SEMANTICA EMERGENTE IN WIKIPEDIA

Tesi di laurea di:
Fabio Colzada, matricola 713990
Mattia Di Vitto, matricola 714583

Relatore: Prof. Marco Colombetti
Correlatori: Ing. David Laniado, Ing Riccardo Tasso

Anno accademico 2010-2011

Obiettivi

- Sviluppare una metrica di "similarità sociale" fra pagine, basandosi sull'attività degli utenti di Wikipedia

- Estrarre una semantica emergente dall’enciclopedia

Approccio

- 1) Costruzione di una rete bipartita utente-pagina
- 2) Creazione di una rete di pagine
- 3) Applicazione alla rete ottenuta di algoritmi non supervisionati di identificazione di comunità
- 4) Analisi dell'attinenza semantica delle pagine di ciascun cluster mediante l'albero delle categorie di Wikipedia

Costruzione della rete bipartita

- Edit longevity: misura quantità e qualità degli interventi degli autori sugli articoli considerando quanto del testo originale inserito da un utente sopravvive nelle 10 revisioni successive della pagina

- Rispetto ad altre metriche tiene conto anche del testo che è stato cancellato, spostato, sostituito ecc...

Costruzione della rete bipartita

- Tf-Idf: metrica molto utilizzata nei campi dell'information retrieval e del text mining

- Adattamento di Tf-Idf al caso in esame

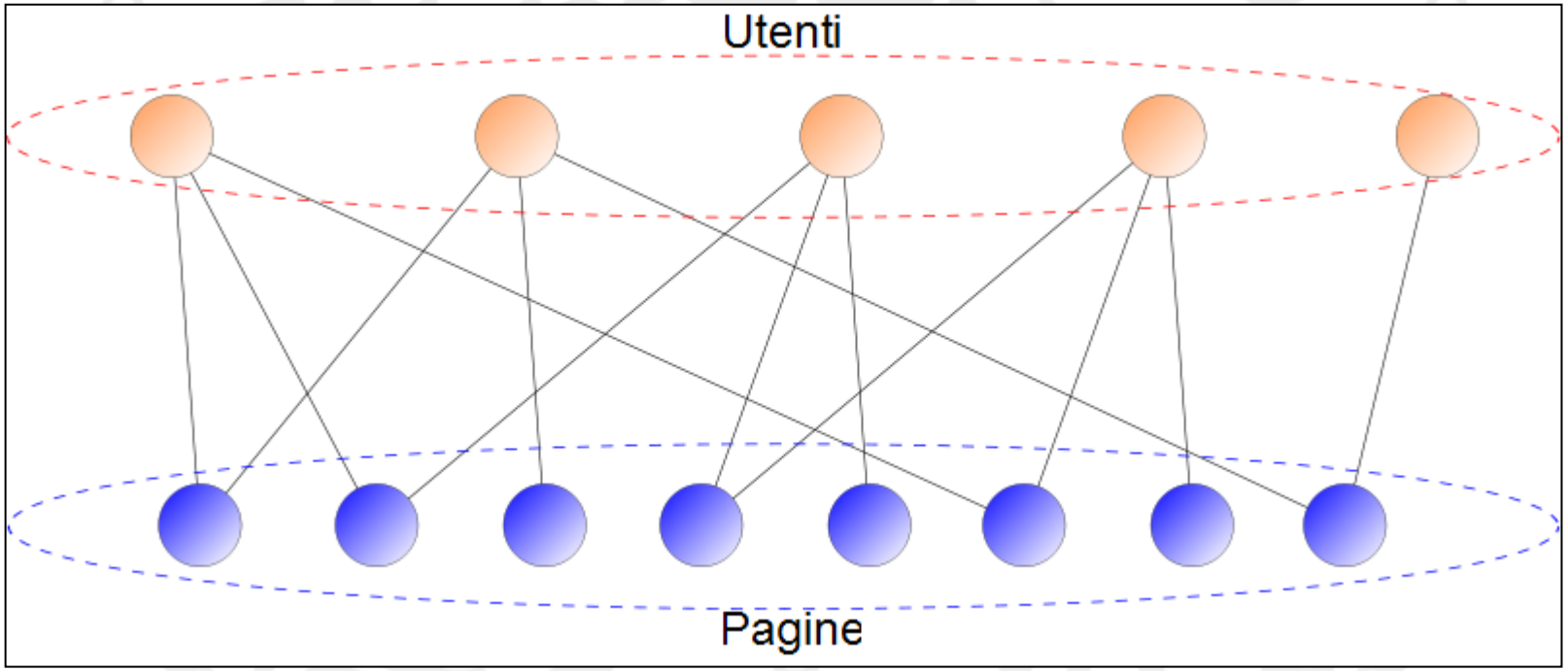
$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

