notare in questo semplice esempio è come l'utente anonimo, che rappresenta cioè il dato aggregato di tutti i contributi degli utenti non registrati, sia tra le prime 10 posizioni per la metrica di longevità dell'intervento valutata rispetto alla sua versione precedente.

Tabella 5.1: Classifica dei primi 10 utenti secondo la metrica di longevità dell'intervento (EL) per la pagina Martin Lutero della versione di Wikipedia in italiano del 2008

Rank	**User Name**	**EL**
1	Piero Montesacro	3542.7388
2	Mkromer	3281.5781
3	Senpai	2236.8657
4	Gierre	1771.6506
5	.snoopy.	1564.5494
6	Rickenbacker	1420.3544
7	Barbarian	1292.1781
8	Pall Mall	1276.7683
9	Atzanteotl	1015.82825
10	Luisa	808.1222

Tabella 5.2: Classifica dei primi 10 utenti secondo la metrica di longevità dell'intervento valutata rispetto alla sua versione più simile (ELS) per la pagina Martin Lutero della versione di Wikipedia in italiano del 2008

Rank	**User Name**	**ELS**
1	Mkromer	3278.2173
2	Gierre	1327.0996
3	anonymous	1323.3905
4	Pall Mall	1073.6687
5	Atzanteotl	1012.8343
6	Gagio	609.9067
7	Acis	247.05849
8	Lp	244.54083
9	Microsoikos	155.93376
10	Hillmann	124.36364

Questo significa che alcuni anonimi hanno dato un buon contributo alla pagina. Tuttavia questo non è rilevato dalla metrica di longevità dell'intervento semplice, che per via di atti vandalici di grandi dimensioni come quello visto nell'esempio penalizza tutti gli altri contributi anonimi.

Ecco spiegati i due grandi motivi per cui si ritiene la metrica proposta in questo lavoro più adatta per caratterizzare il contributo degli utenti all'interno di una singola pagina.

5.3 Selezione dei coautori

Il sottoprocesso di selezione dei coautori valuta quali utenti per ogni pagina includere nell'insieme dei Top User in base al criterio spiegato nel Capitolo 3. Il suo input è dunque l'insieme dei dati riguardanti le classifiche di contributo calcolate per tutti gli utenti di ciascuna pagina secondo le differenti metriche. Rimangono da decidere ancora due parametri del processo. Il primo è la soglia minima di contributo W, al di sotto della quale un utente non può essere considerato un Top User. Il valore di W scelto è pari a 10. Il secondo parametro invece è la soglia percentuale T da considerarsi in relazione al contributo totale positivo di tutti gli utenti all'interno della pagina. Quando l'insieme dei Top User avrà superato la soglia T esso potrà considerarsi completo e nessun'altro utente potrà esservi ulteriormente aggiunto. Il valore scelto per T è pari al 50% del totale dei contributi positivi. In questo modo l'insieme dei Top User sarà caratterizzato dall'aver contributo ad almeno metà del lavoro considerato utile per la pagina. Questo tuttavia sarà vero solo nel caso in cui la pagina non abbia contributi anonimi. Nel caso in cui essi siano presenti questa soglia verrà abbassata in funzione della percentuale del contributo anonimo, come spiegato nel Capitolo 3.

5.3.1 Considerazioni locali

Il processo di selezione dei coautori viene svolto in automatico su tutte le pagine. Tuttavia si vuole presentare un esempio in grado di far comprendere meglio come pagine apparentemente uguali risultino invece differenti dal punto di vista della ripartizione del contributo positivo da parte dei suoi partecipanti. Si considerino le due pagine della versione inglese di Wikipedia del 2007 relative al linguaggio di programmazione *C++* e al virus *HIV*. Entrambe possono considerarsi molto simili sia per numero di interventi che per numero di utenti distinti che hanno partecipato alla loro stesura. La prima infatti conta 1225 interventi, mente la seconda 1282. Il numero di utenti distinti della prima è invece 434, mentre quelli della seconda sono 415. Dal

punto di vista del contributo degli utenti anonimi si è voluto scegliere un caso semplice nel quale, per entrambe le pagine, esso è negativo e di conseguenza non va a diminuire la soglia del 50% di contributo positivo oltre la quale fermarsi nella selezione dei Top User. Si noti ancora una volta che un valore nullo del contributo anonimo positivo non vuole necessariamente dire che in quella pagina sono intervenuti pochi utenti non registrati, bensì che essi hanno prodotto tanto contenuto considerato utile dagli interventi successivi quanto quello considerato inopportuno. Infatti le statistiche dicono che le due pagine hanno un numero di interventi anonimi pari rispettivamente al 28% e al 31% del totale.

L'insieme dei Top User per le due pagine è tuttavia decisamente diverso, come si può osservare nelle Figure 5.5 e 5.6. La prima ha un utente che da solo totalizza ben più del 50%, precisamente il 58.83%. La seconda invece ha bisogno dei primi 10 utenti per raggiungere la soglia del 52% del contributo totale.

Figura 5.5: Diagramma a settori circolari rappresentante la porzione di contributo di ogni utente alla pagina del linguaggio di programmazione C++

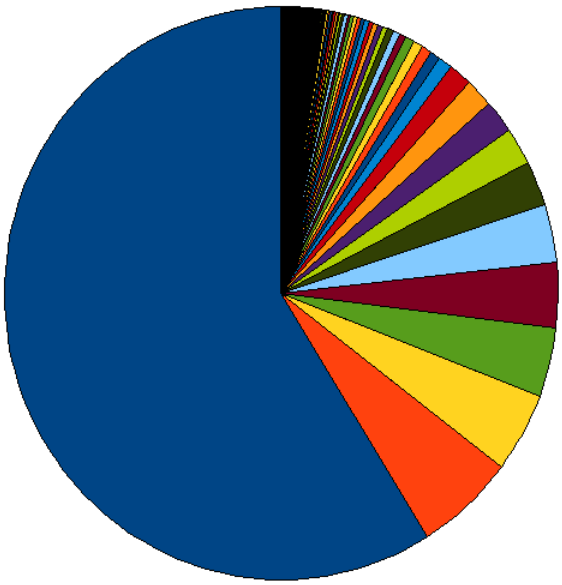
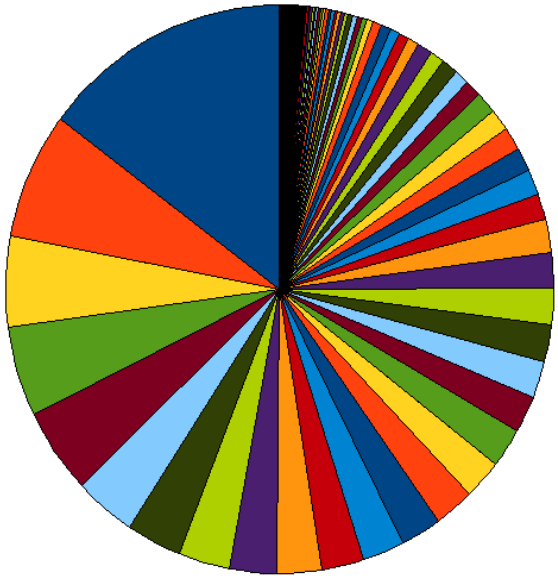


Figura 5.6: Diagramma a settori circolari rappresentante la porzione di contributo di ogni utente alla pagina del virus HIV



5.3.2 Considerazioni globali

Analisi sulle statistiche del processo di selezione dei coautori

Avendo a disposizione le statistiche sul processo di selezione dei coautori descritto nel capitolo 3 risulta a questo punto interessante fare qualche considerazione su di esse. Per non allungare eccesivamente la trattazione e per il fatto che la versione del 2008 è stata ottenuta campionando casualmente delle pagine, si è scelto di descrivere in questa sede solamente le versioni di Wikipedia del 2005 per le due metriche differenti e del 2007 nelle sue due versioni in italiano e in inglese. Innanzitutto, facendo riferimento alle Tabelle 5.3 e 5.4, si può notare come le differenti versioni di Wikipedia analizzate, quella italiana del 2005 e quella inglese e italiana del 2007, abbiano differenti valori nel numero di interventi medio e nel numero medio di utenti distinti per pagina. Questi valori evidenziano la crescita di Wikipedia nella sua versione in italiano e la maggiore dimensione, a parità di anno, di quella in inglese. Anche i valori massimi del numero di interventi per pagina hanno simile andamento, risultando pari a 447 per la versione del 2005, a 1310 e a 28525, rispettivamente per la versione italiana e inglese del 2007. Per quanto riguarda il contributo totale positivo si osserva come anche questo possa es-

Tabella 5.3: Confronto tra le statistiche relative alle pagine delle versioni del 2005 di Wikipedia in italiano usando le metriche di longevità dell'intervento (EL) e di longevità dell'intervento valutata rispetto alla sua versione più simile (ELS).

	IT2005EL	**IT2005ELS**
Edit Count	6.64	6.64
Distinct Users	4.9	4.9
Total Positive Contribution	156.53	184.7
Top Users	0.40	0.56
Top User Edits %	11.85%	9.23%
Top Users %	6.75%	7.72%
Top User Contribution %	26.39%	35.5%
Anonymous Edit Count %	9.94%	9.94%
Anonymous Contribution %	6.86%	9.88%

Tabella 5.4: Confronto tra le statistiche relative alle pagine delle versioni del 2007 di Wikipedia in italiano e in inglese (usando la longevità di un intervento come metrica).

	IT2007EL	**EN2007EL**
Edit Count	10.48	20.81
Distinct Users	7.5	11.98
Total Positive Contribution	265.1	953.2
Top Users	0.68	0.8
Top User Edits %	17.02%	13.42%
Top Users %	10.14%	7.81%
Top User Contribution %	40.67%	36.23%
Anonymous Edit Count %	8.13%	15.25%
Anonymous Contribution %	7.08%	11.27%

sere, a parità di metrica, un buon indicatore di dimensione del lavoro svolto all'interno di un wiki.

Confrontando le due metriche invece si noti come, sul medesimo wiki, la longevità dell'intervento conti un numero di punti leggermente inferiore alla sua variante. Questo è sicuramente dovuto al fatto che questa metrica conta anche delle penalità per chi interviene in modo non opportuno, il che mantiene più basso il valore totale del contributo positivo. D'altro canto si ricordi che la metrica di longevità di un intervento valutata rispetto alla sua versione più simile tende a minimizzare la stima delle dimensioni di un contributo, proprio perché considera sempre le differenze con la versione più simile a esso. Di conseguenza i valori della metrica sono mantenuti tutto sommato vicini a quelli della longevità dell'intervento.

Vedendo un così basso numero medio di Top User per pagina si capisce come molte pagine ne abbiano zero, secondo l'algoritmo impostato con i valori di soglia spiegati nella sezione precedente. Questo renderà interessante un'ulteriore analisi, su quelle pagine che hanno almeno un Top User.

Anche le altre statistiche relative ai Top User assumono quindi un significato fortemente influenzato dal gran numero di pagine che secondo il processo di selezione non hanno utenti molto influenti. Si ricorda che una pagina può essere priva di Top User nel caso in cui nessuno dei suoi partecipanti abbia oltrepassato la soglia minima di contributo W oppure nel caso limite in cui tutti i suoi contributi siano anonimi. Quest'ultimo caso si è verificato per la maggior parte in pagine di dimensioni molto ridotte.

Risulta invece interessante fare delle osservazioni sul contributo anonimo per pagina nelle differenti versioni analizzate. Per la versione di Wikipedia in italiano si vede come all'aumentare delle dimensioni, e probabilmente della popolarità del wiki, aumenti leggermente anche il contributo anonimo, nonostante il numero di interventi sia di poco inferiore. Si osserva anche come la metrica di longevità di un intervento semplice, confrontata con la sua variante, assegni un contributo inferiore agli utenti anonimi dello stesso wiki, quello della Wikipedia italiana del 2005. Anche questo risultato è spiegabile dal fatto che la metrica di longevità di un intervento valutata rispetto alla sua versione più simile non penalizza gli autori di interventi che non verranno mantenuti dagli utenti successivi.

Si osserva infine che la versione in inglese di Wikipedia, rispetto a quella italiana dello stesso anno, riceve parecchi interventi anonimi per pagina in più, che si traducono anche in un maggior contributo positivo.

Tabella 5.5: Confronto tra le statistiche relative alle pagine con almeno un Top User nelle versioni del 2005 di Wikipedia in italiano usando e due metriche di longevità di un intervento (EL) e di longevità di un intervento valutata rispetto alla sua versione più simile (ELS).

