Venture Capital Recommendation System
SI650 Final Report

Tong Yin
tongyin@umich.edu

Mei Fu
meif@umich.edu

Hui(Phoebe) Liang
lianghui@umich.edu

ABSTRACT

How to find the investment companies is always a crucial problem for a start-up. In this paper, we use PageRank model and collaborative filtering model to retrieve which companies are possible to invest. Our result shows that unlike Internet network, it is hard to rank the companies based only on the network information itself. Collaborative filtering could do a nice job instead.

KEYWORDS

Information Retrieval, PageRank, Collaborative filtering.

1 INTRODUCTION AND RELATED WORKS

In the United States, the average number of new businesses established each year is approximately 500,000 ([1]Long 2016). However, less than 0.3% of new companies receive their venture capital deals annually ([2]Carey 2014). Even before a startup applies for venture capital funding, much preparation work is required, and it can take months to years because a startup company needs to prove that it's profitable and scalable.

While the chance of a startup company securing investment from venture capitals is slim, it can be greatly increased if a startup company is personally connected to venture capitalists early on. According to a research study, it is also important to do research on the venture capitalists and understand their funding policies in the sector(s) which the startup company is in ([3]Hall and Woodward 2010). Additionally, there is not an existing website to provide customized recommendations to startup companies although some online databases provide information about past investment history of venture capitalists and their investees. The information is neither personalized nor publicly available. Therefore, we are interested in how to identify and recommend potential investors to startup companies based on their own funding need and attributes.

1.1 RESEARCH QUESTIONS

In particular, we want to utilize networks' concepts and theories to identify and rank the most likely investors. The research questions we are trying to answer are as follows:

1. What are the current characteristics and interesting observations of the startup-investor network?
2. What kind of attributes of a startup company or an investor can help predict the most likely investors?
3. What are the different approaches available for us to develop an effective ranking algorithm based on the limited information we have in the current dataset?
4. What are the most influential investors that are likely to invest in any startup company when we don't have enough information about a startup company?

1.2 RELATED WORK

The relationship between investors and startups is always a hot topic in both academia and industry. Although most of the papers are focused on the U.S. market, it is reasonable to assume that similar conclusions are valid in lot of different areas.

William found that venture capital firms are likely to link with each other to share information and spread financial risk([4]Bygrave, W. D. 1988).The higher investing risk, the more compact investing network would be. Also, the information flow within joint invests is swift and investing companies want as many links as possible to other investing companies. This is in align with our assumption that the edge built through joint investing could facility network grow.

Yael H. finds out that if an investing company is within a better network structure, it is more likely to performance better([5]Hochberg, Ljungqvist, 2007). On the other hand, the portfolio start-ups connecting with better-networked VCs are more likely to live longer. This finding is a strong support for our method of understanding the start-ups and investing relationship in network settings.

Our basic algorithms come from Google's PageRank([6]Page, Motwani,1999). When a node points to another, this connection is regarded as recommendation or voting. The nodes that many other nodes point to should value more. Starting from any arbitrary initial value, we could result in a stable ranking result through finite iteration. This method could be applied to our network structures. When an investing company chooses to support one specific start-up, it is a vote for that start-up and vice versa.

There is also a paper from Qiaozhu M. about how to use hitting time to find similar queries based on a query input([7]Mei, Guo, & Radev 2010). Based on the bipartite graph of query and URL, they firstly find the subgraph of the whole giant graph using depth-first search. Then the transition probabilities between queries are defined through the common URL, weighted by the co-occurrence of query and URL tuples. This paper inspires us to use subgraph and transition probability to find possible investors.

Qiaozhu also proposed a ranking algorithm called DivRank for nodes prestige([8]Mei, Zhou, & Church, 2008) The PageRank assumes that nodes referred by many others should be more prestige. Also, an important node should contribute more compared with many unimportant ones. The original PageRank is based on the random walk assumption, in which the transition probability matrix remains fixed the whole time. Compared with random walk, DivRank is a reinforced random walk which takes historical visiting into consideration. In this way, the investing companies gain more prestige when investing more companies and this give them more exposure to more investing opportunities in the future.

