POLITECNICO DI MILANO

Corso di Laurea in Ingegneria Informatica

Dipartimento di Elettronica e Informazione

SEMANTICA EMERGENTE IN WIKIPEDIA

Relatore: Prof. Marco Colombetti

Correlatori: Ing. David Laniado, Ing. Riccardo Tasso

Tesi di laurea di:

Fabio Colzada, matricola 713990

Mattia Di Vitto, matricola 714583

Anno accademico 2010-2011

Ai nostri genitori

Nel 90% dei casi il problema è tra la sedia e la tastiera

(Anonimo)

Indice

1. Introduzione

L'evoluzione del web verificatasi nell' ultimo decennio, ha privilegiato (e continua tutt'oggi a farlo) un sistema di creazione di contenuti basato principalmente sul contributo di molti piuttosto che sull'intervento di pochi. L'esempio più significativo di questa tendenza è la tecnologia wiki: in un sito wiki i contenuti vengono aggiornati dai suoi utenti che li sviluppano attraverso un'attiva collaborazione.
La modifica di una pagina può essere aperta a tutti o ai soli utenti registrati, consentendo non solo l'aggiunta di nuovi contenuti ma anche la modifica e la cancellazione di quelli già esistenti inseriti dagli autori precedenti. Il sistema prevede la memorizzazione dell'evoluzione di ogni contenuto in modo tale da consentirne il ripristino a una versione cronologicamente precedente qualora si ritenesse necessario (casi di vandalismo, contributi di scarsa qualità...).

L'esempio più emblematico e di successo dell'utilizzo della tecnologia wiki è Wikipedia: un'enciclopedia online tra i siti più visitati del web (il quinto al mondo nel 2011, secondo i dati raccolti da Google), multilingue, collaborativa e gratuita, istituita da Wikimedia Foundation (organizzazione statunitense senza fini di lucro), la cui caratteristica principale è il sistema aperto di modifica e pubblicazione dei contenuti che consente a chiunque di collaborare nella realizzazione di un articolo. La possibilità di aggregare i contributi, e quindi le conoscenze, di centinaia di migliaia di utenti, con dibattiti interni sui contenuti di ciascuna pagina, ha portato la qualità di questo progetto al pari di quella raggiunta dalle più importanti enciclopedie commerciali. Anche se in un determinato istante temporale possono esistere errori, imprecisioni o veri e propri vandalismi tra gli articoli dell'enciclopedia,

la peculiarità di avere spesso decine di utenti attivi su ogni pagina porta i difetti a scomparire molto rapidamente o comunque a essere in numero trascurabile rispetto alla totalità di informazioni disponibili.

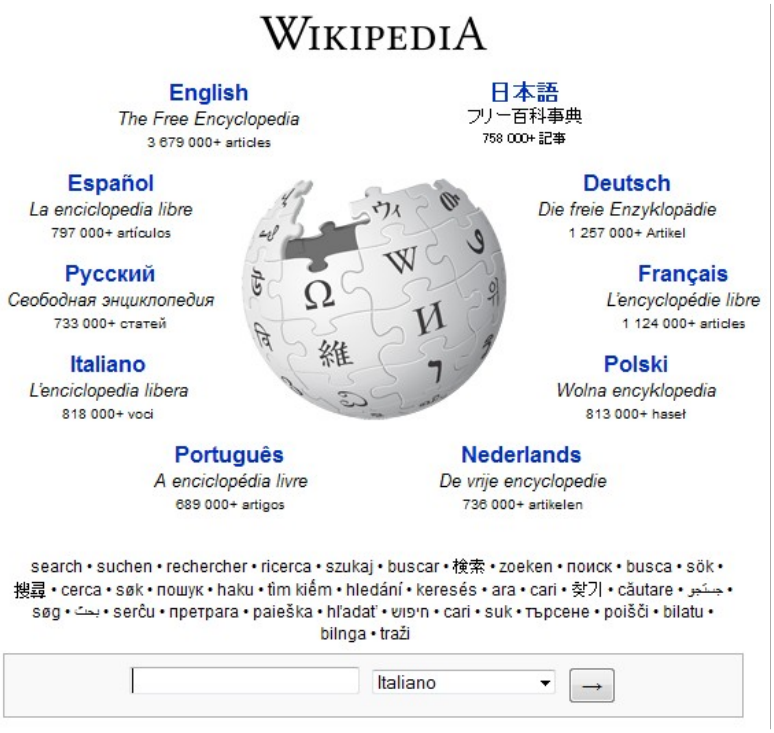

Fig. 1: Home page di Wikipedia. L'enciclopedia è disponibile in molteplici lingue. Di ciascuna versione vengono forniti regolari dump dei database, fonte di informazione per ricerche e progetti.

Il progetto Wikipedia mette a disposizione un'enorme mole di dati riguardanti non solo i contenuti delle sue pagine, ma anche meta-dati e informazioni sull'evoluzione nel tempo degli articoli ad opera degli utenti. Queste informazioni vengono spesso usate per compiere studi statistici di vario genere diventando anche punto di partenza di molteplici progetti di ricerca.

1.1 Obiettivi

Lo scopo di questo lavoro è verificare se l'attività e la collaborazione degli utenti di Wikipedia sui vari articoli possa portare all'emergere di una struttura di comunità tra le pagine dell'enciclopedia e se questa costituisca una semantica emergente.

Si andrà quindi a costruire una rete di pagine in cui ciascuna di queste risulterà collegata a quelle ad essa più simili dal punto di vista dei contributori e verranno applicati ad essa algoritmi non supervisionati di identificazione di comunità opportunamente selezionati. Per giungere a tale risultato è necessario passare per una rete bipartita utente-pagina dove gli utenti sono collegati alle pagine con un determinato peso che dipende dalla qualità e quantità del contributo prodotto dall'utente su quella pagina.
Successivamente si esaminerà la posizione degli elementi dei gruppi identificati all'interno della struttura delle categorie in Wikipedia, al fine di identificare eventuali correlazioni semantiche delle pagine di uno stesso cluster.

2. Stato dell'arte

La possibilità di estrarre una semantica emergente da una piattaforma collaborativa è stata molto studiata con l'avvento delle prime folksonomy.

Il termine folksonomy – da folk (persone) e taxonomy (tassonomia) – è un neologismo che indica una categorizzazione effettuata automaticamente tramite l'attività di tagging collaborativa degli utenti di una comunità, offrendo così a ciascuno la possibilità di descrivere un inseme di contenuti usando parole chiave di propria scelta ([1]).

Una delle più importanti e studiate piattaforme basate sul principio della folksonomia è sicuramente il sito di social bookmarking deli.cio.us, nato nel 2004, dove gli utenti scelgono liberamente tag da associare a una data pagina web.

In [1] si suggerisce di vedere una folksonomy come un grafo tripartito costituito da tre set di dati: attori, concetti e istanze. In del.icio.us gli attori sono gli utenti, che assegnano alle pagine web (istanze) dei tag (concetti). Attori, concetti e istanze sono quindi organizzati in tre set di dati $A=\{a_1,\ldots,a_k\}$, $C=\{c_1,\ldots,c_l\}$, $I=\{i_1,\ldots,i_m\}$ tra i quali si instaurano relazioni di tipo ternario $T \subseteq A\times C\times I$. Il grafo tripartito sarà quindi definito come $H(T)=\langle V,E\rangle$, dove l'insieme dei vertici è $V = A\cup C\cup I$ e l'insieme degli archi è $E=\{\{a,c,i\}\,|\,(a,c,i)\in T\}$. La difficoltà di lavorare con tale tipologia di grafo è affrontata scomponendolo in tre diverse reti bipartite (two-mode graph): una rete attori-concetti (AC), una attori-istanze (AI) e una concetti-istanze (CI). Le reti bipartite possono poi essere portare in reti di tipo one-mode (ossia con una sola tipologia di vertici) mediante proiezione su uno dei due tipi di vertici.

Con una tale varietà di dati è quindi possibile andare ad analizzare un totale di sei reti diverse, a seconda della modalità scelta per giungere

ai vertici e agli archi finali.

L'autore si concentra sulla rete di tipo AC e CI: nella prima due tag sono collegati da un arco qualora un utente li abbia scelti entrambi in una qualche pagina, nella seconda due tag vengono collegati da un arco qualora siano stati utilizzati in una stessa pagina (non necessariamente dallo stesso utente).

I risultati mostrano già l'emergere di una struttura di comunità in cui i tag sono associati in gruppi inerenti a un determinato argomento (Fig. 2). Nonostante nessuna specifica analisi semantica sia stata effettuata sui risultati (ci si è limitati a lavorare su un ristretto set di elementi analizzando manualmente i cluster ottenuti) questi non perdono di validità e mostrano come esista la possibilità di individuare una vera e propria ontologia emergente grazie all'interazione e ai contributi forniti dai singoli utenti in una rete collaborativa.

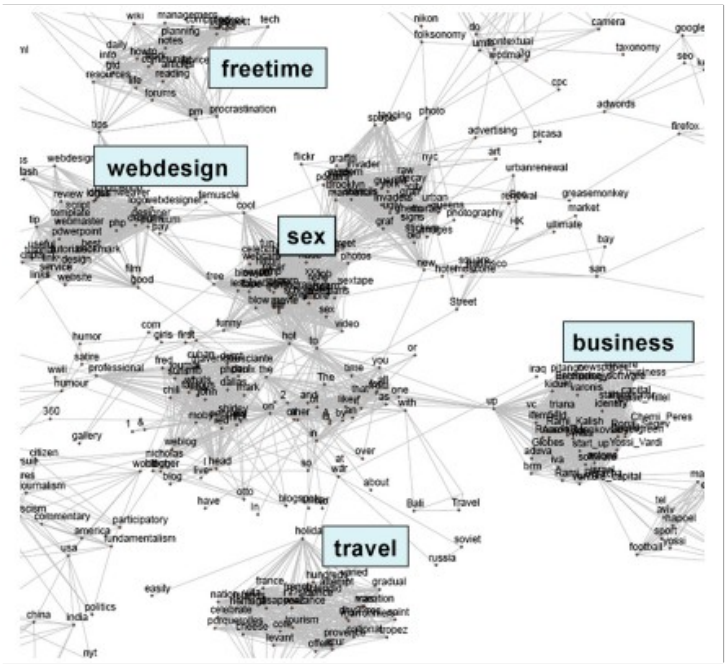

Fig. 2: Risultati presentati in [1]. In evidenza i cluster emersi e il nome dell'argomento coperto dai vari tag.

Un lavoro analogo viene presentato in [2], dove si è portata l'attenzione

sulle differenze nei risultati in base a diverse modalità utilizzare per il calcolo della relatedness tra due pagine.
Gli autori identificano i tag più usati della piattaforma (per avere un set limitato di dati da analizzare) e successivamente analizzano il tipo di relazione semantica esistente tra i termini utilizzando un database semantico-lessicale, WordNet. I vari tag oggetto di studio vengono identificati nel database e se ne calcola la distanza.

Anche Wikipedia del resto si può intendere come folksonomia, in maniera analoga a del.icio.us, e in maniera analoga si possono tradurre le modalità di rilevamento delle semantiche emergenti. In Wikipedia gli utenti producono edit sulle pagine, e inseriscono risposte ai topic di discussione sui vari argomenti dell'enciclopedia. Mentre anche qui gli attori restano gli utenti della piattaforma, possiamo identificare i concetti come i vari articoli o le relative pagine di discussione (sempre tuttavia riferiti agli articoli). Si è dunque considerato sufficiente un modello bipartito, benché siano state anche realizzate reti tripartite basate sull'enciclopedia.
La varietà di dati e le possibilità di analisi sono ora molto maggiori rispetto al sito di social bookmarking: qui possiamo infatti realizzare molteplici reti bipartite i cui nodi potrebbero essere scelti tra utenti, pagine, categorie, discussioni, edit, … .
Il fatto che i concetti non siano più una sola parola, ma veri e propri testi, con possibilità di modificare, cancellare e aggiungere i contributi propri e di altri utenti, rende possibile l'applicazione di vari algoritmi di misura del peso di un contributo, non solo basandosi sulla quantità del testo inserito, ma anche, e soprattutto, sulla qualità. Come si vedrà infatti in 3.1.1, sono stati sviluppati diversi sistemi di valutazione di un edit di un utente, ciascuno in grado di porre risalto su un particolare aspetto.

La possibilità di approfondire tutti questi molteplici aspetti ha portato alla realizzazione di progetti molto vari e interessanti.
Ad esempio uno studio che verte sulle relazioni tra gli utenti di Wikipedia, e che in parte si identifica come il duale di questa tesi, è [6] in cui viene realizzata una rete degli utenti che hanno contribuito alle medesime pagine e dove viene proposta una versione migliorata della metrica "Edit Longevity", necessaria per misurare la qualità degli interventi degli autori, della quale si tratterà in seguito.
Nello stesso lavoro viene anche proposto un sistema per identificare come autori di una pagina quegli utenti che hanno fornito la maggior parte del suo contenuto rilevante, sfruttato in [4] per costruire una rete di coautori al fine di poter identificare sotto-comunità attive su specifici argomenti.

Un altro aspetto interessante di Wikipedia è legato alle pagine di discussione degli articoli (talk pages), i cui commenti vengono utilizzati in [5] per creare molteplici reti corrispondenti a diversi tipi di interazione tra gli utenti. Si possono così identificare schemi di relazioni mediante l'analisi della "directed degree assortativity" (ossia una misura di diversità nelle reti, che quantifica la tendenza dei nodi a collegarsi con altri che possiedono un numero simile di archi) e studiare le differenze tra discussioni relative ad articoli riguardanti diverse aree semantiche.

