Company Investment Recommendation based on Data Mining Techniques*

Svetla Boytcheva[1,2][0000−0002−5542−9168] and
Andrey Tagarev[1]

[1] Sirma AI trading as Ontotext, Sofia, Bulgaria
{svetla.boytcheva, andrey.tagarev}@ontotext.com
[2] Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Bulgaria
svetla.boytcheva@gmail.com

Abstract. There are about seventy thousand companies listed on various stock markets worldwide and there is public information on about three hundred thousand companies on Wikipedia but that is only a small fraction of all companies. Among the millions others are hiding the future technological innovators, market disruptors and best possible investments. So, if an investors has an example of the kind of company they are interested in, how can they successfully find other such investment options without sifting through millions of options?
We propose non-personalized recommendation approach for alternatives of company investments. This method is based on data mining techniques for investment behaviour modelling. The investment opportunities are discovered using the idea of transfer learning of indirectly associated company investments. This allows companies to diversify their investment portfolio. Experiments are run over a dataset of 7.5 million companies, of which the model focuses on startups and investments in the last 3 years. This allows us to investigate most recent investment trends. The recommendation model identifies top-N investment opportunities. The evaluation of the proposed investment strategies show high accuracy of the recommendation system.

Keywords: Knowledge-based models· Data Mining · Investment Recommendation System .

1 Motivation

There are millions of companies worldwide and thousands of new ones are being created on any given day. The pace of innovation means that in many cases, the most interesting companies that utilize novel approaches and technologies are going to be among the ones created recently and those are also the companies

* This research is partially supported by projects that received funding from the European Union Horizon 2020 Research and Innovation Programme – euBusinessGraph (Grant Agreement no.732003) and InnoRate (Grant Agreement no.821518)

that are most in need of capital and expert support. There are many important incubators and groups of angel investors who focus on following these fledgling companies and identifying the most promising ones but due to the sheer volume of potential candidates, in each of these cases the investors are only looking at companies at a very specific stage of development, in a limited geographical location and focused on a specific technology or problem. This means that the number of investment option being considered by any such investor is limited by these factors but that is not a benefit, just a natural limitation on the number of companies that human experts can analyze and consider. In reality any given opportunity is most likely being tackled by multiple companies, probably utilizing different tools or based in different locations. This means that a better way to identify potentially interesting investment opportunities than personal knowledge of a company would be vastly beneficial.

Beyond the sheer number of companies that need to be considered, a further challenge is the very sparse information available for the smaller companies that present the best investment opportunities. Generally speaking, the amount of information available on a company lags behind its importance and waiting for complete detailed information to become available before even considering a company as a candidate will exclude many of the best investment opportunities. This means that any automated approach to the problem not only needs to drastically narrow down the number of potential candidates but must be robust enough to work with only incomplete information about a company.

In this paper we will present an approach to identify promising investment alternatives. Our approach will focus on working with startups and newly created companies with only sparsely available data and the selection methods will be based on statistical analysis of historical investment behavior. The complexity of the problem is high enough, that complete automation of the recommendations isn't a viable option. In our experiments, we focused on the pre-selection step i.e. given a company, we aim to return a list that contains some interesting investment alternatives. This means that a human expert will still go through each candidate in the list to select the relevant ones, but the task is reduced from working with millions of candidates to mere dozens.

2 Related Work

Prediction, forecasting and recommendation systems are widely used in the area of business and finance. Zibriczky [22] presents domain-based review of recommendation systems in Finance, where he investigates applications in online-banking and multi-domain solutions, loans, stocks, real estate, insurance polices and riders, assets allocation and portfolio management, investment opportunities and business plans. Variety of methods are used for recommendation systems like collaborative filtering [13], content–based filtering [11], knowledge–based recommendation [14], case–based recommendation [9], hybrid methods [18], association rules mining [10], fuzzy methods [6], artificial neural networks [12], and support vector machines (SVM) [8]. Investigation of the Venture capital's (VC) invest-

