An Empirical Study of Top-N Recommendation for Venture Finance

Thomas Stone, Weinan Zhang, Xiaoxue Zhao
Department of Computer Science, University College London
Gower Street, London, WC1E 6BT
{t.stone, w.zhang, x.zhao}@cs.ucl.ac.uk

ABSTRACT

This paper concerns the task of top-N investment opportunity recommendation in the domain of venture finance. By venture finance, specifically, we are interested in the investment activity of venture capital (VC) firms and their investment partners. We have access to a dataset of recorded venture financings (i.e., investments) by VCs and their investment partners in private US companies. This research was undertaken in partnership with Correlation Ventures, a venture capital firm who are pioneering the use of predictive analytics in order to better inform investment decision making. This paper undertakes a detailed empirical study and data analysis then demonstrates the efficacy of recommender systems in this novel application domain.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Information Filtering*

Keywords

Information Retrieval, Recommender Systems, Venture Finance, Industry Classification

1. INTRODUCTION

Early-stage investment is a key driving force of technological innovation and is vitally important to the wider economy, especially in high-growth and hi-tech industries (such as Life Sciences, Clean-tech, Information Technology). Venture finance refers to the financing of private companies through the use of venture capital. Venture capital (VC) is a form of private equity, a medium to long-term form of finance provided in return for an equity stake in potentially high growth companies. Early-stage investment is typified by venture capital firms (VCs) who deploy capital towards high-risk ventures. Venture capital has five main characteristics [9]: is a financial intermediary; invests only in private companies; takes an active role in monitoring and helping portfolio companies; primary goal is to maximise financial return by exiting investments through sale or an initial public offering (IPO); invests to fund the internal growth of companies. Whilst there have been some applications of recommender systems to the broader domain of finance, including micro-finance [3], there has seemingly been no previous academic research in applying such techniques directly to venture finance.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507882.

Traditionally, investment opportunities are either referred or identified through technology scans [13], however, modern information retrieval techniques such as recommender systems have emerged in the past several years as an effective way to help people cope with the problem of information overload [12, 10]. The VC investment process involves several main stages: deal origination, screening, evaluation, structuring, and post investment activities [6]. Our intention is to apply recommender systems with the goal of recommending top-N relevant investment opportunities to VC firms and their investments partners. In recent years the traditional venture financing landscape has also shown signs of evolving [2] plus the emergence of entirely new funding sources such as "crowdfunding" which generally operate through online platforms (e.g., AngelList). Such shifts create new opportunities and provide additional impetus and scope for applying information retrieval techniques to this domain. This new domain is quite distinct from existing applications of recommender systems (e.g., Movies) and, as such, represents unique challenges (see Section 2).

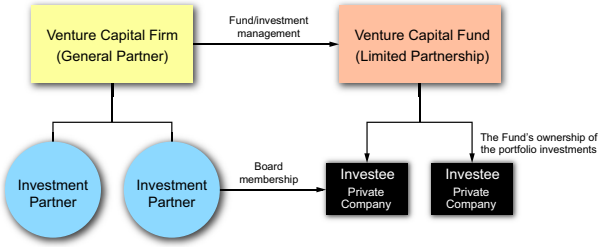


Figure 1: Venture capital (VC) fund structure.

Figure 1 illustrates the typical structure of a VC fund. Although with some variations a typical VC fund is managed by a VC firm (legally referred to as a General Partnership) consisting of several investment partners. The VC fund is essentially an investment fund raised from various institutional investors (legally referred to as Limited Partners, not shown in our figure) such as pension funds, university endowments and family offices. Beyond fundraising the main responsibilities of investment partners (also referred to as General Partners) are sourcing investment opportunities, making investment decisions and taking board membership to assist the management of investee private companies (also referred to as portfolio companies). We have access to a dataset of historical recorded venture financings (i.e., investments) by VCs and their constituent investment partners in private US companies.

Beyond screening prospective investment opportunities and assessing their "fit" for a particular investor several other tasks in venture finance are reliant on some form of company classification such as identifying peers for competitor analysis or comparables for valuation purposes. With advances in information retrieval, particularly text mining and related techniques, it is possible to envision an improved form of industry classification for describing the activities and relationships of private companies. We are interested in resolving the shortcomings of existing classification schemes (i.e., out-of-date, misrepresentation, misinterpretation). Furthermore, an alternative representation of private companies activities (see Section 3.1) offers the potential for improved utility in applying techniques such as recommender systems. Through our empirical study we observe investment strategies, user-item interactions and attempt to improve upon existing industry classification schemes ultimately seeking to improve the top-N recommendation of investment opportunities.

