Using Data Mining Methods for Investment Opportunities Identification

Authors: Svetla Boytcheva, Andrey Tagarev

Abstract: Investors range from angel investors interested in early-stage start-ups to large companies that wish to expand their expertise by acquiring established businesses but a single challenge faces them all- how to identify the technological innovators and market disruptors among the millions of companies worldwide. In this paper, we explore an approach that takes an example company and sifts through the sea of options to identify a manageable number of investment alternatives. The proposed approach for company investment recommendations is non-personalized and is based on data mining and machine learning techniques. We use indirect association rules to generate investment alternatives and to identify companies that belong to similar investment portfolios. We identify several investment behavior models using density-based clustering. For matching companies to different investment types is used classification, based on the JRip algorithm. The inductive logic programming method CN2 is used for learning patterns of investment strategies. The CN2 method is used also in the core of the investment recommendation system, which for a given company generates top-N investment opportunities. Several experiments were performed with a big company knowledge graph in RDF-format with data about 7.5 million companies. The evaluation results show high accuracy of the proposed method and show a promising practical application in the financial sector, allowing companies to diversify their investment portfolio.

Keywords: Investment Opportunities; Data Mining; Machine Learning; Knowledge-based models

JEL: M42

1. INTRODUCTION

There are millions of active companies throughout the globe, many of them created in just the last few years. The rapid rate of technological progress means that we can expect most novel approaches and interesting innovative activity is carried out by recently founded companies that are interested in receiving investments and expert support. Regional incubators and groups of angel investors focus on discovering these startups and identifying the most promising ones but these investors typically look at a very specific stage of development, a single technology and task or a limited geographical region- the sheer scale of the tasks prevents experts from considering all potential candidates. This means that the investment opportunities considered by these investors are limited by this consideration which is certainly not a benefit, just a natural limitation of the amount of information that can be processed by a human expert. The crucial consideration is that any given task and technology is likely being explored by multiple startups around the world, each utilizing a unique approach or based in a separate location. There is a clear need for

a better way to identify potential alternative investment opportunities that do not rely on a human expert with personal knowledge of all companies.

There is an additional aspect to the challenge beyond the sheer number of companies to be considered, namely the sparsity of the available information on newer companies that present the most interesting opportunities. As a general rule, a company needs to first become important and successful before detailed information becomes available on it in even the largest data sources, therefore, waiting on a complete detailed description of a company to become available before considering it as an investment candidate is not acceptable as it will exclude many of the best investment opportunities. The task of the automated approach to the task then becomes to drastically narrow down the selection of potential investment candidates based on incomplete information with a very limited number of features.

This paper will present an automated approach for identifying viable investment alternatives. This approach will focus on working with recently founded companies that have limited and sparsely available data. The selection methods will be based on a statistical analysis of historical investment choices by a large number of investors. A full analysis of an investment option is exceptionally complex and subjective so the aim is not to fully replace human experts but rather assist them through initial pre-selection. This aim will become obvious in the discussion of the experiments where the goal is to generate a manageable list of companies to examine that contains some appropriate alternatives to a specified company. This reduces the load on the human expert from being aware of all possible options to just examining a list of some few dozen opportunities.

2. RELATED WORK

A comprehensive domain-based review of recommendation systems in Finance is published by David Zibriczky (2016). He presents research on the application of various recommendation algorithms and data mining techniques in a broad range of domains like loans, stocks, real estate, assets allocation, insurance policies, and riders, online-banking, and multi-domain solutions, portfolio management, investment opportunities, and business plans.