2 METHOD

Simply saying, our main goal is to build a recommendation model for startup companies to know which investors are their potential angels. We tried to solve this problem by the following approaches: PageRank model, and collaborative filtering. Given a specific companies, we use the PageRank to measure how the startup and the VC are directly matched as an information retrieval task. We also measure how many similar startups have received funding from VC as an collaborative filtering.

2.1 DATA COLLECTION AND EXTRACTION

The dataset currently used in our project contains historical investment history, which includes acquisitions, initial public offering, venture capital, and exit data of each startup company. This dataset is originally from crunchbase.com, a website that

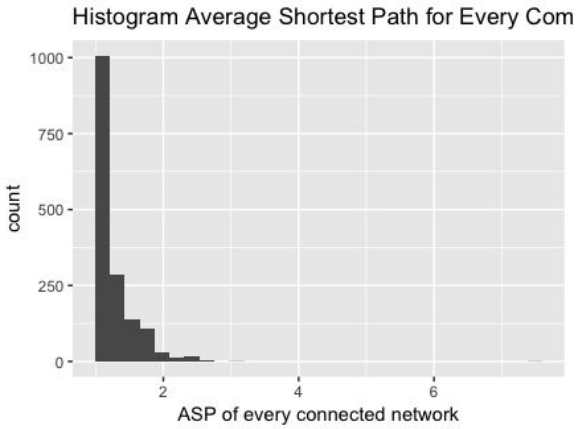
provides information on both startup and Fortune 500 companies. Since we do not have the subscription to the database, we searched and downloaded a public dataset on Github that's available for non-commercial use (https://github.com/notpeter/crunchbase-data), which includes transaction data from 1977 to 2015 for both U.S. and foreign companies and investors. The files used in our analysis are investment.csv, acquisition.csv, and companies.csv. They are all stored in csv format. For the scope of the project, we also pre-processed the data and combined the investment.csv and acquisition.csv. A summary of the dataset is shown below:
Data Size: 168,648 rows. If we filtered out locations outside of the U.S., there are 82,919 rows; Countries: 126. Regions: 984. Cities: 3,638; 18 variables (total)

To provide more details, each company can fall into multiple categories. Each startup company may have several investors and even more than one investment from the same investor. Lastly, an investor may be a startup company itself. On average, an investor invests in 3.34 startup companies yearly, and a startup company has 2.26 investors.

2.2 PAGERANK ALGORITHM

At first for each given start-up, we tried to return the investors with top Pagerank scores based on the start-ups all reachable nodes. It didn't work out for there is a giant connected component for the whole graph.

Inspired by the idea from Dr. Mei's hitting time, we decide to only calculate the PageRank scores for a reasonable smaller subgroup. It is reasonable if we consider the real word instances. It is possible for a start-up to reach all its reachable node but it is more pervasive to assume that a start-up only could reach limited nodes with time and budget constraints.



So, next question is how to choose the optimal subgraph. Above graph is the histogram for an average shortest path for each connected graph. We could tell from this graph that majority connected component has an average shortest path below 3. This shortest path is calculated based on the whole graph, but the way in which we construct our network implies that all the edges are between start-up and investors.
Steps:

1. For each given node N0, find all its investors List1

2. For each investing company found in List1, find all start-ups it has invested in and mark the whole start-ups set as list2.

Step 3. For each start-up in list2 from step2, find all its investing companies and record them as list3.

Step 4. Find the subgroup only contains the original nodes N0, list1, list2 and list3 and run the Pagerank.

Since we saw that most of the average shortest path is under 3, the above steps are actually is all the nodes a start-up could reach within 3 steps.

2.3 COLLABORATIVE FILTERING

Because of the nature of the project, our research questions can be treated as collaborative filtering,

which also referred as recommendation systems. Traditionally, the approach of collaborative filtering is to compute the similarity between users and rank the users based on the similarity to a particular user, and then make a recommendation to this particular user. The intuition is that a user will like a product/service if his or her similar users also like it.

