Recommendation System for Venture Capital Investors

1st Akshata Deo
Department of Computer Engineering
San Jose State University
Student-Id : 012565761
San Jose, CA

2nd Nivetha Jayakumar
Department of Computer Engineering
San Jose State University
Student-Id : 013758667
San Jose, CA

3rd Keerthi Akella
Department of Software Engineering
San Jose State University
Student-Id : 013858819
San Jose, CA

4th Uma Nataraj
Department of Computer Engineering
San Jose State University
Student-Id : 013727259
San Jose, CA

5th Sibirajan Sadhasivam
Department of Computer Engineering
San Jose State University
Student-Id : 013785369
San Jose, CA

***Abstract*—An efficient investment decision is a highly crucial phase for investors. A lot of factors should be kept in mind while making such decisions. Generally, characteristics of the entrepreneur, dedication, market value for their idea and strategies behind planning so that the investment can result in maximum gain. Risk factors should be carefully analyzed while making investments: for a potential startup, a VC needs to specifically estimate how well this new investment can fit into its holding investment portfolio in such a way that investment risk can be hedged. Second, The investment behaviors are not dense enough than conventional recommendation applications and a VCs investments are usually restricted to a countable industry categories, making it impossible to use a topic-diversification method to hedge the risk. The main stages of an investors decision making process involve deal origination, screening, evaluation, structuring and post investment activities. In this paper we have addressed these challenges using collaborative and content based filtering recommendation systems. The goal of this project is to provide analytical guidance to venture capital investors by using recommendation systems.**

***Keywords*—** Collaborative filtering, Content based filtering, KNN, Recommender Systems, Venture Capitalists, Investors

I. Introduction

The huge amount involved in business start-up costs; founders are always concerned with financing their venture. This is one among the most difficult challenge faced by the founder. There are many ways to finance a start-up or small businesses, such as, crowd funding, angel investors etc. This paper represents the solution using recommendation system for venture capital investors[1]. The start-up companies work on various domains. Start-up companies is not restricted to tech industry alone, but it will also explore other fields such as pharmaceuticals, agriculture etc. For the investors to pick the right start-up to invest various recommendation approach will be useful.

Recommender systems has become an integral part of daily lives. A person will use recommendation system in many aspects of everyday routine such as watching TV shows, shopping, social networking websites, reading books etc. The necessity of this recommendation system is increasing day by day with the number of increases in media content on internet and TV is increased[6]. The amount of information that we retrieve, and use has rapidly increased[2]. Data mining is an important aspect of it which helps us retrieve relevant data from a huge pile of data available. Basically, there are three important types of recommendation systems namely, Collaborative filtering, Content based filtering and Knowledge based filtering which helps in predicting the user preferences, thereby helps in recommendation.

As the title of our paper suggests, this paper recommends start-up companies for venture capital investors. By definition, Venture capital is financing that investors provide to start-up companies and small businesses that are believed to have long-term growth potential. Venture capitalists are well off investors who will finance the start-ups or small businesses. It is vital for the venture capitalists to mine data about the companies and strategies to choose companies for financing. Venture capital is one of the ways of funding a start-up or a small business[2][3]. Venture capital has given rise to thousands of startups in the recent years. It is also observed that start-ups financed by venture capitalists have much lower failure rate compared to those companies financed by other means. The goal of this project is to provide analytical guidance to venture capital investors by using recommender systems.

The goal of our project is achieved by exploring various approach to recommendation system. In this project we have

approached collaborative and content-based filtering methods. All methods in data preparation is being followed in this project. Data exploration is accomplished by collecting and analyzing data from mattermark and crunchbase website[1]. Data cleaning is done by removing missing value rows. Hence data pre-processing is achieved. This is followed by data visualization, report generation and making decisions.

II. Factors considered to approach the problem

To approach this problem, we had certain factors to be considered. Before investing in a company, venture capitalist should look for the following,

1) Management: The venture capitalist look for a solid team with a strong management who are capable of handling problems effectively. They would rather invest on a bad idea with a good management than investing on a bad team with good idea[5].