There is a wide range of knowledge-based and machine learning approaches that are successfully used for the design of recommendation systems in Finances. Shiue et al (2008) propose a knowledge-based recommendation system that aims to assist expert's decisions in the process of financial health evaluation of enterprises. Musto et al (2015) introduce a novel case-based recommendation system for financial product recommendations, based on a greedy diversification algorithm that can diversify different investment strategies over time. Chiu--Che Tseng presents an application in investment recommendations of a hybrid method (Tseng, 2004) that combines two decision modules -- influence diagram (a special type of Bayesian network) and decision tree. Paranjape--Voditel and Deshpande (2013) propose a recommendation system for stock market portfolio, based on association rules mining. Tong-Seng Quaha and Bobby Srinivasan (2006) present the ability of artificial neural networks (ANN) to solve the task for stock selection. Huang et al (2005) propose a method based on support vector machines (SVM)

for prediction of financial movement direction. The authors compare the proposed method with other techniques and demonstrate that SVM performance predominates other methods in the task for forecasting weekly movement direction of NIKKEI 225 Index.

Yingsaeree et al (2010) define computation finance taxonomy that shows which analytical method and the corresponding programming techniques are more appropriate for which finance application. The authors investigate both statistical approaches and Artificial Intelligence (AI) techniques. The most popular solutions for the task of Financial forecasting are classification and regression. The classification-based forecasting methods were identified as the best solution for the investment opportunity identification task.

The majority of the proposed solutions are personalized, but we aim to develop a method that is non-personalized, data-driven and unsupervised. Thus we will use data mining techniques to identify patterns of investment opportunities.

3. DATASETS

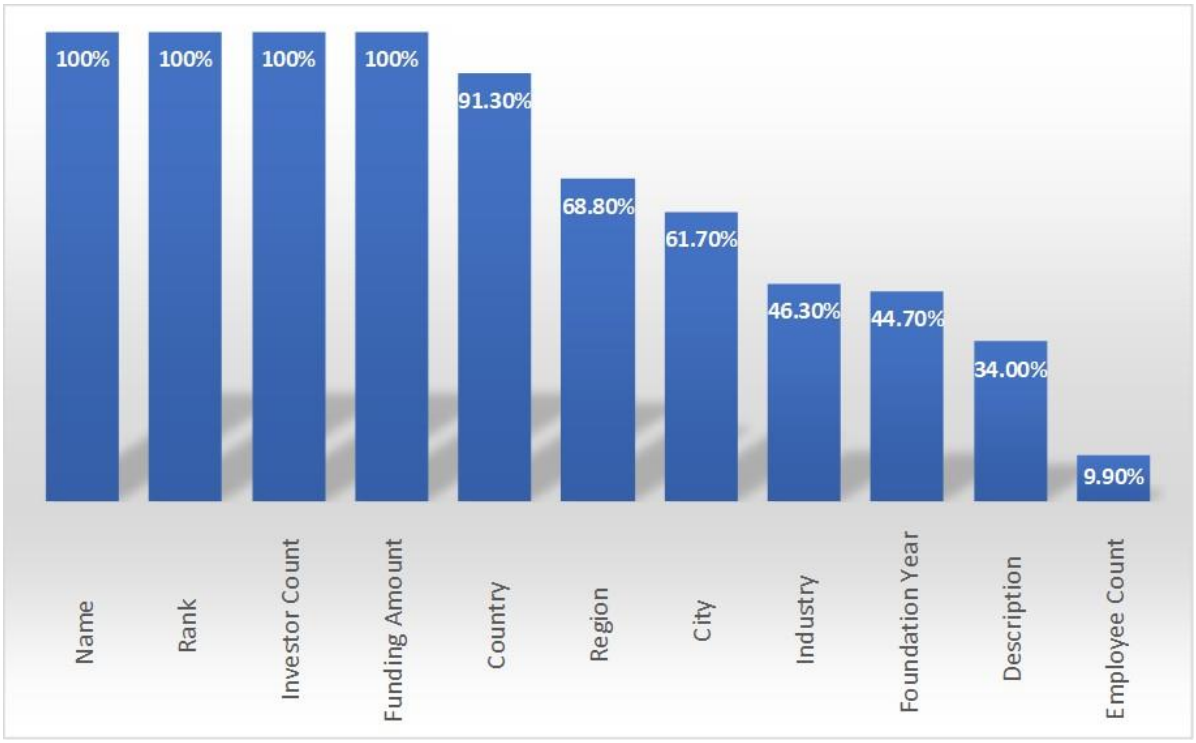
All experiments are carried out on the data from a large-scale proprietary dataset fused from several sources represented in a Linked Data RDF format. This section is dedicated to examining the data contained in the dataset, its format in the form of a unifying model and the RDF datastore that houses it.