Costruzione della rete bipartita

- Tf indica la qualità dell'attività di un utente rispetto alla qualità complessiva dei contributi esistenti per quella pagina

- Idf penalizza tutti quegli utenti che hanno contribuito a un elevato numero di pagine

Costruzione della rete bipartita



Costruzione della rete bipartita

- Costruita la rete bipartita si può procedere alla fase di creazione del grafo di pagine

- Adozione di una metrica di similarità che consenta la transizione da semi-archi utente-pagina ad archi pagina-pagina

Costruzione della rete di pagine

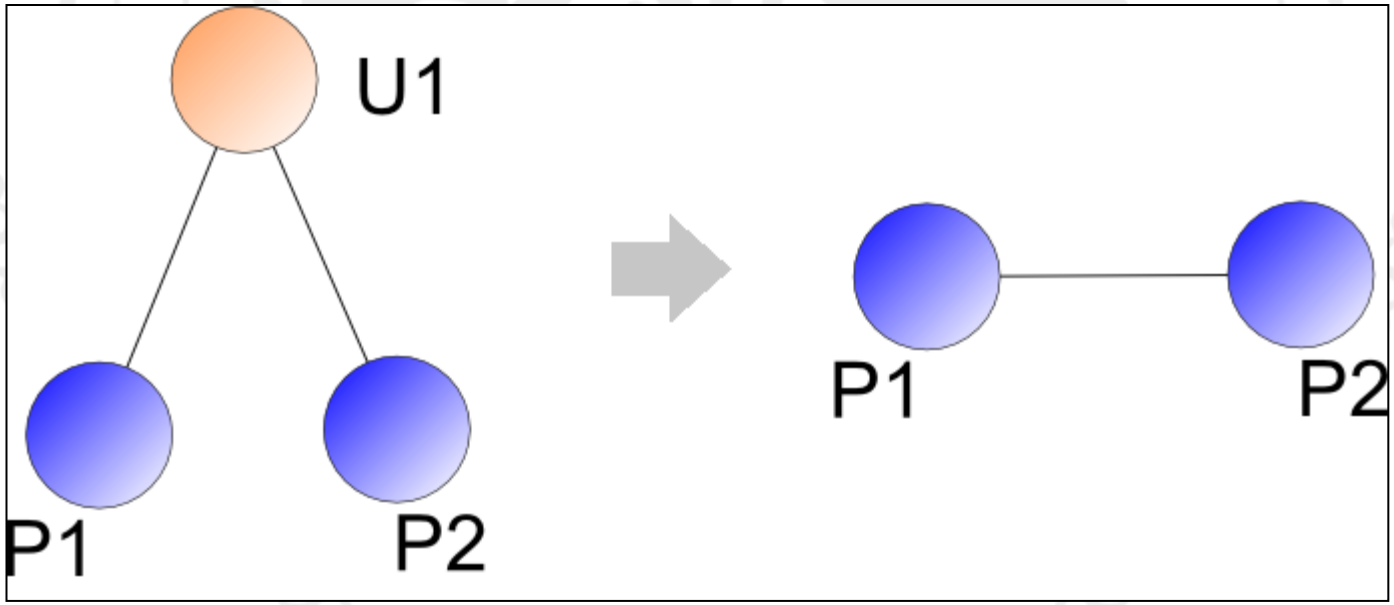
- Coseno di similarità:

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

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

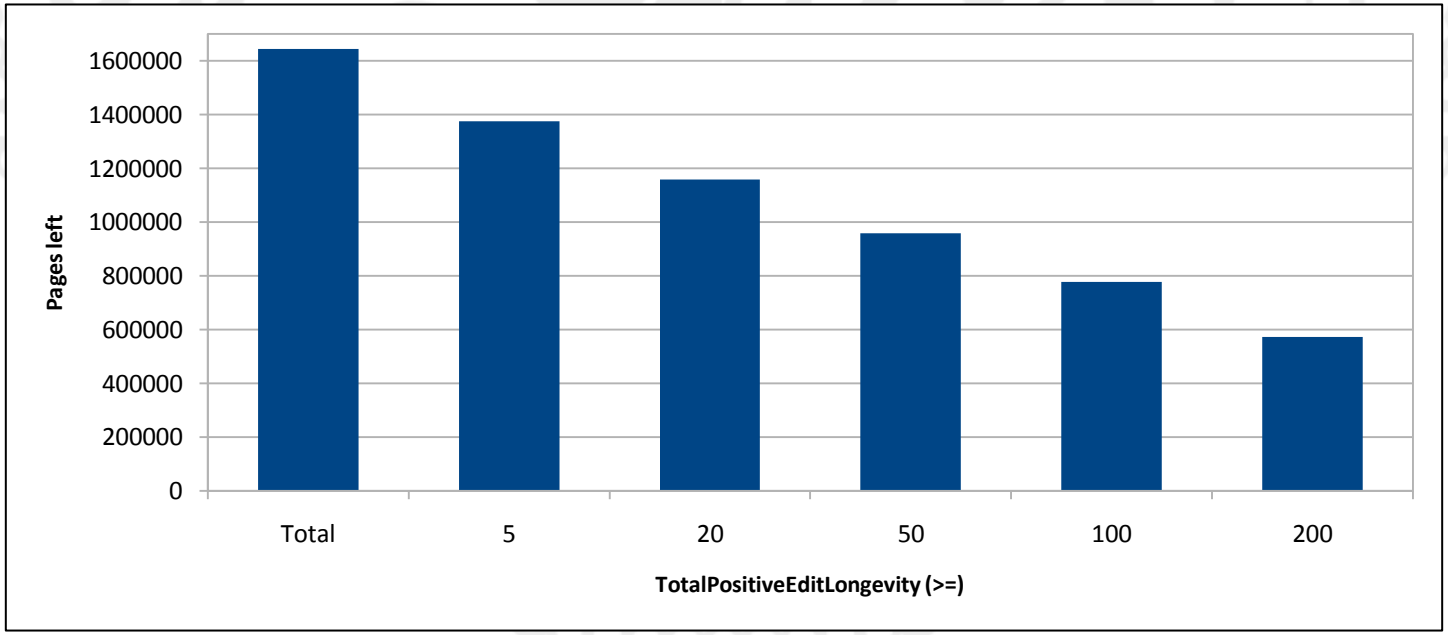
Costruzione della rete bipartita

- Proiezione della rete bipartita sulle pagine