	IT2005EL	**IT2005ELS**
Pages	33708 (52.16%)	31252 (48.36%)
Edit Count	13.23	13.23
Distinct Users	8.58	8.58
Total Positive Contribution	420.76	367.58
Top Users	1.13	1.16
Top User Edits %	24.65%	19.09%
Top Users %	18.77%	15.97%
Top User Contribution %	73.33%	73.40%
Anonymous Edit Count %	11.93%	11.93%
Anonymous Contribution %	8.91%	9.7%

Analisi sulle statistiche filtrate del processo di selezione dei coautori

Si vogliono ora analizzare le statistiche di quelle pagine con almeno un Top User, perché proprio queste saranno quelle analizzate dal sottoprocesso successivo. I risultati sono riassunti nelle Tabelle 5.5 e 5.6. Filtrando dalle statistiche tutte le pagine per le quali non è stato possibile trovare un Top User, il primo fenomeno che si può osservare è proprio quello per il quale esse sono all'incirca metà del totale. Questa perdita non è da considerarsi di poco conto, tuttavia si ritiene più grave cercare di trovare forzatamente degli utenti principali anche in quei casi dove la decisione sarebbe difficilmente valutabile, piuttosto che escludere dai conteggi quelle pagine di ridotte dimensioni o totalmente scritte da utenti anonimi.

Senza di esse si nota come non solo il numero medio di interventi, ma anche il numero medio di utenti distinti per pagina, siano più alti. Stesso dicasi per il contributo totale positivo, il quale però mostra una differenza nel confronto tra le due metriche per la versione italiana di Wikipedia del 2005. Nelle statistiche "filtrate" il contributo totale calcolato con la metrica di longevità dell'intervento è più alto rispetto alla sua variante, in opposizione a quanto accade con le stesse statistiche complete. Questo fa vedere come nelle pagine con almeno un Top User la prima metrica faccia raggiungere valori in media più alti.

Nonostante la rimozione delle pagine senza utenti principali, il numero

Tabella 5.6: Confronto tra le statistiche relative alle pagine con almeno un Top User nelle versioni del 2007 di Wikipedia in italiano e in inglese (usando la longevità di un intervento come metrica).

	IT2007EL	**EN2007EL**
Pages	173173 (57.42%)	1123316 (56.48%)
Edit Count	16.12	34.69
Distinct Users	10.33	18.11
Total Positive Contribution	454	1679
Top Users	1.196	1.42
Top User Edits %	25.18%	19.74%
Top Users %	18.36%	14.74%
Top User Contribution %	70.82%	64.13%
Anonymous Edit Count %	8.55%	16.44%
Anonymous Contribution %	7.67%	12.59%

dei Top User rimane molto basso. Questo significa che la maggior parte di queste pagine ha un solo utente che ha contribuito per la maggior parte a essa.

Il ruolo di rilevanza dei Top User si riconosce tuttavia osservando le tre statistiche del numero di interventi che hanno fatto sul totale, sul loro numero rispetto al numero di utenti distinti e sul contributo effettivo che hanno collezionato rispetto a quello totale positivo. A prescindere dalla versione di Wikipedia considerata essi rappresentano un piccolo gruppo di utenti (in media inferiore al 20%) che con un moderato numero di interventi (mai più del 25.18% del totale) è responsabile della gran parte del contributo positivo di una pagina (almeno del 64%).

Si noti infine come la percentuale di interventi anonimi non cambia particolarmente tra la versione filtrata e quella non filtrata delle statistiche di una pagina. Anche per il contributo anonimo vale la stessa considerazione, anche se in questo insieme di pagine la differenza tra le due metriche diventa molto inferiore rispetto a quelle della versione non filtrata.

Analisi sulle statistiche delle pagine di qualità

Per concludere questa sezione si vuole ora usare il ricco insieme di dati a disposizione su ciascuna pagina di differenti versioni di Wikipedia per cercare di trarre qualche semplice considerazione su un particolare sottoinsieme di esse, quelle Featured. Esse sono considerate le migliori pagine di Wikipedia e per ricevere questa qualifica devono superare un processo di valutazione che

Tabella 5.7: Confronto tra le statistiche relative alle pagine Featured e non per la versione italiana di Wikipedia del 2007.

	IT2007EL FA	**IT2007EL ¬FA**
Edit Count	135.6	43.64
Distinct Users	49.11	24.25
Total Positive Contribution	8520	1425
Top Users	1.45	1.57
Top User Edits %	17%	12.52%
Top Users %	3.8%	7%
Top User Contribution %	65.71%	57.89%
Anonymous Edit Count %	16.45%	15.2%
Anonymous Contribution %	8.39%	13.09%

il più delle volte comincia proprio con l'integrazione delle carenze riscontrate in esse.

Per questo sono state estratte solo quelle statistiche sull'insieme delle pagine Featured per ciascun wiki in analisi e di esse sono stati calcolati i valori medi. Lo stesso processo è stato poi applicato per le pagine che non appartengono a questa categoria, in modo tale da poter confrontare i due distinti sottoinsiemi. Tra queste ultime sono state però rimosse le pagine con meno di 20 interventi, a causa del fatto che non sarebbe stato corretto confrontare le pagine Featured, le quali sono note avere un elevato numero di edit, con un insieme di pagine così diverso da loro. Per semplicità verranno mostrati solo i risultati relativi alle versioni italiana e inglese del 2007 e quindi calcolati con la metrica di longevità di un intervento. La prima ha un totale di 449 articoli Featured, pari allo 0.25% del totale, mentre la seconda conta 1569 articoli Featured, rappresentanti lo 0.07% del totale. Gli altri dati d'interesse sono riassunti nelle Tabelle 5.7 e 5.8. La prima osservazione sulle pagine Featured riguarda il fatto che in media esse sono di maggiori dimensioni rispetto a quelle non appartenenti alla loro categoria, nonostante il filtro di quelle pagine con meno di 20 edit. Esse hanno tre volte il numero degli interventi e il doppio degli utenti distinti. Il contributo totale positivo è poi molto più grande nelle pagine scelte come le migliori di Wikipedia, ma questo è dato dal fatto che si tratta di pagine di dimensioni molto più elevate. Se infatti si confronta questo valore (TPC) con il numero di interventi (EC), si scopre che:

$$\left(\frac{TPC}{EC}\right)_{FA} = 63$$

$$\left(\frac{TPC}{EC}\right)_{\neg FA} = 61$$

Ben più interessante è invece notare che il rapporto tra contributo totale positivo (TPC) e numero di utenti distinti (DU) evidenzia delle differenze tra pagine Featured e non:

$$\left(\frac{TPC}{DU}\right)_{FA} = 170$$

$$\left(\frac{TPC}{DU}\right)_{\neg FA} = 57$$

Questo significa che gli utenti che scrivono su pagine di qualità hanno una media di contributo decisamente più alta rispetto al normale. Tuttavia i dati non consentono di dire se questo fenomeno è causa della qualità superiore o un suo effetto.

Un'altra osservazione è quella del fatto che il numero di Top User rimane molto simile. Di conseguenza la loro percentuale risulta più bassa per le pagine Featured, le quali hanno mediamente un più alto numero di contributori. Le percentuali di interventi e di contributi collezionati dai Top User sono invece più alti per le pagine Featured, a dimostrazione del fatto che mediamente questo gruppo di utenti ha maggiore influenza nelle pagine di qualità.

In ultimo si osservi che gli utenti anonimi contribuiscono in maniera più limitata alle pagine Featured, nonostante il loro numero di interventi sia di poco superiore in questo tipo di pagine. Il motivo può essere ricercato nella maggiore attenzione che esse ricevono da parte degli utenti registrati che quindi finiscono per avere percentuali di contributo più alte. Considerazioni molto simili possono essere fatte anche per la versione inglese di Wikipedia. Una grande differenza rispetto alla versione italiana però è nella percentuale di contributo anonimo. Nel caso inglese le pagine appartenenti alla categoria Featured hanno una percentuale estremamente bassa di contributo anonimo positivo, sintomo probabilmente di un ancor più stretto controllo di tipo editoriale da parte degli utenti registrati.

5.4 Costruzione di una Rete Sociale

Quest'ultimo sottoprocesso ha lo scopo di costruire una Rete Sociale di utenti di Wikipedia considerando come coautori del medesimo articolo l'insieme dei Top User per esso. Di conseguenza il suo unico ingresso è l'output del sottoprocesso precedente.

Tabella 5.8: Confronto tra le statistiche relative alle pagine Featured e non per la versione inglese di Wikipedia del 2007.

	EN2007EL FA	**EN2007EL ¬FA**
Edit Count	719.6	79.57
Distinct Users	216.1	38.36
Total Positive Contribution	$1330 \cdot 10^3$	$3.94 \cdot 10^3$
Top Users	3.98	1.93
Top User Edits %	11.05%	11.08%
Top Users %	2.53%	6.8%
Top User Contribution %	60.07%	53.22%
Anonymous Edit Count %	24.65%	22.89%
Anonymous Contribution %	0.93%	16.6%

Prima di cominciare l'analisi quantitativa delle differenti reti oggetto di questo studio si vuole, a titolo d'esempio, mostrare la rappresentazione grafica della più piccola tra quelle costruite. Questa rete è relativa agli utenti più importanti della versione italiana di Wikipedia del 2005 e, come spiegato nel Capitolo 4, i suoi nodi sono stati colorati secondo il seguente criterio. Si possono distinguere gli utenti amministratori colorati in verde, i Bot in blu, i presunti Bot colorati in marrone, e gli utenti non appartenenti a queste categorie in rosso. L'algoritmo utilizzato per la disposizione dei nodi è quello di *spring embedding*, implementato dalla libreria *Igraph*. Esso si basa sulla metafora fisica secondo la quale ogni nodo è rappresentato come una massa e ogni arco come una molla. In diverse iterazioni è quindi simulata l'attrazione e la repulsione tra nodi sino al raggiungimento di un equilibrio. Si è scelto di impostare la gravità di ciascun nodo in funzione del suo grado, in modo che nodi più centrali secondo questa misura risultino più vicini al centro del piano di visualizzazione. Il risultato della visualizzazione è quello rappresentato in Figura 5.7.

La prima osservazione riguarda il fatto che gli amministratori hanno per la maggior parte un ruolo centrale all'interno della rete. Lo stesso si osserva per i Bot, anche se in misura minore. Non sono pochi nemmeno gli utenti registrati centrali nel grafo. Nonostante la rappresentazione grafica possa dare immediatamente l'idea di alcune proprietà macroscopiche della rete, risulta ovvio accorgersi di come all'aumentare del numero di nodi sia sempre più difficile trarre delle conclusioni precise su di essa. In realtà un metodo molto più comodo rispetto alla visualizzazione su carta risulta essere quello di visualizzare la rete su un calcolatore in modo interattivo. Si hanno così

Figura 5.7: Rete Sociale degli utenti della versione italiana di Wikipedia al 2005, prodotta a partire dalla metrica di longevità dell'intervento valutata rispetto alla sua versione più simile

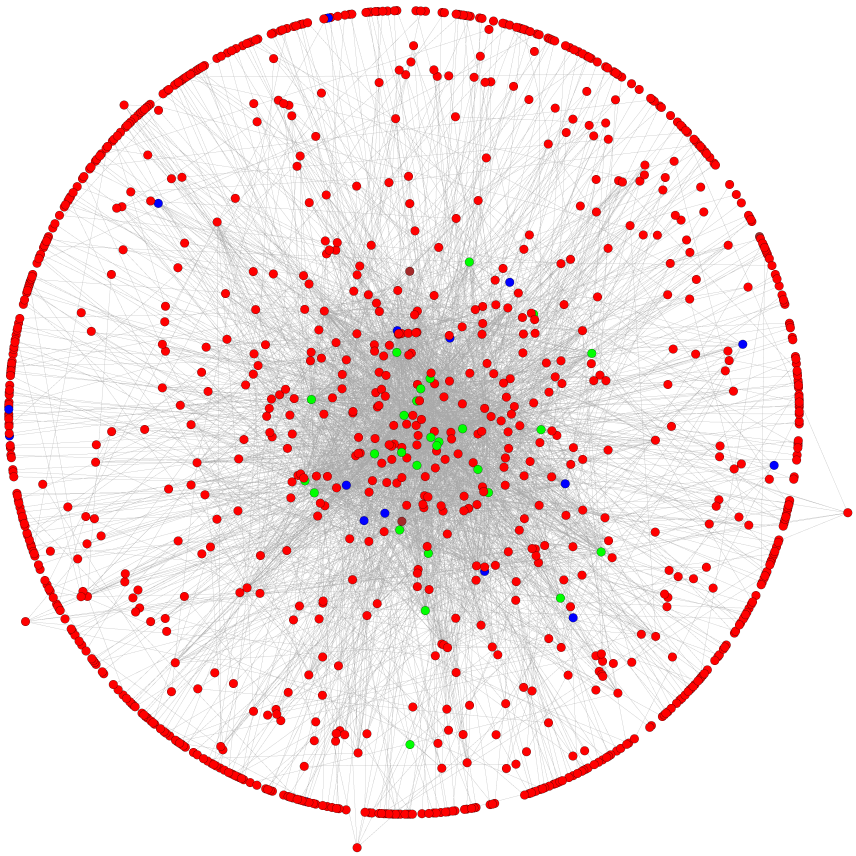


Tabella 5.9: Confronto tra l'utilizzo della metrica di longevità di un intervento (EL) e di longevità di un intervento valutata rispetto alla sua versione più simile (ELS) nella costruzione della Rete Sociale dei coautori della versione in italiano di Wikipedia del 2005.