3.1. The Data

The custom dataset is created through the fusion of several large commercial datasets of investment information carried out at Sirma AI. Each of the constituent datasets contains various aspects of information on companies, investors, and histories of financial transactions between these entities. The data fusion process produces the final dataset by converting each dataset to linked data conforming to the data model described in subsection 3.2 then identifying and merging instances of the same object present in multiple datasets. The fusion process produces a single finalized dataset that contains over 7.5 million unique agents (companies and investors) and 1.5 million financial transactions between agents.

Figure 1 lists the coverage of the larger company features over the whole dataset. As the figure demonstrates, there is full coverage for only two features- company name (not used for suggestion decisions) and RDF rank (a measure of a node's prominence in the knowledge graph). While investor count and funding amount are also available for each agent, in cases where no information on investments is available (very common) both numbers are just set to zero. Of course, the lack of investment is still relevant information, but it makes the full coverage claim not exactly accurate. Location, industry, employee size and foundation year become progressively sparser for startups, but they are nonetheless present in for a significant percentage of the agents. It is worth noting that the agent descriptions are potentially the most powerful single feature but its coverage for new companies is even worse than the 34\% coverage figure suggests because the existing descriptions are often quite terse or uninformative. Additionally, making full use of it requires a deep Natural Language Processing approach which will not be discussed in this paper.

Figure 1 Company feature coverage in the dataset



3.2. The Knowledge Graph

The Linked Data Knowledge Graph in the triplestore represents a unified fully-fused dataset. All triples in the dataset conform to a single unified data model that describes the format and shape of the fully fused dataset represented in the graph.