2. VENTURE FINANCE DATA ANALYSIS

2.1 VentureSource Dataset

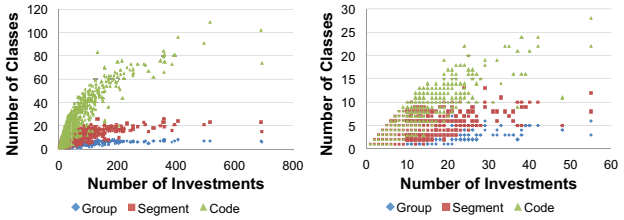
Our dataset was provided by Dow Jones VentureSource, a leading data provider to the venture capital industry, courtesy of Correlation Ventures, a venture capital firm who are pioneering the use of predictive analytics in order to better inform investment decision making. In total we have 21,610 investee private companies (i.e., items), 7,560 venture capital firms and 32,710 investment partners (i.e., two distinct sets of users). In regards to investment relationships VC firms have 83,264 and investment partners have 82,897 relationships with an average of 11.01 and 2.53 relations respectively. Our most prolific VC firm and investment partner have each, respectively, made 600+ and 60+ past investments. On average an investee private company has distinct relationships with 3.85 VC firms and 4.41 investment partners.

Comparing the sparsity directly to other datasets, such as MovieLens1M [8] (95.5%) and Netflix (99.8%), VentureSource is extremely sparse (over 99.9%) and long-tailed which will prove challenging for generating relevant recommendations using existing IR techniques.

Figure 2: Network graph showing industry hierarchy of VentureSource.

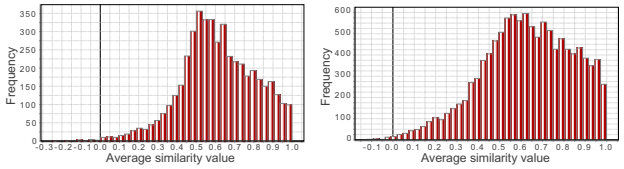
2.2 Industry Hierarchy

The VentureSource dataset includes historical venture financings in the US. Beyond investment relationships we also have a principal component representation of the investee private companies' descriptions and an industry hierarchy, shown in Figure 2, with three tiers (Group, Segment and Code). Group describes broad industry sectors (e.g., "Information Technology") and the subsequent tiers Segment (e.g., "Software") and Code (e.g., "Enterprise Software") provide further granularity. An initial analysis in Figure 3 shows the number of investments by VC firms and investment partners. Investments are plotted against number of distinct industry classes covered at different levels of the hierarchy. We see a concentration of investments in a small number of industry classes, particular for investment partners, even at the lowest level of the industry hierarchy (i.e., Code). This industry classification is similar to other public (e.g., SIC, NAICS) or private (e.g., Capital IQ) classification schemes. It offers a more sophisticated classification scheme compared to similar datasets, such as CrunchBase, which uses a simple category code (e.g., "Games, Video & Entertainment", "Mobile"). As noted in studies on capital market research [4], despite the widespread use of industry classification schemes by academic researchers, few studies directly test their efficacy. Our intention is to utilise VentureSource's industry hierarchy to improve our recommendation performance.



(a) VC firm (b) Investment partner

Figure 3: Number of investments against number of classes by user.