ment behaviour is quite challenging task, due to its sparsity, thus application of classical recommendation methods for venture capital investments are limited. Yingsaeree et al [19] define computation finance taxonomy that shows which method is more appropriate for which domain application and which research task. Usually VCs invest in quite few companies from not so diverse industries. Stone et al [15] propose Top-N recommendation system for venture finance using supervised learning approaches, textual description, fixed set of industry classes, and industry hierarchy. This method alongside with other methods is integrated in NVANA platform that aims to assist in the appraisal of early-stage venture [16]. Zhao et al [21] present five portfolio-based risk-aware recommendation algorithms for predicting new investments, by using CrunchBase dataset. The authors in [20] propose utility-based recommendation algorithm based, on the idea of transfer learning. This approach allows to cope with the problem of personalized recommendation system usage for VCs that lack a history of investment portfolio by profiling investors and using equity funds information. The majority of the proposed solutions are personalized, but we aim to develop method that is non-personalized, data-driven and unsupervised. Thus we will use data mining techniques to identify patterns of investment opportunities.

3 Datasets

Our experimentation is based on a large custom fused dataset available in an RDF triple store. We will now examine the constituent parts of the dataset, the shape of the unifying model and the database used to store it.

3.1 The Data

Our experiments are carried on a custom dataset created at Sirma AI that was created by the data fusion of five large commercial datasets. These datasets contain information on companies, investors and historical information on financial transactions between these entities. As part of the data fusion process, instances of the same entity or event present in multiple datasets were identified and merged in the final dataset. After the data fusion process, the finalized custom dataset we are working with contains 7.5 million companies and investors and 1.5 million investment events.

Table 1 lists the feature counts and their coverage over the dataset. The only two features that we can always rely on are company name (not actually used for suggestions) and RDF rank which is a measure of a company's importance in the overall graph. Investor count and funding amount are also calculated for every company but in cases where the company has received no investment yet, both numbers are zero. This is still useful information, of course, but it makes the 100% coverage number not quite correct. The rest of the features deal with investment, industry, size and foundation year which get progressively less common for newer companies but they are still present in a useful number of cases. Finally, the company description is potentially the most valuable single

Feature	Coverage
Name	100%
Rank	100%
Investor Count	100%
Funding Amount	100%
Country	91.3%
Region	68.8%
City	61.7%
Industry	46.3%
Foundation Year	44.7%
Description	34.0%
Employee Count	9.9%

Table 1. Company feature coverage in the dataset

feature but its coverage for new companies is even worse than the 34% coverage figure suggests and using it in a useful manner requires some serious Natural Language Processing which will not be discussed in this paper.

3.2 The Knowledge Graph

The fully-fused dataset is in the form of a Linked Data graph represented as RDF triples. All triples in the dataset conform to the unified data model that defines the shape and types of data available in the graph.

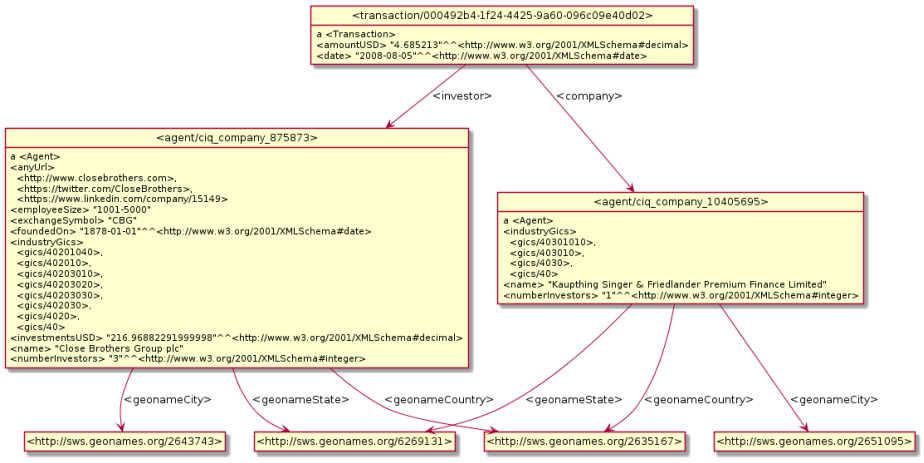

Fig. 1. Part of the Knowledge Graph Model

Figure 1 shows a relevant portion of the total Knowledge Graph model, specifically the connection between a company and an investor through an investment

event. As we can see, there are a variety of features available for the different entities in the model, notably the features we discussed in subsection 3.1.

