# Crawler Technology Applied to Douban

Study on Social Networks based on Interest Interaction

Yeni Liang 201573656

## 1 Introduction

When it comes to Online Social Networks (OSNs) there are plenty of subjects to be studied, one of them being the study of user relationships. With more and more aspects in different people's lives linked together via social networks, this subject becomes a real most-talked-about topic today. We are living in a world overwhelmed with information from different types of connections[1]. As for users on Facebook and other similar networks, most interactions lie within direct communications.

Nevertheless, there are other networks where people prefer to communicate through mediums, for example, by sharing or reviewing a book, recommending a motion picture, promoting a local event or even putting up their own groceries to sell. This is called the "interest interaction" network in China, and Douban is the most successful one. The secret for Douban lies within the ways users interact with each other, and to put it into words, that would be "interest marketing".

Therefore, the attempt to study user relationships on Douban would be more efficient to focus on"interests and hobbies". Such information is available because we can always read what a user likes from his/her homepage. We will start from acquiring basic data and analyze user behaviors, with the help of crawler's specific algorithms and logics.

Interest focused as it may be, Douban is still a social network with tens of million registered users and vast information. So we mainly adopt the breadth-first strategy, with indirect assistance in analyzing and de-emphasizing of URLs by focused crawler. This project consists of generally 2 phases. Phase one is to uncover the features of Douban by searching and crawling information on the key contents of Douban - books, movies, etc.; Phase two is to narrow down to the users sharing the same or similar interests by using clustering analysis, and to grasp the inner-relationship among popular figures.

Same as every other social network system, Douban is anti-crawler enabled, based on user info protection protocol. Thus, during the crawling process we encountered some challenges. For example, as the crawl reached one thousand users, the webpage returned Error 403. To solve this, we set up user agent and time-sleep in the crawler system. One other issue is that many of the users' pages are limited for browsing and would require simulated login for viewing authority.

This project report is composed by six parts. Part 1 states the background. Part 2 introduces the design for the crawling process. Part 3 is the course of crawling followed by the analyses of results in part 4. Part 5 is Challenges Encountered in the Project and Solutions. The findings of the entire project is concluded in part 6.

# 2 Background

## 2.1 Related Work

In this project we use Python to create the web crawler owing to its following features. First, Python is multi-threaded and has mature process models. Crawler is a typical multi-task architecture and so time could be delayed when requesting page. Multi-thread and multi-process will optimize system efficiency and improve its capability of downloading and analyzing. Second, Python has light languages and strong crawler libraries. Scripting language is dominant language in many aspects when developing a small-sized crawler system[2].

The targets for this project are relatively limited in quantity, including about 10,000 users, top 250 movies and about 10,000 published books, and it would be feasible for Rstudio to cope with and visualize downloaded info. In the meantime, with the use of focused crawler, which is different from general crawlers, we can stay relatively focused onto webpages relating to certain subjects or keywords. The key-word library defined in the system of the focused crawler would filter irrelevant information so that it would trace and expand on the wanted information to discover inner relationships targeted by this project.

## 2.2 Analysis on Douban OSNs

In the network topology, we may translate the relational structure of the Douban network into a diagraph G=(V, E), where V represents every vertex and E represents every edge. It can be learnt that a vertex is symbolic of a user, and an edge is symbolic of the relationship that user has with another user. Take Twitter or Chinese Weibo for example, there are basically three types of relationships between two different users. This is to say, a user can either have no relationship with another user, a "friend" relationship in one way or both way.

As is known to us all, the Six Degrees of Separation indicates everyone and everything is six or fewer steps away, by way of introduction, from any other person in the world, so that a chain of "a friend of a friend" statements can be made to connect any two people in a maximum of six steps[3].

In this project, we start from user V1, the source node using Breadth First Search strategy, on to the users that have direct connections with V1, or "degree-one connection", and then "degree-two connection". The degree of intimacy of the two vertexes, or users, could be learnt by calculating the edge. The edge between the vertexes can be calculated by taking into account factors like whether they are both way friends, the similarity of their interests, etc., thus could lead to the degree of relationship. In addition, the intimacy of degree-two connections are often under the influence of the first degree.