Volendo è possibile vedere Wikipedia come folksonomia sotto un altro aspetto interessante e meno evidente: l'assegnamento di categorie alle varie pagine del portale viene effettuato manualmente dagli utenti. Questa attività è di fatto molto simile al sistema di tagging, e porta a una catalogazione e gerarchizzazione naturale delle categorie grazie al

contributo di migliaia di singoli utenti. Questo aspetto viene analizzato in [7], dove si esplora la possibilità di assegnare automaticamente una pagina alla categoria per essa più rappresentativa sfruttando il grafo degli assegnamenti alle categorie (di fatto categorie e pagine vengono viste come lo stesso tipo di vertice in questa rete, gli archi sono invece di due tipi: uno rappresentante l'appartenenza di una categoria a un'altra e uno che indica l'appartenenza di una pagina a una categoria). Successivamente viene comparato l'assegnamento automatico a quello manuale rilevando il grado di qualità raggiunto a seconda del criterio di assegnamento usato.

3. Metodologia

Viene di seguito illustrata la metodologia adottata per la costruzione del grafo e le analisi successive.

3.1 Costruzione del grafo

Il primo obiettivo da raggiungere è arrivare a rappresentare i dati in un grafo pesato e non orientato che abbia come vertici le pagine di Wikipedia. Si è deciso di usare come modello di partenza una rete bipartita utente-pagina che mettesse in relazione gli utenti dell'enciclopedia con gli articoli a cui hanno contribuito. Eseguendo successivamente una proiezione della rete bipartita sulle pagine si otterrà il grafo desiderato.

3.1.1 Metrica di valutazione dei contributi

Tappa preliminare alla creazione della rete di pagine è la costruzione di una rete bipartita utente-pagina, in cui gli utenti sono connessi agli articoli di Wikipedia a cui hanno contribuito mediante un peso.
Prima di poter calcolare quest'ultimo è però necessario definire una misura della qualità degli interventi degli autori (che può anche essere interpretata come il contributo fornito da un utente in un articolo). Tra le molteplici misure disponibili, la scelta è ricaduta su "Edit Longevity". Come già mostrato in [8] questo sistema non soffre dei difetti propri della maggior parte dei sistemi di misurazione della qualità di un intervento (viene qui adottata la stessa terminologia

proposta in [8]): “Number of Edits”, “Text Only”, “Edit Only”, “Text Longevity”, “Ten Revisions”.
La metrica di Number of Edits consiste solamente nel numero di edit effettuati da un autore su una pagina, un sistema troppo semplificato che non tiene conto né della qualità né della dimensione di un edit.
Text Only tiene semplicemente conto del numero di parole scritte da un autore in tutti gli interventi effettuati sulla pagina, senza verificare l'effettiva qualità del contributo aggiunto.
Edit Only consiste nel misurare la differenza tra una revision dell'articolo e quella precedente semplicemente in termini di quantità di testo aggiunto, modificato o cancellato.
Ten Revisions controlla quanto del testo originale sopravvive nelle successive dieci revisioni, anche questa metrica risulta essere troppo semplicistica.

Edit Logevity è invece un alternativa della sopra citata metrica Text Longevity che corrisponde alla quantità di testo inserita da un autore in una revisione al netto di un fattore che indica il decadimento del testo dalle precedenti revisioni, assumendo un range di valori tra 0 (che indica una completa rimozione del testo) e 1 (qualora il testo venga completamente preservato). Rispetto a quest'ultima però, Edit Longevity non tiene solo conto del testo aggiunto che è sopravvissuto attraverso le successive revisioni, ma anche del testo che è stato cancellato, spostato, sostituito ecc...
Si è ritenuto che questa metrica sia la più adatta per valutare al meglio il contributo che un utente ha dato a una determinata pagina e che sia la più efficace per penalizzare le azioni dei vandali.
In realtà sono stati apportati ulteriori miglioramenti a questa metrica: in [6] viene presentata la misura “ELS”, che corrisponde alla longevità di un intervento valutata rispetto alla sua versione più simile, andando

a ridurre di molto l'impatto dei revert. Questa metrica consente, a differenza della normale Edit Longevity, di considerare anche il primo e l'ultimo edit effettuati su una pagina che verrebbero altrimenti ignorati, in quanto il primo non dispone di edit a lui precedenti e l'ultimo non ha edit a lui successivi che lo giudichino.

Si è tuttavia deciso di optare per la metrica Edit Longevity in quanto già presente nel dataset a disposizione e in quanto comunque risulti essere una delle più complete e affidabili tra quelle esistenti. Inoltre il beneficio apportato dall'utilizzo di ELS è probabilmente più rilevante in pagine di piccole dimensioni, che sono qui state ignorate a seguito di soglie applicate ai dati di partenza, volte a liberarci da informazioni poco significative che porterebbero alla creazione di connessioni tra pagine molto deboli (o addirittura fuorvianti) al momento della creazione della rete.

3.1.2 Rete bipartita e normalizzazione Tf-Idf

Si può passare dunque alle fase successiva di costruzione della rete bipartita tra utenti e pagine.

Qualora un utente *U1* contribuisca ad una pagina *P1* si è scelto di modellare un semi-arco tra *U1* e *P1*. Se ora *U1* scrivesse anche sulla pagina *P2*, seguendo il precedente ragionamento si andrebbe a generare un semi-arco anche tra *U1* e *P2*. Le pagine *P1* e *P2* sono ora indirettamente connesse tramite l'utente *U1* (Fig. 3).

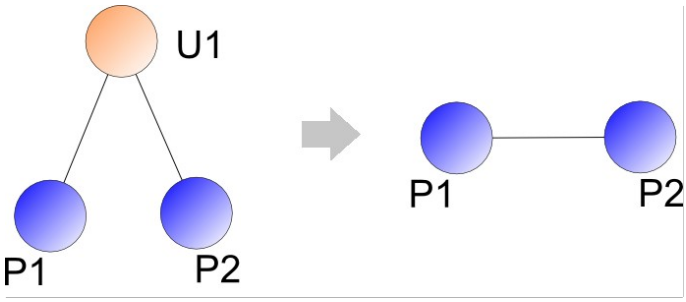

Fig. 3: Proiezione sulle pagine della rete bipartita utenti-pagine. Un utente che ha contribuito a due pagine genera tra queste un arco pesato.

Questa è l'unità base di quella che sarà una rete bipartita utente-pagina che presenta da una parte gli utenti e dall'altra gli articoli dell'enciclopedia (Fig. 4).

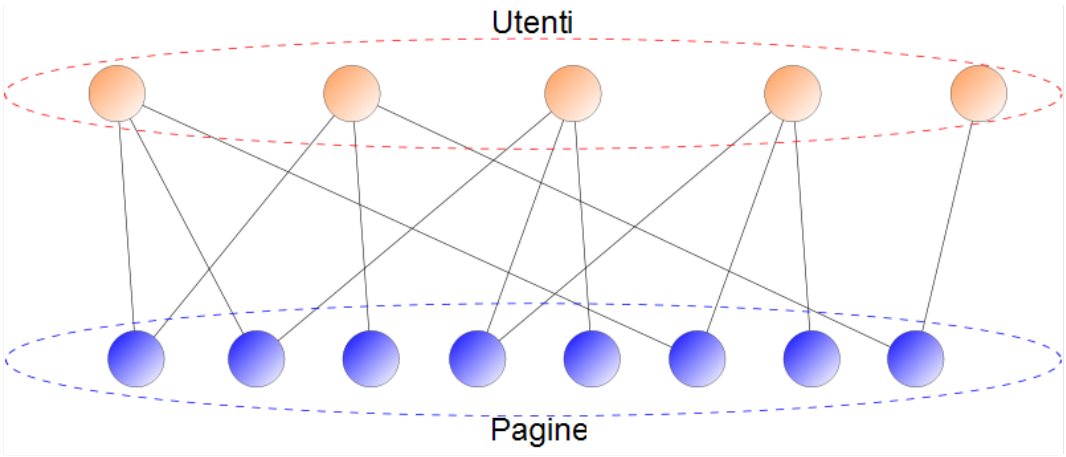

Fig. 4: Schematizzazione della rete bipartita utenti-pagine.

Prima di passare dalla rete bipartita alla rete di pagine si è selezionato un algoritmo per il calcolo dei pesi dei semi-archi (i quali successivamente andranno a costruire gli archi tra le pagine dell'enciclopedia), in maniera tale da valorizzare i collegamenti in base all'attività di un utente su una data pagina. Facendo altrimenti si

otterrebbe una rete bipartita in cui si darebbe la stessa importanza a tutti i semi-archi e dunque a tutti gli interventi degli autori sulle pagine: un utente che genera un contributo marginale sarebbe trattato allo stesso modo di uno che crea un apporto considerevole.
Per raggiungere l'obiettivo proposto si è dunque scelto di appoggiarsi alle metriche *tf* (term frequency) e *idf* (inverse document frequency), molto comuni nei campi dell'information retrieval e del text mining:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

All'interno di un documento, assegnare a ciascuna parola P un determinato valore di *tf-idf* consente di capire l'"importanza" di questa parola rispetto a tutte le altre (all'interno di una collezione di documenti).
In tf, $n_{i,j}$ rappresenta il numero di occorrenze del termine i-esimo considerato (t_i) nel documento j-esimo (d_j), mentre il denominatore corrisponde alla somma del numero di occorrenze di tutti i termini nel documento stesso. In idf invece, $|D|$ è la cardinalità di D (o numero totale di documenti) e $\{d: t_i \in d\}$ è l'insieme costituito da tutti i documenti in cui compare il termine t_i (quindi è $n_{i,j} \neq 0$). Se il termine non dovesse essere presente in nessun documento, ci troveremmo di fronte all'eventualità di una divisione per zero, è dunque pratica comune utilizzare come denominatore $1 + |\{d : t_i \in d\}|$.
Infine moltiplicando tra loro tf ed idf otteniamo il valore tf-idf, utilizzato come peso degli archi della rete bipartita.

L'algoritmo tf-idf si è dimostrato utile anche in campi diversi dall'information retrieval, come mostrato ad esempio in [18], [19] e [20], ed è quindi stato riadattato al nostro caso: in tf il termine $n_{i,j}$

corrisponde ora alla edit longevity dell'utente *i* nella pagina *j*, mentre il denominatore è dato dalla somma di tutte le edit longevity dei *k* contributori della pagina *j*; in idf invece è stato posto $|D|$ come numero totale di pagine e il denominatore come numero di pagine in cui l'utente *i* ha scritto. tf dunque è una caratteristica propria di un determinato utente in relazione a una determinata pagina e indica la qualità dell'attività di quell'utente rispetto alla qualità complessiva dei contributi esistenti per quella pagina. Con idf, invece, sono stati penalizzati tutti quegli utenti che hanno contribuito a un elevato numero di pagine. Questo sia per penalizzare i bot (purtroppo presenti nei dati disponibili) sia gli utenti che hanno scritto su un numero irragionevolmente alto di pagine. Considerando inoltre l'andamento della funzione $\log(x)$, utilizzata nel calcolo dell'idf, si nota che se un utente ha scritto su un numero relativamente basso di pagine (su un totale di 100,000 pagine, si siano editate da poche unità a qualche centinaio di pagine), ci si trova nella parte della curva a minor derivata, il che porta a una variazione molto piccola dell'idf. Se invece, come nel caso dei bot e di alcuni user, ci si trova ad avere contribuito a migliaia, se non decine di migliaia di pagine, allora l'argomento del logaritmo tenderà all'unità e ci si troverà in prossimità della parte della curva a maggior pendenza, ossia il valore totale dell'idf sarà più penalizzante.

3.1.3 Metriche di similarità

Costruita la rete bipartita si può procedere alla fase di creazione del grafo di pagine. Per fare ciò è necessario adottare una metrica di similarità che consenta la transizione da semi-archi utente-pagina ad

archi pagina-pagina.

Tra le metriche che si possono utilizzare per tale scopo troviamo il semplice prodotto scalare:

$$x \cdot y = \sum (x[i] * y[i])$$

dove x e y sono due vettori di uguale lunghezza.

Questa prima metrica è tuttavia troppo semplicistica in quanto si limita a sommare tra loro i prodotti dei contributi degli utenti su due pagine, senza tenere conto, tra le altre cose, dell'importanza relativa dei contributi all'interno degli articoli. Questo aspetto si spiega facilmente con un esempio: assumiamo che *a* e *b* siano due pagine di Wikipedia. In totale 200 utenti hanno scritto su *a* e 400 su *b* e 100 di questi hanno contribuito ad entrambe le pagine. Consideriamo ora le pagine *c* e *d*: alla prima hanno contribuito 3000 utenti, mentre invece alla seconda 5000 utenti e, di nuovo, 100 di questi sono coautori di entrambe le pagine. Si nota subito che se scegliessimo il prodotto scalare come metrica di similarità, verrebbero a generarsi dei collegamenti tra pagine i cui pesi terrebbero in considerazione solamente i semi-archi utente-pagina degli autori che hanno contribuito ad entrambe le pagine, ignorando tutti gli altri contributi. In questo modo l'arco tra le pagine *c* e *d* assumerebbe la stessa importanza di quello esistente tra le pagine *a* e *b*, mentre invece è immediato osservare come il secondo (ovvero il collegamento tra *a* e *b*) sia più rilevante del primo (cioè quello tra *c* e *d*) dal momento che una frazione maggiore del numero totale dei coautori delle due pagine ha contribuito ad entrambe.

Sussiste quindi la necessità di considerare non solo i semi-archi che contribuiscono direttamente alla creazione dell'arco tra due pagine, ma anche tutti gli altri contributi esterni. Una metrica semplice ma molto efficace che risolve questo limite è il *coseno di similarità*:

$$cosine(x, y) = \frac{dot(x, y)}{\sqrt{(dot(x, x) * dot(y, y))}}$$

dove x e y sono sempre due vettori di uguale lunghezza e *dot(x, y)* corrisponde al prodotto scalare tra x e y.
Il valore di similarità risultante sarà compreso nell'intervallo [-1, +1] dove -1 significa che le pagine sono esattamente opposte, 0 che sono indipendenti l'una dall'altra e +1 indica la similarità massima, con i valori intermedi che stanno ad indicare livelli intermedi di similarità o dissimilarità.

A questo punto si ottiene la rete di pagine effettuando la proiezione della rete bipartita sulle pagine utilizzando una metrica di similarità.