This connection between company and investor through an investment event is going to form the basis for training our algorithms. The idea is, in essence, to examine the different portfolios of investments chosen by a particular investor and search for certain patterns within them, thus concluding what makes a certain company a good fit for a given investment portfolio.

It is also worth bearing in mind that the various steps of the algorithms described are going to affect some changes on the contents of the knowledge graph itself in the form of certain features. Firstly, we will mark all potential candidates, excluding companies that were founded before January 1st 2014 or that have gone bankrupt. Secondly, we are going to cluster all potential candidates into a number of classes depending on their features. These steps are not reflected in Figure 1 but they do not change the relevant part of the model in any major way. They are, however, crucial in order to translate the candidate selection rules generated by the algorithms into SPARQL queries that select the actual candidates from among all available companies.

3.3 The Database

The Knowledge Graph is stored in GraphDB [3] – a highly-efficient, robust and scalable RDF database. It allows the incorporation of clustering results through reasoning based on forward-chaining of entailment rules and the retrieval of candidates through the use of graph pattern matching rules translated into the powerful SPARQL language.

4 Methods

We propose an unsupervised data-driven method for non-personalized recommendations for company investments. The learning method is based on three main steps (Fig. 2) - investigation of the investment behavior, identification of investment type and generation of investment strategy.

The main idea behind the investment opportunities is to investigate direct and indirect associations in company investments. Direct association, also called frequent patterns, represent different sets of companies that appear together in the investment portfolios of multiple companies. In contrast, indirectly associated companies (Fig. 3) are seldom found in the investment portfolio of the same company, but they co-occur with common a set of companies (called the mediator set) in a large number of investment portfolios.

4.1 Indirect Association Rules Mining

Companies in our dataset S will be called *items* $V = \{v_1, v_2, ..., v_n\}$. For the collection S we extract the set of all different companies' investment portfolios

[3] GraphDB web page. https://www.ontotext.com/products/graphdb/ Accessed 12 Jun 2019

$P = \{p_1, p_2, ..., p_N\}$, where $p_i \subseteq V$. This set S corresponds to transactions and for each of them is associated unique transaction identifier (tid).

Given a set S of tids, the support of an itemset I is the number of tids in S that contain I. We denote it as $supp(I)$. We define a threshold called $minsup$ (minimum support). Frequent itemset (FI) F is one with at least minimum support count, i.e. $supp(F) \geq minsup$. The task of frequent pattern mining (FPM) of S is to find all possible frequent itemsets in S.

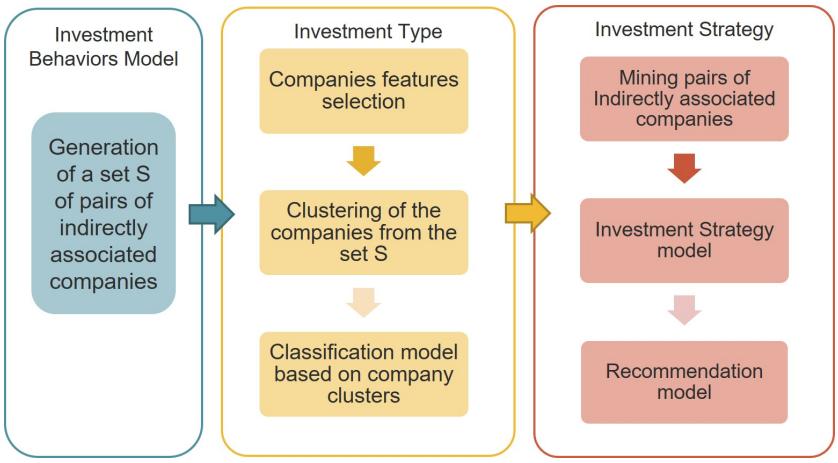

Fig. 2. Investment Strategy model

The following definition for indirect association rules was proposed by Tan and Vipin [17]:

Definition 1. *(Indirect associated pair) An itempair* $\{A; B\}$ *is indirectly associated via a mediator set* $C = \{C_1, ..., C_n\}$ *if the following conditions hold :*

1. $sup(A; B) < minsup$ (*Itempair Support Condition*)
2. There exists a non-empty set C such that $\forall C_i \in C$:
a) $sup(A; C_i) \geq ts$; $sup(B; C_i) \geq ts$ (*Mediator Support Condition*).
b) $d(A; C_i) \geq conf$; $d(B; C_i) \geq conf$ where $d(p; Q)$ is a measure of the dependence between p and Q (*Dependence Condition*).

Condition (1) is needed because an indirect association is significant only if both items seldom occur in the same company's investment portfolio, i.e. they are negatively correlated. Condition (2a) is needed to guarantee statistical significance of the mediator set C. Condition (2b) is needed to guarantee that only items highly dependent on both A and B are used to form the mediator set C. Items in C form close neighborhood.

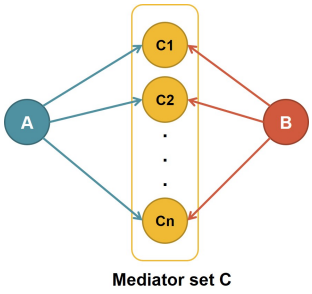

Fig. 3. Indirect association of companies A and B via mediator set $C = \{C_1, ..., C_n\}$

4.2 Investment Behaviours Model

This task starts with preprocessing of the transactions data by converting raw data into item sets by applying hashing – replacing each item (company) with a unique ID and removing duplicates. Each item set is stored in ascending order by id in order to hasten the data mining process. In the initial step we create a model of investments behaviour based on data mining methods for indirect association rules mining (IARM), FPM and Association Rules (ARs). The experiments use Java implementations of the algorithms IndirectRules[17], FPMax [5], and FPGrowthARL [7] from SPMF[4] (Open–Source Data Mining Library). A set $A = \{(I_1, J_1), ..., (I_m, J_m)\}$ of pairs of indirectly associated companies and set $U = \{X | \exists(X, Y) \in A \vee \exists(Y, X) \in A\}$ of startups involved in some pair in A are generated in result.

4.3 Investment Type Identification

This module starts with company features selection $\Phi = \{f_1, ..., f_t\}$. The set of startups U is clustered by density based clustering method [4] into clusters $K = \{K_1, ..., K_M\}$. The JRip classification algorithm [3] is applied to the clusters in order to generate classification rules. JRip was selected because this algorithm results in a small number of ordered rules with high accuracy. The algorithm runs through 4 stages: Growing a rule, Pruning, Optimization and Selection. It has a high time complexity and is considered relatively slow. In our case the execution time is not significant because it is applied just once during the model creation. The precision and number of rules generated are most important as they will be applied multiple times over big dataset and the overall decision and recommendation process relies on them.

In addition the generated classification rules are applied to the entire datasets S of companies.

[4] Open–Source Data Mining Library SPMF. http://www.philippe-fournier-viger.com/spmf/index.php Accessed 12 Jun 2019

4.4 Investment Strategy

An indirectly associated pair (IAP) of companies are symmetric, thus $\forall(I,J) \in A, \exists(J,I) \in A$. Then for each ordered pair $(I,J) \in A, \exists$ a vector with company's features and the corresponding clusters:

$$(f_1(I), ..., f_t(I), cluster(I), cluster(J))$$

Inductive logic method CN2 [1] [2] is applied to learn patterns of investment strategies. For each cluster $K_i \in K$ are selected all companies $I \in U$ such that $cluster(I) = K_i$. Their corresponding vectors are marked as positive, and all remaining vectors are marked as negative. The target value is the cluster of the second company J in the IAP (I,J). The CN2 rule induction algorithm, applied over these vectors, generates ordered classification rules in the form:

$$rule_{il} : \textit{if (condition) then } cluster = K_j.$$

where *condition* is a conjunction of attribute-value pairs of company's features. The main objective of this step is to produce generalized rules, based on the common features of IAPs. Thus for a company X from cluster K_i the most appropriate cluster K_j can be recommended and investment opportunities can be selected from it. The generated CN2 classifier is applied for each company $X \in K_i$ and a rule $rule_{il}$: applicable to the features of X is identified. Some additional equivalence relations are added for part of the features - *same_as*. A new vector is generated for each IAP (X,Y):

$$(rule_{il}, f_1(Y), ..., f_t(Y), same\_f_i(X,Y), ..., same\_f_k(X,Y))$$

CN2 is applied again with the rule as a target value and ordered classification rules in the form are generated:

$$R_{ab} : \textit{if (condition) then } rule = rule_{pq}.$$

where *condition* is a conjunction of attribute-value pairs of company's Y features. The later rules R_{ab} for restriction of the investment opportunity companies features by

4.5 Investment Recommendation

For a given company X classification rules are applied in order to associate the corresponding cluster K_i 4. Then we apply transfer rule $rule_{ij}$ to identify the most appropriate investment strategy. The investment strategy ranks the possible alternative investments clusters. The associated cluster K_j with highest rank is selected. Additional restrictions for the required features of companies from K_j are applied from the rule R_{ab} that corresponds to $rule_{ij}$. Based on these restrictions the possible investment opportunities $\{Y_1, Y_2, ...\}$ from K_j are filtered and ranked. From the investment alternatives for company X, the top-N companies $\{Y_1, Y_2, ..., Y_N\}$ are presented by the recommendation system to the user 5.

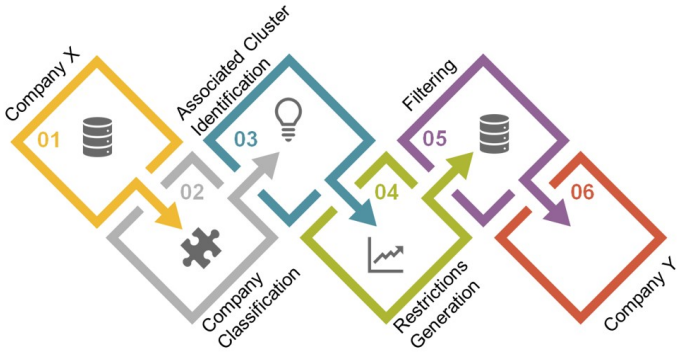

Fig. 4. For given company X recommendation process for investment opportunities

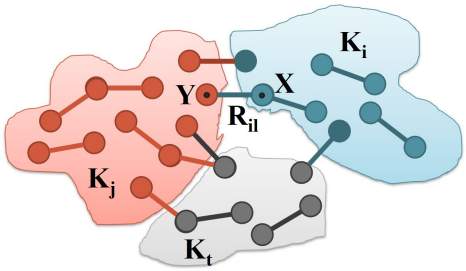

Fig. 5. Investment opportunity

5 Experiments and Results

From the original dataset of 7.5 million of companies the subset of investors who invested recently in startups (last 3 years) was selected. This produced a training dataset containing 112,062 tids and 322,445 companies which was used for experiments. For creating the investment behavior model 2,078,271 indirect association rules were generated using $minsup = 0.000025$ (about 3 investments per startup), $minconf = 0.5$ and $minlift = 1.0$ and 135,717 direct associations (frequent itemsets). There were 1,203 companies in total involved in some indirect association.