# 3 Design for the Crawling Process

## 3.1 Crawling Douban Books and Top 250 Movies

### 3.1.1 Book Crawling

Figures on alexa.com shows that users search on Douban mostly for books and movies[4]. In order to learn the categories of books searched the most, we need to first categorize

the books and acquire the reviewing marks on them. We started by crawling books under the Computer category labelled "machine learning", "Linux", "android", "database", "data structure", "algorithm" for the quantity and reviewing marks. The information then would be stored in Excel worksheets using libraries of NumPy and Openpyxl of Python. Then, we performed the same process for books categorized as "literature", "culture", "life", "economic", "management" and "technology" by Douban.

During the first attempt when the count of crawling reached approximately one thousand, the Douban website returned Error 403 owing to its anti-crawler mechanisms. In order to solve this, we camouflaged the crawler into an internet browser by setting up user agent. We set the request frequency to every 0.1 seconds. By doing so, we have avoided Error 403 in the following attempts.

Through the final analysis on the result of the book crawling, we can learn the No.1 label of books under the Computer category that is the largest in number and highest in terms of reviewing marks. Similarly, the No.1 category of books that is the largest in number and most highly reviewed.

### 3.1.2 Top 250 Movie Crawling

The Crawl of the top 250 movies was performed in the same way as the crawl of books. By sorting out the interpretation of the information on the html, we can extract relevant information by converting the title, ranking, genre and the number of reviews of the movie into regular expressions. Through the analysis on top 250 movies we can find out which movie(s) receives the most audience on Douban.

The results of book and movie crawling demonstrate preferences of the target users as well as the features, or distinctiveness of Douban.

## 3.2 Crawling OSNs

### 3.2.1 Targeting User Information

Prior to the crawl of Douban users we need to first target what kind of information is relevant. In this project we chose the users' profile pictures, locations and movies and books they have seen and read as targets to compare with each other. At the same time, we also target certain individual users for their degree-one and degree-two relationships to study the degree of intimacy between one another.

### 3.2.2 Choosing Initial Node

To start with, I chose myself as the initial node, as my friend list is a combination of popular figures and acquaintances or real-life friends. One simple way to obtain user list and target information is by clicking on the users from my interest list, getting their user IDs, downloading information on their homepages and tracking the users on their interest lists. This is a continuous process and we might even get to every user on this network, as indicated by the Six Degree of Separation. So, as a matter of capacity and time, we limit our search to degree-two relationships.

The search of the second degree of users is based on breadth first strategy. Figure 1 illustrates the principle of access of BFS, which is, choose an initial node, push it onto the stack, and each time the stack pops out an element, search every elements on the next level,

then push them onto the stack[5]. Also we need to mark the elements as the precursors of those on the next level. The systems stops once we find the target elements, or if the whole search ends up empty-handed.

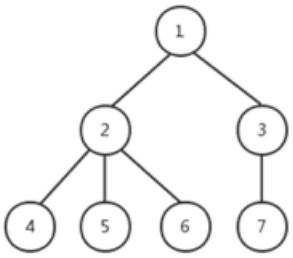


Figure 1: Order of nodes to be visited by the crawler

### 3.2.3 Simulating Login

Because of the limitation on user privacy, user related info is only accessible in a logged-in status. Hence the first task for our crawler is to login. Through the code for its login process we know Douban takes three steps to authenticate its users. First, login requests from users will be sent to the servers. Second, servers receive these requests and generate login keys respectively and return to users. Finally, users submit their user IDs together with login keys to the server, and in return they receive verification codes to complete the process. When users successfully login onto the website, servers will return the current login status and correct user information. In view of the above, we used Urllib modules to obtain url, and cookies to encrypt the requests of user IDs and login keys.

### 3.2.4 Crawling Process

When simulated login is complete, the crawler will start crawling target information. Firstly, crawler enters the friend list of the initial node to extract the target users from parsed HTML source codes. The friend list will be downloaded as txt. file in order to be imported to Rstudio. Then, crawler searches the location, movie and book preferences of the users that the initial node takes interest in. Information will also be stored as txt. Files. This process will be repeated on the degree-one nodes to acquire their friend lists and other relevant info.