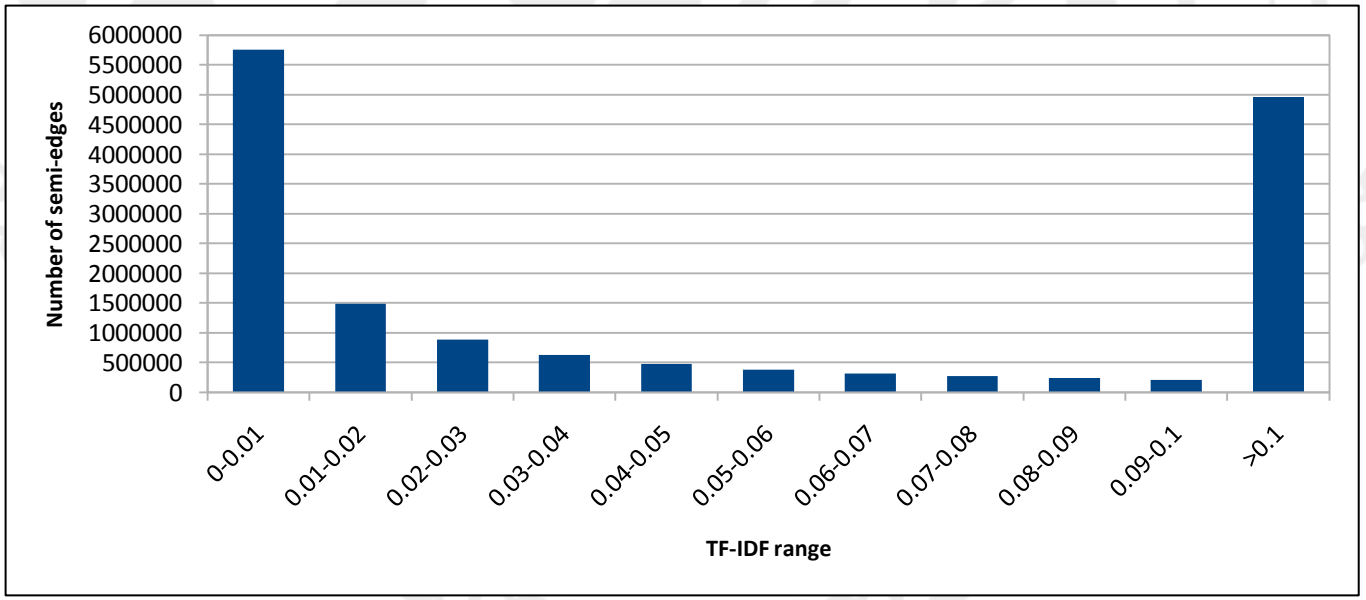
Costruzione della rete di pagine (soglie)

- Eliminazione degli articoli troppo piccoli: imposizione di un valore limite alla edit longevity complessiva totalizzata da ciascuna pagina (soglia scelta: 20)