2) Assessment of risks: The job of the venture capitalists involve a lot of risks. Hence, the venture capitalist would want the company to show them how efficient they are going to use their money[3]. It is important for them to know what the company has accomplished and what the company is going to accomplish.

3) Size of the market: The venture capitalists will want to know how big the market is for the product and what the company is going to accomplish and how unique is the product from rest of the other companies.

4) Competitive product edge: The company seeking for funds must establish a product for which there is no prior solutions provided by other companies to make it different from others and to show the differentiation in software[4].

III. DataSet Overview

Different sources of data were investigated based on parameters like efficiency of data, type and details of data and availability to provide the data in low cost. Based on these criterion two websites were selected for this project which contain efficient and comprehensive dataset with easy extraction policies. Following are the two datasets considered for this project,

- Mattermark (https://mattermark.com/): This website provides data about companies , investors and 5,000+ investment events. It allows data extraction using rest API protocol based calls. Python request library used to obtain data from Mattermark API. Though it only allows hundred calls per private API key for free and then the key expires, yet it was enough for this academic project. More data was available only on paid subscription. Once the hundred calls were over, the data were downloaded and saved as a CSV file for further project execution. A snapshot of raw data is shown in Fig1
- Crunchbase (https://www.crunchbase.com/): This website provides a huge dataset with 12,000 companies, 11,000 investors and 52,000+ investment events. API access has been restricted on this website since 2016. Last available data in mass quantity - 2013 Snapshot, officially provided by the company. We have taken the CSV (comma separated file) from this website. A snapshot of raw data is shown in Fig2

```
mattermark_dataset.shape
(28827, 12)
mattermark_dataset.columns
Index(['Unnamed: 0', 'company_category_code', 'company_city',
       'company_country_code', 'company_name', 'company_region',
       'company_state_code', 'funded_at', 'funded_year', 'funding_round_type',
       'investor_name', 'raised_amount_usd'],
      dtype='object')
mattermark_dataset.head()
```

	Unnamed: 0	company_category_code	company_city	company_country_code	company_name	company_region	company_state_code	funded_a
0	0	Education	Boston	USA	Examity	Boston	MA	2019-05-01
1	1	NaN	Austin	NaN	KERV Interactive	NaN	TX	2019-04-3(
2	2	NaN	NaN	NaN	zhiyun health	NaN	NaN	2019-01-1(
3	3	NaN	NaN	NaN	HealthMetrics	NaN	NaN	2019-01-1(
4	4	NaN	Stony Brook	NaN	Black Diamond Therapeutics	NaN	NY	2019-01-1(

Fig. 1. mattermark raw dataset

```
investments_data.shape
(52870, 20)
investments_data.columns
Index(['company_permalink', 'company_name', 'company_category_code',
       'company_country_code', 'company_state_code', 'company_region',
       'company_city', 'investor_permalink', 'investor_name',
       'investor_category_code', 'investor_country_code',
       'investor_state_code', 'investor_region', 'investor_city',
       'funding_round_type', 'funded_at', 'funded_month', 'funded_quarter',
       'funded_year', 'raised_amount_usd'],
      dtype='object')
investments_data.head()
```

	company_permalink	company_name	company_category_code	company_country_code	company_state_code	company_region	company_city
0	/company/advercar	AdverCar	advertising	USA	CA	SF Bay	San Francisco
1	/company/launchgram	LaunchGram	news	USA	CA	SF Bay	Mountain View
2	/company/utap	uTaP	messaging	USA	NaN	United States - Other	NaN
3	/company/zoopshop	ZoopShop	software	USA	OH	Columbus	columbus

Fig. 2. Crunchbase raw dataset

IV. Data Preprocessing

Data preprocessing is a process of cleansing data for acquiring more knowledge about it. We are surrounded by immense data for which we lack knowledge as there exists a lot of inconsistencies in data. The data is available in various formats such as files, databases or comma-separated values. The data obtained can have missing values, incorrect data or misspellings during data entry. These inconsistencies are reviewed and the outliers are removed before the data is exposed to further processing. Data preprocessing certainly increases the quality and readability of the data. Every model requires its own kind of data preprocessing, although a basic steps should be performed before applying any model. Following are the steps followed to preprocess the data in this project:

- Combined Dataset: In this paper, we have considered two datasets: mattermark and crunchbase. So in this step we combined the two dataset to make it one big dataset. During combining name and type of attributes were taken care of. For an example, mattermark had series attribute with values a,b,c etc, whereas that of crunchbase was searies-a, searies-b, searies-c etc. So all the values were normalised in this step.
- Data Cleaning: For feature selection some attributes were dropped and some were added. Those attributes which hold less weightage and just overfitting our model were dropped. Some attributed like data-funded, month-funded, year-funded were combined in just one attribute

named as date-funded. All the missing values were removed. Since, we had huge dataset, methods to address these missing values were not required. Then, all the duplicate values were removed. Normalised the whole dataset to make it look like even.

- Instances Selection and Partitioning: The whole dataset were divided into two parts: test data and training data. 80 percent were reserved for training data and 20 percent were reserved for testing data. Models were applied on training data and then to check the performance of recommendation model, it was executed considering the test data.
- Representation Transformation: Label encoding was performed for some of the attributes which had categorical values. Applying models on attributes with non-numeric data becomes difficult. So to resolve this problem we used label encoding. It made calculations comparatively easy.

After completing data preprocessing the dataset looked liked as shown in fig3

```
total_dataset.shape
(81697, 11)
total_dataset.columns
Index(['company_category_code', 'company_city', 'company_country_code',
       'company_name', 'company_region', 'company_state_code', 'funded_at',
       'funded_year', 'funding_round_type', 'investor_name',
       'raised_amount_usd'],
      dtype='object')
total_dataset.head()
```

	company_category_code	company_city	company_country_code	company_name	company_region	company_state_code	funded_at	fur
0	advertising	San Francisco	USA	AdverCar	SF Bay	CA	2012-10-30	
1	news	Mountain View	USA	LaunchGram	SF Bay	CA	2012-01-23	
2	messaging	NaN	USA	uTaP	United States - Other	NaN	2012-01-01	
3	software	columbus	USA	ZoopShop	Columbus	OH	2012-02-15	
4	web	Cleveland	USA	eFuneral	Cleveland	OH	2011-09-08	

Fig. 3. Dataset after preprocessing

V. Data Visualization

The next step in preprocessing is data visualization. "A picture is worth a thousand words" and once the dataset is refined, it is exposed to visualize the patterns, trends and correlations. Our data was visualised to find the relationship between certain features for a better understanding of the dataset before building a model.

The following are the relationships observed between few attributes,

1) Trends over the years: The dataset was visualised to find the maximum investments for a year in the list of years specified. The data was plotted to look for the maximum invested year and found to be the year 2012 with the highest investments on the company by the investors.

From the above, its clear that the year 2012 has the maximum and second highest as 2011. It is also inferred that the years 1997 to 2002 did not have much variation in the investments on the company by the investors.

2) Breakdown of the Industries: The Investor to Company ratio was plotted to see the relation between them. From the stacked bar graph we obtained, it depicts that the software category among all the other categories has highest number of investors.

The investor to company ratio for software category is observed to be nearly equal while the ratio of the same when

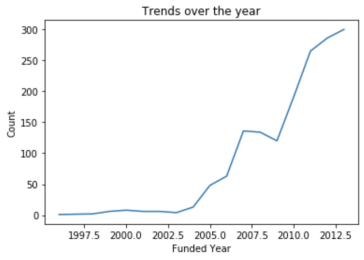

Fig. 4. Trends over the years

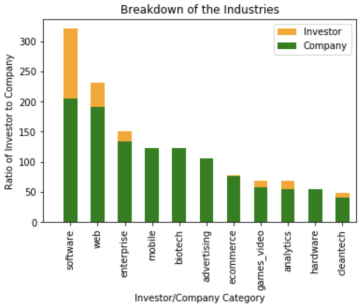

Fig. 5. Breakdown of the Industries

considered for the other categories seems to have a negligible amount of investors for the number of companies that have developed.

