Scalable Community Detection using Label Propagation & Map-Reduce

---

Akshay U. Bhat
Cornell University
aub3 [at] cornell.edu

We describe a scalable implementation of label propagation algorithm, using map-reduce formalism for detecting communities in networks with 10-100 million nodes. Using a 55 node Hadoop cluster, we test our implementation on a Twitter network containing 40 million users and 1.4 billion edges. The algorithm is capable of finding meaningful communities in reasonable amount of time also we found it to be capable of finding communities in different iterations at various levels such as nationality, special interests, location, and organization.

Results

Following are results using Twitter network, where all users with more than 900 followers (as of June 2009) have been removed:

- After 15th iteration
- After 7th iteration

Example of communities from above runs:

- Top 15 communities detected after 15th iteration (Each community approximately represents a nation)
- A small community of mostly linkedin employees detected using the algorithm after 7th iteration
- A small community of users associated with MIT Media lab detected using the algorithm after 7th iteration
- A small community of users associated with Semantic Web / W3C detected using the algorithm after 7th iteration

Following are results using the complete Twitter network:

- After 10th iteration

Each line in above files (except the top 15 communities file) has following format
Label \t Count \t User \t Score \t User \t Score \t User \t Score \t .....
Each line represents a single community and contains upto 10,000 users, if more than 10,000 users are present within a community then that community occurs over multiples lines and has same label. In case there is no screen_name for a given userid, then the userid is present.

To explore above resutls, you can use this python script. Just save above files in the same directory as the script.

Dataset & Code

This Network dataset collected by researchers from KAIST is used for testing the code, Currently it has follower information for 40 Million users collected in June 2009. Also for purpose of converting userids to screen names we

use dataset provided by Prof. Jure Leskovec and J. Yang.
Both datasets can be obtained from following link http://an.kaist.ac.kr/traces/WWW2010.html

Please have a look at the terms under which the data is being distributed by the researchers, It might NOT be suitable for commercial or non-research applications.

You can obtain code used to generate the result following github repository:
https://github.com/AKSHAYUBHAT/TwitterCommunityDetection

Note: The code was used to generate the resutls, and it is NOT designed to be used as a library. Still if you wish to use it with your own dataset, please drop me a line, I might be able to help. The code is designed to be used with hadoop installation at cornell and the twitter dataset, however it can be easily modified for use on a local hadoop installation.

Applications

The results from this algorithm can be used for numerous applications, such as link prediction, builiding better models for user behaviour, understanding language in context of community. An interesting application is valuation of tweets produced by users. Since tweets from users in some communities might be more valuable than others.

Acknowledgment

This work is funded in part by National Science Foundation grants CNS-0403340, SES-0537606, IIS 0634677, and IIS 0705774. We thank Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon and J. Yang. & Jure Leskovec for providing the dataset used in this paper.

References:

1. Usha Nandini Raghavan, Réka Albert, and Soundar Kumara, Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106 (2007)
2. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, What is Twitter, a Social Network or a News Media? WWW 2010, April 26–30, 2010
3. Akshay Bhat, Analogy Engines for Semantic Web. ISWC 2010