However, there are several main challenges we face in developing an appropriate and effective algorithm for the collaborative filtering task. First, the information we have about each individual company and investor is limited. For example, we have only information about the location, investment date, funding amount, funding round of each startup, and its investor. Secondly, each startup company usually has only a couple transactions in the dataset. The majority of the companies have two or three records. The total number of unduplicated investors is slightly higher than the number of unduplicated companies. Thirdly, direct investor rating information is not available to us, considering this is a collaborative filtering problem.

Although the challenges are nontrivial, we have found ways to cope the challenges and achieved considerably high accuracy rate, given the drawback of the dataset. In developing an algorithm to identify potential investors, we adopt the ranking-based approach and retrieve the most likely investors.

Before developing the algorithm, we first created two relatively simple baseline models to use as benchmark. The first baseline model built is solely based on how similar of two startup companies. To do this, we compare the categories, state, location, funding round, and funding need (raised_amount_usd).

An simple example in Table 1

	Company A	Company B	Match_Ratio
Category	Crowdfunding |Finance	Crowdfunding|Games| Non Profit	1/3
State	CA	CA	1
Region	Los Angeles	SF Bay Area	0
Fund Type	Angel | Venture A	Seed | Angel	1/2
Yearly funding amt	1M-10M	100K-1M	0
Investors	500 Startups, Tim Draper, K5 ventures	500 Startups, New Associates	1/4
CF Score			2.08

For each startup company, we calculate the match_ratio for each existing attribute in the dataset. As demonstrated above, a company A has one category overlapped with company B's categories. The match_ratio is then the count of overlapped categories divided by the total number of company B's categories. The same calculation apply to all the attributes. The sum of the match_ratio is the match_score. After finding similar companies, we retrieved all the similar companies' investors.

The second baseline model we built is based on bipartite graph, which edges are represented by two groups of nodes, one for startup companies, and the

other one for investors. Then we make a projected one-node network for each group. For each startup company, we identify its first-degree neighbors in the network and find all of the neighbors' direct investors.

To evaluate these two baseline models, we used 2014's data for training and 2015's data for testing. We then test the companies, which exist in both years and have investment transactions, by measuring the intersection of investors retrieved from the baseline models and the actual investors in 2015. Overall, the accuracy of the simple similarity-based model is 27.9% on average, which means we were able to identify at least one investor in 2015 for almost one third of the companies. On the other hand, using only information from the bipartite graph, the accuracy is about 10.8%.

To increase the accuracy rate, we develop a hybrid approach that combines both similarity and the first-degree connected neighbors in the projected graph. Additionally, we take into consideration of whether an investor has invested in the startup company before. From our initial analysis, we realize an investor is likely to invest in the same startup company in the same year or the following year. We also rank and return the top 10 investors of each startup. As a result, we create with the following algorithm:

Steps:

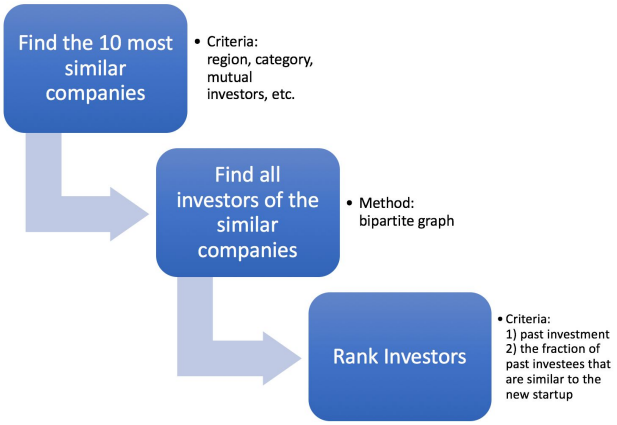
1. For each startup company, find the top 10 similar companies based on cf score of each existing attribute. We also calculate and add to the cf score the ratio of mutual investors over the total number of investors of both companies being compared. However, the investor ratio is only added if the company in comparison is a first-degree neighbor in the one-node projected graph. If not, we don't add the score. A startup company may be an investor itself. To handle this type of companies, we treat it as two nodes in both sets.