Example 1: Some indirect associations that are generated in this step, where a and b are indirectly associated items, i.e. investment alternatives:

```
(a=27 b=37|mediator=26) #sup(a,mediator)=3 #sup(b,mediator)=3
#conf(a,mediator)=1.0 #conf(b,mediator)=0.75

(a=155843 b=155844|mediator=155837 155839 155840 155850)
#sup(a,mediator)=3 #sup(b,mediator)=3
#conf(a,mediator)=1.0 #conf(b,mediator)=1.0
```

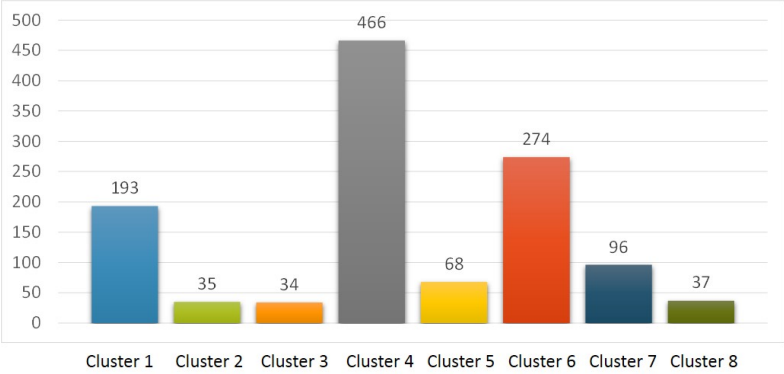

Fig. 6. Startups from the training set grouped in 8 clusters

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.772	0.008	0.949	0.772	0.851	0.832	0.938	0.839	cluster1
	1.000	0.000	1	1.000	1	1.000	1	1	cluster2
	1.000	0.000	1	1.000	1	1.000	1	1	cluster3
	0.987	0.049	0.927	0.987	0.956	0.929	0.975	0.927	cluster4
	0.956	0.008	0.878	0.956	0.915	0.911	0.983	0.891	cluster5
	0.945	0.013	0.956	0.945	0.95	0.936	0.987	0.963	cluster6
	0.979	0.007	0.922	0.979	0.949	0.945	0.994	0.954	cluster7
	0.919	0.000	1	0.919	0.958	0.957	0.988	0.946	cluster8
Weighted Avg.	**0.939**	**0.024**	**0.94**	**0.939**	**0.938**	**0.920**	**0.976**	**0.926**	

Fig. 7. JRip classification rules accuracy for 8 clusters

Applying density based clustering these companies are grouped into 8 clusters (Fig. 6). The largest cluster (cluster4) contains US-based companies from technological industries that predominate in startups datasets and have common investment model. Despite this imbalance, the classification method JRip generated 39 rules with high accuracy (Fig. 7). In all generated rules the industry feature values were used as a condition. There were 5 rules that used the rank feature value as additional criterion and a few rules used some of the other features like funding, number of the investors and foundation year. The CN2 algorithm generated 215 rules for associated cluster identification and 99 rules for investment opportunity recommendation for 201 features. Weighted Relative Accuracy (WRAcc) was used as evaluation measure of rules search. Beam width was set to 20, and the learning mode to exclusive, the maximal length for rules was set to 15, and statistical significance – 1.0. Evaluation results for 10–fold cross validation with training set size 66% show high precision – 0.959, recall – 0.958 and F1–measure – 0.958.

In Example 1 the company $a = 27$ is classified in $cluster6$, and the company $b = 37$ is in $cluster1$. The features vectors (rank, investors, funding, foundationYear, location, numberEmployees, industry, cluster) for IAP of companies with IDs 27 and 37 are:

```
27:(0.00032,2,0.09,?,3175395,?,45103010;451030;45;4510,cluster6)
37:(0.00042,3,0.0,?,3175395,1-10,202010;20;20201070;2020,cluster1)
```

where "?" denotes missing value. We can see that both companies have comparable ranks, number of investors, same country location, but operate in different industry sectors.

For example for the company "Even Financial, Inc." the top 5 investment alternative recommendations generated were startup companies with the same location, comparable rank, similar number of employees and investors and funding, but from different industries – software, electronics, finance, technologies, merchandise. This shows that the experimental results support the main objective of the investment recommendation system to diversify the investment portfolio.

6 Conclusion and Further Work

We set ourselves the task of attempting to identify potential investment alternatives based on the historical data of investment events contained in our Knowledge Graph with over 7.5 million companies and 1.5 million financial transactions. We explored a variety of statistical approaches to the problem and evaluated their performance, finally identifying the most promising combination of algorithms for the task. Some initial feedback from financial experts is that there are some useful leads in the investment candidates suggested by the final algorithm although this is very much a pre-selection step and serious analysis by a human expert is required.