Figure 2 Part of the Knowledge Graph Model

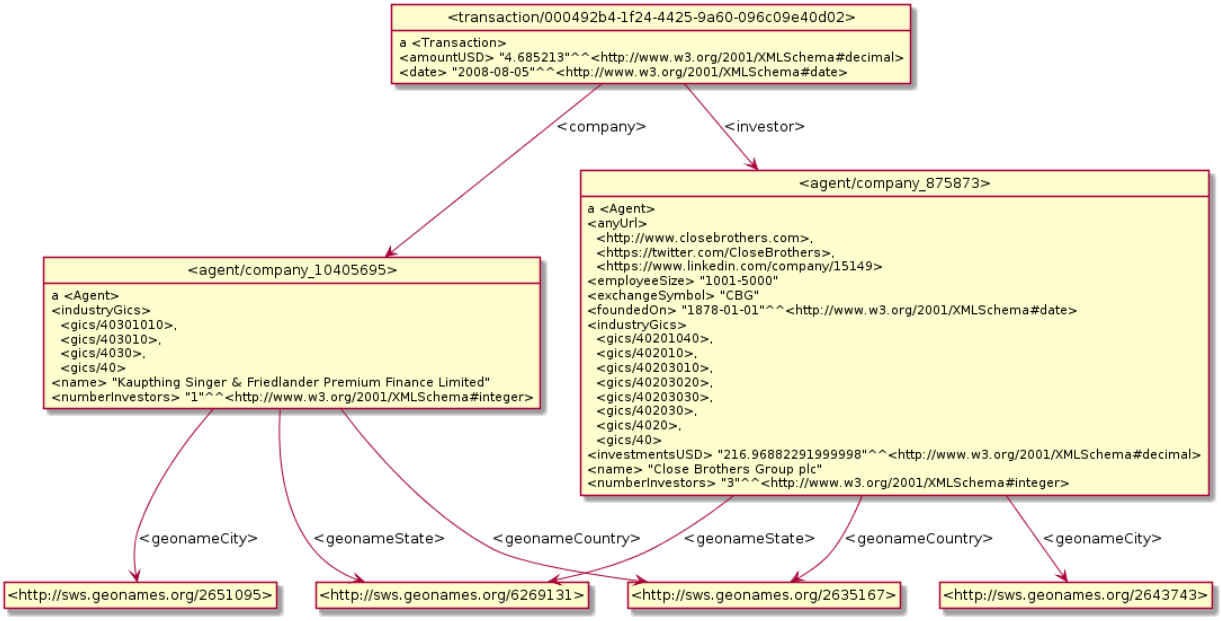


Figure 2 describes the core portion of the Knowledge Graph ontology model, specifically the contents of and relations between a company and an investor through an investment event. The figure is an example of the variety of features available for the various agents in the model, specifically the major features listed in subsection 3.1.

The investment events that create a connection between two actors (company and investor) are going to be the source of training data for our algorithms. Conceptually, we will be examining the hundreds of thousands of unique investment portfolios in search of useful patterns that would allow the argument to infer the combination of properties that would make a certain company a good addition to a specific investment portfolio.

The algorithms we examine will have multiple stages that communicate with each other through the addition of information into the knowledge graph itself. Initially, all potential investment candidates will be marked by selecting active companies that were founded after January 1st, 2014. Next, investment candidates will be clustered into a variable number of classes based on their features. These features were not included in Figure 2 but they do not meaningfully change the relevant part of the model. Nevertheless, their existence is absolutely crucial for the final step which will be selecting final candidates from among all options by running the selection rules on the datastore as SPARQL queries.

3.3. The Database

Sirma's GraphDB[1] -- a highly-efficient, robust and scalable RDF database, is used to store and access the Knowledge Graph. GraphDB 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 represented as graph queries in the powerful SPARQL language.

We propose an unsupervised data-driven method for non-personalized recommendations for company investments. The training method is based on five main stages (Figure 3) - selection of training data, investor behavior research, determining the type of investment, investment strategy and creating rules for investment recommendations.