The above process has in fact treated the initial interface and the directly linked interfaces as father node and child nodes, with subsequent links being the child nodes of previous links. The crawler will maintain two queues during the crawling process, one being the queue for completed searches, the other being the queue for searches in progress. Under the original status, the queue for completed searches is empty and the queue for searches in progress contains the initial node only. When the search on all of the target user's information is finished, that user will be queued up in the queue for completed searches. In circumstances where new users are encountered during the search on a certain user, crawler will run the user's ID through the completed and in-progress queues to check if searches on the same user has been performed. If not, then this new user will be queued up at the bottom of the in-progress queue[6]. The search will be terminated when the maximum capacity of the completed queue is reached, rather than searches on all users is done.

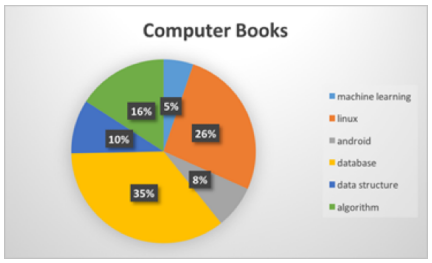


Figure 2: Computer Books

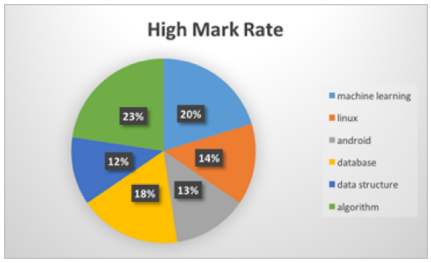


Figure 3: High Mark Rate

# 4 Analyzing the Results

## 4.1 Results of Book Crawling

The crawling of the category of Computer Books on Douban has acquired the following information.

Books labelled Machine Learning totaled 645 in volume, and 129 among which received reviewing marks above 9, taking up 20% of the entire label; Books labelled Linux totaled 3,255 in volume, and 455 among which received reviewing marks above 9, taking up 14% of the entire label; Books labelled Android totaled 930 in volume, and 121 among which received reviewing marks above 9, taking up 13% of the entire label; Books labelled Database totaled 4,350 in volume, and 783 among which received reviewing marks above 9, taking up 18% of the entire label; Books labelled Data Structure totaled 1,170 in volume, and 140 among which received reviewing marks above 9, taking up 12% of the entire label; Books labelled Algorithm totaled 1,950 in volume, and 499 among which received reviewing marks above 9, taking up 23% of the entire label.

Based on these findings, we can generate two pie charts in figure 2 and figure 3.

A rough conclusion could thus be made that the book on database is the largest in volume yet generally not highly reviewed. On the other hand, the book on machine learning may be the smallest in volume but achieved 20% in terms of good remarks. The rest of labels did not show much significant discrepancy in the reviewing marks.

Similarly, a pie chart as below shows the results of the crawling of five categories of books in Douban.

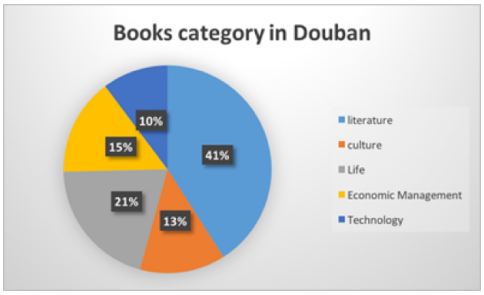


Figure 4: Book Category in Douban

From figure 4 we find that literature books take up nearly half of the entire collection of Douban while technology books account for the smallest proportion, which could be attributed to the strategy of marketing of Douban as it targeted the minorities for whom

cultural consumption plays a major role. Therefore, Douban is actually an "literature's website".

## 4.2 Results of Top 250 Movies Crawling

Through the search and categorization of Douban Top 250 movies, we discovered the fact that feature movies receive the most audience, and the romance movie and crime movie seemed to be the very distant second and third. But we should note that one movie could be probably classified into several genres. For example, the Top 1 movie, the Shawshank Redemption is classified as "feature/crime". Hence, the figure 5 generated from these findings could reflect the popularity of feature movies in a relative way as most movies are all tale-telling and can be classified as feature movies basically.