2. Find all the investors that are one degree away from these 10 similar companies in the bipartite graph.

3. Rank the investors using the combination of two scores below:

a) past_investor_score: whether it invested in the startup company before. If it is a past investor of the startup company, we calculate the ratio by dividing 1 by the total number of the testing company's past investors in 2014.

b) sum the similarity score of each investee that that investor has directly invested and also in the top 10 similar companies.



Before we come up with this final ranking algorithm, we also tried other methods such as directly comparing the similarity between an investor and a startup company and see whether we would be able to achieve reasonable accuracy. However, this method performs the worst out of all the methods we tried. This suggests that an investor's attributes are not as important when recommending investors to a startup. Additionally, we tried to weight the different attributes when calculating similarity between any

two startup companies, but this process is heuristic rather than systematic because of two reasons. First, there is not a good method to determine the weight and the number of transactions is limited for each startup that can be used to train any model. Lastly, this approach does not improve the performance either. For this exercise, we also tried including 2013 in the training data, but the result does not vary much compared with only 2014 data in the training data.

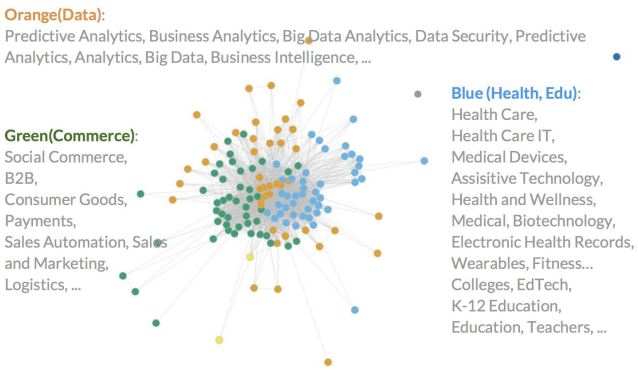
2.4 Query Suggestion

There are mainly 6 important variables in our collaborative filtering matrix. From the actual usage perspective, variables like region, funding type, funding amount, etc. can be easily inputted or selected to match with the data in our dataset. However, there exist many diverse categories in our dataset, and it is hard to let users choose all the categories in our dataset that match with their company.

To solve this, we first tried k-means clustering method to find categories related to user's first input. However, there are too many overlaps among the clusters so the result is not very good.

```
Top terms per cluster:
Cluster 0: mobile media social commerce advertising analytics web marketing internet curated
Cluster 1: technology clean software information services web energy finance analytics mobile
Cluster 2: biotechnology health care software medical diagnostics pharmaceuticals hardware technology clean
Cluster 3: software enterprise hardware mobile analytics security saas web data technology
Cluster 4: health care wellness biotechnology medical software technology devices mobile diagnostics
Cluster 5: web curated media_social hosting commerce advertising analytics marketing software
```

We then tried network technique to cluster the categories, which is to maximize the modularity score for different communities([9]Clauset, Newman, Moore, 2004). We first constructed a bipartite network graph between investors and the categories they invested. Then the category network was constructed upon this bipartite graph. We used a greedy algorithm to find subcommunities in the category graph by maximizing the modularity. The higher modularity score means the links within the subcommunity is denser than its random graph. The final modularity score is around 0.14 which is not quite high, because some investors invest in multiple industries. However, the result successfully clustered similar categories together, and also the potentially related categories that may fit a cross-industry company.



3 RESULTS AND EVALUATION

3.1 PAGERANK MODEL

There are close to 2000 start-ups appears both in 2014 and 2015 so our evaluation is based on these entries. We could see from the graph that on average the modified PageRank perform not bad with mean accuracy precision of 0.3206029.