[1] https://www.ontotext.com/products/graphdb/

Figure 3 Pipeline of investment recommendation system

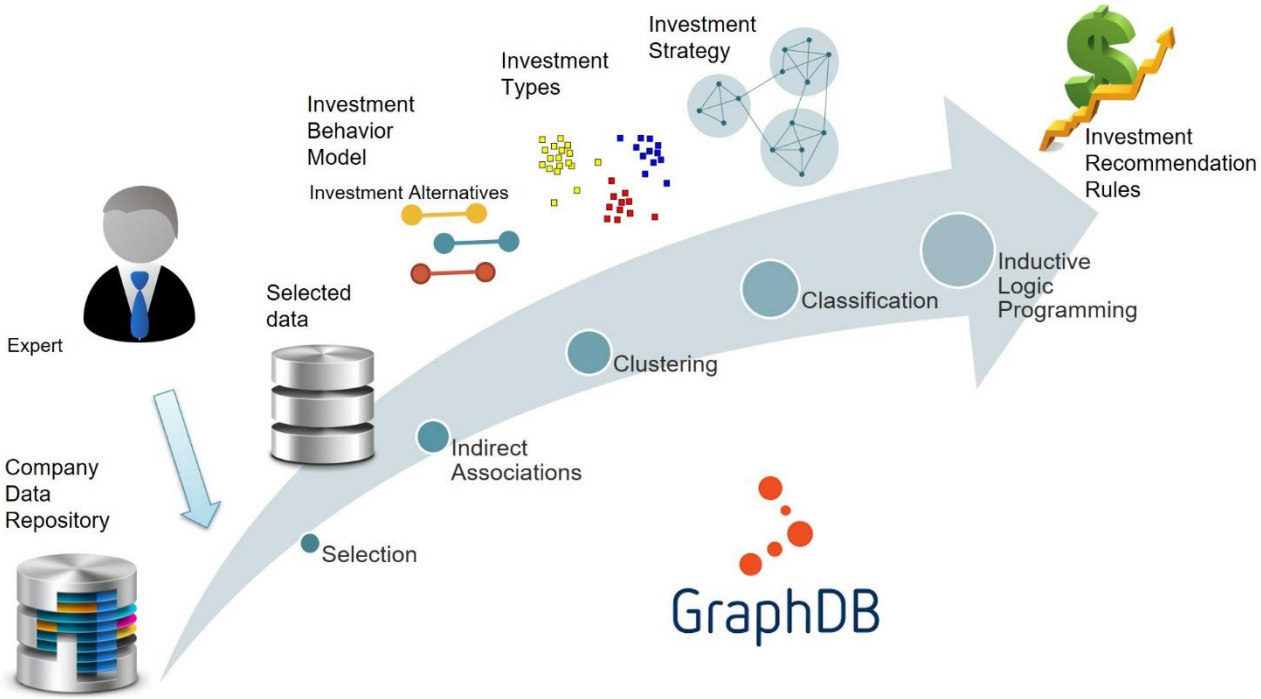
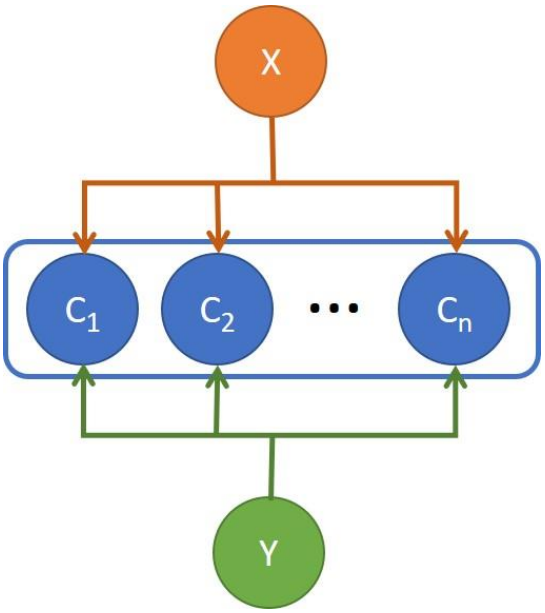


Figure 4

The basic idea of investment opportunities is to explore direct and indirect investment associations in company investments. Direct associations, also called frequent patterns, are well-known models, represent various groups of companies that appear together in the investment portfolios of several companies. Unlike indirectly associated companies (Figure 5), colocation in the same investment portfolio of these companies is rare, but they are found together with a common set of companies (called an intermediary set) in a huge amount of investment briefcases.

Figure 5 Indirectly associated companies X and Y with mediator set $C = \{C_1, C_2, \ldots, C_n\}$



4. SELECTION

The financial domain is quite dynamic and investment trends quickly change in the time. In order to be able to model adequate recommendations, we select from the company graph are selected only such investors that made investments recently (for example in the last 3 years). The resulting dataset contains vectors of such investors and companies in which they invested.

4.1. Indirect Association Rules Mining

Companies in our dataset S will be called ***items*** $V = \{ v_1, v_2, \ldots, v_n\}$. For the collection S we extract the set of all different companies' investment portfolios $R = \{ r_1, r_2, \ldots, r_n\}$, where $r_i \subseteq V$. 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) I is one with at least minimum support count, i.e. $supp(I) \geq minsup$. The task of frequent pattern mining (FPM) of S is to find all possible frequent itemsets in S.

The following definition for indirect association rules was proposed by Tan and Vipin (2000):

Definition: Indirect associated pair

An itempair $\{X; Y\}$ is indirectly associated via a mediator set M if the following conditions hold:

1. $sup(X; Y) < minsup$ (*Itempair Support Condition*)

2. There exists a non-empty set M such that $\forall M_i \in M$:

   a) $sup(X; M_i) \geq ts$; $sup(Y; M_i) \geq ts$ (*Mediator Support Condition*).

   b) $d(X; M_i) \geq conf$; $d(Y; M_i) \geq conf$ $where$ $d(p; Q)$ is a measure of the dependence between p and Q (*Dependence Condition*).