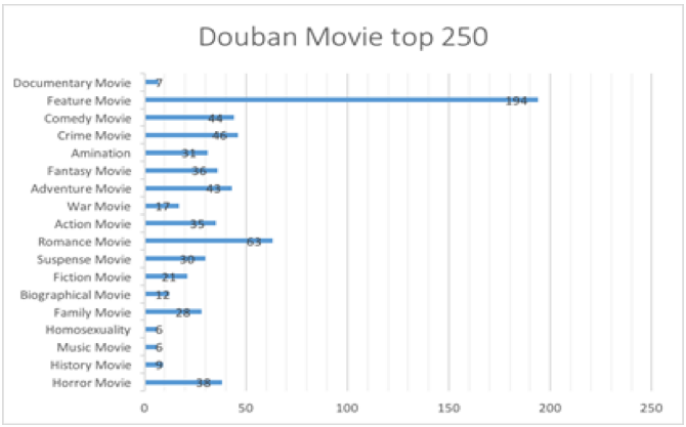


Figure 5: Douban Movie Top 250

## 4.3 Analysis of Crawling Results of OSNs

Based on the results provided by our crawler, we employed Rstudio to analyze the data and generate network diagrams of user relationships. Because the search is originated from the friend list of the author and expanded to the network among the friends' friends, a network diagram will more explicitly demonstrate the form of this relationship network. First, a network diagram can be drawn based on the degree-one relationship screened from the data sets, which consists of all the users directly linked to September, the source node. From figure 6 we can also see the types of relationship by the directions of arrows.

As it shows, we refer to the 5 users that are in mutual relationships with September as Friends, and the rest as Targets. Next, a more complex version of this diagram can be drawn by setting the size and color of the nodes judging by the number of friends and targets of the captured users, on the second degree.

The colored nodes in figure 7 represent the degree-one nodes with the initial node in the middle. Nodes with more friends and targets are bigger in the size of the circles and represent users that are more active. In Rstudio we can also separate the circles of each node by unifying the node and its circle into the same color, as illustrated in figure 8.

From the colored diagram we can discover the fact that the branches of the relationship networks originated from one node are linked to each other. In other words, intersections

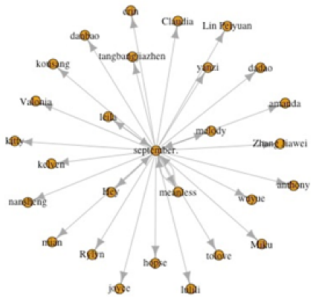

Figure 6: Relationship of the Source Node

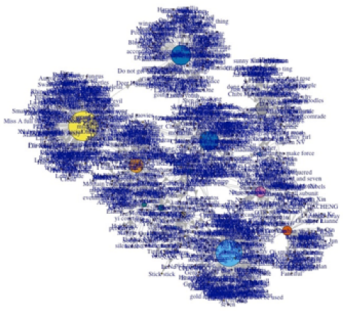

Figure 7: Degree-two Connection

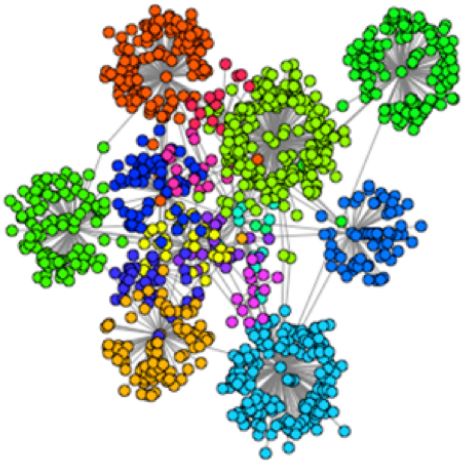

Figure 8: Separate the Circles of Each Node

exist among different relationship networks with a common source node. Now, the question is that who among these users play the role of an intermediate degree that links different networks together. Based on the premise that the bonding function of an individual user correlates positively with his/her intermediate degree, we define the intermediate degree: $\sum(\frac{g_{ivj}}{g_{ij}}$, i≠j, i≠v, j≠v), where $g_{ij}$ represents the number of shortcut between nodes i and j, and $g_{ivj}$ represents the number of shortcuts between nodes i and j that pass the node v (shortcut is the shortest path between two nodes)[7]. By using function betweenness(), we calculated the intermediate degree of each node and came to a scatter diagram as follows, where the vertical axis represents the quantity value of the intermediate degree.