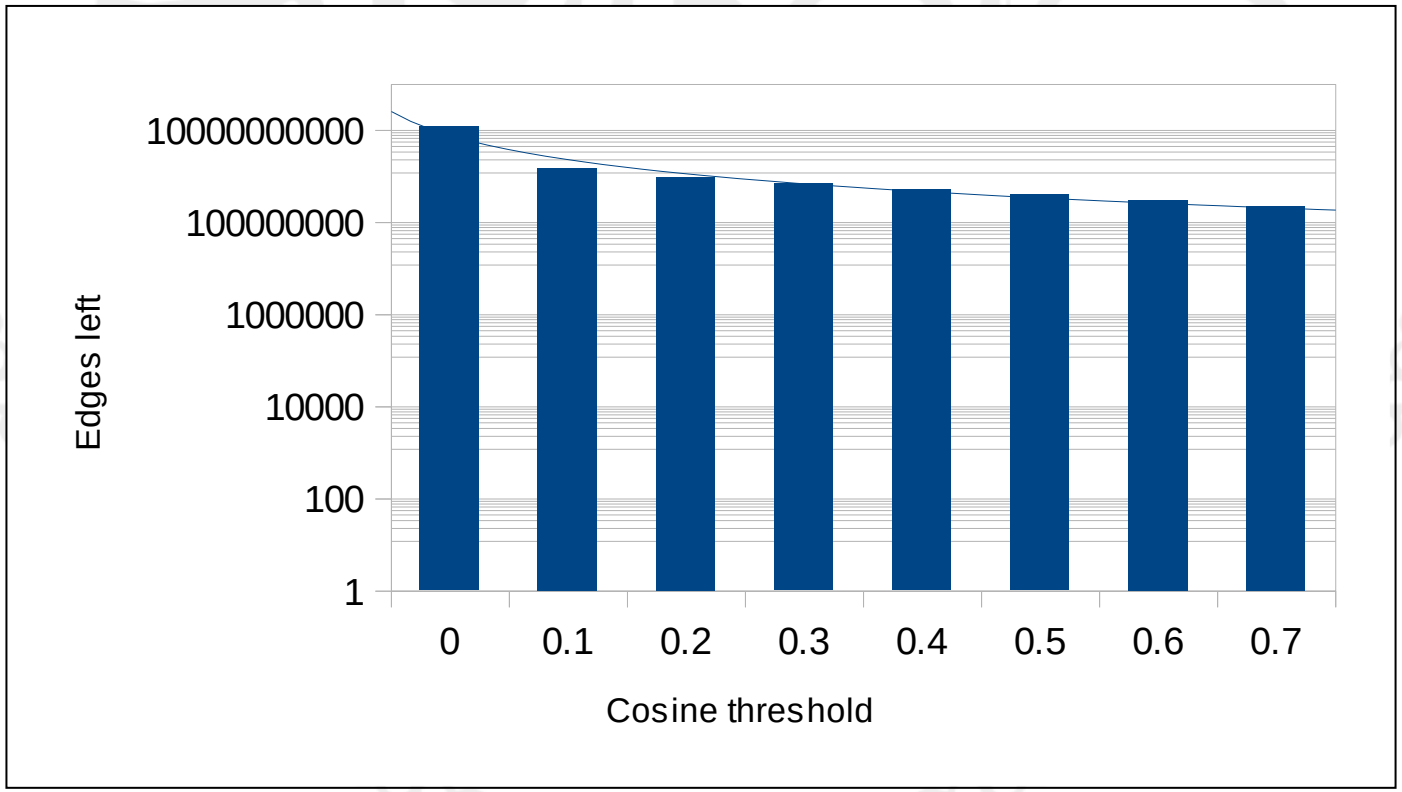
Costruzione della rete di pagine (soglie)

- Eliminazione dei semi-archi con basso Tf-Idf: rimozione dei semi-archi che porterebbero ad avere archi di scarsa importanza (soglia scelta: 0.03)