3.1.4 Dualismo con la rete di utenti

Uno dei problemi affrontati in [6] è stato quello di individuare le possibili relazioni tra gli utenti dell'enciclopedia tramite l'utilizzo di una rete di utenti. La realizzazione di questa rete prevede però un importante passaggio preliminare di selezione degli autori di ciascuna pagina che, fra tutti gli altri, possono essere visti come i suoi contributori più importanti: questi autori sono i cosiddetti "Top Users".
Lo step di selezione dei coautori assume un'importanza critica nel caso

di Wikipedia in quanto si è ritenuto che le relazioni create tra due utenti qualunque dell'enciclopedia che hanno contribuito allo stesso articolo, non siano sufficientemente significative per giungere a reali conclusioni sulla sua struttura di comunità.
Una volta selezionati i coautori principali degli articoli tramite l'algoritmo di selezione dei top users, si può procedere alla creazione della rete degli utenti di Wikipedia. L'ipotesi di fondo è stata quella di considerare in relazione tra di loro gli autori più importanti di uno stesso articolo; questa intuizione in parte nasce dagli studi sulla collaborazione tra autori di articoli scientifici e tra gli sviluppatori di software open source, ovvero due tipologie di comunità che, proprio come Wikipedia, cercano di migliorare la qualità dei propri prodotti tramite la collaborazione.
Analizzando infine la rete ottenuta si arriva alla conclusione che lo scopo principale degli utenti più importanti dell'enciclopedia sembra essere proprio quello di cercare di intervenire il più possibile in ciascuna delle sue aree.

Il lavoro qui presentato può essere visto come il duale di quello esposto in [6]: in entrambe i lavori si è infatti optato per un modello di rappresentazione dei dati (il grafo) molto noto e studiato, in modo tale da potersi appoggiare ad algoritmi di analisi testati ed affidabili. Il nostro obiettivo è stato invece quello di realizzare una rete di pagine partendo dall'ipotesi che i collegamenti tra pagine generati dai contributi degli utenti possano portare alla formazione di una struttura di comunità tra gli articoli dell'enciclopedia analizzabile tramite algoritmi di clustering in modo da verificarne l'esistenza e la qualità.

3.2 Identificazione di comunità all'interno della rete

Una volta realizzato un grafo pesato rappresentante le connessioni esistenti tra le pagine è possibile proseguire con numerosi tipi di analisi differenti. Studiarne caratteristiche di tipo assoluto, quali densità o diametro, è parso in questo frangente poco significativo e di scarso interesse scientifico.
Si è preferito infatti indagare, all'interno della rete, l'eventuale esistenza di una struttura di comunità e valutare, nei limiti delle possibilità, la pertinenza semantica all'interno dei vari gruppi di pagine ottenuti.

3.2.1 Community detection – algoritmi

Si è verificato come molte reti realizzate su modelli reali non siano omogenee e composte da un unico blocco indistinto di nodi, piuttosto si manifestano strutture di comunità, ossia raggruppamenti di vertici dove si registrano alte densità di archi all'interno di ciascun gruppo, e un relativamente basso numero di collegamenti tra i vari gruppi.
Negli ultimi anni si è esplorato in modo approfondito la possibilità di identificare in modo automatico comunità all'interno di reti, siano esse sociali o non (ad esempio reti bio-chimiche), e svolgere così un'analisi della formazione di raggruppamenti più o meno spontanea a seguito dell'intreccio di relazioni tra gli elementi stessi della rete.

3.2.1.1 Primi approcci

Come suggerisce anche Newman in [9] un primo semplice sistema per

identificare una struttura di comunità può essere quello di applicare un algoritmo di partizionamento della rete. Tuttavia, se questi sistemi possono funzionare su grafi di piccole dimensioni, soffrono di importanti problematiche che ne impediscono l'utilizzo in questo nostro frangente. Risulta sufficiente analizzare i limiti dei due principali algoritmi di partizionamento (la bisezione spettrale e l'algoritmo di Kernighan-Lin) per concludere che questo approccio proposto in un primo momento dalla letteratura scientifica non è adatto a identificare comunità in una rete come quella in analisi: per utilizzare la bisezione spettrale risulta necessario conoscere a priori il numero di comunità che si vogliono identificare, un problema simile si manifesta nell'algoritmo di Kernighan-Lin, dove è necessario specificare la dimensione delle comunità che si intende trovare. I due algoritmi, inoltre, eseguono una bisezione del grafo, ossia una divisione in due gruppi; l'identificazione di un numero superiore di comunità si raggiunge iterando l'algoritmo, che, tuttavia, in nessuno dei due casi fornisce un'indicazione sul numero di iterazioni da eseguire per raggiungere una divisione ottima della rete.

3.2.1.2 Algoritmi gerarchici

Lo studio di reti sociali ha portato in primo piano l'utilizzo di algoritmi di clustering gerarchico volti a definire una struttura ad albero o dendrogramma, dove le foglie sono i vertici del grafo e i nodi a ciascun livello identificano un particolare raggruppamento in comunità.

Gli algoritmi di questa tipologia possono seguire due metodologie operative differenti e diametralmente opposte, entrambe tuttavia prevedono che tra ogni coppia di vertici $\langle i, j \rangle$ sia presente un collegamento con un dato peso: in un primo caso (algoritmo divisivo) è possibile considerare una situazione di partenza in cui tutti i vertici

appartengono a un'unica comunità, da cui viene eliminato l'arco di peso minore, e così in maniera ricorsiva, ottenendo gruppi di dimensione sempre inferiore, fino ad ottenere comunità di un solo elemento ciascuna. Il secondo caso (algoritmo agglomerativo) prevede invece di iniziare considerando ogni vertice come unico membro di una data comunità, e a ogni passo unire due elementi scegliendo di volta in volta il link di peso maggiore non ancora selezionato, così da ottenere comunità di dimensione sempre crescente.

Nonostante questa metodologia di clustering non necessiti di specificare il numero di comunità o il loro numero di elementi, risulta ancora impossibile identificare la suddivisione ottima tra quelle ottenute da ogni iterazione dell'algoritmo.

Inoltre, come verificato anche sperimentalmente in [9] gli algoritmi di tipo gerarchico riescono spesso a identificare in maniera soddisfacente solo parti di comunità, corrispondenti a quegli elementi con una forte similarità, lasciando scoperti altri membri.

3.2.1.3 L'algoritmo di Girvan-Newman

Una prima evoluzione dei semplici algoritmi di clustering gerarchico si ha con l'algoritmo di Girvan-Newman proposto in [10], dove per elaborare i pesi dei vari archi si sfrutta il concetto di edge betweenness, definita come "il numero di shortest path che collegano due vertici tra loro". Archi con una betweenness molto alta indicano quindi una connessione lontana dal centro di una community.

La peculiarità di questo algoritmo sta nel fatto che dopo aver calcolato una prima volta la betweenness per ciascuna coppia di vertici ed aver eliminando l'arco con betweenness maggiore, si procede a un ricalcolo della betweenness di tutti gli archi prima di ripetere il procedimento di taglio. Se il ricalcolo non venisse effettuato e si fosse in presenza di due

comunità connesse tra loro da più di un arco, allora non vi sarebbe la certezza che tutti gli archi di connessione tra i due gruppi abbiano una betweenness molto alta, ricalcolando invece il peso degli archi abbiamo la sicurezza che almeno uno di questi archi avrà un peso, ovvero una betweenness, molto alto.

Benché questo meccanismo incrementi notevolmente la qualità dei risultati finali, il ricalcolo del peso dei link a ogni iterazione risulta, nel nostro caso specifico, un problema di difficile risoluzione considerato l'elevata dimensione della rete (come sottolinea lo stesso autore, l'algoritmo non è adatto per reti di grandi dimensioni in quanto presenta, in grafi sparsi, una complessità di $O(n^3)$, con n corrispondente al numero di nodi nel grafo in analisi). Inoltre va considerato che questo algoritmo non prevede ancora un sistema per determinare quando si è raggiunta la suddivisione ottima della rete in comunità, ovvero viene ancora generato un dendrogramma senza un'indicazione di quale livello corrisponda alle comunità effettive della rete.

3.2.1.4 Modularity

Sono gli stessi Newman e Girvan in [11] a indicare un sistema da implementare nel loro algoritmo per avere un'indicazione su dove si sia raggiunta la divisione in clusters ottimale all'interno del dendrogramma. L'indice della qualità raggiunta da una divisione in comunità prende il nome di "modularity" e viene definita come il numero di archi che cadono all'interno di ciascun gruppo di una determinata divisione sotto analisi, meno il numero di archi ottenuti da una rete random con lo stesso valore di degree medio dei vertici in ciascuna delle stesse comunità. In questo modo è possibile stabilire se effettivamente il numero di archi interni a un gruppo sia maggiore di

quanto ci si aspetterebbe da un grafo casuale.

3.2.1.5 Implementazione con betweenness

L'algoritmo basato sull'edge betweenness sviluppato da Girvan-Newman è stato allora migliorato implementando il sistema di ottimizzazione della modularity in [11], consentendo in questo modo di rilevare quale, tra le varie divisioni ottenute, genera un insieme di comunità corrispondente a quelle realmente presenti nella rete. Tramite l'algoritmo nella sua forma vista in precedenza, si ottiene un dendrogramma che si comincia a esplorare in verticale, calcolando per ogni divisione in community rilevata il valore della modularity.
Il valore massimo della modularity riscontrato corrisponde quindi alla divisione ottima della rete in comunità.
Sebbene quindi adesso vi sia la possibilità di identificare la migliore divisione in comunità, rimane tuttavia il problema del ricalcolo della betweenness a ogni iterazione dell'algoritmo. Una operazione che con le dimensioni della rete in analisi non è eseguibile in tempi ragionevoli.

3.2.1.6 Fastgreedy

Un'interessante implementazione della massimizzazione della modularity è quella del fastgreedy [12], un algoritmo che si propone di raggiungere ottimi risultati con reti di grandi dimensioni. L'algoritmo è di tipo agglomerativo, quindi inizialmente ogni vertice viene considerato come una comunità a sé stante e vengono inizializzate tre strutture dati per poter operare:
- una matrice $\Delta Q_{i,j}$ dove ogni elemento *(i, j)* contiene il delta della modularity che si verrebbe a generare se le due comunità *i* e *j* venissero unite tra loro;
- un max-heap *H* contenente il più grande elemento di ciascuna riga in

$\Delta Q_{i,j}$, insieme ai valori di *i* e *j* corrispondenti;

- un array ordinato di elementi a_i, corrispondente alla frazione di archi legati a vertici nella comunità *i*.

Il fastgreedy prevede, come definito in [12], tre passaggi:

1. calcolare i valori iniziali di $\Delta Q_{i,j}$ e di a_i, e riempire H con il valore più alto trovato su ciascuna riga della matrice.
2. si procede poi selezionando il primo elemento in H (che sarà il maggiore) e unendone le comunità corrispondenti, aggiornando le corrispettive righe di ΔQ (ossia vengono unite le due righe sommandone gli elementi), il valore di a_i e il contenuto di *H*.
3. si ripete il punto precedente fino ad avere una sola comunità.

Per come è definita l'equazione che identifica ciascun ΔQ si avrà un solo massimo della funzione, ossia basterà, a ogni iterazione del punto 2 dell'algoritmo, verificare se il ΔQ corrente è ancora maggiore di zero, altrimenti la divisione a cui si era precedentemente arrivati risulterà essere quella ottima, e ogni successiva iterazione porterà a un decadimento della modularity.

La miglior complessità temporale dell'algoritmo, rispetto al sistema che sfrutta la edge-betweenness, risiede nel fatto che a ogni iterazione non è necessario eseguire un ricalcolo di tutti i pesi del grafo, dovendo così aggiornare anche gli elementi della matrice ΔQ. Con il nuovo procedimento dobbiamo solamente, a ogni iterazione, unire due righe *i* e *j*, ossia aggiornare i valori di *i* ed eliminare la riga *j*; l'algoritmo fastgreedy risulta così avere una complessità, in reti reali di tipo sparso e gerarchico, pari a $O(n \log^2(n))$, con n numero di vertici. Rispetto alle complessità temporali dei principali algoritmi concorrenti al fastgreedy, questo si pone come un ottimo strumento di rilevamento di comunità all'interno di grafi di dimensione molto elevata.

Va detto che questo sistema è stato inizialmente sviluppato per grafi non pesati, ma in [13] ne è stata implementata una versione che fa uso

di archi pesati per realizzare l'identificazione delle comunità.

3.2.1.7 Walktrap

Un altro algoritmo di interesse che sfrutta il sistema della modularity è il walktrap [14], il quale si basa sull'intuizione che "cammini random di breve lunghezza in un grafo tendono a essere "intrappolati" in zone densamente connesse, corrispondenti alle comunità". Viene in tal modo definita una nuova tipologia di distanza che consente di terminare l'algoritmo in un tempo $O(n^2 \log(n))$.

Nonostante dai dati forniti dagli autori si prospetti una precisione nell'identificare comunità perfino maggiore di quella del fastgreedy, l'assenza di un'adeguata implementazione e la complessità temporale ancora proibitiva hanno portato a non scegliere questo algoritmo nelle analisi svolte sulla nostra rete.

3.2.1.8 Limiti Modularity

Un importante limite ai sistemi basati sull'ottimizzazione della modularity è stato portato all'attenzione della comunità scientifica da Barthelemy e Fortunato in [15]. Gli autori mostrano come il valore della modularity di un cluster, ottenuto in una data divisione della rete, sia strettamente dipendente dal numero di archi contenuti nella rete stessa. In particolare si dimostra come in molte situazioni reali cluster contenenti un numero di edges pari a $l_s < \sqrt{2L}$ (con L pari al numero totale di collegamenti nella rete) siano in realtà un agglomerato di più sotto-gruppi, dotati di propria identità, interconnessi tra loro. Inoltre al crescere delle connessioni tra i sotto-gruppi si verifica questo fenomeno indesiderato anche per valori più alti di l_s, fino al caso limite in cui vengono riconosciuti in maniera

errata come cluster interi blocchi di sotto-cluster aventi complessivamente $l_s < \frac{L}{4}$.