Condition (1) is needed because an indirect association is significant only if there are seldom occurrences of both items in the same company financial transactions portfolio, i.e. negatively correlated. Condition (2a) is needed to guarantee the statistical significance of the mediator set M. Condition (2b) is needed to guarantee that only items highly dependent on both X and Y are used to form the mediator set M. Items in M form a close neighborhood.

The task of Model Investment Behaviors begins with the preliminary processing of transaction data by converting the raw data into itemsets by applying hashing -- replacing each item (company) with a unique identifier and removing duplicates. The itemset is stored in ascending order of the items identifier in order to speed up the data mining process. At the initial stage, we create an investment behavior model based on the data extraction methods used for indirect association rules mining (IARM), FPM and association rules (AR). For experiments are used Java implementations of the algorithms

IndirectRules (Tan, 2000}, FPMax (Grahne, 2003), and FPGrowthARL (Han, 2004) from Open-Source Data Mining Library SPMF[1].

As a result, it is generated set P of pairs of indirectly associated companies and set J of startups involved in some pair in P.

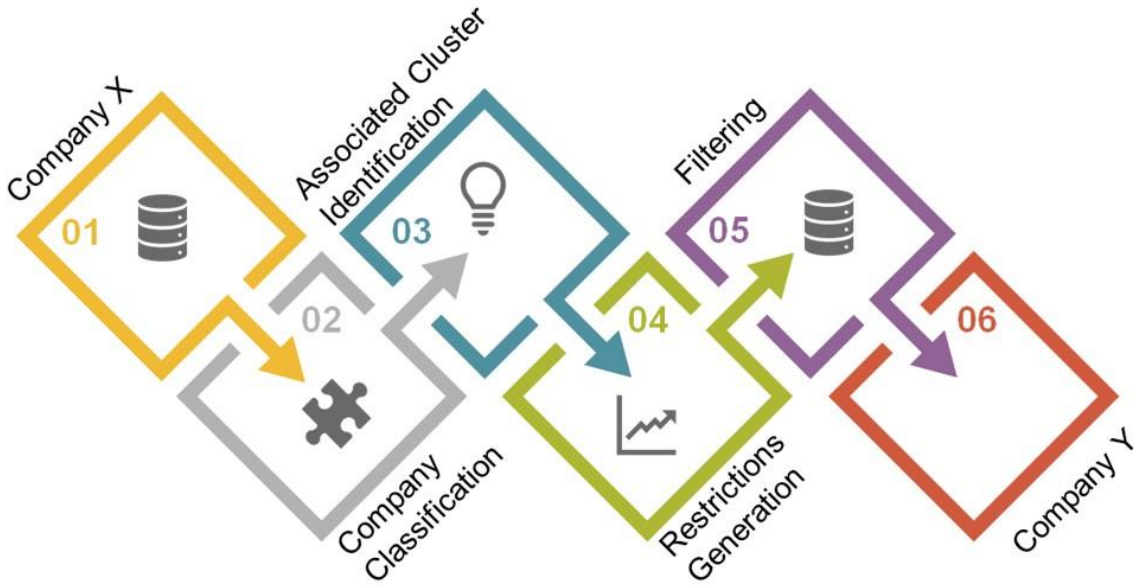
4.2. Clustering

The Investment Type Identification task starts with company features selection. The set of startups J is clustered by density based clustering method (Ester, 1996) into clusters.