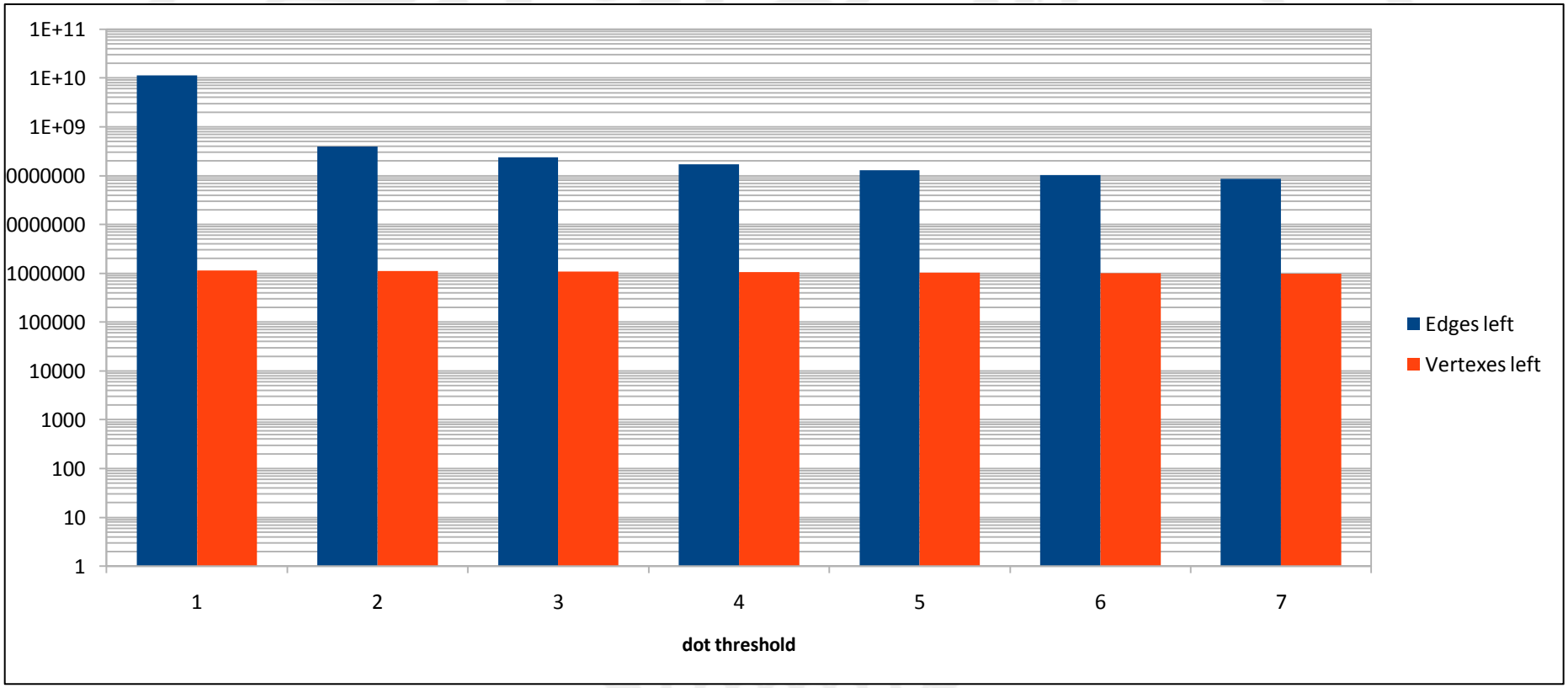
Costruzione della rete di pagine (soglie)

- Eliminazione degli archi con basso peso: soglia impostata a 0.5



Costruzione della rete di pagine (soglie)

- Eliminazione degli archi generati da pochi utenti comuni: soglia impostata a 5



Costruzione della rete di pagine

- Il grafo ottenuto non è connesso
- Presenza di numerosi componenti, molti dei quali di piccole dimensioni (spesso inferiori a 4 elementi)
- Presenza di due componenti di grandi dimensioni: 26 629 (23% del totale delle pagine) e 20 498 (18%)

Identificazione di comunità all'interno della rete

- Di solito le reti reali non sono omogenee
- Manifestazione di strutture di comunità
- Struttura di comunità: alta densità di archi all'interno di ciascun gruppo di vertici e numero relativamente basso di collegamenti tra i vari gruppi

Identificazione di comunità all'interno della rete (modularity)

- Modularity: indice della qualità raggiunta da una divisione in cluster

- Numero di archi che cadono all'interno dei gruppi di una determinata divisione, meno numero di archi che si otterrebbero in quei gruppi in una rete random con lo stesso valore di degree medio dei vertici

Identificazione di comunità all'interno della rete (Fastgreedy)

- Resolution limit: difficoltà nella rilevazione delle comunità formate da pochi elementi