Il limite di risoluzione ("resolution limit") si manifesta in maniera più evidente in grafi di grandi dimensioni e in presenza di componenti fortemente connessi tra loro.

La soluzione proposta a questo inconveniente consiste nell'eseguire una ottimizzazione della modularity successiva su ciascun elemento trovato dopo una prima esecuzione dell'algoritmo, controllando che i nuovi gruppi trovati siano realmente delle comunità. Questo accertamento consiste nell'appurare che il valore della modularity per ciascuna sotto-comunità sia maggiore di zero. Viene comunque fatto notare come spesso una modularity semplicemente maggiore di zero possa portare alla divisione non voluta di cluster già corrispondenti a comunità; a tal proposito, considerando che spesso l'effettiva presenza di comunità è caratterizzata da una modularity di almeno 0.4, portarsi a un valore prossimo a questa soglia migliora i risultati ottenuti. Anche in questo frangente si evidenzia comunque il limite dell'ottimizzazione della modularity, in quanto a un passaggio successivo dell'algoritmo non necessariamente corrisponde un aumento del valore totale della modularity per quella data divisione come ci si aspetterebbe.

3.2.1.9 Louvain method

Un altro algoritmo di tipo agglomerativo che si propone come capace di una complessità temporale addirittura lineare su grafi reali e una qualità dei risultati maggiore rispetto ai precedenti algoritmi analizzati è quello proposto in [16]. Il procedimento si divide in due fasi, ripetute poi iterativamente: in un primo momento ogni nodo del grafo è assegnato a una propria comunità. Per ogni nodo *i* si considera

ciascun vicino *j* e si sposta *i* nella comunità *j* che permette un guadagno massimo e positivo di modularità, qualora ciò non fosse possibile si lascia *i* nella propria comunità. Si ripete ciò per ogni nodo fintanto che è possibile rilevare un miglioramento della modularity. La fase successiva prevede che ogni comunità rilevata venga considerata un nodo di una nuova rete, dove il peso della connessione tra due nodi è dato dalla somma dei pesi degli archi tra i nodi delle due comunità e archi interni contribuiscono alla formazione di un autoanello sul gruppo stesso.
Si ripetono così i due passi fino a che non si rilevano più cambiamenti a fronte di nuove iterazioni e la massimizzazione della modularity è raggiunta.
Questo algoritmo non soffre del problema del resolution limit in quanto esegue un'ottimizzazione locale e non globale, identificando fin da subito comunità di piccola dimensione.
La velocità di esecuzione e la precisione dei risultati generati fa del Louvain method uno degli algortmi più usati negli ultimi anni per l'identificazione non supervisionata di comunità in reti di grandi dimensioni (per la prima volta si parla della possibilità di analizzare reti con milioni di nodi e miliardi di archi in tempi molto brevi).

Le dimensioni della rete in analisi, in termini non solo di vertici ma anche e soprattutto di archi, impone la scelta di un algoritmo con un'ottima complessità temporale senza sacrificare la qualità dei risultati finali.
Tra gli algoritmi analizzati si è scelta l'implementazione del fastgreedy nella suite Igraph ([26]) e l'implementazione ufficiale del Louvain method ([28]) per via delle loro caratteristiche precedentemente analizzate.

3.3 Analisi semantica

A questo punto ogni comunità individuata da uno degli algoritmi scelti corrisponde a un gruppo di pagine interconnesse tra loro da un rilevante numero di archi. Sarebbe ora interessante poter stabilire se questi cluster calcolati abbiano al loro interno una qualche valenza semantica, ossia se pagine che con questo sistema sono state associate allo stesso gruppo cadano tutte dentro una determinata categoria o a una selezione di categorie correlate tra loro.

In [2] si è cercato di eseguire un'analisi simile a questa su dati raccolti in del.icio.us sfruttando il database semantico-lessicale WordNet. Tuttavia, in questo caso, non è sembrato possibile utilizzare questa sorgente per la mancanza di una buona corrispondenza dei termini e per una mancanza di molti termini enciclopedici.
Il modo migliore di procedere è sembrato ricorrere a un sistema che in qualche modo consenta una mappatura uno a uno con i nomi originali degli articoli dell'enciclopedia utilizzati nel corso delle nostre analisi, visto che questi cambiano più spesso di quanto ci si aspetterebbe anche nel giro di soli pochi mesi.
L'interesse è quindi ricaduto su uno strumento già presente in Wikipedia, ossia un albero di tutte le categorie dell'enciclopedia cui le pagine appartengono.
Avendo dunque un'organizzazione topologica abbastanza specifica delle categorie si può fare l'assunzione che spostandosi di poco all'interno dell'albero si rimanga all'interno di uno stesso argomento, e che quindi le categorie identificate durante questo spostamento contengano pagine molto probabilmente connesse tra loro da uno stretto legame semantico.
L'analisi intrapresa mira a verificare se, con uno spostamento molto

circoscritto all'interno dell'albero, si incontrino abbastanza categorie da andare a includere il maggior numero possibile di pagine del particolare cluster.

4. Implementazioni e risultati

4.1 Grafo

4.1.1 Analisi dei dati di partenza

Il primo passo verso la realizzazione della rete bipartita consiste nell'analisi dei dati disponibili, ricavati in parte direttamente dai dump della Wikipedia inglese aggiornati al 2007, organizzati in tabelle all'interno di un database e in parte derivanti dai risultati ottenuti in [6].

I contenuti delle tabelle non sono immediatamente utilizzabili per realizzare il grafo. Questi devono essere infatti sfoltiti, rimuovendo informazioni poco utili o che addirittura porterebbero a risultati errati.

Si consideri innanzitutto che l'utente identificato dall'ID "0" è in realtà fittizio e racchiude tutte le attività degli utenti anonimi; è perciò inutile sfruttarne i dati relativi, sarebbe anzi dannoso e fuorviante.

Bisogna inoltre prestare attenzione a tutti quegli utenti che hanno contribuito ad un solo articolo dell'enciclopedia. Essi infatti non sono di alcuna utilità in quanto non genererebbero alcun collegamento all'interno del grafo finale perché nella rete bipartita sarebbero collegati ad una sola pagina. Dunque l'utente "0" e quest'ultimo insieme di utenti devono essere scartati.

4.1.2 Sistema di indicizzazione

L'applicazione allpairs della suite SimilarityEngine compie un'importante procedura di indicizzazione del file di input, preliminare al calcolo della similarità tra le pagine di Wikipedia. Una volta fornito in ingresso il file contenente i semi-archi utente-pagina, viene creato in memoria un indice che associa a ciascuna pagina dell'enciclopedia un vettore rappresentante gli utenti che hanno contribuito a essa con i relativi valori del peso tf-idf. All'interno di questo vettore gli utenti saranno rappresentati sotto forma di indici in corrispondenza dei quali verranno memorizzati i valori dei pesi dei semi-archi che li legano ad una determinata pagina.
Se ad esempio gli utenti 1, 34, 543, 5678 scrivessero sulla pagina "computer" con un valore di tf-idf rispettivamente pari a 2.35, 12, 7.8 e 44, all'interno dell'indice verrà associato alla pagina "computer" un vettore in cui gli indici 1, 34, 543, 5678 memorizzeranno rispettivamente i valori 2.35, 12, 7.8, 44.
Una volta ultimato l'indice, allpairs provvederà a confrontare tra loro, a seconda della metrica di similarità scelta, le pagine di tutte le possibili coppie formabili a partire dal file fornito in input al programma, tramite il confronto tra i vettori dell'indice associati ad esse. Si ottiene così un file contenente il livello di similarità per ogni coppia di pagine reciprocamente connesse.

Per i motivi precedentemente esposti la metrica di similarità scelta in questo lavoro è il coseno di similarità. I vettori x e y che compaiono nelle formule presentate in 3.1.3 corrispondono ora ai vettori dell'indice precedentemente creato.

Adottando questo sistema siamo dunque in grado di associare il giusto grado di importanza a ciascun arco che si verrà a generare tra due pagine. Un arco avrà dunque peso tanto maggiore quanto maggiore

sarà il peso dei semi-archi che lo hanno generato e quanto minore sarà il numero e l'importanza dei contributi degli utenti presenti tra i contributori di una delle due pagine ma non di entrambe.

4.1.3 Processo di creazione del grafo

Passiamo ora alla fase di creazione del grafo vera e propria, analizzando separatamente ciascuno dei passi necessari alla costruzione della rete di pagine. Il primo passo trattato sarà la scrittura del file che sarà usato come input per l'applicazione allpairs della suite SimilarityEngine ([25]), grazie alla quale si otterrà in output il file contenente il livello di similarità tra le varie coppie di pagine. La fase successiva comporta la sfoltimento del file di similarità ottenuto; verranno eliminati gli archi generati da un numero di utenti troppo basso appoggiandosi al prodotto scalare, per poi arrivare alla parte finale del processo che corrisponde alla scrittura della rete precedentemente costruita.

4.1.3.1 Creazione del file di input

Il file testuale fornito in input all'applicazione allpairs contiene la lista di tutti i semi-archi utente-pagina formattati come segue:

"Pagina Utente Peso" separati da tabulazione, un semi-arco per linea, raggruppati per pagina.

Viene qui proposto un esempio dove le pagine e gli utenti sono rappresentati da interi (rispettivamente nella prima e seconda colonna)

con le pagine in ordine crescente:

1	25	0.2311
1	782	0.0523
1	535	3.2319
2	782	0.8231
2	25	1.9324
2	535	0.0132
3	25	0.0921
3	13	4.9428

Nella creazione di questo file testuale vengono ignorati tutti gli utenti che hanno contribuito ad una sola pagina di Wikipedia in quanto, dal momento che costoro genererebbero un solo semi-arco utente-pagina, non porterebbero alla costruzione di nessun arco tra pagine.
Un'altra doverosa scrematura consiste nel tralasciare tutte le pagine troppo piccole, che darebbero quindi luogo a un gran numero di archi tra pagine di scarso significato: per fare ciò imponiamo una soglia sulla somma dei valori di edit longevity totalizzati da tutti i contributori di una pagina all'interno di essa, in modo tale da scartare tutti gli articoli per i quali questa somma è al di sotto della soglia, liberandoci di tutte le pagine a cui hanno contribuito solo pochi utenti.
Un'ultima necessaria selezione va ad interessare il peso dei semi-archi utente-pagina in quanto preservare semi-archi con un basso valore di tf-idf significherebbe dare luogo ad archi tra pagine di scarsa importanza il cui unico effetto sarebbe quello di appesantire la quantità di dati su cui lavorare successivamente.

Dall'esecuzione di allpairs si ottiene infine in output un file testuale in cui ogni pagina è legata a una serie di altre pagine con un determinato peso.

4.1.3.2 Una soglia sui contributori degli archi

Arrivati a questo punto non disponiamo ancora di dati sufficientemente significativi perché possano essere direttamente impiegati nella costruzione del grafo. Il file di output generato al passo precedente contiene ancora infatti un gran numero di archi tra pagine di scarso valore (se non addirittura fuorvianti) in quanto generati da un basso numero di utenti. Si è infatti voluto dare importanza non solo al peso degli archi ma anche al numero di contributori che hanno portato alla loro creazione. Questo dato non è stato utilizzato direttamente nel processo di passaggio dalla rete bipartita alla rete di pagine, ma è stato impiegato come ulteriore soglia di accettazione degli archi.

Dall'insieme di archi ottenuto si è quindi deciso di rimuovere tutti quelli che sono stati creati da un numero di contributori inferiore a una data soglia, ritenendo che non fossero abbastanza efficaci per poter stabilire una connessione veramente significativa tra due pagine.

Ad esempio, se un astronomo che si è occupato della pagina "Legge di Hubble" fosse un appassionato di ceramiche orientali, ma anche l'unico tra gli altri astronomi con questo interesse, si avrebbe una forte connessione tra il primo articolo e la pagina "Ceramiche vietnamiti", il che porterebbe a un accostamento poco significativo. Anche se vi fosse più di un contributore ad aver generato quest'arco, ma questi continuassero a essere un numero esiguo rispetto a quanto ci si aspetterebbe a fronte della lunghezza del periodo di attività considerato, l'arco stesso continuerebbe ad essere poco importante.

Per questa ragione è stato creato un secondo file di input per l'applicazione allpairs identico all'originale, sostituendo però a tutti i valori dei pesi dei semi-archi il valore "1" e, successivamente, è stata lanciata nuovamente l'applicazione allpairs usando il prodotto scalare

come metrica di similarità.

In questo modo si viene a creare un file di output, la cui formattazione rimane la stessa di quella di un normale file di output di allpairs, dove però il peso di ogni arco tra pagine corrisponde al numero di utenti che hanno contribuito alla creazione di quest'ultimo. Sfruttando questo risultato come una sorta di whitelist è possibile filtrare qualsiasi collegamento tra articoli generato da un numero di contributori al di sotto del valore imposto per la soglia, che non saranno presenti all'interno della whitelist stessa.

Otteniamo così un file equivalente al precedente, ma contenente soltanto gli archi generati da un numero di contributori maggiore della soglia desiderata ed esente da vertici isolati, della cui rimozione ci si preoccupa durante il processo di eliminazione di questi collegamenti indesiderati.

4.1.3.3 Scrittura del grafo

Dal passo precedente si ottiene come risultato un file di similarità tra pagine sfoltito di tutti i suoi collegamenti poco significativi. Rimane ora da compiere solamente la fase di scrittura della rete in uno dei formati standard per la rappresentazione dei grafi: il formato Pajek ([29]). Terminato quest'ultimo passaggio il grafo sarà rappresentato da un file .net nel formato:

```
*Vertices 3
1  "CPU"
2  "Java"
3  "Ram"
*edges
1  2
```

```
3  1
```

Poiché nel file di whitelist, così come nel file di similarità, è presente una riga per ogni pagina, durante la scrittura del file .net si presenteranno problemi di simmetria e riflessività. Infatti, se nella riga riferita alla pagina *P0* compare la pagina *P1* (immaginiamo con peso *w*), successivamente, nella riga riferita a *P1*, comparirà *P0* (sempre con peso *w*). Bisognerà dunque prestare attenzione a queste ripetizioni insite nei file di similarità e di whitelist perché non dovranno certamente comparire all'interno della rappresentazione finale della rete, costituita appunto dal file .net.