	IT2005EL	**IT2005ELS**
Articoli esaminati	64617	64617
Articoli con almeno un autore	33708	31252
Articoli con più di un autore	3886	3980
Numero di autori	2281 (37.33%)	2296 (37.57%)
Numero di autori nella rete	1052 (17.21%)	1168 (19.11%)
Articoli per autore	14.78	13.61
Autori per articolo	1.13	1.16
Collaboratori per autore	6.12	7.59
Componente più grande	1006 (95.62%)	1126 (96.4%)
Seconda componente più grande	2 (0.2%)	2 (0.2%)
Coefficiente di clustering	0.16	0.17
Distanza media	3.22	3.11
Diametro	8	7

a disposizione funzioni di zoom che consentono di aumentare o diminuire il grado di dettaglio della rappresentazione e di spostare l'attenzione solo su alcune parti della rete. Ciò nonostante è opportuno approfondire l'analisi utilizzando misure più precise di tipo quantitativo.

5.4.1 Considerazioni a livello macroscopico

Confronto tra due metriche

Nella Tabella 5.9 sono confrontate le due reti prodotte per la versione in italiano di Wikipedia del 2005. Le due reti prodotte dalle due metriche di longevità dell'intervento semplice e di quella valutata rispetto alla versione più simile non mostrano grandi differenze a livello macroscopico. Gli articoli con più di un autore risultano essere, come già osservato nella sezione precedente, molto pochi rispetto a quelli con un solo autore. Questo significa che tra gli articoli esaminati in questo wiki è molto difficile trovarne uno in cui due o più autori abbiano contribuito in maniera importante a esso. Quest'osservazione trova conferma anche nel numero medio di autori per articolo, che risulta essere prossimo all'unità. Anche per la versione italiana di Wikipedia del 2008, che ricordiamo essere analizzata su un campione casuale delle pagine totali, valgono considerazioni molto simili a quelle fatte

Tabella 5.10: Confronto tra l'utilizzo della metrica di longevità di un intervento (EL) e di longevità di un intervento valutata rispetto alla sua versione più simile (ELS) nella costruzione della Rete Sociale dei coautori della versione in italiano di Wikipedia del 2008.

	IT2008EL	**IT2008ELS**
Articoli esaminati	121216	121216
Articoli con almeno un autore	56646	54869
Articoli con più di un autore	15853	15753
Numero di autori	8038 (23.98%)	8811 (26.28%)
Numero di autori nella rete	5309 (15.83%)	6150 (18.35%)
Articoli per autore	7.05	6.23
Autori per articolo	1.35	1.4
Collaboratori per autore	6.99	8.94
Componente più grande	4957 (93.37%)	5765 (93.74%)
Seconda componente più grande	5 (0.09%)	4 (0.06%)
Coefficiente di clustering	0.099	0.121
Distanza media	3.55	3.45
Diametro	9	8

per la versione del 2005. Come si può notare guardando la Tabella 5.10, anche l'analisi macroscopica delle due reti prodotte con le due metriche porta a risultati decisamente simili. Questo è da interpretarsi con il fatto che a livello globale le differenze tra le due metriche non sono visibili. Tuttavia ciò non significa che i singoli utenti considerati dalla rete siano gli stessi.

A questo punto si vogliono quindi confrontare le proprietà macroscopiche delle due Reti Sociali dei coautori di Wikipedia nelle sue versioni italiana e inglese del 2007. Si osservi la Tabella 5.11 per avere un prospetto dei dati studiati.

Si approfitterà anche della disponibilità in letteratura di studi analoghi su Reti di coautori in comunità di ricercatori scientifici e di sviluppatori di software open source.

Per la prima tipologia di rete gli studi sono molto numerosi e in particolare si sono scelti quelli di (Cotta and Merelo, 2006, Newman, 2001a,b). Le comunità analizzate da questi studi sono: quella dell'*Evolutionary Computation* (EC), un'area della computer science; quella dei *ricercatori biomedici* (Medline); quella estratta dagli archivi del *Physics E-print* (Physics), relativa alla ricerca nel campo della fisica; quella estratta dal *High-Energy Physics Literature Database* (SPIRES); quella dell'archivio della *Networked Computer Science Technical Reference Library* (NCSTRL), relativo a differenti aree

Tabella 5.11: Confronto tra le due Reti Sociali dei coautori della versione in italiano e in inglese di Wikipedia del 2007 (utilizzando la metrica di longevità dell'intervento).

	IT2007EL	**EN2007EL**
Articoli esaminati	301586	1988629
Articoli con almeno un autore	173173	1123316
Articoli con più di un autore	27905	339338
Numero di autori	11073 (31.17%)	169999 (20.19%)
Numero di autori nella rete	6352 (18.46%)	106979 (12.7%)
Articoli per autore	15.64	6.61
Autori per articolo	1.2	1.43
Collaboratori per autore	8.27	10.49
Componente più grande	6092 (95.91%)	99886 (93.37%)
Seconda componente più grande	5 (0.07%)	6 (0.005%)
Coefficiente di clustering	0.109	0.037
Distanza media	3.37	3.77
Diametro	8	12

della computer science.

Ben più complicato risulta trovare studi su grandi reti di sviluppatori di software open source. Lo studio più interessante per il confronto macroscopico risulta essere quello di (Gao et al., 2003), i quali analizzano la rete degli sviluppatori della comunità di SourceForge (OSS-SF). La prima osservazione da fare riguarda il fatto che i dati a disposizione su questa rete sono meno ricchi rispetto alle altre. Ciò nonostante si è ritenuto molto interessante utilizzarli per il confronto. In secondo luogo va ricordato che, mentre per gli autori di Wikipedia e della comunità scientifica il frutto della collaborazione è rappresentato dall'articolo, per la comunità degli sviluppatori è il progetto software a imporre le relazioni tra gli individui.

Le Tabelle 5.12 e 5.13 riassumono i risultati delle ricerche appena descritte.

Le due reti di coautori della versione italiana e inglese di Wikipedia sono molto diverse per dimensioni, come prevedibile. In particolare la rete della versione in inglese conta un numero di nodi 15 volte più grande rispetto a quella in italiano. Tuttavia, rispetto al numero totale di utenti registrati, la percentuale di quelli che vengono inclusi nella rete italiana è superiore a quelli nella rete inglese (18.46% contro 12.7%). Questo può significare che, all'aumentare degli utenti in un wiki, non necessariamente gli utenti importanti aumentano con lo stesso ritmo.

Tabella 5.12: Confronto tra Reti Sociali dei coautori di comunità scientifiche come analizzate in (Cotta and Merelo, 2006, Newman, 2001a,b) (prima parte).

	EC	**Medline**	**Physics**
Articoli esaminati	6199	2163932	98502
Numero di autori	5492	1520251	52909
Articoli per autore	2.9	6.4	5.1
Autori per articolo	2.56	3.75	2.53
Collaboratori per autore	4.2	18.1	9.7
Componente più grande	$3.6 \cdot 10^3$ (67.1%)	$1395 \cdot 10^3$ (92.6%)	$44 \cdot 10^3$ (85.4%)
Seconda componente più grande	36 (0.65%)	49 ($3 \cdot 10^{-5}$)	18 $3 \cdot 10{-}4$
Coefficiente di clustering	0.808	0.066	0.43
Distanza media	6.1	4.6	5.9
Diametro	18	24	20

Tabella 5.13: Confronto tra Reti Sociali dei coautori di comunità scientifiche come analizzate in (Newman, 2001a,b, Gao et al., 2003) (seconda parte).

	SPIRES	**NCSTRL**	**OSS-SF**
Articoli/Progetti esaminati	66652	13169	50000
Numero di autori	56627	11994	70000
Articoli per autore	11.6	2.6	n. p.
Autori per articolo	8.96	2.22	n. p.
Collaboratori per autore	173	3.6	n. p.
Componente più grande	$49 \cdot 10^3$ (88.7%)	$6.39 \cdot 10^3$ (57.2%)	n. p. (35%)
Seconda componente più grande	69 (0.1%)	42 (0.3%)	n. p.
Coefficiente di clustering	0.726	0.496	0.7
Distanza media	4	9.7	n. p.
Diametro	19	31	8

L'aspetto più sorprendente è quello per il quale la versione inglese di Wikipedia risulta avere molti più articoli con più di un autore rispetto a quella italiana. Considerati in percentuale rispetto al numero totale di articoli la prima risulta averne il 17%, mentre la seconda solo il 9%. Questo dato dice sicuramente che le pagine di ridotte dimensioni della versione di Wikipedia in italiano sono molte di più, in proporzione, di quelle della versione in inglese.

Sempre dal punto di vista del numero di nodi, non indifferenti per molte delle metriche macroscopiche, si noti come la rete inglese sia confrontabile con le reti Physics, SPIRES, NCSTRL e OSS-SF. La rete con numero di nodi più vicino a quella italiana risulta essere invece quella della comunità EC.

Nel numero medio di articoli per autore la rete inglese è molto simile a Medline, mentre quella italiana ha il massimo valore rispetto a tutte le altre reti. Questo è significativo del fatto che la maggior parte del lavoro, in questa versione italiana di Wikipedia, è portato avanti da una ristretta cerchia di persone.

Se già nella sezione precedente il numero di autori per articolo era sembrato molto basso, a maggior ragione esso appare molto più basso rispetto a quello delle altre comunità. D'altra parte, per quanto riguarda gli articoli scientifici, è molto difficile che essi vengano scritti da una sola persona. Per questo motivo si è deciso di calcolare questo parametro solo per il sottoinsieme di articoli di Wikipedia con più di un autore, tenendo bene a mente il fatto che questi sono un piccolissimo campione del totale. Il valore medio di autori per articolo di questo sottoinsieme è pari a 4.73 per la versione inglese di Wikipedia e pari a 2.21 per quella italiana. In questo la comunità inglese risulta essere più simile ancora una volta a quella degli autori di articoli medici, mentre quella italiana è paragonabile alla comunità di EC, Physics e NCSTRL.

Il numero medio di collaboratori per autore è una proprietà macroscopica che dipende dal numero di nodi della rete. Perciò è opportuno confrontare solo quelle reti con numero di nodi paragonabile tra loro. La rete italiana, rispetto alla comunità di EC, mostra ben il doppio numero di collaboratori. Anche quella inglese, rispetto alla comunità di Physics e NCSTRL, ha un valor medio più elevato di autori per articolo. Un po' particolare da questo punto di vista la comunità SPIRES, viziata dal fatto che spesso molti articoli al suo interno sono di tipo sperimentale e per questo motivo con un numero molto alto di autori.

Si è verificato inoltre che le distribuzioni del numero di collaboratori per autore, mostrate per le due versioni di Wikipedia nelle figure 5.8 e 5.9, hanno

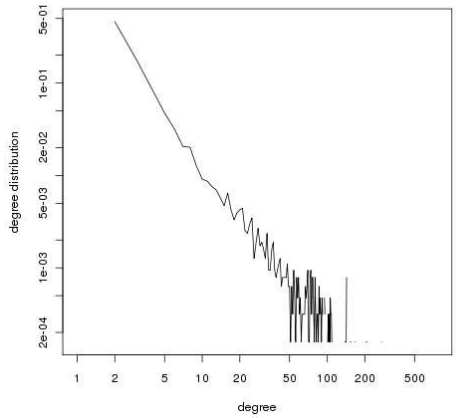


Figura 5.8: Degree distribution della Rete Sociale di Wikipedia nella sua versione in italiano del 2007

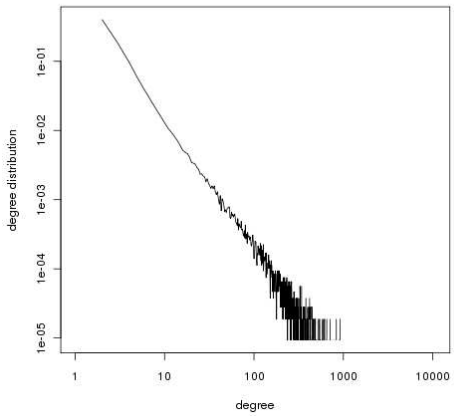


Figura 5.9: Degree distribution della Rete Sociale di Wikipedia nella sua versione in inglese del 2007

un andamento di tipo power-law. Entrambe, disegnate su scala bilogaritmica disegnano una retta con coefficiente angolare pari a -2.1 per la versione italiana e a -2 per quella inglese. Si può quindi dire che per queste reti vale il modello di preferential attachment e cioè che i nuovi autori tendono a collegarsi a quelli che hanno un maggior numero di collaboratori. Questo, all'interno di Wikipedia, si può spiegare con il fatto che gli utenti più attivi, come si vedrà successivamente, hanno l'obiettivo di migliorare l'enciclopedia nella sua globalità e raramente intervengono dove c'è già un altro utente con il loro comportamento.

Lo studio delle componenti connesse mostra invece come le reti costruite siano molto poco frammentate. Per entrambe oltre il 90% dei nodi appartiene alla stessa componente. Oltre a questo si osserva che sia la distanza media tra due nodi qualsiasi che il diametro delle due reti (la lunghezza del percorso più lungo tra quelli minimi tra due nodi) hanno valori molto bassi rispetto a tutte le altre reti di coautori. Questa particolarità riguarda ancora una volta l'esistenza di alcuni utenti che si occupano di rendere Wikipedia tanto più possibile uniforme e coerente al suo interno. Essi, lavorando indistintamente su molte pagine, finiscono per diventare loro grandi contributori e quindi ad ampliare il loro numero di contatti all'interno della rete.