On the other hand, if we remove the investing companies who invested in the same start-up both in 2014 and 2015 in our forecast, the average correct rate drops to 0.01046548. Our findings support the beliefs that a startup is likely to gain money from its old investors friends. This is in accordance with the findings from random walk approach.

3.2 COLLABORATIVE FILTERING

For the testing dataset: we used U.S. companies that occur in the transaction data in both years 2014 and 2015. The size of the testing companies is 1732. Since each startup has an average of only 2.5 investors in a year, we consider it a success if at least one out of the top 10 ranked investors is the actual investor in 2015.

Average precision: 8.3%
Precision = retrieved and relevant investors / total number of retrieved investors

Average precision = sum of all precision scores / total number of testing companies

*Relevant investors are investors that actually invested in the testing company in 2015.

Average recall: 28.9%
Recall = retrieved and relevant companies/ total number of relevant investors
Average recall = sum of all recall scores / total number of testing companies

*Accuracy: 51.3%
We defined accuracy differently than a regular ranking measure. Since each startup is funded by an average of 2.5 investors in a year, we consider it a success if at least one out of the top 10 ranked investors is an actual investor in 2015. To calculate the accuracy, we count the number of startups that we retrieved at least one investor that actually invested in 2015. Next, we divide the count by the total number of testing companies.

4 DISCUSSIONS AND CONCLUSIONS

In this project we used 2 algorithms to recommend startups with the top 10 investors with highest investment likelihood, and the accuracy is about 51.3%. It is hard to say who would be the most possible to invest based on only the network relationship. We also have a main takeaway is that investors like to invest in similar companies and re-invest in past investees in the following year.

We use the PageRank also to rank the possible investing companies using NetworkX package. The default damp factor is 0.85. We may explore different damp factors in the future for the specific network. Another improvement is that we only use the data from 2014 to train our model for the forecast because of time constraint. It is possible to use all the data we have from 2009 with some time adjustment factor. It is also possible to take into consideration of the weight, i.e, how many times one specific company invested in a startup if we expand our training dataset. Also, we use a similar idea of getting subgraph in the paper "hitting time", it is possible for us to define our own possibility and run iteration until converge.

A.1 Introduction and Related Works

A.1.1 Research Questions

A.1.2 Related Works

A.2 Method

A.2.1 Data Collection and Extraction

A.2.2 PageRank Algorithm

A.2.3 Collaborative Filtering

A.2.4 Query Suggestions

A.3 Results and Evaluation

A.3.1 PageRank Model

A.3.2 Collaborative Filtering

A.4 Discussions and Conclusions

A.5 References

ACKNOWLEDGMENTS

This work is for the SI650. We thank Qiaozhu Mei for his excellent guidance this whole semester and Ai Wei for his useful feedback.

REFERENCES

[1]Bygrave, W. D. (1988). The structure of the investment networks of venture capital firms. Journal of Business Venturing, 3(2), 137-157.

[2]Carey, R. (2014, June 25). The payoff and probability of obtaining venture capital - 80,000 Hours. Retrieved April 17, 2017, from https://80000hours.org/2014/06/the-payoff-and-probability-of-obtaining-venture-capital/#fn-14061-10

[3]Hall, R. E., & Woodward, S. E. (2010). The Burden of the Nondiversifiable Risk of Entrepreneurship. The American Economic Review, 100(3), 1163–1194.

[4]Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. The Journal of Finance, 62(1), 251-301.

[5]Long, H. (2016, September 8). U.S. startups near a 40-year low. Retrieved April 17, 2017, from http://money.cnn.com/2016/09/08/news/economy/us-startups-near-40-year-low/index.html

[6]Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.

[7]Mei, Q., Guo, J., & Radev, D. (2010, July). Divrank: the interplay of prestige and diversity in information networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1009-1018). Acm.

[8]Mei, Q., Zhou, D., & Church, K. (2008, October). Query suggestion using hitting time. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 469-478). ACM.

[9]Clauset A, Newman ME, & Moore C. (2004, December). Finding community structure in very large networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics E70,066111