Si è così giunti al termine del processo che consente di ottenere una rete di pagine partendo da una rete bipartita composta da collegamenti utente-pagina. Gli archi del grafo dunque non danno più indicazioni sulla qualità di ciascun intervento fatto da un utente su una pagina, ma rappresentano invece il grado di similarità tra i nodi della nuova rete ottenuta.

Rimane ora un ultimo punto da affrontare perché la trattazione possa considerarsi completa: l'impostazione delle varie soglie lungo tutto il percorso di creazione del grafo di pagine. È infatti di cruciale importanza scegliere questi valori in modo tale da raggiungere un efficace compromesso tra quantità e qualità dei dati, in quanto optare per soglie troppo alte porterebbe sì a lavorare con grafi di piccole dimensioni e facilmente gestibili, ma con un esiguo numero di collegamenti, il che sarebbe indice di uno scarso significato globale della rete ottenuta. D'altro canto però non risulta neppure praticabile la strada inversa, dal momento che, viste le dimensioni di Wikipedia, si avrebbe a che fare con dei grafi di dimensioni talmente esagerate da

risultare totalmente ingestibili da parte degli strumenti hardware e software a nostra disposizione per l'analisi e lo studio dei grafi.

4.1.4 Soglie

Vengono ora presentati i valori impostati per le soglie durante le varie fasi della costruzione del grafo, mostrando inoltre per ogni tipo di soglia l'impatto che questa può avere nella realizzazione della rete in termini di archi, semi-archi e vertici rimossi/preservati in funzione dei valori per essa scelti.
Analizziamo dunque i valori delle soglie nell'ordine in cui compaiono durante la creazione della rete di pagine.

Il primo vincolo che si incontra nella costruzione del grafo è la soglia utilizzata per scartare tutti gli articoli troppo piccoli. Questa selezione viene condotta imponendo un valore limite alla edit longevity complessiva totalizzata da ciascuna pagina e andrà a pesare sul numero di semi-archi utente-pagina che verranno preservati durante il processo di creazione del file di input per l'applicazione allpairs. In Fig. 5 viene mostrato il numero di semi-archi mantenuti in corrispondenza dei vari valori associati alla soglia.

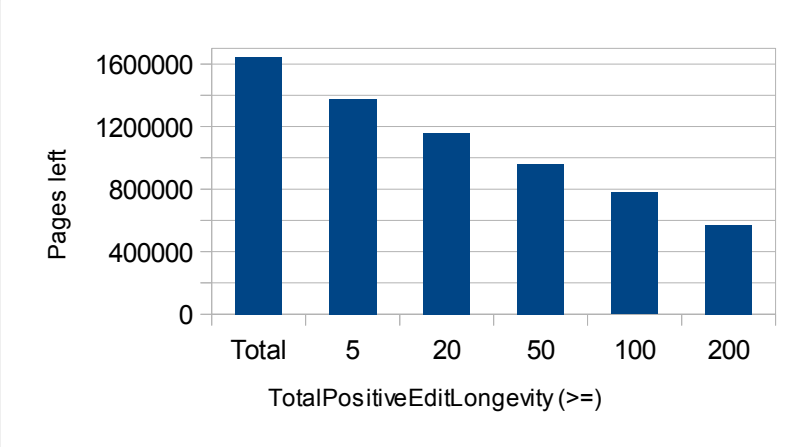

Fig. 5: Pagine rimaste a fronte di una data soglia sulla edit longevity complessiva (TotalPositiveEditLongevity).

Al fine di eliminare le pagine più piccole e meno importanti si è scelto un valore di soglia pari a 20, consentendoci di lavorare con le prime 1,157,964 pagine su un totale iniziale di 1,644,902.

Il secondo parametro da impostare è il livello di accettazione del peso dei semi-archi, in modo tale da escludere tutti quei collegamenti utente-pagina con un valore di tf-idf sufficientemente basso da portare alla creazione di archi di scarsa importanza.
In Fig. 6 è illustrata la distribuzione dei semi-archi in funzione del valore del loro peso.

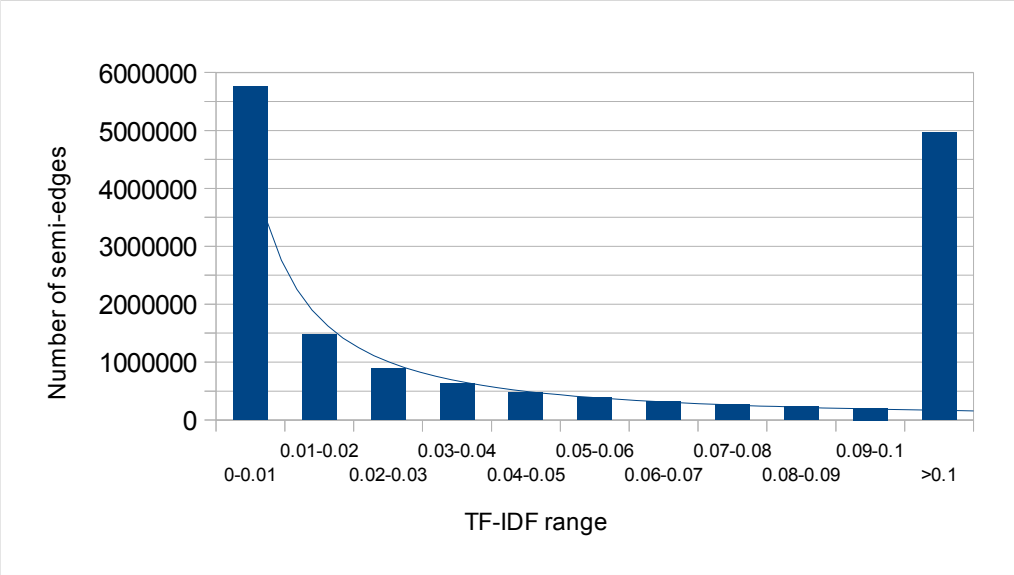

Fig. 6: Distribuzione dei semi-archi in base al valore di TF-IDF. Si nota come una gran parte dei semi-archi si attesti su valori del peso molto bassi.

Risulta evidente come la funzione presenti una distribuzione heavy-tailed, e come la stragrande maggioranza dei semi-archi presenti valori molto bassi, prossimi allo 0. Seguendo il sistema di valutazione indicato la maggior parte dei semi-archi risulta quindi irrilevante.

In Fig. 7 sono riportati i semi-archi che vengono preservati in corrispondenza dei vari livelli di soglia.

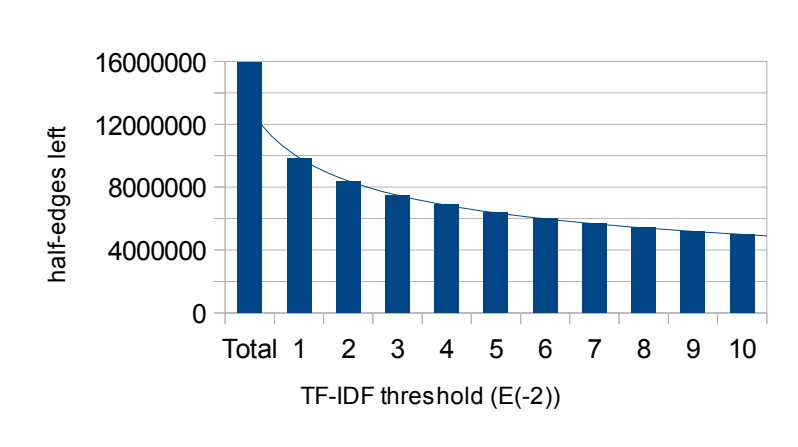

Fig. 7: Numero di semi-archi rimasti a seguito di una data soglia sul valore di TF-IDF. Basta un basso valore sulla soglia per eliminare una buona parte dei semi-archi.

Si osserva come l'andamento del numero di semi-archi in funzione della soglia sul tf-idf presenti ancora una distribuzione heavy-tailed. Da questa considerazione e dai dati mostrati in Fig. 6 ne consegue che la maggior parte dei semi-archi viene eliminata per valori molto bassi della soglia. Si è impostato un valore minimo sui pesi di 0.03 al fine di eliminare tutti quei contributi poco significativi, ottenendo in tal modo, nella fase successiva, solo gli archi più importanti.

Arriviamo quindi a dover scegliere un valore limite per la metrica di similarità coseno: tutti gli archi tra pagine con peso inferiore alla soglia di similarità verranno perciò ignorati. L'istogramma in Fig. 8 mostra quanti archi vengono accettati per ognuno dei vari valori soglia.

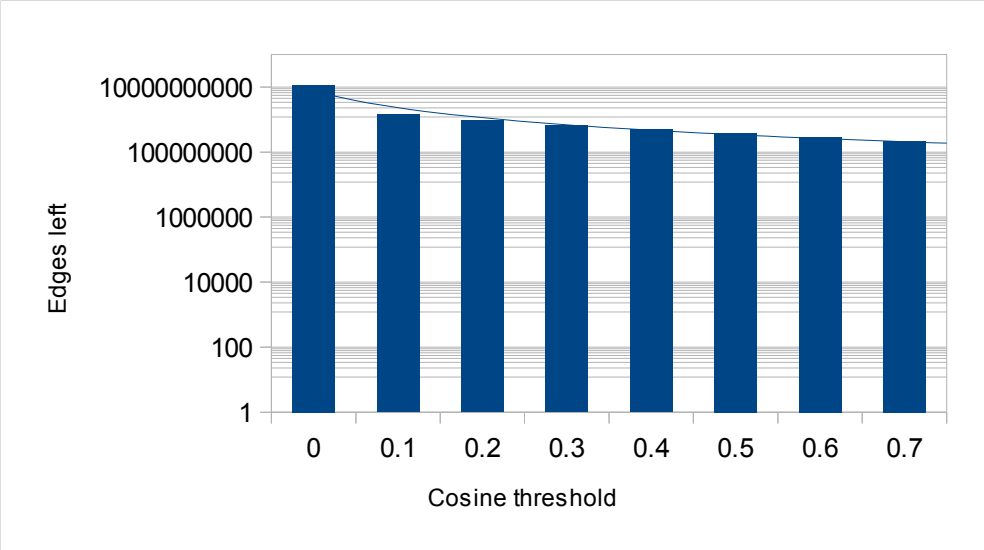

Fig. 8: Numero di archi rimasti a seguito di una soglia sul coseno di similarità. Una buona parte degli archi presenta un peso finale prossimo allo zero e viene quindi eliminata già per valori molto bassi della soglia.

In maniera simile ai risultati ottenuti per i pesi sui semi-archi (Fig. 7), anche in questo caso vediamo come la maggior parte dei pesi degli archi si attesti su valori molto bassi: senza alcuna soglia il grafo andrebbe a essere costituito da più di 11,4 miliardi di archi, ma basta introdurre una soglia pari a 0.1 per vedere il numero di elementi scendere drasticamente a 1,4 miliardi.
Si è optato per un valore di soglia di 0.5, in modo da mantenere le connessioni più significative e contenere la quantità di informazioni da elaborare.

Rimane infine da selezionare quale debba essere il numero minimo di utenti che contribuiscono alla creazione di un arco tra due pagine affinché quest'ultimo possa essere considerato veramente rilevante. L'istogramma in Fig. 9 mostra il numero di vertici e di archi presenti nel grafo al variare di quest'ultima soglia, dopo aver applicato tutte le altre soglie viste eccetto quella sul coseno.

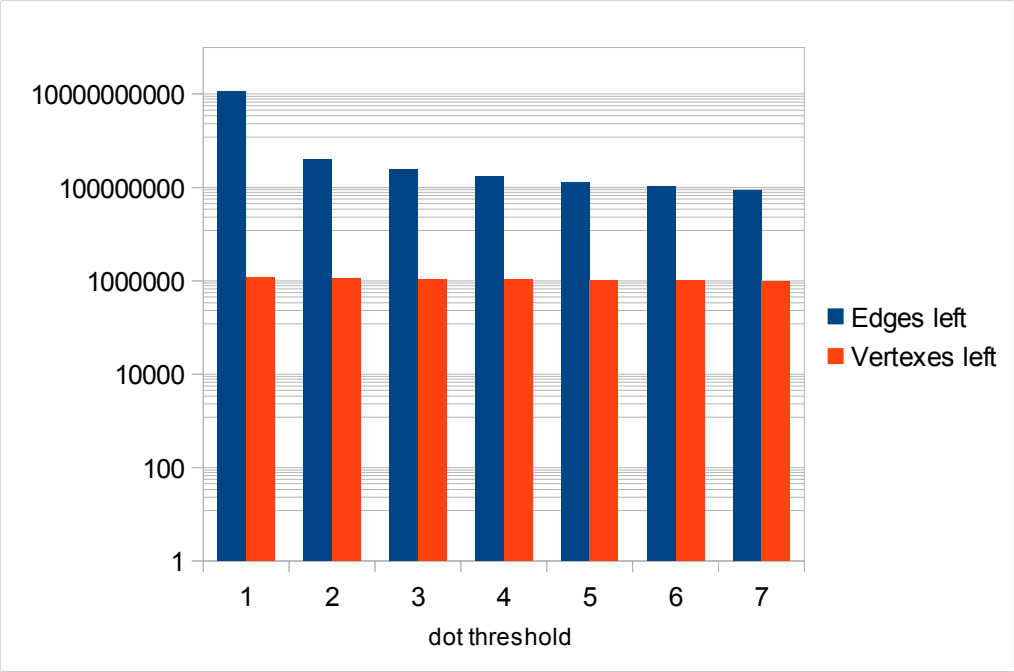

Fig. 9: Archi e vertici rimasti a fronte di una data soglia sul numero di contributori degli archi. Si noti come il 96,5% degli archi sia generato da un solo contributore.