Suppose users with intermediate degree above 150,000 are major bonding media, we used V(g)[bte≥150000] to identify them and tracked four users. These four users are September (the author), Wuyue, Anthony and Mian. Except for the author, the other three users are all popular figures on Douban and are all degree-one targets of the author. Therefore, apart from the source node, the intermediary role is mostly played by the popular figures on the network, which is not difficult to understand because the popular figures can be friends to each other and targets of normal users at the same time.

In the following paragraphs we introduce the results of the analysis on similarities between the source node and other users. There are totally 8,890 users whose information was downloaded by the crawler. Firstly we can locate theses users by drawing a diagram of

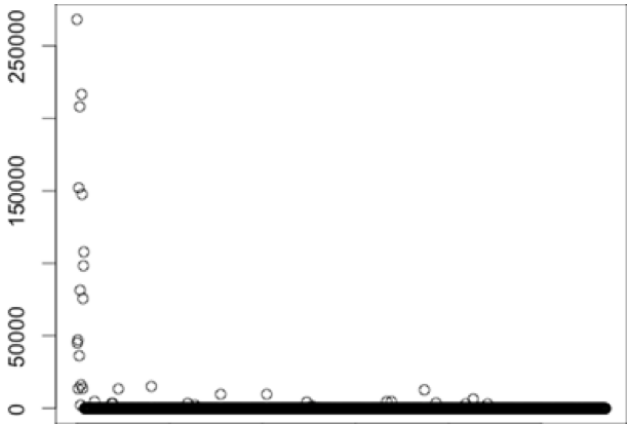


Figure 9: Scatter Diagram of Intermediate Degree

geographical position.

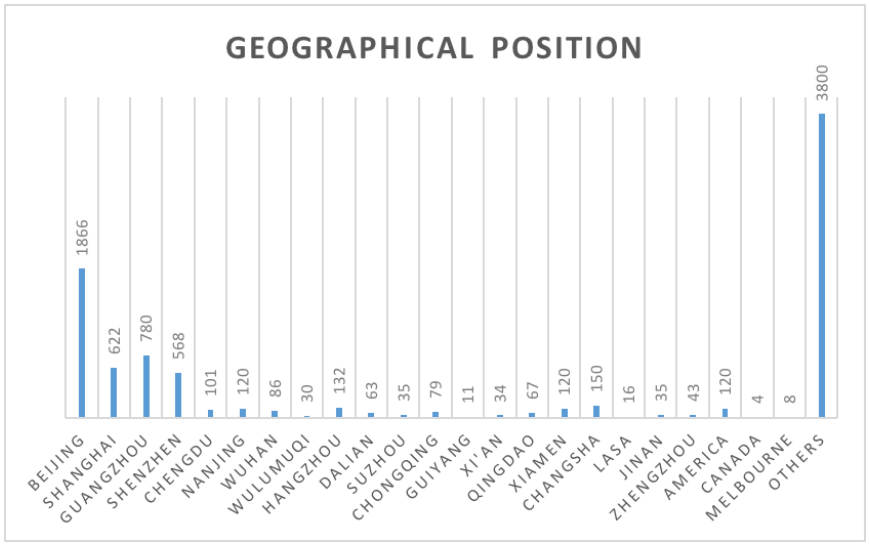


Figure 10: Geographical Position of the Users

From the geographical positions of the users we can see there are 780 users located in Guangzhou, which is the same location as the author, and Beijing ranks first with 1,866 of user volume. The next question is who among these users share the most interests with the author. The finding of the similarity of interests was limited within the area of books and movies. Although Douban has offered a "Common Likes" feature that tells the number of common interests between one user and another, the number cannot be divided into specific area of interests. And more often what we would like to find out is that when we share the same reading preferences with A, what would our music and movie tastes look like? Starting from the source node, there are 26 books and 78 movies on the author's interest list. We then sorted each person's preferences on books and movies with Rstudio and compared with the author in terms of the title of the movies and books. The comparing process continued until the users are ranked in order of similarity.