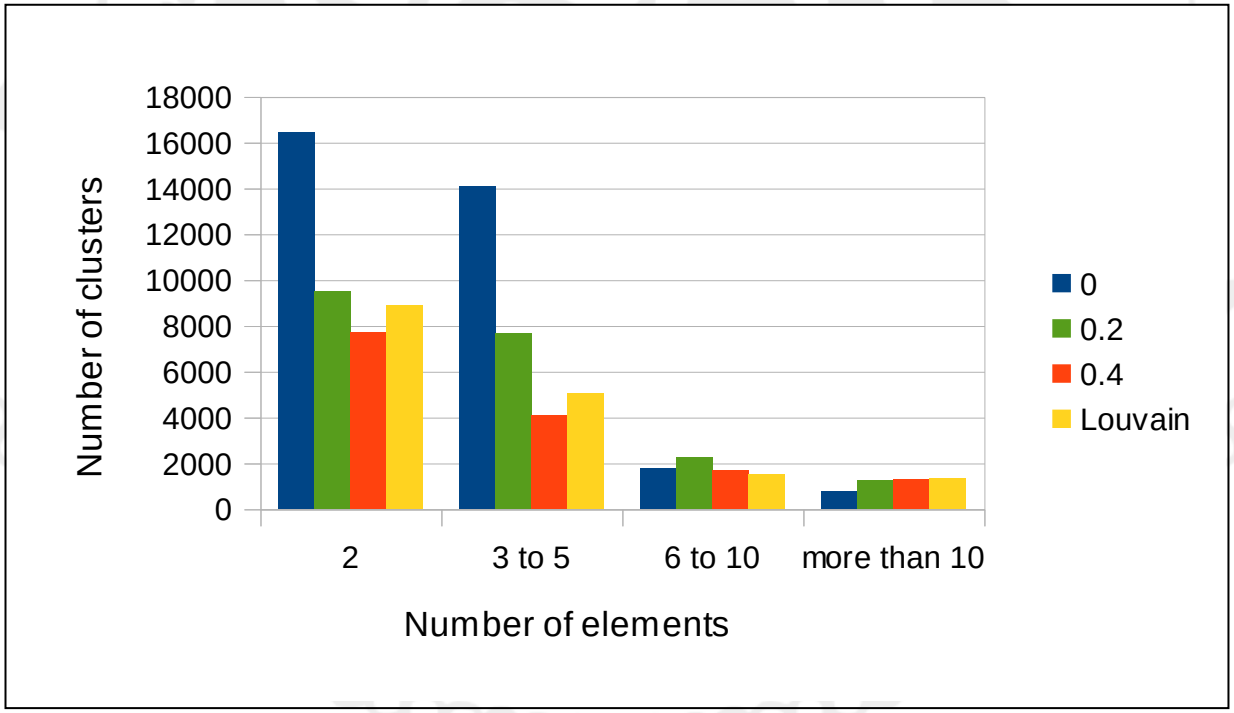
- Effetti: presenza di gruppi di pagine con più di 1000 elementi nel componente formato da 26629 pagine

Identificazione di comunità all'interno della rete (Fastgreedy)

- Soluzione: molteplici iterazioni del fastgreedy su tutti i cluster trovati a seguito di una prima iterazione (interrompendo la ricorsione se la modularity scende al di sotto di 0.4)
- Risultati: il 52% dei cluster presenta solo due elementi
- Tuttavia restano presenti numerosi cluster di dimensione maggiore che verranno considerati nelle successive analisi

Identificazione di comunità all'interno della rete (Louvain)

- Esente dal problema del resolution limit
- Più cluster di due soli elementi (70% del totale)
- Più cluster sopra i 50 elementi (da 126 nel caso del fastgreedy, con soglia sulla modularity a 0.4, a 249)
- Risultati migliori rispetto al fastgreedy: non c'è necessità di successive iterazioni



Analisi semantica

- Le pagine all'interno di ciascuno dei cluster precedentemente individuati cadono all'interno di un ristretto numero di categorie tra loro attinenti?

Analisi semantica

- Albero delle categorie di Wikipedia: un albero di tutte le categorie dell'enciclopedia, cui le pagine appartengono

- Ipotesi: spostandosi di poco all'interno dell'albero si rimane all'interno di uno stesso argomento

Analisi semantica

- Analisi limitata ai cluster di almeno 5 pagine
- Spostamento circoscritto nell'albero di un livello verso l'alto e di due verso il basso
- Risultati: l'83% dei cluster presenta almeno il 95% degli elementi al suo interno semanticamente simili tra loro (sia eseguendo l'algoritmo Fastgreedy che il Louvain method)

Conclusioni

- **Nella quasi totalità dei casi i cluster generati presentano elementi inerenti a un determinato argomento**

- **Dai contributi e dai comportamenti degli utenti è possibile identificare una semantica emergente dall'enciclopedia**