Osserviamo come un rilevante numero di archi risulti generato dall'intervento di un solo utente (addirittura il 96,5% del totale). Seguendo la nostra interpretazione di questa quantità ne possiamo dedurre che questi archi possano essere esclusi dalla trattazione, indipendentemente dal peso, perché poco rappresentativi del comportamento di una folta comunità di individui. Il numero di pagine disponibili, invece, diminuisce in maniera molto minore rispetto agli archi; nella trattazione, entro certi limiti, si potrà quindi ignorare questo aspetto.

La perdita di archi diminuisce sensibilmente per valori successivi della soglia, consentendo di scegliere un valore maggiore senza sacrificare eccessivamente la quantità di informazioni disponibili per la trattazione. Si è pertanto ritenuto che dopo 25 mesi di attività di Wikipedia il minimo numero di contributori necessari affinché un arco

possa essere considerato significativo sia pari a 5.

Il grafo ottenuto scegliendo in sequenza tutte le soglie come indicato è composto da 113,831 vertici e da 8,186,706 archi.

4.2 Clustering

Una volta in possesso di un grafo pesato di pagine, si sono applicati i due algotirmi di identificazione di comunità precedentemente scelti e analizzati: Fastgreedy e Louvain method. Come si nota in Fig. 10 il grafo non è connesso, ma risulta invece composto da numerosi componenti, molti dei quali di piccole dimensioni (spesso inferiori a 4 elementi) e due, i maggiori e unici con tale dimensione, che contengono rispettivamente 26629 e 20498 pagine.

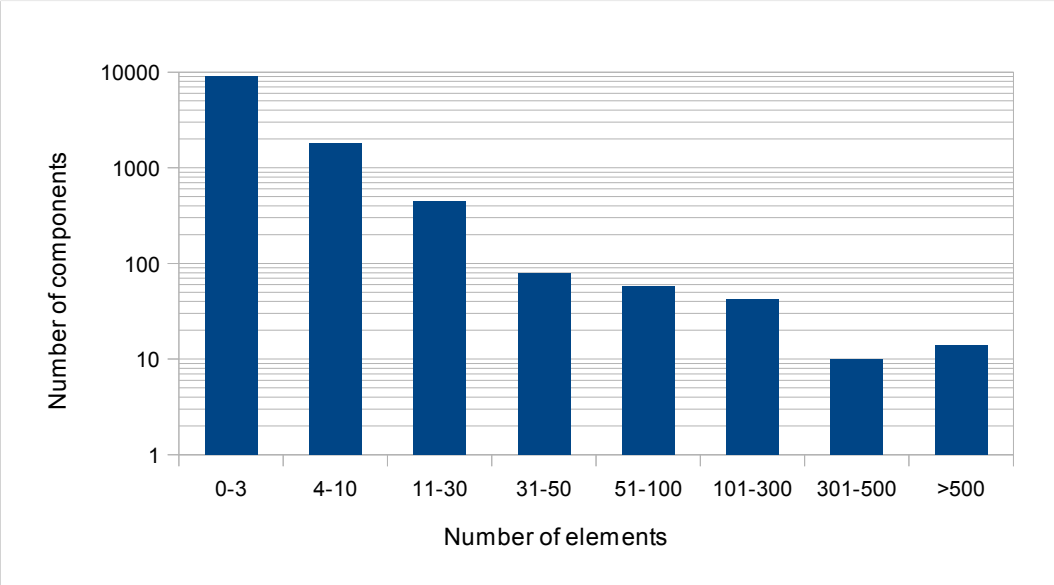

Fig. 10: Distribuzione dei componenti della rete di pagine in base al numero di nodi.

Mentre per il Louvain method si è deciso di procedere eseguendo

l'algoritmo sull'intero grafo, ignorando la suddivisione in componenti, per il Fastgreedy si è preferito operare sui singoli blocchi.

4.2.1 Fastgreedy

Si è voluto analizzare il problema della resolution limit nel fastgreedy a seguito della prima iterazione dell'algoritmo sul componente di maggior dimensione[1] (che subisce perciò il maggiore impatto del limite di risoluzione). L'identificazione di comunità ha portato a una divisione in cluster come da Fig. 11.

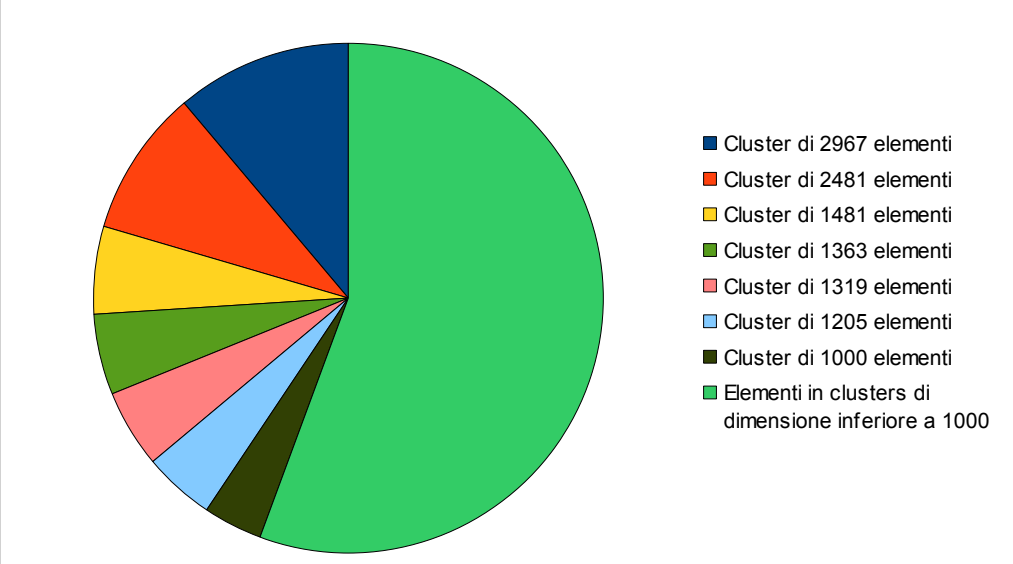

Fig. 11: Distribuzione delle pagine nei vari clusters, in evidenza i 7 maggiori gruppi di nodi.

Si noti la presenza di gruppi di pagine di dimensione superiore a 1000

1 Benchè questo tipo di analisi si sarebbe potuto effettuare anche sul componente di dimensione quasi analoga a quello maggiore (20498 elementi), si è ritenuto più significativo lavorare solo su quello da 26629 elementi, in quanto, osservando gli elementi del primo, si riconoscono solo articoli di città degli USA. La maggior varietà di elementi si è ritenuto potesse portare quindi a risultati più interessanti.

elementi, i quali vanno a coprire il 44% del totale delle pagine del componente; andando a esaminarli manualmente si nota come le pagine al loro interno possano rientrare in una categoria ben definita (eccezion fatta per il maggiore di questi gruppi, di cui si tratterà in 4.3.3). Troviamo un cluster con personaggi della politica inglese, uno sullo sport, uno sul calcio, uno contenente date e uno riguardante la politica canadese. Questi potrebbero a primo avviso essere considerate come comunità ben definite; non avviene, come ci si sarebbe aspettato, che comunità molto diverse tra loro risultino accorpate in un grande gruppo. Tuttavia si deve ritenere che il resolution limit abbia influenzato la generazione di questi clusters, poiché questa problematica è intrinsecamente presente nell'ottimizzazione della modularity del fastgreedy (come peraltro riportato in [15]).

Tornando quindi alla trattazione sull'intera rete di pagine, supposto che esistano gruppi frutto di un'agglomerazione di comunità più specifiche, si procede a molteplici iterazioni del fastgreedy su tutti i sottografi corrispondenti ai cluster trovati a seguito della prima iterazione, interrompendo il processo di ricorsione e scartando l'ultima suddivisione trovata, qualora questa porti a una modularity minore di 0.4. Al termine del processo la distribuzione dei cluster in base alla loro dimensione risulta essere quella in Fig. 12: come si vede, a seguito della ricorsione, la maggior parte delle comunità presenta dimensioni irrisorie, nel 66% dei casi si tratta di gruppi costituiti da due sole pagine; questi ultimi sono di scarso interesse e derivati per una buona parte dai componenti del grafo già in partenza di piccole dimensioni (Fig. 10); per l'analisi che si andrà a condurre, restano tuttavia presenti numerosi cluster di dimensione maggiore che andremo ad utilizzare.

È interessante osservare come, spostando la soglia della modularity a 0 o a 0.2, diminuisca la presenza di cluster di grandi dimensioni, a favore dei piccoli gruppi di poche unità di elementi.

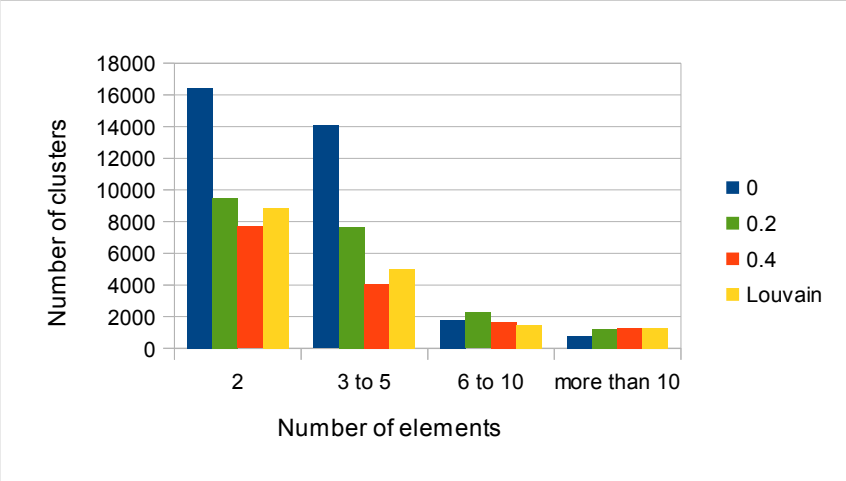

Fig. 12: Distribuzione dei clusters in base alla loro dimensione. Confronto dei risultati ottenuti con Louvain method e ricorsione del Fastgreedy con soglie sulla modularity a 0, 0.2 e 0.4.

4.2.2 Louvain method

Esente dal problema del resolution limit, l'algoritmo esegue una divisione in cluster come da Fig. 12. La percentuale di cluster di soli due elementi è salita qui al 70% rispetto al totale ed è anche presente un unico cluster di un singolo elemento, di interesse nullo per i nostri scopi di analisi. Tuttavia sono notevolmente aumentati i gruppi di dimensione maggiore, in particolare quelli sopra i 50 elementi (da 126 nel caso del fastgreedy, con soglia sulla modularity a 0.4, a 249).
Benché con questo algoritmo si ottengano risultati simili ad alcune applicazioni del fastgreedy, è da ritenere che il Louvain method abbia dei risultati più rappresentativi della realtà, in quanto non necessità di successive iterazioni e la scelta di impostare la soglia per la modularity

a 0.4, nella ricorsione del fastgreedy, è arbitraria e troppo statica per portare a risultati ottimi.

Risulta ora possibile comparare quella che è stata la suddivisione in comunità, a seguito della prima iterazione del Fastgreedy sul componente del grafo di maggior dimensione, con il risultato ottenuto con il Louvain method, che non necessita di iterazioni. A tal proposito si confrontino Fig. 11 e Fig. 13:

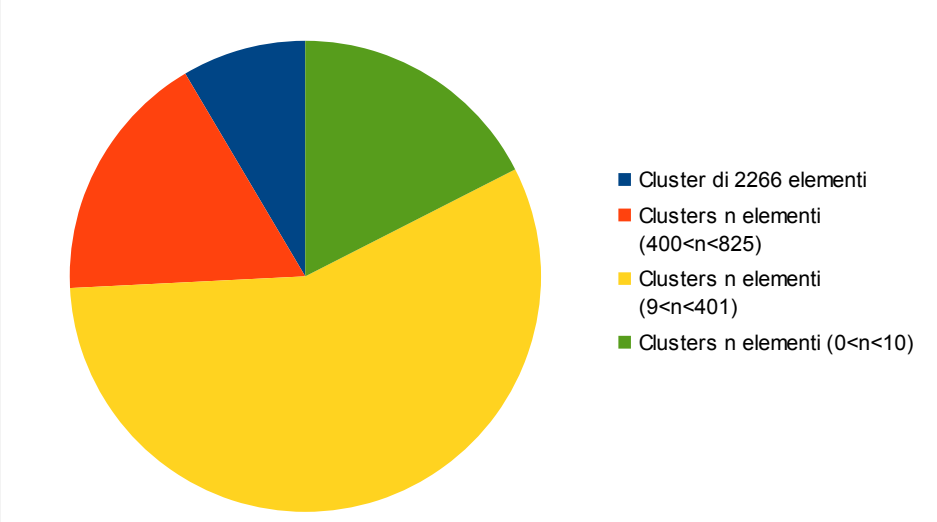

Fig. 13: Distribuzione delle pagine nei vari clusters, in evidenza il gruppo con il maggior numero di elementi.

L'assenza del resolution limit è evidente in quanto mancano tutte quelle comunità di grande dimensione ritrovate a seguito della prima iterazione dell'algoritmo Fastgreedy. Qui infatti l'unico gruppo di dimensione veramente elevata si attesta sui 2266 elementi (raggruppamento di particolare significato, analogo di quello di 2967 pagine identificato in Fig. 11 e di cui si rimanda la trattazione a 4.3.3).
In questo caso il numero di elementi contenuti in clusters di grande dimensione si attesta al 22% del totale degli elementi del componente

esaminato (consideriamo qui i cluster maggiori di 500 elementi in quanto vi è un solo gruppo più grande di 1000 unità), ben lontano dal 44% ottenuto con il Fastgreedy, probabilmente proprio a seguito all'accorpamento di comunità, le quali saranno identificate solo a seguito di altre iterazioni dell'algoritmo.

Il numero di pagine appartenenti a gruppi di dimensione irrisoria (meno di 5 elementi) si attesta, sempre limitatamente ai risultati sul componente principale del grafo, al 3% del totale. Nel caso del Fastgreedy, invece, il numero di comunità di dimensione inferiore a 5 elementi è soltanto pari a 3 unità. Questo, tuttavia, è dato soprattutto dal fatto che il Louvain method compie una divisione in comunità completa. Perchè si possa dire lo steso nel caso del Fastgreedy servirà aver effettuato le varie ricorsioni sui sotto-grafi individuati.