Anche la rete di sviluppatori open source ha un diametro molto piccolo, ma risulta molto frammentata, poiché solo il 35% dei suoi nodi appartiene alla componente più grande della rete.

L'ultimo confronto delle misure macroscopiche mette in evidenza una

certa differenza delle le reti di utenti di Wikipedia sia con quelle di coautori che con quelle di sviluppatori. Il coefficiente di clustering di entrambe le reti inglese e italiana è molto basso. L'unica rete di coautori che raggiunge valori simili è quella di Medline, che si ricorda essere la più grande per numero di nodi. Il motivo di questa caratteristica è da ricercarsi in due particolarità di Wikipedia.

La prima è ancora una volta la tendenza di certi individui molto attivi all'interno dell'enciclopedia che cercano di controllare uniformemente la totalità delle pagine. Si è visto come un basso coefficiente di clustering possa essere caratteristico di una rete gerarchica. In questo caso la gerarchia può essere considerata quella di questo gruppo di utenti molto attivi ciascuno dei quali è collegato con moltissimi altri utenti che non necessariamente sono collegati tra di loro. Inoltre, come già accennato, se uno degli utenti attivi si accorge del fatto che una data pagina è già seguita da uno del suo gruppo, dedicherà le sue energie a un'altra pagina con maggiore necessità di cure.

La seconda caratteristica di Wikipedia che contribuisce a mantenere basso il coefficiente di clustering è il fatto che la partecipazione alla scrittura di una pagina è completamente volontaria e libera. Ci sono molti fattori che influenzano due ricercatori a collaborare allo stesso articolo: vicinanza geografica e appartenenza alla medesima area di ricerca sono solo alcuni di questi. Perciò in una rete di autori scientifici la probabilità che si formino delle triple transitive è molto più alta rispetto a una rete di autori di Wikipedia. Sarebbe tuttavia interessante verificare il clustering coefficient di altri wiki, magari con utenti appartenenti a gruppi più organizzati.

5.4.2 studio delle sociometric star

In questo paragrafo verranno analizzate le sociometric star delle differenti Reti Sociali appena costruite, con lo scopo di capire quali siano gli individui più importanti per la comunità di Wikipedia e se essi siano o meno gli stessi indicati dalle altre metriche che non tengono conto della topologia della rete di rapporti individuata. Le classifiche delle prime venti posizioni sono consultabili nell'allegato A.

Per **la rete relativa alla versione del 2005 di Wikipedia in italiano** le tre metriche di Betweenness Centrality, Closeness Centrality e Degree Centrality sono molto simili, sia che si tratti della rete prodotta utilizzando la metrica di longevità di un intervento che la sua variante valutata rispetto alla versione più simile. Anche i rapporti tra classi di utenti sono sostanzialmente i medesimi e si può notare, rispetto alla metrica di conteggio degli interventi,

come gli agenti umani prevalgano sui Bot nei primi 100 posti della classifica. Essi infatti non sono mai più di 6.

I primi utenti in classifica, consultabili per le prime 20 posizioni nelle Tabelle A.4, A.6, A.8, A.5, A.7 e A.9, risultano essere, in ordine: l'amministratore *Snowdog*, l'utente registrato *Twice25* e i due amministratori *Marcok* e *Shaka*. Se per i primi due non c'è molto da stupirsi, poiché erano già i primi in classifica per le metriche globali, gli utenti *Marcok* e *Shaka* assumono in queste misure topologiche una maggiore importanza. Probabilmente questo è dovuto al fatto che i loro interventi risultano essere molto omogenei su tutta Wikipedia.

Si può affermare che in un wiki agli inizi, come quello della Wikipedia italiana del 2005, i ruoli degli utenti che dedicano molto del loro tempo a esso non siano molto differenziati.

La classifica delle sociometric star secondo la Eigenvector Centrality, consultabile nelle Tabelle A.10, A.11, è l'unica che mostra notevoli differenze con le altre misure di centralità. Si ricorda tuttavia che questa è l'unica che sfrutta il peso degli archi, il quale indica in quante pagine i due autori sono stati inclusi nello stesso insieme dei Top User.

Si noti come i primi due utenti in classifica siano due Bot, *Gacbot* e *Luki-Bot*, che distanziano di parecchio tutti gli altri utenti. Questo è dovuto sicuramente al fatto che essi si sono trovati coautori della stessa pagina per ben 109 volte, valore molto elevato rispetto a tutti gli altri. Ciò è possibile poiché entrambi i Bot si occupano di compiti molto generali su tutte le pagine della Wikipedia italiana.

Un altro Bot[2], *NTBot*, risulta essere tra le prime posizioni per questa classifica (tra la sesta e la nona posizione per le due metriche di costruzione della rete), mentre per la Closeness e Betweenness Centrality esso risulta essere ben oltre, tra la trentesima e la cinquantesima posizione. Il legame tra Degree Centrality ed Eigenvector Centrality si nota per il fatto che nella prima misura questo Bot si posiziona tra la diciannovesima e la ventitreesima posizione.

I "sovrani" delle altre classifiche, *Snowdog*, *Twice25*, *Shaka* e *Marcok* perdono abbastanza del loro prestigio in questa metrica. Si ritiene che questo sia un aspetto molto particolare di Wikipedia. La Eigenvector Centrality assegna prestigio agli individui che sono collegati ad altri individui di prestigio. Il fatto che questi quattro utenti perdano di importanza in questa metrica significa che probabilmente sono risultati connessi a nodi di scarsa

[2]Non dichiarato come tale nella lista dei Bot ma solamente sulla sua pagina utente.

rilevanza e che quindi essi sono riusciti a portare il loro controllo e la loro esperienza nella maggior parte delle pagine importanti di Wikipedia.

Una considerazione simile può essere fatta, ad esempio, anche per l'utente *Davide*, che si trova nel ruolo opposto di essere in alta posizione (tra la settima e l'ottava) per le classifiche di Eigenvector Centrality, viceversa per quelle di Closeness e Betweenness Centrality (tra la diciottesima e la ventiseiesima posizione) e a metà strada per la Degree Centrality (quindicesima e diciassettesima posizione). La sua variazione di posizione è esattamente dello stesso tipo di quella del Bot *NTBot*. Secondo il significato della Eigenvector centrality si sta parlando di utenti che si legano preferibilmente con altri utenti di prestigio. Questo comportamento per un utente umano indica un atteggiamento opposto rispetto quello descritto poc'anzi. Per un Bot invece questo è in un certo senso prevedibile. Essi infatti possono compiere gravi danni a Wikipedia e dunque il fatto che siano spesso collegati a un utente importante può significare che essi sono spesso controllati. Ad esempio si nota come *Gacbot*, Bot gestito dall'amministratore *Gac*, sia stato coautore ben 7 volte proprio con il suo gestore. Anche il Bot *NTBot* risulta avere un forte legame, pari a 7, con l'amministratore *Snowdog*.

La rete **relativa alla versione del 2008 di Wikipedia in italiano** si trova a rispecchiare le dinamiche di un wiki decisamente più maturo rispetto a quello di tre anni prima. Le sociometric star più importanti per le classifiche di Betweenness, Closeness e Degree Centrality risultano essere sempre gli utenti *Snowdog*, *Twice25*, *Marcok* e *Shaka*, segno del fatto che essi hanno mantenuto il loro impegno costante nel corso degli anni. Le classifiche relative a questi dati sono rappresentate nelle Tabelle A.21, A.23, A.25, A.22, A.24 e A.26. L'aspetto importante di questo fatto è che, se nel 2005 questi individui primeggiavano anche nelle classifiche generate dalle metriche globali come ad esempio il conteggio degli interventi, in questo caso essi risultano essere sempre tra i primi contributori solo nelle misure di centralità della rete. Ad esempio *Snowdog* è quasi sempre primo in classifica (al limite secondo) in entrambe le reti generate dalle due solite metriche, e solo dodicesimo per numero di interventi, classifica nella quale *Shaka* non rientra nemmeno tra i primi venti.

Esiste anche l'esempio contrario di utenti, come ad esempio l'amministratore *.snoopy.*, primi in una delle classifiche globali, in questo caso in quella della longevità dell'intervento, ma meno importanti per quelle che tengono conto della topologia della rete. Stesso dicasi per l'utente *Murray*, al secondo posto per la classifica globale di longevità dell'intervento valutata rispetto alla sua versione più simile ma non presente nelle prime 20 posizioni di nessun'altra classifica basata sulle proprietà topologiche della rete.

Tutto ciò risulta essere molto interessante, poiché significa che una personalità di spicco per contributo in Wikipedia non è necessariamente tra le più centrali all'interno della Rete Sociale.

Le classifiche delle reti generate dalle due metriche di longevità dell'intervento e la sua variante rispecchiano inoltre le differenze concettuali alla loro base. La prima metrica da più importanza agli utenti amministratori, probabilmente perché da molto peso agli interventi di revert. Infatti per la Betweenness Centrality calcolata sulla rete prodotta usando la metrica di longevità dell'intervento ben 42 amministratori rientrano nei primi 100 posti della classifica. Per la medesima centralità, calcolata sulla rete prodotta usando la variante della metrica precedente, questi sono solo 29. Lo stesso si osserva anche per la Closeness Centrality (42 contro 28 amministratori) e per la Degree Centrality (40 contro 26 amministratori).

Confrontando invece le differenti metriche di centralità, ci si rende conto del fatto che le tre di Betweenness, Closeness e Degree Centrality sono molto simili tra loro. Come per la versione del 2005 si può dire che in questa versione di Wikipedia i ruoli degli utenti più importanti non sono particolarmente distinti. L'eccezione è ancora una volta rappresentata dalla Eigenvector Centrality, la quale si riveste dunque di un certo interesse per cogliere particolari comportamenti di utenti come individuati anche nella versione del 2005. Le Tabelle relative ai primi 20 classificati secondo questa metrica sono le numero A.27 e A.28. Come nel caso del 2005 gli utenti più importanti risultano essere tre Bot, i quali distaccano gli altri utenti di parecchi punti. All'interno delle classifiche si notano utenti centrali per la Eigenvector Centrality e molto meno per le altre misure di centralità, come *Moloch981* che nelle prime è in quarta posizione e nelle restanti mai prima della tredicesima. Viceversa l'utente *Pils56* è sempre importante per le misure di Betweenness, Closeness e Degree Centrality ma sempre attorno alla quarantesima posizione per le misure di Eigenvector Centrality. A ciò si aggiunge il fatto che nella versione attuale le differenze della metrica di longevità dell'intervento con la sua variante sono più accentuate.

La rete **relativa alla versione del 2007 di Wikipedia in italiano**, analizzata dal punto di vista delle sociometric star, è sicuramente molto simile a quella del 2008, analizzata poc'anzi. La particolarità sta nel fatto che la versione del 2008 contiene solo u sottoinsieme delle pagine totali. Questo è già di per se un risultato interessante perché significa che gli utenti più importanti della Wikipedia in italiano si distribuiscono in modo tendenzialmente uniforme su tutte le pagine dell'enciclopedia. Non si reputa dunque interessante l'analisi dettagliata dei risultati di questo studio, poiché troppo simili a quelli appena descritti, complice sicuramente la breve distanza

temporale che intercorre tra le due versioni del wiki. Nel caso di uno studio approfondito dei soggetti di un wiki ricorrere a un campionamento delle pagine può considerarsi comunque rischioso, anche perché non è dimostrato che questa dinamica non sia relativa solo alla Wikipedia in italiano. Per approfondire lo studio delle sociometric star di questa versione si consultino le Tabelle A.14, A.15, A.16 e A.17.

Si conclude quindi l'analisi delle sociometric star sulla rete **relativa alla versione del 2007 di Wikipedia in inglese**. Le Tabelle che mostrano i dati dei primi 20 classificati sono le numero A.31, A.32, A.33 e A.34. Anche in questo caso le tre misure di Betweenness, Closeness e Degree centrality indicano il medesimo sottoinsieme di utenti in cima alle classifiche: si tratta dei Bot responsabili dell'individuazione e della riparazione degli atti vandalici *AntiVandalBot*, *TawkerBot2* e *TawkerBot4*, e dei due amministratori *Can't sleep, clown will eat me* e *Wiki alf*. Già si notano delle differenze con le versioni italiane di Wikipedia.

Non c'è infatti un utente come l'italiano *Twice25* che pur non ricoprendo un ruolo di amministratore si ritrova sempre nelle primissime posizioni di ogni classifica basata su topologia della rete. Egli è dunque da considerarsi un caso di utente eccezionale.

Inoltre tra questi utenti molto influenti ci sono ben tre Bot, mentre nel caso italiano tutti questi erano umani. Il motivo è da ricercarsi nella metrica utilizzata per costruire la Rete Sociale di questa versione di Wikipedia. La metrica di longevità dell'intervento da molta importanza a chi ripristina un vandalismo, come i tre Bot che non casualmente finiscono sempre ai vertici delle classifiche. Questo fenomeno non si può notare nella versione in italiano di Wikipedia, perché Bot di questo genere sono considerati dalla comunità di questo wiki troppo pericolosi e solo di recente si stanno sperimentando gli effetti che possono avere sull'enciclopedia. Sarà interessante valutare se, negli anni a venire, l'effetto dei Bot contro i vandalismi quali *YaFKBOT* risaliranno le posizioni in classifica per le reti italiane costruite con metrica di longevità dell'intervento.