The immediate next step would be to define a more rigorous metric for evaluation and engage some domain experts to carry out. This would allow us to identify the specific strengths and weaknesses of the selected approach and provide a numerical evaluation of the performance so it can be used as a baseline in further iterations.

From a feature engineering perspective, the most useful next step would be to tackle the company descriptions with the use of a modern NLP approach that would perform meaningful text processing. This would most likely take the form of a neural network approach that can identify semantic similarities between company descriptions and reduce the similarity to a number that can be input into the existing algorithms.

References

1. Clark, P., Boswell, R.: Rule induction with cn2: Some recent improvements. In: European Working Session on Learning. pp. 151–163. Springer (1991)
2. Clark, P., Niblett, T.: The cn2 induction algorithm. Machine learning **3**(4), 261–283 (1989)
3. Cohen, W.W.: Fast effective rule induction. In: Machine Learning Proceedings 1995, pp. 115–123. Elsevier (1995)

4. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd **96**(34), 226–231 (1996)
5. Grahne, G., Zhu, J.: High performance mining of maximal frequent itemsets. In: 6th International Workshop on High Performance Data Mining. vol. 16, p. 34 (2003)
6. Guo, H., Sun, B., Karimi, H.R., Ge, Y., Jin, W.: Fuzzy investment portfolio selection models based on interval analysis approach. Mathematical Problems in Engineering **2012** (2012)
7. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery **8**(1), 53–87 (2004)
8. Huang, W., Nakamori, Y., Wang, S.Y.: Forecasting stock market movement direction with support vector machine. Computers & Operations Research **32**(10), 2513–2522 (2005)
9. Musto, C., Semeraro, G., Lops, P., De Gemmis, M., Lekkas, G.: Personalized finance advisory through case-based recommender systems and diversification strategies. Decision Support Systems **77**, 100–111 (2015)
10. Paranjape-Voditel, P., Deshpande, U.: A stock market portfolio recommender system based on association rule mining. Applied Soft Computing **13**(2), 1055–1063 (2013)
11. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: The adaptive web, pp. 325–341. Springer (2007)
12. Quah, T.S.: Improving returns on stock investment through neural network selection. In: Artificial Neural Networks in Finance and Manufacturing, pp. 152–164. IGI Global (2006)
13. Sayyed, F., Argiddi, R., Apte, S.: Generating recommendations for stock market using collaborative filtering. Int. J. Comput. Eng. Sci **3**, 46–49 (2013)
14. Shiue, W., Li, S.T., Chen, K.J.: A frame knowledge system for managing financial decision knowledge. Expert Systems with Applications **35**(3), 1068–1079 (2008)
15. Stone, T., Zhang, W., Zhao, X.: An empirical study of top-n recommendation for venture finance. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1865–1868. ACM (2013)
16. Stone, T.R.: Computational analytics for venture finance. Ph.D. thesis, UCL (University College London) (2014)
17. Tan, P.N., Kumar, V., Srivastava, J.: Indirect association: Mining higher order dependencies in data. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 632–637. Springer (2000)
18. Tseng, C.C.: Portfolio management using hybrid recommendation system. In: IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE'04. 2004. pp. 202–206. IEEE (2004)
19. Yingsaeree, C., Nuti, G., Treleaven, P.: Computational finance. Computer **43**(12), 36–43 (2010)
20. Zhang, L., Zhang, H., Hao, S.: An equity fund recommendation system by combing transfer learning and the utility function of the prospect theory. The Journal of Finance and Data Science **4**(4), 223–233 (2018)
21. Zhao, X., Zhang, W., Wang, J.: Risk-hedged venture capital investment recommendation. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 75–82. ACM (2015)
22. Zibriczky, D.: Recommender systems meet finance: A literature review. In: CEUR-WS: Proc. of the 2nd International Workshop on Personalization and Recommender Systems in Financial Services - FINREC 2016. vol. 1606, pp. 3–10 (2016)