3) Preferred Investment location: The preferred investment location of the investors were plotted in order to have a better understanding of how the investment patterns reflected around the states of the United States.

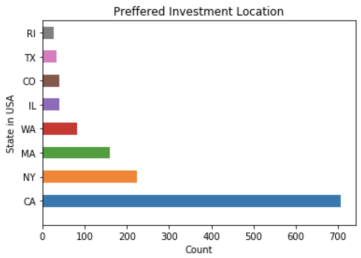

Fig. 6. Preferred Investment location

From the horizontal bar graph, it is inferred that California is the highest among the states of America to invest on a company. California being the maximum is followed by the New York, Massachusetts, Washington and several others. It is clear from the visualization that the preferred location is California.

4) Preferred Funding round type: To categorize the companies based on its round type, the data plotted on the same. It is observed that post-ipo, private-equity and crowdfunding has nearly negligible count of the companies in our dataset.

Angel has the highest number of companies for which the investors had funded in the recent years. From the above all

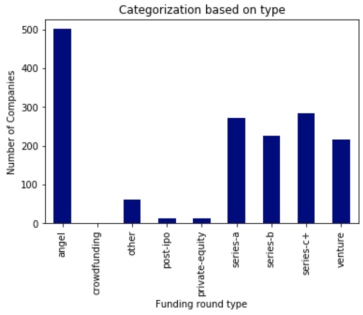

Fig. 7. Preferred Funding round type

visualization, the patterns of the dataset was understood better visualised than having it as just a numeric context.

VI. Recommendation Models

In this paper we have chosen three models: Content based, user based collaborative filtering and K nearest neighbour. Reason behind using first two models is, it is still the most frequently used recommendation method in industry applications. Second, it naturally incorporates the company attributes such as industry hierarchy, which are normally used in traditional screening methods. Following is the detailed discussion on each model from the project point of view:

A. User Based Collaborative Filtering

In user based collaborative filtering approach, recommendation is done on the basis of users past behaviour. Suppose we want to recommend a company to an investor. To address this problem first similarity will be calculated between the given investor and rest of all the other investors based on investor's details. This similarity can be measured by many methods like pearson correlation, cosine similarity, jaccard similarity etc. In this paper, cosine similarity measure has been used. The formula of cosine similarity is shown in fig8

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Fig. 8. cosine similarity formula

Now after calculating the similarity of the given user with each user, top n similar users are considered for further calculation. Now we will predict the ratings for books which have not been read by given user yet but have been nicely rated by most similar users. This approach is based on the belief that an item liked by similar peer will be liked by the user belong to the same peer.

B. Content Based Approach

Content based approach is the item based approach. Suppose we want to recommend a company to a given investor. To do so first we will find out what all companies investor has invested in. Now we will find out the similarity between those companies and other companies based on company's category/country/state code, funding round type, raised amount etc, hence known as item based approach. Again, similarity can be measured by many methods like pearson correlation, cosine similarity, jaccard similarity etc. In this paper, cosine similarity measure has been used. After calculating the similarity measures between these companies, only top n companies will be considered to recommend the given investor. This is the working principle of content based approach. This approach is based on the belief that the companies invested by an investor in past, similar kind of companies will be liked by an investor to invest in future also.

C. K Nearest Neighbour

K-Nearest Neighbour algorithm was made use to segregate the samples from the data set which can then be fed to the filtering model to find recommendations. KNN algorithm classifies the input samples into different classes. Usually it is binary classification, but as the number of output classes increases, the model gets bulky and mo

VII. Recommendation model Results

Precision and Recall are recognized as evaluating indicator about the recommendation effect of the recommended system. Formula for precision and recall shown in fig9

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Fig. 9. Precision Recall formula

where, R(u) is the list of recommendations based on the user's behavior in training set and T(u) is the user's behavior in testing set. Although Precision and Recall is not necessarily related in the calculation formula, it is unlikely to achieve high Precision and high Recall at the same time in an actual recommendation system. Billsus and Pazzani proposed the F indicator to get an equilibrium point between Precision and Recall to evaluate the recommendation system. F indicator can be calculated with the formula shown in fig10

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Fig. 10. F1 indicator formula

All experimentation ran on Windows8.1 with Python3.

