Università degli Studi di Padova

Corso di laurea triennale in
Ingegneria dell'Informazione

Analisi di una rete di autori di pubblicazioni scientifiche

Laureando:
Pietro Maria Nobili

Relatore:
Prof.ssa Cinzia Pizzi

Matricola:
1067941

Correlatore:
Dott. Mattia Samory

18 Luglio 2018
Anno Accademico 2017/2018

Indice

Abstract

È stato generato un grafo delle pubblicazioni del DEI.

Sono stati cercati cluster nel grafo.

Sono stati confrontati con la struttura delle comunità del dipartimento.

Introduzione

Breve descrizione del community detection e della sua importanza, in generale e nel caso particolare delle comunità di autori di pubblicazioni scientifiche.

Presentazione della struttura della tesi.

Capitolo 1

Community detection

1.1 Studi precedenti

Descrizione della community detection.
Descrizione della bibliografia attuale sui grafi di coautori.

1.2 Metodi usati

Che metodi vengono usati per generare le comunità.

1.2.1 Girvan–Newman

Descrizione metodo GN.

1.2.2 Blockmodel

Descrizione metodo blockmodel.

1.2.3 Clauset-Newman-Moore

Descrizione metodo Clauset-Newman-Moore se implementato.

Capitolo 2

Estrazione dati

2.1 Struttura database microsoft

Il database è formato da vari file in plain-text che contengono i record di Autori, Paper e Affiliation.

2.2 Processo di estrazione

Parto da una lista di nomi, in parte etichettati.
Estraggo gli ID autori.
Estraggo terne ID paper-autore-affiliation.
Creo gli edge.
Questo è un grafo, ma ha dei difetti:

- Vengono estratti gli ID di autori omonimi
- Allo stesso autore (una persona fisica) sono accoppiati più ID autore

2.2.1 Filtro di autori per affiliation

Per risolvere il primo problema:
Il set di ID autori estratto viene filtrato, tenendo solo gli ID autore che hanno almeno un paper con affiliation padovana.
Scarto quelli che non hanno mai affiliation padovana.
Molti autori spuri vengono eliminati.

2.2.2 Unione di ID autore in singoli nodi

Per risolvere il secondo problema, si propongono due modi:

Per nome

I nodi con nomi uguali o abbreviazioni l'uno dell'altro sono considerati un unico nodo.

Per distanza

Il metodo per nomi introduce un errore potenzialmente molto grave, nomi come Michele Zorzi e Mattia Zorzi vengono confusi e considerati un unico nodo. Nella lista di nomi considerata questo succede solo nel caso citato, ma in dataset più ampi i falsi positivi aumentano considerevolmente. Si tenta di risolvere questo problema considerando unici due nodi in base alla distanza minima che hanno nel grafo.

I nodi con nomi uguali o abbreviazioni l'uno dell'altro sono considerati un unico nodo solo se sono anche vicini nel grafo: si calcola il cammino minimo tra i nodi e li si unisce se è di lunghezza minore di x

TODO Il valore ottimo di x sarebbe bello scoprirlo sperimentalmente

Capitolo 3

Troubleshooting

Il grafo generato presenta ancora delle carenze.

La lista di partenza include gli afferenti DEI attuali, mentre il database è del 2015. Questo comporta la mancanza di professori, non più a Padova, che avevano ruolo di aggregatore di una comunità.

TODO inserire il nome di Apostolico nella lista di partenza e valutare i cambiamenti

Un modo proposto per includere i nomi mancanti è: estrarre gli ID autore; estrarre i paper-aut-aff; da questa lista di paper estrarre tutti gli ID autore; potenzialmente ridurre i paper (e gli autori) estratti in base alle affiliation (anche solo del DEI e non di tutta Padova); iterare il processo con la nuova lista di autori.

In questo modo si estraggono anche troppe comunità, una singola collaborazione con un dipartimento esterno comporta alle iterazioni successive l'inclusione di molti autori di quel dipartimento. Un modo per risolvere il problema è, alla fine delle iterazioni e della creazione dei cluster, considerare solo quelli che contengono almeno un nome che era nella lista originale: in questo modo dipartimenti esterni, anche se connessi al grafo, non vengono inclusi.

Capitolo 4

Risultati

Descrizione della v-measure
Presentazione dei valori di v-measure per i vari grafi e commenti.

Capitolo 5

Conclusioni

Partendo da una lista di nomi di un dipartimento si può estrarre un grafo che lo rappresenti? Le comunità generate rispecchiano quelle reali?

Bibliografia

[1] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.