Nel complesso le classifiche di Betweenness, Closeness e Degree Centrality risultano ancora molto simili tra di loro, il che mette in luce come effettivamente i ruoli all'interno degli utenti più attivi di Wikipedia siano poco differenziati.

Rimane tuttavia valida la considerazione che queste personalità importanti siano abbastanza differenti da quelle individuate dal semplice conteggio degli interveti, dove ancora una volta prevalgono i Bot, e dalla classifica globale di longevità dell'intervento. Quindi il ruolo all'interno della comunità non risulta essere completamente determinato dalla valutazione

del contributo senza tener conto dei rapporti tra gli utenti.

Rimane anche la decisa differenza tra la classifica prodotta dalla Eigenvector Centrality rispetto a quelle prodotte dalle altre misure di centralità. Nel caso della Wikipedia in inglese questa differenza è ancora più clamorosa. Solo 27 amministratori rientrano nei primi 100 posti in classifica per la Eigenvector Centrality, mentre per le misure di Betweenness, Closeness e Degree Centrality essi sono rispettivamente 67, 76 e 72. Si noti che questi valori non possono essere confrontati direttamente con quelli delle versioni italiane di Wikipedia, poiché essa conta in assoluto molti meno amministratori rispetto alla sua controparte inglese. Si ricorda inoltre che questo risultato è da considerarsi sintomo di una particolarità di Wikipedia. Gli amministratori non si collegano tra di loro ma si ripartiscono in modo omogeneo tra tutti gli utenti con lo scopo di migliorare la qualità globale dell'enciclopedia.

Anche in questa versione di Wikipedia i primi due classificati secondo la Eigenvector Centrality sono due Bot che da soli distaccano di parecchi punti i loro successore. Essi sono *SmackBot* e *Rambot*, due programmi che agiscono su tutte le pagine di Wikipedia svolgendo compiti di correzione generica o di aggiornamento di informazioni. Il valore così alto di centralità è da spiegarsi con il fatto che essi risultano coautori della stessa pagina per più di 10 mila volte.

Capitolo 6

Conclusioni e sviluppi futuri

6.1 Conclusioni

Il problema affrontato da questo lavoro è stato quello di definire e applicare una metodologia per analizzare, in modo automatico, la comunità di utenti che collaborano attraverso un wiki a un progetto di costruzione di una base di conoscenza.

Non trattandosi di un problema banale, esso è stato scomposto in quattro sottoproblemi principali, ognuno dei quali è stato affrontato sotto diversi aspetti. Prima di tutto si è inquadrato dal punto di vista teorico ciascun sottoproblema e si sono proposte delle soluzioni motivate da proprietà formali. Successivamente ci si è occupati della progettazione e dell'implementazione delle soluzioni proposte. Infine si sono applicati i metodi sviluppati per verificarne la fattibilità in casi d'uso reali.

Il primo sottoproblema è stato quello di selezionare, nella moltitudine di dati presenti all'interno di un sistema informativo basato su tecnologia wiki, quelli di maggior interesse per lo studio in questione. Per fare questo è stata necessaria un'ampia ricerca bibliografica nell'area degli studi su Wikipedia che ha portato a una consapevolezza su alcune delle sue dinamiche più importanti e sulle misure maggiormente indicate per coglierle. L'estrazione di questi dati è stata applicata a quattro wiki reali per i quali l'interesse per dei risultati è concreto. Essi sono la versione in lingua italiana di Wikipedia negli anni 2005, 2007 e 2008 e quella in lingua inglese del 2007.

Questi dati sono quindi stati sfruttati per valutare il contributo di ciascun utente all'interno di ogni pagina per i wiki studiati. In particolare è stato ponderato l'utilizzo di tre differenti metriche.

La prima, il conteggio degli interventi, è stata ritenuta troppo poco precisa per gli scopi del lavoro ma è da considerarsi interessante in quanto punto

di riferimento per la valutazione delle metriche più complesse. Il suo svantaggio è quello di non tener conto dei parametri qualitativi e quantitativi degli interventi.

La seconda, la longevità dell'intervento, è stata utilizzata per la prima volta in questo lavoro per questo scopo, pur essendo pensata dai suoi autori per il calcolo del contributo per ogni utente a livello globale. Essa tiene conto sia della qualità che della quantità degli interventi ma mostra le sue debolezze principalmente nella valutazione del contributo di quegli utenti responsabili del ripristino di una versione precedente o autori di modifiche non accettate dalla comunità.

Infine si è proposta una nuova metrica, chiamata longevità dell'intervento valutata rispetto alla sua versione più simile, sia per far fronte alle imprecisioni di quella precedente che per cogliere in modo più significativo il contributo di un utente per questo lavoro.

Le tre metriche sono state implementate in dei moduli software in grado di calcolarle a partire dai dati estratti nella fase precedente.

I risultati su tutte le versioni di Wikipedia considerate hanno rispecchiato le considerazioni attese. In particolare si è verificato come la metrica di conteggio degli interventi metta in risalto i Bot agenti sul wiki, software in grado di fare un numero elevato di modifiche tipicamente di piccola entità e non necessariamente corrette. Quella di longevità dell'intervento vorrebbe invece privilegiare gli autori delle modifiche più durature, ma a causa del suo problema con l'errata valutazione dei ripristini, finisce per favorire eccessivamente i manutentori del wiki.

Quella di longevità dell'intervento valutata rispetto alla sua versione più simile invece attenua questo effetto indesiderato ma dimostra come in realtà la maggior parte dei manutentori dei wiki considerati sia anche responsabile degli interventi più consistenti e duraturi. Si ritiene che essa sia la metrica migliore per un'analisi di questo tipo, anche considerato che rispetto alla sua versione originale non introduce particolari overhead computazionali. Si è anche visto come essa sia la misura più valida per stimare il contributo positivo degli utenti anonimi in un wiki.

Il successivo sottoproblema affrontato è stato quello di trovare gli utenti più importanti per ciascuna pagina a partire dai contributi calcolati con le metriche studiate. Si sono formulati dei vincoli che l'insieme di questi utenti deve rispettare per essere considerato tale.

Si è quindi progettato e implementato un algoritmo in grado di calcolare questo sottoinsieme di utenti per ciascuna pagina di un wiki.

L'applicazione di questo processo ha permesso di osservare che pagine apparentemente simili possono essere state costruite in maniere molto diffe-

renti. Il caso emblematico è quello di due pagine con medesimo numero di contributori nelle quali per la prima uno solo tra essi ha contribuito molto di più degli altri e per la seconda un certo numero di utenti ha contribuito in maniera omogenea.

L'ultimo sottoproblema affrontato è stato quello di cogliere le possibili relazioni tra gli utenti più importanti individuati col metodo di selezione. L'assunzione fatta è stata quella secondo la quale due o più autori considerati tra i più importanti della medesima pagina sono in relazione tra loro. Questo tipo di assunzione è ispirata in una certa misura dagli studi sui collaboratori di articoli scientifici e sugli sviluppatori di software open source, due tipologie di comunità che hanno in comune con Wikipedia l'obiettivo di migliorare i risultati dei propri lavori attraverso la collaborazione.

Si ritiene che l'approccio di selezione sia molto importante nel caso di Wikipedia, poiché le relazioni instaurate tra due utenti che hanno scritto all'interno della stessa pagina sono troppo deboli per trarre reali conclusioni sulla sua comunità.

La semplice implementazione di questa tecnica ha permesso di modellare in una Rete Sociale le complesse dinamiche degli utenti più importanti all'interno dei wiki scelti. La scelta di ricondursi a un modello di rappresentazione dei dati noto e molto studiato in letteratura si è rivelata vincente principalmente perché si sono potute utilizzare tecniche di analisi collaudate.

Non di minor valore è stata la possibilità di confrontare le comunità di Wikipedia con alcune di quelle di coautori e di sviluppatori modellate come reti in altri lavori.

L'analisi a livello macroscopico delle reti ha permesso di quantificare l'aumento di dimensioni della comunità di Wikipedia nell'intervallo di tempo studiato, dal 2005 al 2008.

Si è inoltre avuto modo di vedere come la comunità degli utenti di Wikipedia in italiano sia molto più piccola rispetto a quella degli utenti della versione in inglese. Ciò è vero in termini assoluti, ma in proporzione al numero di utenti si è osservato come il nucleo forte di questa comunità sia più piccolo nella versione inglese. Esse hanno inoltre delle caratteristiche in comune che le contraddistinguono dalle comunità di coautori scientifici e di sviluppatori software. Quelle più interessanti sono la bassa distanza media tra due nodi qualsiasi all'interno della rete e il basso coefficiente di clustering. A quanto pare lo scopo degli utenti più importanti all'interno delle due versioni di Wikipedia sembra essere quello di non lasciare alcune delle sue aree senza il loro intervento.

L'analisi delle sociometric star ha confermato questa tendenza e ha permesso di individuare il gruppo di utenti centrali per le differenti misure

di Degree, Betweenness e Closeness Centrality. Osservare che questi sono praticamente gli stessi nelle differenti misure ha fatto capire quanto sia importante il loro contributo all'interno di ciascuna versione di Wikipedia e quanto le loro mansioni siano poco differenziate. L'unica misura che mostra significative differenze con le altre è quella di Eigenvector Centrality, la quale però mostra come gli utenti più importanti tendano a legarsi con molti altri non appartenenti al loro gruppo sempre nell'ottica di non lasciare l'enciclopedia senza la loro presenza.

È in questo tipo di analisi che cominciano a vedersi le differenze tra la metrica di longevità dell'intervento e quella proposta da questo lavoro. Quest'ultima da meno importanza ai revert e agli interventi non accettati dalla comunità e pertanto mette in risalto gli autori di nuovo contenuto piuttosto che quelli che si occupano di ripristinare gli atti vandalici o di modificare la forma del testo. Poiché il calcolo delle due metriche può considerarsi di complessità confrontabile, si ritiene più adatta per studi di questo tipo quella di longevità dell'intervento valutata rispetto alla sua versione più simile.

6.2 Sviluppi futuri

Il lavoro appena presentato mette alla luce sia possibilità di miglioramento che nuove direzioni verso le quali sviluppare le analisi su Wikipedia e in generale sui wiki.

Un limite del quale sarebbe interessante valutare l'entità ma ancor di più trovare una correzione, è quello che attualmente impedisce di calcolare il contributo del primo intervento. In questo lavoro si è proposta una soluzione per il calcolo del contributo anche nel caso dell'ultimo intervento, ma si ritiene che il primo possa essere considerato di importanza non inferiore, specialmente per quanto riguarda gli studi a livello della singola pagina. In proposito si reputa interessante la possibilità di aggiungere una versione vuota antecedente alla prima, che possa fungere da arbitro per essa evitando quindi che non riceva voti di qualità.

Si è visto come il processo di selezione dei coautori possa calcolare molte caratteristiche per ciascuna pagina che riescono a discriminare anche quelle che a prima vista risulterebbero uguali. Un passo avanti per sfruttare queste informazioni potrebbe essere quello di valutare l'impatto dell'utente con contributo più alto per ciascuna pagina oppure quello del creatore di ogni articolo. Sarebbe ancora più interessante trovare dei modelli in grado di sfruttare queste features per riconoscere alcune tipologie di articoli, come ad esempio quelli di qualità o più controversi. Per queste analisi si ritiene che modelli di apprendimento supervisionato potrebbero risultare efficaci.

In effetti la parte di questo studio che apre più porte per gli sviluppi futuri risulta essere la costruzione della Rete Sociale. Questo perché, come si è già avuto modo di sottolineare, questo tipo di modello è stato molto studiato in letteratura e per esso sono disponibili molti algoritmi di analisi automatica che potrebbero fornire validi risultati anche applicati a dei wiki.

Con il tipo di rete costruita in questo lavoro un primissimo sviluppo potrebbe essere quello che considera solo alcuni sottoinsiemi di pagine e ne disegna la Rete Sociale degli autori. Ad esempio si intravede una strada interessante nella rappresentazione della Rete Sociale relativa agli autori delle pagine Featured. Si potrebbe in questo modo confrontare se questa sottorete ha delle peculiarità sia rispetto alla rete globale che rispetto a quelle di altri sottoinsiemi di pagine. Inoltre sarebbe interessante capire se gli autori centrali in queste sottoreti risulterebbero essere gli stessi della rete globale.

La seconda direzione da intraprendere potrebbe essere quella che sfrutta maggiormente le informazioni contenute nel peso degli archi. Tra le misure di centralità calcolate si è visto che solo la Eigenvector Centrality ne fa uso e che i risultati ottenuti da essa sono decisamente significativi. Ad esempio si potrebbero realizzare delle sottoreti caratterizzate da quegli archi maggiori di una certa soglia con lo scopo di vedere quali utenti interagiscono più spesso nelle stesse pagine.