VIII. Results Analysis

According to the test results, conclusions can be drawn as follow:

- When the number of recommended companies is the same and the number of neighbors is similar, using the Jaccard similarity can get better results
- When the recommended number of companies is the same and using the same similarity calculation rule, the recommended effect is showing an upward trend with the number of neighbors increasing. When the number of neighbors run up to about 100, the recommended effect achieves the best effect. After that, the effect declines slowly and then tends to stabilize
- For three different companies recommending category, recommending 5 companies can get the highest Precision but the lowest Recall. Recommending 10 companies will get the medium effect. Recommending 15 companies will get the highest Recall and best total effect (the highest F value).

IX. Flow Diagram

The flow diagram of our approach is depicted in fig11,

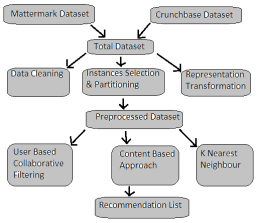

Fig. 11. flow diagram

X. Problems faced

Following are some problems faced during the execution of this project:

- Comparing the sparsity directly to the other datasets, such as MovieLens1M and Netflix, our dataset was extremely sparse (over 99.9 percent) and long-tailed.
- Shortage of numerical data makes it hard to find relation with words (which are not as relating as numerical data)
- The investment behaviours are much sparser than conventional recommendation applications and a venture capital investments are usually limited to a few industry categories, making it impossible to use a topic-diversification method to hedge the risk.
- Industry hierarchy has not been addressed in this paper due to some limitations. However, the recommendation performance can be improved by introducing the existing industry hierarchy information such as Group, Segment, Code.
- Precision value is quite low, which is most likely due to the extreme sparsity of the our dataset

XI. Future work

As a future work for this project risk factors can be cautiously considered when making investments: for a potential startup, a venture capital needs to specifically estimate how well this new investment can fit into its holding investment portfolio in such a way that investment risk can be hedged. Second, ranking algorithm can be used for diverse recommendation. Third, crowdfunding which generally operate through online platforms (e.g., AngelList). This approach is a different ball game and provide additional impetus and scope for applying information retrieval techniques to this domain.

XII. Conclusion

We have successfully implemented User based Collaborative filtering and Content based filtering in this paper for mattermark and crunchbase dataset. This helps in recommending the right list of start-up companies that are believed to have long term growth. Collaborative method helps in exploiting other venture capitalists data in order to invest in a venture, where content based method helps in exploiting their own set of patterns of investments. In this paper, we have performed data preprocessing by combining mattermark and crunchbase datasets. We have also done data cleaning by removing the noisy data. In order to have a clear vision on our preprocessed dataset we have used data visualization methods to observe the data pattern and trend. Achieved prediction and recommendation of start-ups for venture capital investors using K Nearest neighbor approach, Collaborative based approach and Content based approach. For future ranking algorithm can be used for detailed analysis of data for better recommendation.

XIII. Acknowledgment

We would like to thank Professor Shih Yu Chang and our Teaching Assistant Mr. Surya Sonti for guidance and support in successfully executing this project. We would also like to thank our team members who worked long hours and on weekends to make this happen.

References

[1] Jrg Gottschlich, Oliver Hinz, "A decision support system for stock investment recommendations using collective wisdom", Volume 59, March 2014
[2] Mona Taghavi, Kaveh Bakhtiyari, Edgar Scavino, "Agent-based computational investing recommender system", Hong Kong, China, October 2013
[3] Jisun An, Daniele Quercia, Jon Crowcroft, "Recommending investors for crowdfunding projects", Seoul, Korea, April 2014
[4] PA Gompers, J Lerner - 1999 What drives venture capital fundraising?
[5] P Gompers, J Lerner - Journal of economic perspectives, 2001 The venture capital revolution
[6] WA Sahlman - Journal of financial economics, 1990 - Elsevier The structure and governance of venture-capital organizations