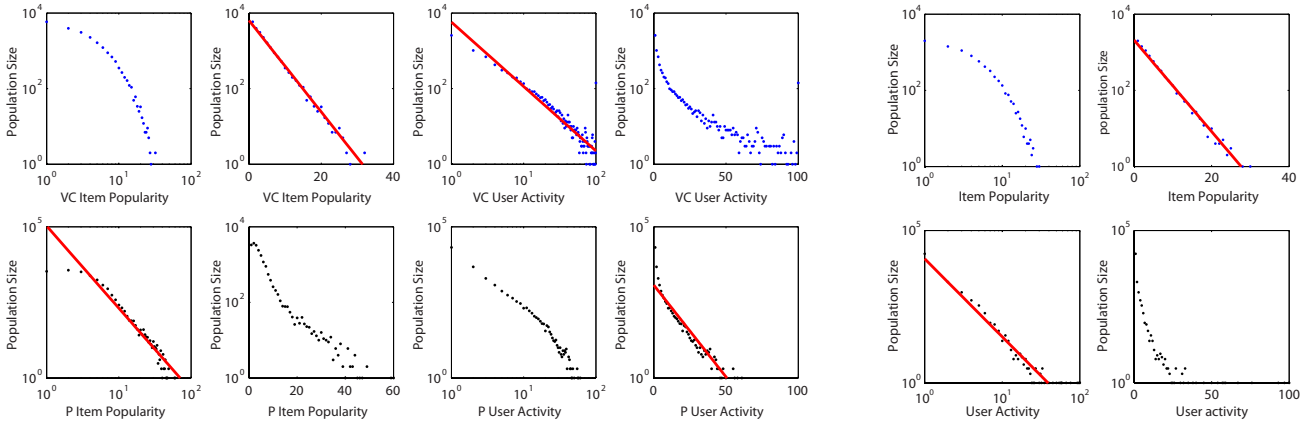
(a) VC firm (b) Investment partner

Figure 4: Average pair-wise cosine similarity for VC and investment partner portfolios.

2.3 Investment Strategies

We are interested in the decision making trade-offs made by investors (e.g., specialize or diversify) under conditions of uncertainty. In particular discovering whether VC firms and their individual investment partners specialize in terms of industries or sub-industries in which they make their investments. Intuitively we would expect individual investment partners, and to a lesser extent VC firms, to specialize in their investment strategies.

Whilst there are is no strict limits an "average" VC firm (i.e., $100 million fund size) will have a small number (i.e., less than 10) of investment partners who will take board seats in the companies in which they chose to make investments. These individual investments constitute the VC firm's overall portfolio of investments, which we would, again only intuitively, expect to be specialized to some degree, at least beyond a random portfolio of private companies. For a portfolio P of n companies we calculate $n(n-1)/2$ similarity



(a) User-item interaction difference between venture capital firms (VC) and investment partners (P).

(b) Simulation of user-item interaction.

Figure 5: User-item interaction analysis.

measures aggregated using an average pair-wise similarity s across the portfolio, based upon the principal components derived from company descriptions. Figure 4 shows the distribution of average pair-wise cosine similarity s for portfolios across all VC firms and all investment partners. The positively skewed distributions, suggest the dominant investment strategies are in favour of specialization, especially for investment partners. This provides the motivation for using the industry hierarchy in generating relevant investment opportunity recommendations for VC firms and investment partners.

2.4 User-Item Interaction Analysis

We conducted an investigation of the user-item interaction data from VentureSource, observing both the VC firm and investment partner interactions with investee private companies. In Figure 5(a), we show histograms with log-log and semi-log of the user and item popularity distributions identifying some interesting characteristics.

For the VC firm and investment partner datasets, the distributions appear to be quite distinct. Observing Figure 5(a), in relation to item popularity, the VC firms follow an exponential distribution (upper second panel), while the investment partners follow a power-law (lower first panel). On the contrary, for user popularity, the VC firms follow a power-law (upper third panel) and investment partners follow an exponential distribution (lower fourth panel).

From this observation we can see that the two datasets have some fundamental differences in their network properties, which may coincide with the algorithm performances on the two datasets. From a complex network perspective, the power-law distribution represents the existence of very active VCs, which is referred to as "the fat tail effect" [1]. In comparison, the exponential distributions on the investee private companies and the investment partner sides indicate that there is no such extreme properties. Since each investee private company or individual partner cannot be involved in such a large number of investment relationships.

In order to explain such network properties, specifically from the VC firm perspective, we ran a simulation (see Figure 5(b)). For items (upper two panels), it displays an exponential distribution. The rule is that the probability that one item gathers one more connection is proportional to its current degree. For users (lower two panels), it displays a power-law distribution. The rule is that the probability that one user creates one more connection is proportional to the second order of its current degree. Hence, the effect that the "rich get richer" is even greater in approximating the power-law distribution (i.e., for the VC firm).