4.3 Confronto con la struttura delle categorie

4.3.1 Creazione dell'albero delle categorie

Benché venga spontaneo immaginare la topologia delle categorie come un albero, capita molto spesso che si verifichino cicli, rendendo la struttura un vero e proprio grafo.
Il problema è stato invece affrontato semplicemente interrompendo l'esplorazione dell'albero nel caso si incontri un nodo, ovvero una categoria, già esplorato. La presenza di categorie vuote, di categorie contenute in sé stesse o di categorie non contenute in nessun'altra sopra-categoria non ha creato problemi, tuttavia questi elementi evidenziano una mancanza di vincoli posti da Wikipedia atti a

mantenere una struttura precisa delle categorie.
Grazie ai dati presenti nei dump dei database di Wikipedia si è dunque potuto risalire per ciascuna pagina alle categorie cui appartiene, e per ogni categoria in quale altra categoria è contenuta.
Come viene evidenziato anche in [17] dall'analisi vanno eliminate tutte quelle categorie "di sistema", ossia, ad esempio, le categorie di pagine da eliminare o da rivedere, categorie rappresentanti elenchi di altre categorie, gruppi di utenti, ecc... Si è tuttavia scelto di mantenere le categorie che contengono la parola "list", in quanto indicano elenchi di pagine che per l'analisi che si vuole condurre possono indicare elementi molto legati tra loro (ne sia un esempio "Lista di Primi Ministri inglesi") a differenza della altre categorie in blacklist. Essendo le categorie assegnate dagli utenti, non è raro che alcune pagine contengano solo categorie di sistema, e che quindi non possano essere collocate all'interno dell'albero definito in precedenza. Tuttavia anche in questo caso si tratta di una relativa minoranza di pagine sul totale, pari al 13%, e i risultati finali non vengono perciò alterati.

4.3.2 Copertura di un cluster

Come già accennato in precedenza vogliamo ora verificare in quale misura le pagine di un dato cluster siano semanticamente vicine tra loro, ossia se pagine con molti contributori in comune appartengano a un dato argomento, o se invece si vengano a creare cluster molto eterogenei con elementi molto diversi tra loro.
Il principale problema incontrato nell'identificazione delle categorie di ciascun articolo è una non sempre precisa corrispondenza uno a uno tra le pagine di Wikipedia e i nomi delle pagine presenti nel database

usato nella creazione di questa topologia (questo è anche dato dal fatto che mentre per la realizzazione dei cluster si sono sfruttati i database del 2007, per la strutturazione delle categorie si è scelto di utilizzare quelli più aggiornati del 2010, per poter avere una topologia delle categorie il più preciso e dettagliato possibile); per risolvere la questione si è cercato di mantenere una mappatura il più possibile completa tra i dati presenti nei cluster e quelli nel database, riuscendo ad eliminare solamente una percentuale minima delle pagine corrispondente allo 0.04% sulle 113831 totali.
L'assunzione che piccoli spostamenti nell'albero delle categorie portino a un altrettanto breve spostamento nell'area semantica pare lecita in quanto il dettaglio con cui le categorie sono strutturate nella tassonomia pare essere, al 2010, sufficientemente alto. L'unico problema potrebbe sorgere in prossimità della radice dell'albero, dove piccoli spostamenti portano a muoversi entro categorie generiche che possono avere anche grande distanza semantica. Tuttavia questo non pare un problema, in quanto difficilmente una pagina è associata direttamente a una di queste "macrocategorie", ed è ancora più difficile che più pagine di uno stesso cluster appartengano alle categorie generiche che possano essere incontrate nelle varie esplorazioni dell'albero.

Nel caso specifico si è deciso di verificare se in un dato cluster vi sia la presenza di una categoria, tra tutte quelle cui le pagine di questa comunità appartengono, che consenta con un piccolo spostamento nell'albero di individuare un elenco di categorie sufficiente a coprire tutte le pagine del cluster, o comunque il numero maggiore possibile (dicendo che una pagina è coperta da una categoria intendiamo che la pagina è contenuta in questa categoria).

4.3.3 Risultati delle analisi

Per poter effettuare un'analisi statistica si è scelto di esaminare solamente le comunità di dimensione sufficiente a produrre un dato percentuale significativo, in particolare piccoli cluster di 2 o 3 elementi rischiano di produrre percentuali non comparabili con quelle di cluster più grandi, infatti in questo caso se anche un solo elemento non venisse coperto porterebbe a una variazione della percentuale di copertura rispettivamente del 50 e del 33%; si sono quindi scartati tutti i gruppi di pagine con meno di 5 membri, portando avanti quindi lo studio su circa il 72% degli articoli di Wikipedia rimasti per quanto riguarda i cluster realizzati con il Louvain method e il 76% seguendo invece i risultati del fastgreedy usando soglia 0.4 sulla modularity (da ora i risultati derivati dai dati del fastgreedy saranno indicati con [F] e quelli derivati dall'utilizzo del Louvain method con [L]).
Va detto che i risultati derivati dall'applicazione del Louvain method saranno molto probabilmente migliori, per via della massimizzazione locale della modularity. Del resto, nel caso del fastgreedy, il problema della resolution limit è stato aggirato imponendo una soglia sulla modularity minima, necessaria per rendere "valida" una partizione, troppo netta, il che potrebbe aver portato alla divisione di alcune comunità o alla mancata identificazione di altre.

Muovendosi nell'albero per ciascuna categoria di un livello verso l'alto e di uno verso il basso e procedendo come descritto si ottengono risultati molto interessanti. Mantenendo la soglia sulla dimensione minima delle comunità a 5 si rileva come i clusters, con almeno il 95% degli elementi al loro interno semanticamente simili tra loro, siano il 73%

nel caso [F] e il 74% nel caso [L] . Questa percentuale sale addirittura all'83% sia in [F] che in [L] nel caso si volesse effettuare uno spostamento nell'albero di un livello verso l'alto e di due verso il basso, spostamento che può ancora considerarsi molto circoscritto rispetto alle dimensioni totali dell'albero e che dimostra come gli elementi dei cluster abbiano di fatto una notevole pertinenza (da questo momento [L] ed [F] si riferiranno a questo particolare caso).

In Fig. 14, Fig. 15 e Fig. 16 sono riportati tre clusters identificati dal Louvain method, è evidente la forte pertinenza semantica degli elementi:

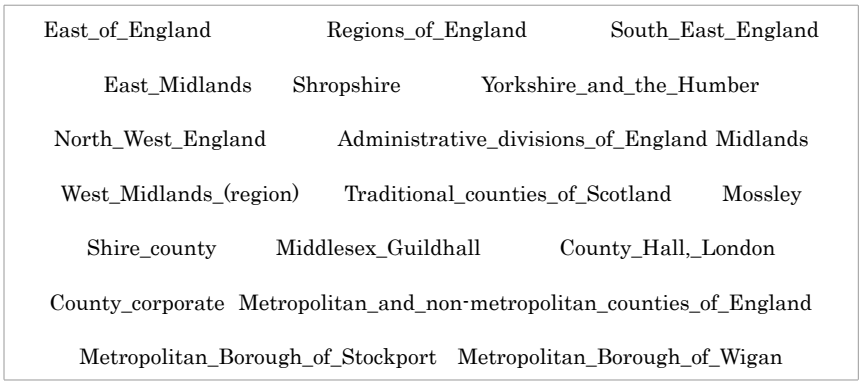

Fig. 14: Cluster numero 453 – Pagine su regioni del Regno Unito

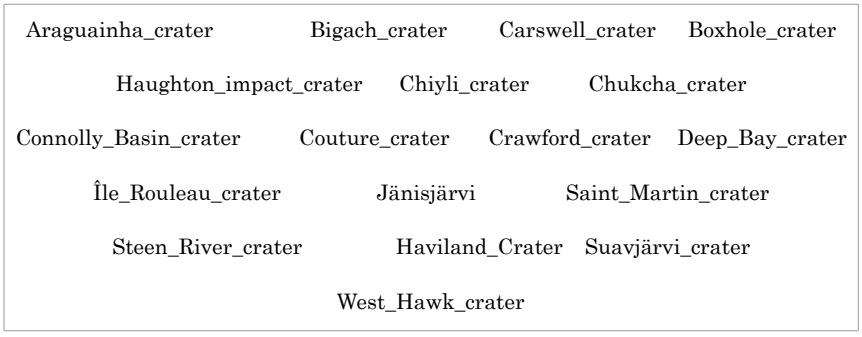

Fig. 15: Cluster numero 6201 – Crateri terrestri.

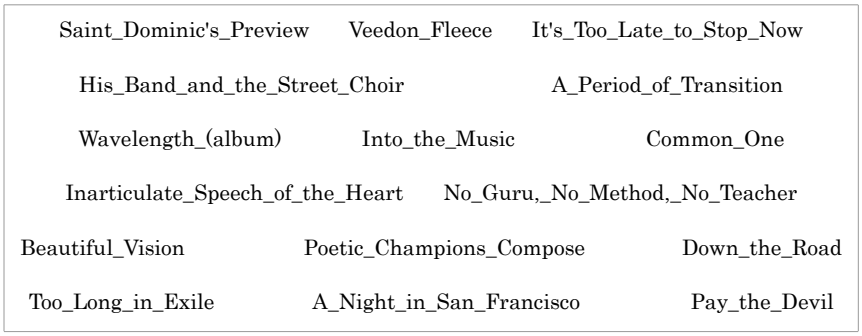

Fig. 16: Cluster numero 11960 – Album di Van Morrison

Un'altra interessante osservazione sui risultati riguarda quei cluster di dimensione considerevole ma che non presentano un'alta percentuale di elementi coperti. Si può verificare come, sia in [L] che in [F], la pressoché assoluta totalità dei clusters di dimensione maggiore di 50 elementi con almeno 10 dei propri elementi non raggiunti da alcuna categoria "di copertura", corrisponda in realtà a gruppi di pagine di argomento ben definito. Questo è uno dei problemi derivanti dal fatto che l'assegnazione delle categorie alle pagine di Wikipedia è un processo facoltativo e poco supervisionato, si devono prendere quindi questi risultati con la dovuta cautela, in quanto, come si è visto, se per clusters risultati completamente coperti si può affermare con discreta certezza che appartengono tutti a una stessa area semantica, non si può con la stessa certezza affermare che clusters non coperti (completamente o in larga parte) siano costituiti da elementi molto diversi tra loro.

A fronte di questa considerazione si può quindi assumere che nella maggior parte dei clusters, se non addirittura nella totalità (ma andrebbero analizzati manualmente anche i cluster più piccoli per verificare ciò), sono presenti pagine sempre inerenti a un determinato argomento.

Esiste tuttavia un grosso cluster di 1931 elementi, nel caso [L], in cui il

68% delle pagine non risulta coperto e, a seguito di un'analisi manuale, si può riscontrare come questo sia l'unico insieme significativo di pagine veramente eterogenee tra loro. Lo stesso avviene in [F], dove questo cluster ha però dimensione 2517 e la percentuale di pagine scoperte è il 59%.
In Fig. 17 sono mostrati alcuni elementi presenti in questo particolare gruppo di pagine.

Il particolare contenuto di questo cluster porterebbe a ritenere che pagine di argomento piuttosto generico, che possono essere modificate da una grande varietà di utenti, non riescano a generare una densità di collegamenti verso quella che sarebbe la loro naturale collocazione. Tuttavia se questo fosse vero sarebbe stato più probabile ritrovare queste pagine sparse tra gli altri clusters o in piccoli clusters indipendenti piuttosto che in un unico grande insieme. Non è dunque da scartare l'ipotesi che questa sia veramente una comunità a sé stante al pari di tutte le altre.

Great_Pyramid_of_Giza Gold Catalyst Gottfried_Leibniz Golf
Gettysburg_Address Chaos_theory Golden_ratio Grizzly_bear Genocide
Glacier Gamma_ray Glass Golgi_apparatus Gothic_fiction
Green_Bay_Packers Giant_Panda Giraffe George_Vancouver
Goth_subculture Glucose Garfield LGBT_social_movements Gordon_Brown
Goldfish Gunpowder_Plot Hades Helium Halogen George_Frideric_Handel
Holland Henry_Ford Hemoglobin Clinical_depression The_Holocaust
Edvard_Munch Hinduism House Daniel_Defoe Bone Herman_Melville
Hindu History_of_the_Internet Halle_Berry Honda Heaven Human_rights
Halloween Hezbollah Hockey Heroin Human_cloning
Henry_VIII_of_England Henrik_Ibsen Hoover_Dam Hair Haggis
Hammurabi Detroit_Tigers Italy India Frank_Sinatra Internet Indonesia
Troll_(Internet) Isaac_Newton Carl_Sagan Iran Italian_language Iron
Iodine Iliad Illinois Insane_Clown_Posse Industrial_Revolution Insect
Ice Imperialism Infinity

Fig. 17: Estratto dell'unico cluster di soli elementi eterogenei.

4.4 Confronto con Wikisuggestion

Nel progetto Wikisuggestion ([24]) si era effettuata una divisione in cluster della rete di pagine riferita alla Wikipedia italiana del 2007, sfruttando soglie diverse per via della minore quantità di dati disponibili (decisamente inferiore rispetto alla Wikipedia inglese aggiornata allo stesso anno).

In particolare in quel frangente si era ricorsi a un'analisi puramente qualitativa dei cluster, ma si era constatato come in molti casi si fossero costituiti gruppi di pagine con una chiara pertinenza semantica, seppur con alcuni elementi estranei e in nessun modo ricollegabili agli altri. Un altro elemento riscontrato è stata la fusione di cluster appartenenti a diverse aree semantiche senza che vi fosse una vera pertinenza tra le aree stesse.