Sulle reti prodotte non sono inoltre ancora stati applicate tecniche di clustering con lo scopo di scoprire comunità tra gli utenti di Wikipedia. L'aspetto interessante da approfondire in questo caso sarebbe quello relativo alla scoperta del fatto che il gruppo degli utenti più importanti tende a spartirsi la rete piuttosto che a collaborare nelle stesse pagine.

Anche se sono state confrontate le stesse versioni di Wikipedia di anni differenti, un nuovo lavoro da affrontare sarebbe quello dello studio dell'evoluzione della sua comunità nel corso del tempo. Per esso si potrebbero considerare degli intervalli temporali di dimensione fissata entro i quali calcolare una rete di autori. Suddividendo quindi la storia di Wikipedia in un certo numero di intervalli si otterrebbero altrettante reti per forza di cose differenti tra loro. Si potrebbe quindi valutare l'evoluzione all'interno della rete di alcuni utenti per scoprire se ci sono delle regolarità nel modo di intervenire o di inserirsi all'interno della comunità.

Si reputa il confronto con le comunità di sviluppatori di software open source ancora ampiamente approfondibile. Purtroppo attualmente gli studi su questo tipo di comunità riguardano ancora reti di modeste dimensioni. Sarebbe invece interessante valutare se un approccio come quello proposto in questo lavoro, in grado cioè di tener conto anche della quantità e la qualità

degli sviluppatori, potrebbe essere applicato anche a queste comunità.

Un ultimo interessante studio realizzabile a partire da questo lavoro è quello di costruzione di una rete bipartita, cioè con due classi distinte di nodi. Oltre agli utenti anche le pagine possono essere rappresentate all'interno della rete e collegate con degli archi ai loro autori. Lo scopo di questo studio è quello di mettere in luce gruppi di pagine scritte da autori molto vicini tra loro. Oltre a ciò si può assegnare a ogni pagina una categoria, grazie alla quale verificare se alcuni autori possono essere considerati esperti di qualche sottodominio di Wikipedia.

Infine, vista la generalità del metodo proposto, si auspica che la metodologia proposta possa essere applicata anche a dei wiki diversi da Wikipedia.

Bibliografia

Sisay Fissaha Adafre and Maarten de Rijke. Discovering missing links in wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, New York, NY, USA, 2005. ACM. ISBN 1-59593-215-1.

B. T. Adler, De L. Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. Technical report, School of Engineering, University of California, May 2008a.

B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. Technical report, School of Engineering, University of California, May 2008b.

Thomas B. Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2007. ACM Press. ISBN 9781595936547.

Almeida, Mozafari, and Cho. On the evolution of wikipedia. ICWSM'2007 Boulder, Colorado, USA, 2007.

A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311 (3-4):3, 2002.

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

D. Barbagallo, C. Francalanci, and F. Merlo. The impact of social networking on software design quality and development effort in open source projects. In *Twenty Ninth International Conference on Information Systems, Paris*, 2008.

F. Bellomi and R. Bonato. Network analysis for wikipedia. In *Wikimania 2005*, 2005.

Joshua E. Blumenstock. Automatically assessing the quality of wikipedia articles. *UCB iSchool Report 2008*, 21, 2008a.

Joshua E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1095–1096, New York, NY, USA, 2008b. ACM. ISBN 978-1-60558-085-2.

Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM. ISBN 1-59593-223-2.

Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. Temporal analysis of the wikigraph. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0769527477.

A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(3):036116+, 2006.

Krishnendu Chatterjee, Luca de Alfaro, and Ian Pye. Robust content-driven reputation. In Dirk Balfanz and Jessica Staddon, editors, *AISec*, pages 33–42. ACM, 2008. ISBN 978-1-60558-291-7.

Carlos Cotta and Juan J. Merelo. Who is the best connected ec researcher? centrality analysis of the complex network of authors in evolutionary computation. *CoRR*, abs/0708.2021, 2007a.

Carlos Cotta and Juan-Julián Merelo. The complex network of ec authors. *SIGEVOlution*, 1(2):2–9, 2006. ISSN 1931-8499.

Carlos Cotta and Juan-Julián Merelo. Where is evolutionary computation going? a temporal analysis of the ec community. *Genetic Programming and Evolvable Machines*, 8(3):239–253, 2007b. ISSN 1389-2576.

Yrjo Engestrom. *Activity theory and individual and social transformation.* Cambridge University Press, 1999.

A. Forte and A. Bruckman. Scaling consensus: Increasing decentralization in wikipedia governance. *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 157–157, Jan. 2008. ISSN 1530-1605. doi: 10.1109/HICSS.2008.383.

Y. Gao, V. Freeh, and G. Madey. Analysis and modeling of open source software community, 2003.

Raymond G. Gordon. *Ethnologue Languages of the World*. SIL International, 2005.

Paul Heckel. A technique for isolating differences between files. *Commun. ACM*, 21(4):264–268, 1978. ISSN 0001-0782.

Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *CoRR*, abs/cs/0512085, 2005.

Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9.

A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*, 2007.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:668–677, 1999.

Peter Kollock and Marc Smith. Managing the virtual commons: Cooperation and conflict in computer communities. In Susan Herring, editor, *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*, pages 109–128. John Benjamins, Amsterdam, 1996.

N. Korfiatis. Evaluating wiki contributions using social networks: A case study on wikipedia. *Online Information Review*, 30:3, 2006.

N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

Jean Lave and Etienne Wenger. *Situated Learning : Legitimate Peripheral Participation*. Cambridge University Press, September 1991. ISBN 0521423740.

Andrew Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *5th International Symposium on Online Journalism*, 2004.

Sonya Lipczynska. Power to the people: the case for wikipedia. *Reference Reviews incorporating ASLIB Book Guide*, 19(2):6–7, February 2005. ISSN 0950-4125.

Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, January 2007. ISBN 3540378812.

Alfred J. Lotka. The frequency distribution of scientific productivity. *J Washington Acad Sci*, 16:317–324, 1926.

G. Madey, V. Freeh, and R. Tynan. The open source software development phenomenon: An analysis based on social network theory. In *Eighth Americas Conference on Information Systems*, 2002.

Deborah L. McGuinness, Honglei Zeng, Paulo Pinheiro da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, May 2006.

M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):016131+, June 2001a.

M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132+, June 2001b.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

Felipe Ortega and Jesus M. Gonzalez Barahona. Quantitative analysis of the wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-861-9.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

Gopal Pandurangan, Prabhakara Raghavan, and Eli Upfal. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.

Laura Rassbach, Trevor Pincock, and Brian Mingus. Exploring the feasibility of automatically rating online article quality, 2007.

Eric S. Raymond. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2001. ISBN 0596001088. Foreword By-Young,, Bob.

Dirk Riehle. How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 3–8, New York, NY, USA, 2006. ACM Press. ISBN 1595934138.

John P. Scott. *Social Network Analysis: A Handbook.* SAGE Publications, January 2000. ISBN 0761963391.

Anselm Spoerri. What is popular on wikipedia and why? *First Monday*, 12 (4), 2007.

B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality discussions in wikipedia. In *International Conference on Knowledge Management (ICKM) 2005. 7 - 9th July 2005*, 2005.

F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. In *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, page 78, 2007.

Fernanda B. Viegas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, New York, NY, USA, 2004. ACM. ISBN 1-58113-702-8.

Georg von Krogh and Sebastian Spaeth. The open source software phenomenon: Characteristics that promote research. *J. Strateg. Inf. Syst.*, 16 (3):236–253, 2007. ISSN 0963-8687.

Jakob Voss. Measuring wikipedia. In *10th International Conference of the International Society for Scientometrics and Informetrics 2005*. International Society for Scientometrics and Informetrics, 2005.

S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, Cambridge, United Kingdom, 1994.

Wikipedia. Wikipedia — wikipedia, l'enciclopedia libera, 2009a. URL `http://it.wikipedia.org/w/index.php?title=Wikipedia&oldid=21914379`. [Online; visualizzata 8-Febbraio-2009].

Wikipedia. Wikipedia:niente ricerche originali — wikipedia, l'enciclopedia libera, 2009b. URL `http://it.wikipedia.org/w/index.php?title=Wikipedia:Niente_ricerche_originali&oldid=20841628`. [Online; visualizzata 10-Febbraio-2009].

Wikipedia. Wikipedia:risoluzione dei conflitti — wikipedia, l'enciclopedia libera, 2009c. URL `http://it.wikipedia.org/w/index.php?title=Wikipedia:Risoluzione_dei_conflitti&oldid=21486218`. [Online; visualizzata 10-Febbraio-2009].

Wikipedia. Aiuto:namespace — wikipedia, l'enciclopedia libera, 2009d. URL `http://it.wikipedia.org/w/index.php?title=Aiuto:Namespace&oldid=19659215`. [Online; visualizzata 10-Febbraio-2009].

Wikipedia. Wikipedia:vetrina — wikipedia, l'enciclopedia libera, 2009e. URL `http://it.wikipedia.org/w/index.php?title=Wikipedia:Vetrina&oldid=22005957`. [Online; visualizzata 11-Febbraio-2009].

Wikipedia. Progetto:coordinamento/vetrina — wikipedia, l'enciclopedia libera, 2009f. URL `http://it.wikipedia.org/w/index.php?title=Progetto:Coordinamento/Vetrina&oldid=20967637#Note`. [Online; visualizzata 11-Febbraio-2009].

Wikipedia. Wikipedia:bot — wikipedia, l'enciclopedia libera, 2009g. URL `http://it.wikipedia.org/w/index.php?title=Wikipedia:Bot&oldid=21977435`. [Online; visualizzata 11-Febbraio-2009].

Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-861-9.

Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. Computing trust from revision history. In *PST*, page 8, 2006.

Allegato A

Classifica dei primi venti utenti di Wikipedia secondo differenti metriche.

Per le analisi delle classifiche globali e di sociometric star sono state prodotte le liste ordinate di tutti gli utenti per ciascuna delle versioni di Wikipedia analizzate. Per ciascuna sono quindi stati estratti i primi 100 utenti ed è stato verificato se essi appartenevano rispettivamente alla classe di Amministratore (A), Bot (B) o di semplice utente registrato (R). In quest'allegato verranno elencate le posizioni, i nomi, la classe di appartenenza e il punteggio totalizzato (a seconda delle differenti metriche) dei primi venti classificati.

Le didascalie delle tabelle conterranno le seguenti abbreviazioni:

- le sigle IT e EN indicheranno rispettivamente una versione italiana e inglese di Wikipedia;
- la sigla EL indicherà che la Rete Sociale per la quale si sta esaminando la classifica è stata costruita utilizzando la metrica di reputazione di longevità dell'intervento;
- la sigla ELS indicherà che la Rete Sociale per la quale si sta esaminando la classifica è stata costruita utilizzando la metrica di reputazione di longevità dell'intervento valutata rispetto alla sua versione più simile.

A.1 Versione di Wikipedia in italiano al 13.12.2005

Tabella A.1: IT2005 - Conteggio degli interventi globale

rank	user	type	value
1	Gacbot	B	37108.0
2	ZeroBot	B	26729.0
3	Snowdog	A	23915.0
4	Luki-Bot	B	19594.0
5	Alfiobot	B	16686.0
6	Twice25	R	16039.0
7	Gac	A	11744.0
8	Frieda	A	11137.0
9	Shaka	A	11072.0
10	YurikBot	B	10328.0
11	Marcok	A	9167.0
12	Hellisp	R	8639.0
13	C1PB8	B	8439.0
14	TierrayLibertad	A	8311.0
15	Civvi	A	7686.0
16	Sbisolo	A	7259.0
17	Alfio	A	7083.0
18	Cruccone	A	6937.0
19	Paginazero	A	6869.0
20	FioriCadenti	B	6780.0

Tabella A.2: IT2005 - Longevità dell'intervento globale

rank	user	type	value
1	Snowdog	A	1153350.0
2	Twice25	R	511495.0
3	Sbisolo	A	271732.0
4	Shaka	A	244924.0
5	Paginazero	A	227595.0
6	Civvi	A	196447.0
7	MM	R	191389.0
8	Renato Caniatti	R	176241.0
9	Marcok	A	172452.0
10	Margherita	R	161968.0
11	Luki-Bot	B	150560.0
12	Gac	A	147734.0
13	Alfio	A	143487.0
14	Frieda	A	143057.0
15	Gianluigi	R	142630.0
16	M7	R	136973.0
17	Kormoran	R	123021.0
18	Alberto da Calvairate	R	122589.0
19	Cosoleto	R	119477.0
20	Hill	A	110431.0

Tabella A.3: IT2005 - Longevità dell'intervento valutata rispetto alla sua versione più simile globale

rank	user	type	value
1	Snowdog	A	821648.0
2	Twice25	R	295126.0
3	Sbisolo	A	267415.0
4	Shaka	A	197893.0
5	Renato Caniatti	R	165655.0
6	Paginazero	A	155948.0
7	MM	R	150389.0
8	Margherita	R	146925.0
9	Luki-Bot	B	146177.0
10	Gianluigi	R	138491.0
11	Marcok	A	134696.0
12	Kormoran	R	118235.0
13	Alberto da Calvairate	R	109655.0
14	Ary29	A	88716.8
15	Pil56	A	83989.2
16	Civvi	A	82542.5
17	Frieda	A	82307.7
18	Alfio	A	81147.5
19	Gacbot	B	78830.2
20	Hill	A	76147.9