In the simulation experiment, if we sample the user/item by their current degrees, both sides will have exponential distributions. In order to model the power-law distribution, we need to be more biased on the node degrees. As a result, when we sample one side by a quadratic form of the nodes' current degree, we can approximate the power-law. An explanation might be that active VC firms (i.e., those that making more investments) become more popular and well known subsequently receiving more investment opportunities and therefore making more investments. From the perspective of the VC firm, if VC firm A has two times as many investments as VC firm B, then A is more than two times popular than B, which makes the investment popularity of VC firms a power-law distribution.

3. METHODOLOGY

Given the extreme sparsity of collaborative data for our particular use case, with investors making only a small number of investments and with limited co-investment, content-based recommender systems utilising industry hierarchy information seem appropriate. Equipped with industry hierarchy information, our recommendation techniques have been developed (in Section 3.2). We are interested in improving the accuracy and relevance of top-N recommendations in our particular use case but also in evaluating the utility of alternative industry classification schemes (in Section 3.1).

3.1 Industry Assignment

Our particular focus is around industry assignment or how companies are assigned to different industry classes (or categories). An inherent limitation of existing industry classification schemes means companies must be fully assigned to a single industry class (i.e., at each tier of the industry hierarchy). There is no notion by which a company may be assigned to more than one single class (i.e., multiple assignment). This is a common limitation amongst several widely adopted industry classification schemes (e.g., CrunchBase) not solely VentureSource.

In order to generate multiple category information for each investee private company, we propose a supervised learning approach whereby we are given a description of a document (i.e., private company descriptions) and a fixed set of labels (i.e., industry classes). Through implementing various learning methods (Naïve Bayes[1], SVM, Random Forest) we learn a classification function for each industry class for all investee private companies with textual descriptions (i.e., multi-label classification). This process allows us to

[1]Naïve Bayes offered superior classification accuracy.

classify new companies against an existing scheme (e.g., VentureSource) but also to generate novel classification schemes (e.g., multiple assignment). By using the confidence level of the classifier for each industry class we can simply define a threshold confidence level (e.g., 0.5) and we can generate the multiple class assignment, essentially industry "tags".

3.2 Recommendation Models

The recommendation models we use are based on the item-based k-Nearest Neighbor (kNN) [5]. The reason for choosing item-based neighborhood models is because (i) it is still the most frequently used recommendation method in industry applications; (ii) it naturally incorporates the company attributes such as industry hierarchy, which are normally used in traditional screening methods; (iii) we have also tried latent factor models but they appear not as effective as the item-based models, which might be caused by the extreme sparsity and unique user-item interaction properties.

The key component in item-based models is the item-item similarity function. Specifically, we have two basic settings of the item-item similarity. The first one is based on the cosine similarity of the industry hierarchy of the companies (i.e., content-based). The second one is based on the overlap of the VC firms or investment partners of investee private companies (i.e., collaborative filtering). We develop various item-based models based on these two item-item similarity functions. Also, we leverage the linear ensemble method [14] to combine the advantages of the different models.

4. EXPERIMENTS

4.1 Experimental Setting

VentureSource is chosen as the test dataset in our experiment. We use the MyMediaLite recommender system library [7] to implement a standard item-based k-Nearest Neighbor collaborative filtering (CF) approach and then we incorporate both existing (Group, Segment, Code) and generated (Multi) item attributes (i.e., industry hierarchy) in order to improve our performance. Our results are benchmarked against the performance of a random recommender system. In order to evaluate the performance of our recommendation model we calculate the following commonly used evaluation metrics: area under the curve (AUC), mean average precision (MAP) and precision (Prec@N).

4.2 Experimental Results

The overall results of the compared algorithms are shown in Table 1 for VC firms and Table 2 for investment partners. From our results we make the following observations: (i) All the algorithms obtain improved performance against the baseline Random, which indicates the efficacy of our item-based and ensemble models. (ii) The general performance on AUC is not as satisfactory as traditional CF datasets, such as 0.92 on Netflix [11], which indicates the difficulty of performing traditional CF algorithms on this investment opportunity recommendation task. (iii) The recommendation performance is improved by introducing the existing industry hierarchy information (Group, Segment, Code). (iv) By combining the item-based kNN CF and industry Code information using the linear ensemble method, we get our empirical best model on the AUC measure. (v) Initially the multiple industry assignment (Multi) leads to some additional improvement, however, it seemingly has no significant impact on the ensemble methods. (vi) The values of MAP and precision are quite low, which is most likely due to the extreme sparsity of the VentureSource dataset.

