CS5232 Project Proposal

Antoine Francois Pascal Creux (A0123427M)

March 18th, 2014

1 Concept Study

Goodreads is a website launched in early 2007, which lets "people find and share books they love... [and] improve the process of reading and learning throughout the world."[4]. Two months after Flixster[2], a website where users share and rate movies they have watched, Goodreads let anyone record and mark the books they have read, and share with everyone their reaction after reading the last best-seller they ́ve just devoured.
Besides, goodreads has developed a social network above this book sharing experience. This platform connects people to books and authors by share ratings and recommendations, and thus gather people together. Readers see what their friends have read, but they can also meet new people sharing the same literature taste. In 2015, Goodreads platform gathers 34 million reviews from 30 million members on 900 million books[4].

2 Problem Description

Even if goodreads already has already developed a social network within their website (You can have a friendship relation with another reader), at first we would like to generate, study and compare all hidden networks we can generate based on book read, book reviews and authors followers. Based for insance on a book-user network which is bipartite, we would like to produce a user-user network. Then, we want to highlight any difference between those, or strenghen relations between readers if those networks share similar patterns. In order to carry out this experiment, we will make use of the best suited clustering algorithms such as K-neighrest in our study. Finally, based on our generated graph, and if we have enough time to implement it, we can study a user's influence over its friends. Indeed, we can see when a user has read a book, and then, see if one user of its circle may have recommended him a book.

3 Underlying Assumptions

4 Requirements

Goodreads data is not available online, we have to acquire all data through their APIs. Stanford students have already studied Goodreads platform by building a product recommendation. [3]. However, they have not published the data they acquired and we think that their data acquisition is wrongly implemented based on erroneous assumptions. Indeed, they acquired data through the *book.shelves* api method. Given a *user_id* and a bookshelf name, they were able to query all the books a user may have put in this virtual bookshelf. The major issue is that goodreads do not list

every user_id. They then decided to query for user_id randomly to query the books people have put to their public 'reading' shelve. Finally, they only gather 4000 users and their 'reading books'.

We have decided to query the data using other APIs, available at [1]:

- friend - Get a user's friend
- group.list - List groups for a given user
- group.members - Return members of a particular group
- group.show - Get info about a group
- fanship.show - Show author fanship information
- owned_books.list - List books owned by a user
- shelves.list - Get shelves owned by a user
- user.show - Get shelves owned by a user
- user.followers - Get user's followers
- user.following - Get people a user is following

Here, we can see that 4 user networks can be generated:

- based on the followed-follower relationship between two users, we can build a user network named the Twitter Network
- based on the groups a user is a member of, we can build a user network named Group Network
- based on the authors a user is fan of, we can build a user named Author Network
- based on the books a user has read, we can build a user network named Book Network

Goodreads have 30 million users which may be too large for our study, but the goal is to gather an acceptable amount of data. Goodreads provide the top user in certain categories such as the top 50 active users, top 50 readers, top 50 most popular reviewers well as the best reviews. These statistics are computed for the last week, last month, last 12 months and all the time for each country(as well as worlwide). Based on the top users, we have a list of user_id from which we can query the data. On the other hand, for each user, we will store the books read and stored on the shelves, the authors the user is fan of, the groups the user belongs to.

5 Project Objectives

The first objective is to query all the data, and store it efficiently. On a second time, we will develop a clustering algorithm, in other words, given a set of users, group them into a group where whose members share similarities. We will try to identify cliques or less precisely, relevant social cirles for each network. Then, we will compare the different patterns we could make out between these cliques.

6 Literature review

We compare n methods for the clustering of our dataset. We give a brief summary of the different clustering methods and state if these methods will be scalable to our own dataset.

Quantitatively, clustering aims at assigning each data point to a cluster.

6.1 K-means

K-means clustering aims at computing k clusters from that will minimize the distance from the data points to the cluster, such as we will split n data points into k groups.

It does not apply only to graphs. The most common application of k-means clustering is the iterative algorithm named Lloyd's algorithm:

1. Initalize (randomly preferably) the center of the k clusters
2. For each point, assign to it the closest cluster
3. Compute the new position of every cluster as the center of all data points assigned to this cluster
4. Repeat 2 and 3 until the clusters do not move (Relative to a ϵ for instance)

7 Proposed Contributions

8 Success Measures

9 Project Plan

10 Deliverables

References

[1] goodreads - APIs documentation. `https://www.goodreads.com/api`.

[2] Flixter. `https://www.flixter.com`, 2007.

[3] Book Recommendations on GoodReads.com. `http://cs229.stanford.edu/proj2008/IsaacsonSebastian-GoodReadsRecommendations.pdf`, 2008.

[4] goodreads - About us. `https://www.goodreads.com/about/us`, 2012.