Si era allora supposto che la rete fosse ancora in fase di definizione, in quanto la Wikipedia italiana ha una frequenza di interventi ovviamente inferiore di quella inglese ed è anche più giovane, e si era prevista la possibilità che risultati migliori si sarebbero potuti ottenere tramite l'analisi di dati più completi e con soglie migliori. Essendo questo il caso che si prospettava nel precedente progetto è possibile ora fare un breve confronto tra i risultati ottenuti nell'analisi dei due grafi. Tuttavia questa comparazione va fatta in modo puramente qualitativo, non avendo condotto una vera analisi sulla semantica del contenuto dei cluster nel corso del progetto "Wikisuggestion".

Notiamo anzitutto che, come si era previsto, sono scomparsi dai cluster gli elementi estranei, e a prima vista i gruppi sono ora ben definiti, senza singoli elementi provenienti da aree semantiche lontane.

Allo stesso modo sono scomparsi i cluster chiaramente derivati dall'unione di due comunità separate o da parti di esse; come visto dall'analisi della semantica dei gruppi di pagine siamo quasi sempre in

presenza di comunità appartenenti ad aree ben distinte tra loro. Questo, ovviamente, a meno dei casi particolari già analizzati e degli eventuali cluster di piccola dimensione non ispezionati manualmente e che potrebbero contenere elementi poco pertinenti, ma questi sarebbero tuttavia una minoranza visti i dati raccolti.

Risulta pertanto lecito supporre che la struttura di Wikipedia sia in continuo divenire, con aggiunta di nuove pagine, rimozione di articoli obsoleti e spostamento di contenuti, una vera e propria fase di assestamento che vede una pagina consolidata all'interno della rete da noi generata solo dopo un certo arco di tempo.
Bisogna dire che in [1] viene portato all'attenzione, almeno nel caso in esame di del.icio.us, come esistano molti fattori che possono intercorrere a cambiare la struttura della rete in esame, come l'arrivo di utenti o il loro abbandono, gli interessi della comunità possono cambiare o potrebbe cambiare il modo, la frequenza, la qualità dei contributi di un utente. Se questo può essere vero nel sito di social bookmarking si ritiene che, viste le dimensioni di un enciclopedia come Wikipedia e il numero di utenti in essa, sia difficile che cambiamenti di questo genere possano generare modifiche della rete apprezzabili nel breve periodo. Il cambiamento dovrebbe essere davvero di massa per poter produrre differenze evidenti, è più probabile che la struttura della rete punti a una stabilizzazione progressiva. Eventuali veri cambiamenti nei comportamenti e nei contenuti possono essere assorbiti in un arco di tempo sufficientemente lungo, o cominciando a eliminare i dati riferiti ai primi mesi di attività di Wikipedia (con le dovute cautele).

Difficile fare una previsione sull'evoluzione di quel gruppo contenente solo pagine semanticamente lontane l'una dall'altra (4.3.3).

All’aumentare degli edit questo gruppo potrebbe andare dividendosi, con ogni elemento ricollocato in un cluster di maggiore pertinenza, nel caso prevalgano in numero e qualità gli edit di quella categoria di utenti veramente esperti e legati a una di quelle pagine. D'altronde questo particolare gruppo potrebbe persistere qualora effettivamente i suoi elementi costituiscano una vera comunità al pari delle altre. Sarebbe tuttavia inaspettata l’esistenza di una comunità tanto eterogenea e lontana da una qualsiasi struttura semantica.

5. Conclusioni e sviluppi futuri

In questa tesi si è mostrato un sistema volto alla generazione di una rete di pagine dell'enciclopedia Wikipedia basata sulla presenza di edit da parte di utenti comuni sui diversi articoli del portale. Sono stati illustrati vari sistemi di valutazione della qualità di un contributo, al fine di poter dare un valore ai semi-archi della rete bipartita creata in un primo momento. Si sono evidenziati i problemi legati alla quantità e alla qualità delle informazioni grezze disponibili inizialmente e le varie soglie e sistemi utilizzati per eliminare i dati meno rilevanti. Si sono mostrati i passi e le convenzioni adottate per il passaggio da rete bipartita pagine-utenti a una rete di sole pagine, eseguendo su quest'ultima due algoritmi di clustering basati entrambi sulla massimizzazione della modularity, dopo aver analizzato i vari algoritmi presenti nel campo dell'identificazione di comunità all'interno di una rete.

In tal modo si è potuto procedere con una semplice ma efficace analisi della semantica interna ai vari gruppi di pagine, sfruttando l'albero delle categorie di Wikipedia, e arrivando così alla conclusione che nella grande maggioranza dei casi i cluster generati presentano elementi inerenti a un determinato argomento. È stato inoltre mostrato il significato di alcuni risultati inattesi, sottolineando analogie e differenze rispetto al precedente progetto Wikisuggestion.

Sono molti gli sviluppi e i miglioramenti applicabili ai vari procedimenti analizzati nel corso di questa tesi.
Un primo incremento della qualità dei risultati finali si potrebbe sicuramente ottenere dall'utilizzo di dump dei database raccolti nello

stesso anno sia per quanto riguarda le informazioni di pagine ed edit degli user, sia per quanto riguarda le categorie; in tal modo si potrebbe ridurre il numero di pagine perse a causa del diverso nome in periodi differenti.

Un altro miglioramento si potrebbe perseguire tenendo conto, come fatto notare in [7], che molte categorie, soprattutto spostandosi verso le categorie più generiche, presentano significati molto simili e potrebbero essere accorpate tra loro. Tuttavia, considerati i risultati già ottenuti senza eseguire accorpamenti di questo genere, l'unico cambiamento sarebbe, eventualmente, una conferma ancora più netta del fatto che in generale i cluster contengono elementi semanticamente molto simili tra loro.

Un interessante elemento da implementare nel modello proposto per l'analisi dei cluster potrebbe considerare non solo quanti utenti hanno scritto su una determinata pagina, ma anche quali utenti vi hanno contribuito, andando così a verificare se i contributi alle pagine di un dato cluster provengono in maggior parte dagli stessi utenti, o se invece non vi sia alcuna predominanza di autori.

Inoltre la possibilità di selezionare un set di utenti che hanno scritto sulle pagine di un certo cluster potrebbe consentire di individuare eventuali altre pagine esterne a quel gruppo di elementi e magari di un'area semantica completamente diversa, ma che in qualche modo è di interesse per gli utenti di quel gruppo. Ad esempio si potrebbe scoprire che una certa percentuale di persone che si interessano di linguaggi di programmazione ascoltano un certo gruppo musicale.

Considerato il confronto tra il precedente progetto Wikisuggestion e la

presente tesi, sarebbe utile esaminare l'evoluzione nel tempo della rete di pagine e dei cluster che la definiscono (sfruttando i periodici aggiornamenti dei dump di Wikipedia), in modo da studiare la reale collocazione di ogni pagina al crescere degli edit, potendo così andare a realizzare anche previsioni sulla futura evoluzione dell'enciclopedia.

All'interno dei singoli clusters, infine, si potrebbe considerare la sotto-rete costituita dai vertici ivi presenti e dalle connessioni che intercorrono tra loro, andando a misurare la betweenness centrality dei nodi e determinare quali siano veramente interni al cluster e quali potrebbero invece essere più "limitrofi" e avere connessioni con altre aree semantiche, come del resto già verificato in [1] su del.icio.us. .

Se si considera poi una rete in cui i nodi siano le comunità identificate e analizzate in questa trattazione e dove i link tra due nodi corrispondano alle somme degli archi tra gli elementi di quelle due comunità, è possibile svolgere un'ulteriore analisi sulle relazioni che intercorrono tra comunità. Un rapido esempio in Fig. 18 mostra come esistano interessanti sviluppi in tal senso: facendo collassare la rete come spiegato, eliminando i loop e i componenti molto piccoli, restano gruppi di comunità debolmente connesse tra loro, di cui si mostra un esempio (i nomi dei nodi corrispondo al primo elemento della comunità in cui questo è contenuto). In questo particolare estratto del grafo di comunità si nota come il componente in analisi contenga solo comunità inerenti al mondo della Formula1.

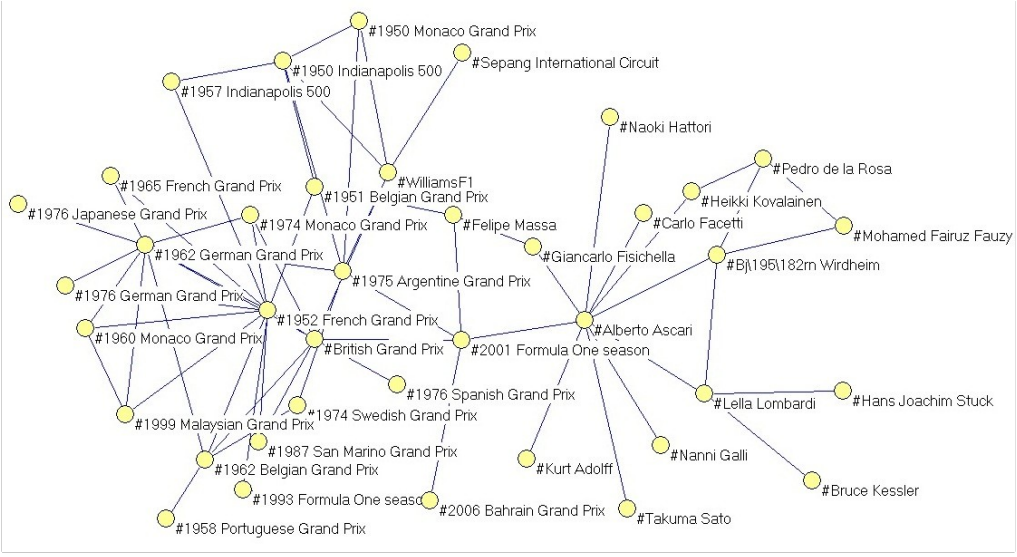

Fig. 18: Componente di una possibile rete di comunità.

Le applicazioni pratiche di questo elaborato presentano diverse applicazioni all'interno della stessa Wikipedia. Sarebbe ad esempio possibile, durante il processo di assegnamento di una categoria a una pagina, suggerire le categorie degli articoli contenuti nello stesso cluster cui quella pagina appartiene. Inoltre, a fronte di un cluster coperto per una percentuale molto alta dei suoi elementi, si potrebbe andare a suggerire per le pagine rimaste escluse una delle categorie di copertura rilevate.

Sempre nell'ambito dell'analisi svolta su Wikipedia, il suggeritore sviluppato in Wikisuggestion potrebbe essere convertito in un add-on per browser, o in una feature implementata direttamente in Wikipedia, che vada a suggerire articoli simili a quello visitato, sganciandosi dalla spesso incompleta suddivisione in categorie.

Altre applicazioni possono riguardare biblioteche dove sarebbe così possibile suggerire libri di possibile interesse basandosi su un volume appena scelto.

6. Bibliografia

[1]: P. Mika. Ontologies are us: A unified model of social networks and semantics. Web Semant. 5 (1), 5-15, 615-631, 2006.

[2]: C. Cattuto, D. Benz, A. Hotho, G. Stumme. Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems, in 'Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)', 39-43, 2008.

[3]: C. Cattuto, D. Benz, A. Hotho, G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In The Semantic Web — ISWC 2008, Volume 5318 of Lecture Notes in Computer Science, Heidelber. Springer Berlin. 615-631, 2008.

[4]: D. Laniado, R. Tasso. Co-authorship 2.0: Patterns of collaboration in Wikipedia. In Proc. of Hypertext 2011, 2011.

[5]: D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians talk: network and tree structure of Wikipedia discussion pages. In ICWSM. The AAAI Press, 2011.

[6]: R. Tasso. Analisi della Costruzione Partecipativa di un Wiki con un'Applicazione a Wikipedia, Politecnico di Milano, 2008.

[7]: J. Farina. Assegnamento automatico di macrocategorie agli articoli di Wikipedia, Politecnico di Milano, 2010.

[8]: B.T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the Wikipedia. In Proceedings of the 4th International Symposium on Wikis. ACM, 1-10 , 2008.

[9]: M. E. J. Newman. Detecting community structure in networks. Eur. Phys. J. B 38, 321-330, 2004.

[10]: M. Girvan, M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99 , 7821–7826, 2002.

[11]: M. Newman, M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113, 2004.

[12]: A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. Physical Review E, 70:066111, 2004.

[13]: Newman MEJ. Analysis of weighted networks. Phys Rev E 70: 056131, 2011.

[14]: M. Latapy, P. Pons. Computing communities in large networks

using random walks. Journal of Graph Algorithms and Applications, 10, 191-218, 2006.

[15]: S. Fortunato, M. Barthelemy. Resolution limit in community detection. Proceedings of the National Academy of Science, 104:36-41, January 2007.

[16]: V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks. J. Stat. Mech., P10008, 2008.

[17]: S. Ponzetto, M. Strube. Deriving a large scale taxonomy from wikipedia. In: Proc. of the 22nd National Conference on Artificial Intelligence (A A A I-07), Vancouver, 1440-1447, 2007.

[18]: D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In AAAI '08, 25-30, 2008.

[19]: C. Mueller-Birn, J. Lehmann, and S. Jeschke. A composite calculation for author activity in wikis: Accuracy needed. Proceedings of Web Intelligence and Intelligent Agent Technology, 2009.

[20]: E. Gabrilovich, S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India, 1606-1611, January 2007.

[21]: J. Leskovec, K. Lang, M. Mahoney. Empirical Comparison of Algorithms for Network Community Detection. ACM WWW International conference on World Wide Web (WWW), 2010.

[22]: B.T. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In Proceedings of the 16th International Conference on World Wide Web, page 270, 2007.

[23]: S. Fortunato. Community detection in graphs. Physics Reports, vol. 486, 75-174, 2010.

[24]: F. Colzada, M. Di Vitto. Wikipedia pages suggestion system (WikiSuggestions). Politecnico di Milano, 2011.

[25]: R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In WWW '07,131-140, 2007.

[26]: Igraph, v. 0.5.4, http://igraph.sourceforge.net/

[27]: Python, 2.7.1, http://www.python.org/

[28]: Louvain method C++ implementation, http://sites.google.com/site/findcommunities/

[29]: Pajek, v. 1.28, http://vlado.fmf.uni-lj.si/pub/networks/pajek/