Table 1: Performance for VC firm.

Model	AUC	Prec@5	Prec@10	Prec@15	MAP
Random	0.4999	0.0000	0.0001	0.0002	0.0006
Group	0.5590	0.0008	0.0007	0.0006	0.0016
Segment	0.5700	0.0012	0.0011	0.0008	0.0016
Code	0.5920	0.0013	0.0013	0.0014	0.0023
Multi Group	0.5727	0.0006	0.0006	0.0006	0.0014
Multi Segment	0.5730	0.0008	0.0008	0.0008	0.0016
Multi Code	0.5841	0.0012	0.0011	0.0096	0.0023
CF	0.6362	0.0127	0.0099	0.0083	0.0169
CF + Group	0.6430	0.0129	0.0101	0.0087	0.0172
CF + Segment	0.6477	0.0131	0.0107	0.0090	0.0185
CF + Code	0.6582	0.0143	0.0108	0.0091	0.0175
CF + Multi Group	0.6440	0.0125	0.0100	0.0083	0.0169
CF + Multi Segment	0.6414	0.0113	0.0094	0.0079	0.0153
CF + Multi Code	0.6478	0.0126	0.0096	0.0081	0.0165

Table 2: Performance for investment partner.

Model	AUC	Prec@5	Prec@10	Prec@15	MAP
Random	0.4947	0.0001	0.0001	0.0001	0.0007
Group	0.5557	0.0001	0.0001	0.0001	0.0009
Segment	0.5687	0.0002	0.0002	0.0002	0.0013
Code	0.5825	0.0004	0.0005	0.0005	0.0018
Multi Group	0.5604	0.0001	0.0001	0.0002	0.0009
Multi Segment	0.5663	0.0004	0.0003	0.0003	0.0015
Multi Code	0.5783	0.0005	0.0004	0.0005	0.0019
CF	0.6163	0.0093	0.0069	0.0057	0.0203
CF + Group	0.6233	0.0094	0.0071	0.0058	0.0210
CF + Segment	0.6283	0.0092	0.0069	0.0056	0.0197
CF + Code	0.6312	0.0087	0.0063	0.0051	0.0188
CF + Multi Group	0.6216	0.0089	0.0067	0.0055	0.0199
CF + Multi Segment	0.6227	0.0085	0.0062	0.0051	0.0183
CF + Multi Code	0.6234	0.0071	0.0054	0.0045	0.0161

5. CONCLUSION AND FUTURE WORK

In this paper we demonstrate the efficacy of recommendation techniques in relation to the novel application domain of venture finance. Through our venture finance data analysis, we discover fundamental differences in user-item interaction patterns between VC firms and their individual investment partners. Our methodology takes advantage of our access to venture financing data to improve the investment opportunity recommendation quality on the VentureSource dataset.

In future work, we plan to take more investment factors into consideration, such as the investment amount, stage and location. In addition, we hope to conduct a user study within such a venture finance scenario to further assess the real-world applicability of our models beyond offline evaluation metrics.

6. REFERENCES

[1] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 2000.

[2] S. Anthony. Is Venture Capital Broken?, 2012.

[3] T. Bhaskar and G. Subramanian. Loan recommender system for microfinance loans: Increasing efficiency to assist growth. *Journal of Financial Services Marketing*, 15(4), 2011.

[4] S. Bhojraj, C. M. C. Lee, and D. K. Oler. *Journal of Accounting Research.*

[5] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM TOIS*, 2004.

[6] V. H. Fried and R. D. Hisrich. Toward a Model of Venture Capital Investment Decision Making. *Financial Management*, 23(3):28–37, 1994.

[7] Z. Gantner and S. Rendle. MyMediaLite: A free recommender system library. 2011.

[8] GroupLens Research. MovieLens Data Sets, 2013.

[9] A. Metrick. *Venture Capital and the Finance of Innovation.* 2007.

[10] M. Montaner. A Taxonomy of Recommender Agents on the Internet. pages 285–330, 2003.

[11] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

[12] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme. Fast context-aware recommendations with factorization machines. *SIGIR*, 2011.

[13] T. T. Tyebjee and A. V. Bruno. A Model of Venture Capitalist Investment Activity. *Management Science*, 1984.

[14] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms.* Chapman & Hall, 2012.