Tabella A.4: IT2005 - Betweenness Centrality (EL)

rank	user	type	value
1	Snowdog	A	97137.4
2	Twice25	R	72980.2
3	Marcok	A	54170.0
4	Shaka	A	43664.0
5	Frieda	A	35136.5
6	Alfio	A	31745.0
7	Gac	A	31489.4
8	M7	R	27774.3
9	Luki-Bot	B	25945.4
10	Ary29	A	24590.4
11	Sbisolo	A	22869.4
12	Paginazero	A	22442.9
13	Cruccone	A	20806.5
14	Civvi	A	19726.0
15	Blakwolf	R	15623.1
16	Hill	A	15474.6
17	Alberto da Calvairate	R	13605.3
18	Pil56	A	13074.8
19	Renato Caniatti	R	11418.0
20	Urby2004	A	10889.7

Tabella A.5: IT2005 - Betweenness Centrality (ELS)

rank	user	type	value
1	Snowdog	A	101099.02
2	Twice25	R	91499.55
3	Marcok	A	69952.25
4	Shaka	A	60752.42
5	Frieda	A	46670.49
6	Alfio	A	40590.44
7	Luki-Bot	B	38808.47
8	Sbisolo	A	30240.99
9	Ary29	A	26689.53
10	Gac	A	23256.63
11	Hill	A	20911.34
12	M7	R	20522.29
13	Blakwolf	R	20306.47
14	Civvi	A	19260.86
15	Renato Caniatti	R	17440.73
16	Paginazero	A	17440.49
17	Pil56	A	14713.42
18	Cruccone	A	14555.02
19	Hellisp	R	14402.07
20	Alberto da Calvairate	R	14174.4

Tabella A.6: IT2005 - Closeness Centrality (EL)

rank	user	type	value
1	Snowdog	A	0.020878
2	Twice25	R	0.020854
3	Marcok	A	0.0207971
4	Shaka	A	0.0207958
5	M7	R	0.0207905
6	Gac	A	0.0207843
7	Alfio	A	0.0207802
8	Civvi	A	0.0207794
9	Frieda	A	0.0207769
10	Ary29	A	0.0207765
11	Sbisolo	A	0.0207749
12	Luki-Bot	B	0.0207543
13	Blakwolf	R	0.0207531
14	Renato Caniatti	R	0.0207482
15	Suisui	R	0.0207408
16	Tomi	A	0.0207335
17	Paginazero	A	0.0207298
18	Cruccone	A	0.0207265
19	Lucius	R	0.0207167
20	Alberto da Calvairate	R	0.0207163

Tabella A.7: IT2005 - Closeness Centrality (ELS)

rank	user	type	value
1	Snowdog	A	0.022790743
2	Twice25	R	0.022788517
3	Shaka	A	0.022726387
4	Marcok	A	0.022722404
5	Luki-Bot	B	0.022705164
6	Frieda	A	0.022702955
7	Ary29	A	0.022696331
8	Sbisolo	A	0.022687508
9	Alfio	A	0.022682216
10	Renato Caniatti	R	0.022680452
11	Suisui	R	0.022677368
12	Blakwolf	R	0.022677368
13	M7	R	0.022675605
14	Gac	A	0.022653157
15	Civvi	A	0.02264876
16	TierrayLibertad	A	0.022633824
17	Alberto da Calvairate	R	0.022632947
18	Davide	R	0.022623733
19	Urby2004	A	0.022621978
20	Paginazero	A	0.022614526

Tabella A.8: IT2005 - Degree Centrality (EL)

rank	user	type	value
1	Snowdog	A	195.0
2	Twice25	R	170.0
3	Marcok	A	137.0
4	Shaka	A	119.0
5	Frieda	A	104.0
6	Gac	A	98.0
7	Ary29	A	97.0
8	M7	R	94.0
9	Alfio	A	93.0
10	Sbisolo	A	89.0
11	Luki-Bot	B	88.0
12	Civvi	A	84.0
13	Blakwolf	R	67.0
14	Suisui	R	67.0
15	Renato Caniatti	R	65.0
16	Paginazero	A	64.0
17	Davide	R	62.0
18	Cruccone	A	60.0
19	NTBot	R	59.0
20	Alberto da Calvairate	R	58.0

Tabella A.9: IT2005 - Degree Centrality (ELS)

rank	user	type	value
1	Snowdog	A	234.0
2	Twice25	R	226.0
3	Marcok	A	177.0
4	Shaka	A	158.0
5	Frieda	A	134.0
6	Luki-Bot	B	128.0
7	Sbisolo	A	120.0
8	Ary29	A	119.0
9	Alfio	A	117.0
10	Blakwolf	R	102.0
11	Renato Caniatti	R	100.0
12	Suisui	R	95.0
13	M7	R	93.0
14	Gac	A	91.0
15	Davide	R	86.0
16	Civvi	A	85.0
17	Alberto da Calvairate	R	77.0
18	TierrayLibertad	A	75.0
19	Pil56	A	73.0
20	Hill	A	70.0

Tabella A.10: IT2005 - Eigenvector Centrality (EL)

rank	user	type	value
1	Gacbot	B	1.0
2	Luki-Bot	B	0.958058
3	Snowdog	A	0.321956
4	Ary29	A	0.279109
5	Nuno Tavares	R	0.277106
6	NTBot	R	0.23325
7	Suisui	R	0.204787
8	Davide	R	0.195715
9	Twice25	R	0.172338
10	YurikBot	B	0.127765
11	Frieda	A	0.115216
12	Gac	A	0.113018
13	Robbot	B	0.107178
14	M7	R	0.105374
15	Sentruper	R	0.103843
16	Lucius	R	0.103574
17	Civvi	A	0.0926033
18	Shaka	A	0.0918019
19	Tomi	A	0.0814234
20	GiorgioPro	R	0.0783213

Tabella A.11: IT2005 - Eigenvector Centrality (ELS)

rank	user	type	value
1	Luki-Bot	B	1.0
2	Gacbot	B	0.96
3	Snowdog	A	0.84
4	Twice25	R	0.77
5	Ary29	A	0.77
6	Suisui	R	0.58
7	Davide	R	0.57
8	NTBot	R	0.51
9	Blakwolf	R	0.36
10	Shaka	A	0.34
11	Lucius	R	0.32
12	Marcok	A	0.3
13	Sbisolo	A	0.29
14	Robbot	B	0.29
15	M7	R	0.29
16	Frieda	A	0.28
17	Renato Caniatti	R	0.27
18	Civvi	A	0.25
19	TierrayLibertad	A	0.25
20	Gac	A	0.25

A.2 Versione di Wikipedia in italiano al 22.05.2007

Tabella A.12: IT2007 - Conteggio degli interventi globale

rank	user	type	value
1	YurikBot	B	136509.0
2	ZeroBot	B	129455.0
3	Thijs!bot	B	99925.0
4	Gacbot	B	61110.0
5	FlaBot	R	56432.0
6	Eskimbot	R	53727.0
7	.anacondabot	B	52035.0
8	Alfiobot	B	50196.0
9	TekBot	B	42841.0
10	SashatoBot	R	38337.0
11	Snowdog	A	37779.0
12	SunBot	B	26328.0
13	CruccoBot	B	26232.0
14	Gac	A	24257.0
15	Hellis	A	23003.0
16	Marcok	A	22801.0
17	Ary29	A	21848.0
18	Senpai	A	21623.0
19	Paginazero	A	20864.0
20	Twice25	R	20608.0

Tabella A.13: IT2007 - Longevità dell'intervento globale

rank	user	type	value
1	Gacbot	B	3748880.0
2	Snowdog	A	1649590.0
3	ZeroBot	B	1089270.0
4	.snoopy.	R	964410.0
5	Paginazero	A	862554.0
6	Civvì	R	788019.0
7	Senpai	A	768220.0
8	Gac	A	709439.0
9	.anaconda	R	697830.0
10	M7	R	662629.0
11	Luisa	A	647740.0
12	Twice25	R	630500.0
13	MM	A	626376.0
14	Valepert	A	545159.0
15	FioriCadenti	B	541089.0
16	Al Pereira	A	536347.0
17	Shaka	A	471502.0
18	Pil56	A	459400.0
19	Retaggio	A	433274.0
20	Marcok	A	433072.0

Tabella A.14: IT2007 - Betweenness Centrality (EL)

rank	user	type	value
1	Gacbot	B	2284970.0
2	Snowdog	A	1338200.0
3	Marcok	A	1007540.0
4	Twice25	R	974866.0
5	Shaka	A	919819.0
6	Civvì	R	750756.0
7	Senpai	A	655877.0
8	Pil56	A	638937.0
9	Gac	A	616122.0
10	Ary29	A	605788.0
11	YurikBot	B	592948.0
12	Hellis	A	531450.0
13	Luisa	A	525848.0
14	Al Pereira	A	519121.0
15	Paginazero	A	492864.0
16	ZeroBot	B	478797.0
17	Moongateclimber	R	459102.0
18	FlaBot	R	423287.0
19	Kal-El	A	412721.0
20	M7	R	391371.0

Tabella A.15: IT2007 - Closeness Centrality (EL)

rank	user	type	value
1	Snowdog	A	0.00381602
2	Gacbot	B	0.00381587
3	Twice25	R	0.00381523
4	Civvì	R	0.00381482
5	Ary29	A	0.00381476
6	YurikBot	B	0.00381474
7	Shaka	A	0.00381453
8	Marcok	A	0.00381447
9	Luisa	A	0.00381443
10	Senpai	A	0.00381438
11	Gac	A	0.00381417
12	Paginazero	A	0.00381415
13	Al Pereira	A	0.00381397
14	Pil56	A	0.00381382
15	Luki-Bot	B	0.00381349
16	M7	R	0.00381348
17	.snoopy.	R	0.00381344
18	FlaBot	R	0.00381342
19	Moongateclimber	R	0.0038134
20	ZeroBot	B	0.00381338

Tabella A.16: IT2007 - Degree Centrality (EL)

rank	user	type	value
1	Gacbot	B	762.0
2	Snowdog	A	597.0
3	Twice25	R	471.0
4	Marcok	A	430.0
5	Shaka	A	411.0
6	Civvì	R	392.0
7	YurikBot	B	384.0
8	Ary29	A	366.0
9	Gac	A	342.0
10	Senpai	A	339.0
11	Luisa	A	317.0
12	Luki-Bot	B	312.0
13	Pil56	A	307.0
14	Paginazero	A	301.0
15	Al Pereira	A	293.0
16	ZeroBot	B	288.0
17	Moongateclimber	R	272.0
18	Hellis	A	271.0
19	M7	R	267.0
20	Moloch981	R	261.0

Tabella A.17: IT2007 - Eigenvector Centrality (EL)

rank	user	type	value
1	Gacbot	B	1.0
2	Luki-Bot	B	0.921799
3	Escarbot	B	0.564668
4	Thijs!bot	B	0.492745
5	Snowdog	A	0.435403
6	Moloch981	R	0.32365
7	C1PB8	B	0.313277
8	CruccoBot	B	0.256777
9	Gac	A	0.252704
10	YurikBot	B	0.22286
11	Chlewbot	B	0.199643
12	Biobot	B	0.177099
13	Ary29	A	0.133014
14	NTBot	R	0.115506
15	TuvicBot	R	0.105556
16	FioriCadenti	B	0.102278
17	Alexale	R	0.101641
18	ZeroBot	B	0.0968776
19	Eskimbot	R	0.0854848
20	Sentruper	R	0.08334

A.3 Versione di Wikipedia in italiano al 17.03.2008

Tabella A.18: IT2008 - Conteggio degli interventi globale

rank	user	type	value
1	ZeroBot	B	127191.0
2	YurikBot	B	112235.0
3	SieBot	B	79216.0
4	SunBot	B	58952.0
5	Gacbot	B	53730.0
6	Eskimbot	B	46062.0
7	FlaBot	R	40443.0
8	SashatoBot	B	40020.0
9	Alfiobot	B	39121.0
10	CruccoBot	B	29989.0
11	Escarbot	B	28294.0
12	Snowdog	A	28059.0
13	.anacondabot	B	21828.0
14	Paginazero	A	17735.0
15	Luki-Bot	B	17648.0
16	Gac	R	17302.0
17	Twice25	R	17080.0
18	Marcok	A	16908.0
19	Abbot	B	16593.0
20	TekBot	B	16019.0

Tabella A.19: IT2008 - Longevità dell'intervento globale

rank	user	type	value
1	Snowdog	A	1115310.0
2	.snoopy.	A	649072.0
3	.anaconda	A	638408.0
4	M7	R	578562.0
5	Civvì	R	511210.0
6	Paginazero	A	485987.0
7	Twice25	R	477092.0
8	Senpai	A	473881.0
9	Giovannigobbin	R	471163.0
10	Retaggio	A	440501.0
11	Murray	R	423904.0
12	Luisa	R	399705.0
13	Gac	R	383654.0
14	MM	A	383650.0
15	Marcok	A	343745.0
16	Al Pereira	A	343180.0
17	Moroboshi	A	336634.0
18	Phantomas	A	333954.0
19	Shaka	R	293063.0
20	ZeroBot	B	285405.0