The result shows that in terms of movies, the author shares the most similar tastes (33 commonly enjoyed movies) with Mian from Beijing, on the degree-one level of the author's

relationship network. On the other hand, in terms of books, the author shares the most similar tastes (8 commonly enjoyed books) with Spur also from Beijing, but on the second degree of the author's relationship network (with the intermediate degree being Zhang Jiawei).

# 5 Challenges Encountered in the Project and Solutions

## 5.1 Duplication of User ID

Because Douban allows multiple occupation of the same user ID, the problem of ID duplication was discovered when conducting analysis on user lists and thus might cause error in the statistical results.

## 5.2 False Geographical Position

When sorting the downloaded data we found a number of users did not disclose their locations. And because Douban does not verify a user's real location, there might also be a number of false information that may cause statistical error.

## 5.3 Incorrectly Formatted Characters

Because the user ID on Douban does not necessarily have to be letters or characters, there are plenty of special customized symbols or signs. Firstly these symbols and signs could not successfully write, so we then replaced them with regular characters.

## 5.4 Large Intersection with Popular Figures

In the analysis on similarity of interests, we realized that due to the large base number of movie and book collection of the popular figures, the intersection could be significantly large. Even if for some popular figures with moderate intersections, they would still rank high on the comparing list. For example, the Target user of the author, Mian, is a famous film critic and has seen 5,262 movies. Therefore, we could use weight allocation to reduce the centrality.

## 5.5 Error 403

Due to the sensitivity of both the anti-crawler mechanisms and API interface of Douban, even if this project did not aim at large data volume, the website returned Error 403 from time to time. We made modifications to our searching process such as browser simulation and time interval setups, but did not thoroughly solve Error 403 or rejection of connection by remote host.

# 6 Conclusion

In this project we performed crawling and analysis on basic data and information. This project allows us to see that on the one hand, Douban is a website with specific position and target user groups. People from a certain age group in China tend to enjoy the cultural

interaction on this website, such as the author. Douban's success lie in the good user experience in the design of both the page layouts and the contents. Based on the unique "Interest Marketing" strategy, Douban has become a platform for people to socialize with the oxidant of interests, creating a highly adherent user group. However we might also imagine the possible future paths for such websites with strong user viscosity to attract new, multilayered users. On the other hand, similar to all the other social networks, Douban is actually a huge net of relationship. Admittedly this net is weaved densely in some parts and sparsely in other parts, but if we look inside, connections can always be found, and there is rarely any group existing on a stand-alone basis. The coincidence of the sociability and homogeneity of Douban users, therefore, is relatively high to some extent. In the meantime, we can have more features like friend recommendation based on the similarity of interests or relationship connections.

This project has encountered some challenges and could be improved in many aspects. In times of big data, the acquisition and analysis of data has become a key factor in decision making. Through this project of retrieving and analyzing data, I have had a better understanding of data mining and hope to prepare for more sophisticated higher dimensional data mining. Also, the study of social networks has been of great education since through the statistical analysis we are actually approaching the nature of human behaviors, which in essence is the study of human beings themselves.

# References


[1] C. Haythornthwaite L. Garton and B.Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3, 1997.

[2] Allen Downey. *Think Python*. Green Tea Press, 2003.

[3] *Six degrees of separation' theory tested on Facebook*. Telegraph, 2012.

[4] http://www.alexa.com/siteinfo/douban.com.

[5] Mark Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th international conference on World Wide Web*, pages 114–118, 2001.

[6] Mike Thelwall. Results from a web impact factor crawler. *Journal of Documentation*, pages 177–191, 2001.

[7] J. Lang S. Ye and F. Wu. Crawling online social graphs. In *Proceedings of the 12th International Asia-Pacific Web Conference, IEEE*, pages 236–242, 2010.