Tabella A.20: IT2008 - Longevità dell'intervento globale valutata rispetto alla sua versione più simile

rank	user	type	value
1	Snowdog	A	796649.0
2	Murray	R	403788.0
3	MM	A	303436.0
4	Twice25	R	303174.0
5	ZeroBot	B	268880.0
6	Shaka	R	246293.0
7	Sailko	R	222438.0
8	Marcok	A	218630.0
9	Paginazero	A	212970.0
10	Dapa19	R	205075.0
11	Civvì	R	185802.0
12	Ary29	A	180562.0
13	Margherita	R	176412.0
14	Moroboshi	A	163484.0
15	Pil56	A	162163.0
16	YurikBot	B	161879.0
17	Moongateclimber	A	147346.0
18	Kanchelskis	R	146770.0
19	Luki-Bot	B	135436.0
20	Actarux	R	134560.0

Tabella A.21: IT2008 - Betweenness Centrality (EL)

rank	user	type	value
1	Snowdog	A	1144440.0
2	Marcok	A	979666.0
3	Twice25	R	822636.0
4	Shaka	R	741030.0
5	Civvì	R	576855.0
6	Ary29	A	558301.0
7	Gac	R	558089.0
8	M7	R	532594.0
9	Paginazero	A	491720.0
10	Pil56	A	387816.0
11	Luki-Bot	B	377995.0
12	.snoopy.	A	370091.0
13	Luisa	R	362241.0
14	Senpai	A	351133.0
15	Moongateclimber	A	327258.0
16	MM	A	320738.0
17	Alfio	R	293011.0
18	Retaggio	A	287199.0
19	Cruccone	A	286468.0
20	Frieda	A	279551.0

Tabella A.22: IT2008 - Betweenness Centrality (ELS)

rank	user	type	value
1	Snowdog	A	1617660.9
2	Twice25	R	1578227.2
3	Marcok	A	1469164.9
4	Shaka	R	1215362.6
5	Ary29	A	673039.4
6	Luki-Bot	B	659141.75
7	Pil56	A	503805.2
8	Civvì	R	502930.8
9	Alfio	R	502596.44
10	Paginazero	A	478342.2
11	YurikBot	B	457426.12
12	MM	A	448764.9
13	Moongateclimber	A	425835.12
14	Frieda	A	424663.75
15	Biopresto	A	403092.34
16	Gac	R	399661.8
17	Sbisolo	A	398191.44
18	Hellis	A	390998.06
19	Sailko	R	331020.72
20	Moloch981	R	321067.53

Tabella A.23: IT2008 - Closeness Centrality (EL)

rank	user	type	value
1	Snowdog	A	0.00282402
2	Twice25	R	0.00282352
3	Civvì	R	0.00282341
4	Marcok	A	0.00282336
5	Gac	R	0.0028232
6	Shaka	R	0.00282318
7	M7	R	0.00282315
8	Ary29	A	0.00282299
9	Paginazero	A	0.00282286
10	.snoopy.	A	0.00282282
11	Retaggio	A	0.00282255
12	Gacbot	B	0.0028225
13	Senpai	A	0.00282249
14	Luisa	R	0.00282236
15	Luki-Bot	B	0.00282235
16	Moloch981	R	0.00282225
17	TierrayLibertad	R	0.00282222
18	.anaconda	A	0.00282221
19	Frieda	A	0.00282221
20	Pil56	A	0.00282215

Tabella A.24: IT2008 - Closeness Centrality (ELS)

rank	user	type	value
1	Twice25	R	0.002584
2	Snowdog	A	0.002584
3	Marcok	A	0.002583
4	Shaka	R	0.002583
5	Luki-Bot	B	0.002583
6	Ary29	A	0.002583
7	YurikBot	B	0.002582
8	Civvì	R	0.002582
9	Alfio	R	0.002582
10	Sbisolo	A	0.002582
11	Gacbot	B	0.002582
12	Frieda	A	0.002582
13	Moloch981	R	0.002582
14	Pil56	A	0.002582
15	MM	A	0.002582
16	Hellis	A	0.002582
17	TierrayLibertad	R	0.002582
18	Gac	R	0.002582
19	Blakwolf	R	0.002582
20	Paginazero	A	0.002582

Tabella A.25: IT2008 - Degree Centrality (EL)

rank	user	type	value
1	Snowdog	A	450.0
2	Marcok	A	369.0
3	Twice25	R	358.0
4	Shaka	R	309.0
5	Civvì	R	303.0
6	Ary29	A	291.0
7	M7	R	280.0
8	Gac	R	277.0
9	Paginazero	A	261.0
10	Luki-Bot	B	259.0
11	.snoopy.	A	231.0
12	Gacbot	B	228.0
13	Luisa	R	213.0
14	Moloch981	R	206.0
15	Senpai	A	194.0
16	Pil56	A	192.0
17	Retaggio	A	191.0
18	Moongateclimber	A	179.0
19	Alfio	R	175.0
20	MM	A	174.0

Tabella A.26: IT2008 - Degree Centrality (ELS)

rank	user	type	value
1	Twice25	R	630.0
2	Snowdog	A	616.0
3	Marcok	A	552.0
4	Shaka	R	455.0
5	Luki-Bot	B	404.0
6	Ary29	A	359.0
7	YurikBot	B	314.0
8	Civvì	R	294.0
9	Alfio	R	280.0
10	Gacbot	B	262.0
11	Paginazero	A	251.0
12	Frieda	A	251.0
13	Pil56	A	248.0
14	Moloch981	R	247.0
15	Gac	R	237.0
16	Sbisolo	A	234.0
17	MM	A	232.0
18	Hellis	A	227.0
19	Moongateclimber	A	225.0
20	Cruccone	A	214.0

Tabella A.27: IT2008 - Eigenvector Centrality (EL)

rank	user	type	value
1	Luki-Bot	B	1.0
2	Gacbot	B	0.998392
3	Escarbot	B	0.912338
4	Moloch981	R	0.292444
5	Chlewbot	B	0.292057
6	CruccoBot	B	0.174115
7	Ary29	A	0.146896
8	KocjoBot	R	0.144943
9	YurikBot	B	0.126076
10	NTBot	B	0.0968164
11	Caulfield	A	0.0807674
12	Panairjdde	R	0.078821
13	Sentruper	R	0.0739949
14	Snowdog	A	0.07348
15	Davide	R	0.0699253
16	DorianaV.	R	0.0502988
17	Cloj	R	0.0429489
18	Twice25	R	0.0379827
19	Rdocb	A	0.034815
20	Gac	R	0.0342502

Tabella A.28: IT2008 - Eigenvector Centrality (ELS)

rank	user	type	value
1	Luki-Bot	B	1.0
2	Escarbot	B	0.954
3	Gacbot	B	0.794
4	Moloch981	R	0.328
5	Chlewbot	B	0.275
6	CruccoBot	B	0.2
7	Ary29	A	0.163
8	KocjoBot	R	0.148
9	YurikBot	B	0.145
10	NTBot	B	0.103
11	Snowdog	A	0.1
12	Panairjdde	R	0.1
13	Davide	R	0.09
14	Caulfield	A	0.081
15	Sentruper	R	0.077
16	Twice25	R	0.068
17	Cloj	R	0.055
18	DorianaV.	R	0.054
19	Rdocb	A	0.04
20	Civvì	R	0.039

A.4 Versione di Wikipedia in inglese al 06.02.2007

Tabella A.29: EN2007 - Conteggio degli interventi globale

rank	user	type	value
1	Bluebot	B	354471.0
2	SmackBot	B	318115.0
3	YurikBot	B	263161.0
4	D6	B	174427.0
5	AntiVandalBot	B	135886.0
6	Rich Farmbrough	A	121659.0
7	FlaBot	B	120526.0
8	RussBot	B	110884.0
9	Rambot	B	110556.0
10	Cydebot	B	106720.0
11	Alaibot	B	101046.0
12	CmdrObot	B	97552.0
13	Thijs!bot	R	80847.0
14	Pearle	B	77939.0
15	Tawkerbot2	B	70477.0
16	SimonP	A	65359.0
17	Everyking	A	64495.0
18	Bobblewik	R	62032.0
19	OrphanBot	B	60051.0
20	Bryan Derksen	A	58715.0

Tabella A.30: EN2007 - Longevità dell'intervento globale

1	AntiVandalBot	B	1.54982E8
2	Tawkerbot2	B	7.9321504E7
3	Curps	A	1.73712E7
4	RexNL	A	1.6521E7
5	Wiki alf	A	1.55372E7
6	Tawkerbot4	B	1.51391E7
7	Nakon	R	1.50512E7
8	Antandrus	A	1.44912E7
9	MER-C	R	1.14126E7
10	Luna Santin	A	1.09964E7
11	Everyking	A	1.04266E7
12	Can't sleep, clown will eat me	A	9831660.0
13	Pgk	A	8121540.0
14	Ahoerstemeier	A	8037990.0
15	Ryulong	R	7406540.0
16	Wayward	A	7067020.0
17	GraemeL	A	6947110.0
18	Peruvianllama	A	6854080.0
19	Shanes	A	6724360.0
20	Titoxd	A	6690800.0

Tabella A.31: EN2007 - Betweenness Centrality (EL)

rank	user	type	value
1	AntiVandalBot	B	9.2356301E8
2	Tawkerbot2	B	5.4576301E8
3	Can't sleep, clown will eat me	A	1.14059E8
4	Tawkerbot4	B	8.51678E7
5	Wiki alf	A	8.2322096E7
6	SmackBot	B	7.94404E7
7	RexNL	A	7.58974E7
8	Rambot	B	7.4174496E7
9	MER-C	R	7.1054096E7
10	Antandrus	A	6.82658E7
11	Ahoerstemeier	A	6.59763E7
12	Everyking	A	5.58109E7
13	Luna Santin	A	5.21519E7
14	Nakon	R	4.98052E7
15	SimonP	A	4.79159E7
16	Dale Arnett	A	4.69095E7
17	Wetman	R	4.65524E7
18	RoyBoy	A	4.60353E7
19	Bryan Derksen	A	4.48202E7
20	Olivier	A	4.40904E7

Tabella A.32: EN2007 - Closeness Centrality (EL)

rank	user	type	value
1	AntiVandalBot	B	1.40944E-4
2	Tawkerbot2	B	1.40943E-4
3	Can't sleep, clown will eat me	A	1.40938E-4
4	Tawkerbot4	B	1.40938E-4
5	Wiki alf	A	1.40938E-4
6	RexNL	A	1.40938E-4
7	Antandrus	A	1.40937E-4
8	Ahoerstemeier	A	1.40937E-4
9	MER-C	R	1.40937E-4
10	Everyking	A	1.40937E-4
11	Nakon	R	1.40937E-4
12	Luna Santin	A	1.40937E-4
13	RoyBoy	A	1.40936E-4
14	Curps	A	1.40936E-4
15	VoABot II	B	1.40936E-4
16	SimonP	A	1.40936E-4
17	Pgk	A	1.40936E-4
18	Bryan Derksen	A	1.40935E-4
19	Wetman	R	1.40935E-4
20	Hadal	A	1.40935E-4

Tabella A.33: EN2007 - Degree Centrality (EL)

rank	user	type	value
1	AntiVandalBot	B	10730.0
2	Tawkerbot2	B	7842.0
3	Can't sleep, clown will eat me	A	2729.0
4	Tawkerbot4	B	2685.0
5	Wiki alf	A	2570.0
6	RexNL	A	2437.0
7	Antandrus	A	2235.0
8	MER-C	R	2168.0
9	Luna Santin	A	2008.0
10	Ahoerstemeier	A	1980.0
11	Nakon	R	1896.0
12	Everyking	A	1838.0
13	SmackBot	B	1823.0
14	VoABot II	B	1781.0
15	Rambot	B	1779.0
16	RoyBoy	A	1650.0
17	Curps	A	1473.0
18	Wetman	R	1455.0
19	Pgk	A	1436.0
20	Ryulong	R	1390.0

Tabella A.34: EN2007 - Eigenvector Centrality (EL)

rank	user	type	value
1	SmackBot	B	1.0
2	Rambot	B	0.986681
3	KevinBot	B	0.386164
4	Rich Farmbrough	A	0.0614892
5	LouI	A	0.0546235
6	Bkonrad	A	0.0352375
7	Stepp-Wulf	R	0.0281709
8	AshyLarry	R	0.0225722
9	Marvin01	R	0.0180855
10	Theduckman1763	R	0.0166762
11	WhisperToMe	A	0.0152289
12	TooPotato	R	0.0100468
13	Dual Freq	R	0.00939417
14	Seth Ilys	A	0.0088015
15	CapitalR	R	0.00781116
16	Badbilltucker	R	0.00764774
17	Epolk	R	0.00763356
18	Swid	R	0.00752088
19	DragonflySixtyseven	A	0.0068484
20	ArkansasTraveler	R	0.0068102