4.3. Classification

The classification algorithm JRip (Cohen,1995) is applied to the generated clusters to generate classification rules. The generated rules model various investment strategies. JRip was chosen because it allows the creation of a compact set of ordered rules with high accuracy. The algorithm goes 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, since it is applied only once when creating the model. The accuracy and number of rules generated are paramount, as they will be repeatedly applied to a large data set and the entire decision-making process and recommendations depend on them. In addition, the generated classification rules are populated to the entire datasets of companies as a preliminary processing step to expedite the investment recommendation process. At this stage, the pairs in the set P of the corresponding clusters and features are filled.

Figure 6 For the given company X recommendation process for investment opportunities



[1] http://www.philippe-fournier-viger.com/spmf/index.php

4.4. Inductive logic programming

The most important task of generating investment recommendation rules is based on the inductive logical programming (ILP) method CN2 (Clark,1989). It is used to study investment strategy models in each cluster. Guideline rules are generated by the investment strategy model. The CN2 algorithm is chosen because it can work well even with imperfect data. In our case, many of the selected features are missing.

4.5. Investment Recommendation System

For a given company X has applied classification rules for associating the cluster. We then apply the transfer rule to determine the most appropriate investment strategy. The investment strategy ranks the possible alternative investment clusters. The highest-ranked associated cluster is selected. Additional restrictions apply to the required features of the companies in this cluster. Based on these constraints, potential investment opportunities are filtered and ranked. From the investment alternatives of company X were selected the top-N companies, which are presented through a system of recommendations to the user.

5. EXPERIMENTS AND RESULTS

The initial step is the data selection - we filter from the original dataset of 7.5 million, only information about the investors, who invested recently (last 3 years) in startups. The resulting dataset contains 112,062 tids and 322,445 companies. This dataset will be used as a training set for experiments.

Figure 7 Startups from the training set grouped in 8 clusters

The next step is the investment behavior model, based on generated 2,078,271 IAR indirect association rules using minsup = 0.000025 (about 3 investments per startup for 3 year period, i.e. one investment per year), minconf = 0.5 and minlift = 1.0 and 135,717 direct associations (frequent itemsets). The total amount of different companies involved in any IAR is 1,203.

Example 1:

Some indirect associations that are generated on 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=27 b=44 | mediator=26 )
#sup(a,mediator)= 3 #sup(b,mediator)= 3
#conf(a,mediator)= 1.0 #conf(b,mediator)= 1.0

(a=922 b=18116 | mediator=1249 )
#sup(a,mediator)= 24 #sup(b,mediator)= 6
#conf(a,mediator)= 0.9230769230769231
#conf(b,mediator)= 0.8571428571428571

(a=155843 b=155844 |
mediator=155837 155839 155840 155845 155847 155850 )
#sup(a,mediator)= 3 #sup(b,mediator)= 3
#conf(a,mediator)= 1.0 #conf(b,mediator)= 1.0
\end{verbatim}
```

The best clustering result was achieved by applying the density-based clustering. The companies are grouped into 8 clusters (Figure 7). The largest cluster (cluster4) contains companies located in the USA from technology industries that predominate in startups datasets and have a common investment model.

In Example 1 the investment alternatives belong to the following clusters:

```
27 (cluster6) -  37 (cluster1)
27 (cluster6)  - 44 (cluster5)
922 (cluster1) -  18116 (cluster1)
155843 (cluster5) -  155844 (cluster6)
```

The features vectors for companies with IDs 27 and 37 are:

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

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.

Figure 8 JRip classification rules accuracy for 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**	

Although this disbalance, the classification method JRip generates 39 rules with high accuracy (Figure 8). In all generated rules were used the industry feature values as condition, there were 5 rules that used the rank feature value as additional criteria and few rules used some of the other features like received funding, number of the investors and foundation year.

Example 2: Some classification rules generated by JRip:

```
(351030 >= 1) =>  cluster=cluster3 (30.0/0.0)
(35101010 >= 1) => cluster=cluster3 (4.0/0.0)
(352020 >= 1) => cluster=cluster2 (31.0/0.0)
(303020 >= 1) and (funding >= 105.64) =>
                    cluster=cluster2 (2.0/0.0)
(303020 >= 1) and (rank <= 0.00024) and
(investors >= 6) => cluster=cluster2 (2.0/0.0)
(40 >= 1) and (402010 >= 1) =>
                    cluster=cluster8 (16.0/0.0)
(40 >= 1) and (4030 >= 1) =>
                    cluster=cluster8 (5.0/0.0)
(201040 >= 1) and (foundationYear >= 2002) and
(20104020 <= 0) => cluster=cluster8 (5.0/0.0)
```

The algorithm CN2 generated 215 rules for associated cluster identification and 99 rules for investment opportunity recommendation.

Example 3:Rules generated by CN2:

Figure 9 Rules for associated cluster identification, generated by CN2

	IF conditions	THEN class	Distribution	Probabilities [%]	Quality	Length
0	rank_1≥0.00063 AND countryStr_1≠6252001 AND funding_1≥3.9 AND 50_1=0 AND countryStr_1≠1861060 AND countryStr_1≠2635167 AND countryStr_1≠294640 AND countryStr_1≠1814991 AND 30101030_1=0 AND 101010_1=0	→ comp_2=cluster1	[168, 19, 0, 7, 0, 8, 5, 0]	79 : 9 : 0 : 4 : 0 : 4 : 3 : 0	0.148	10
1	countryStr_1≠6252001 AND countryStr_1≠6251999 AND funding_1≤102.33 AND 452030_1=0 AND 251020_1=0 AND 30101020_1=0 AND 30101030_1=0 AND countryStr_1≠2510769 AND countryStr_1≠2921044 AND 25504040_1=0 AND 403010_1=0 AND 151030_1=0	→ comp_2=cluster4	[6, 0, 2, 61, 6, 26, 4, 1]	6 : 1 : 3 : 54 : 6 : 24 : 4 : 2	0.069	12
2	countryStr_1≠6251999 AND funding_1≥14.82 AND countryStr_1≠2635167 AND 2520_1=0 AND 30_1=0 AND 45205010_1=0 AND investors_1≤11.0	→ comp_2=cluster4	[11, 2, 1, 42, 8, 21, 11, 7]	11 : 3 : 2 : 39 : 8 : 20 : 11 : 7	0.065	7
3	foundationYear_1≤2011.0 AND 351020_1=0 AND 151010_1=0 AND 10_1=0 AND 45_1=0 AND countryStr_1≠294640 AND countryStr_1≠3017382 AND investors_1≥3.0 AND countryStr_1≠2921044 AND countryStr_1≠6251999 AND 201040_1=0 AND 2520_1=0 AND 25504040_1=0 AND 201010_1=0	→ comp_2=cluster1	[20, 0, 1, 1, 0, 2, 1, 4]	57 : 3 : 5 : 5 : 3 : 8 : 5 : 14	0.071	14
4	investors_1≤4.0 AND foundationYear_1≤2009.0 AND 2010_1=0 AND countryStr_1≠3017382 AND funding_1≥0.34	→ comp_2=cluster5	[2, 0, 0, 0, 10, 2, 2, 0]	12 : 4 : 4 : 4 : 46 : 12 : 12 : 4	0.057	5
5	countryStr_1≠6252001 AND funding_1≥0.45 AND countryStr_1≠2635167 AND 2530_1=0	→ comp_2=cluster6	[4, 1, 0, 0, 0, 14, 2, 0]	17 : 7 : 3 : 3 : 3 : 52 : 10 : 3	0.058	4
6	funding_1≥1.5 AND 45_1=0 AND 151010_1=0 AND 25302010_1=0 AND funding_1≤79.1 AND 2520_1=0 AND foundationYear_1≥1987.0 AND 2540_1=0	→ comp_2=cluster4	[13, 1, 12, 21, 1, 17, 9, 0]	17 : 2 : 16 : 27 : 2 : 22 : 12 : 1	0.059	8
7	rank_1≤0.00067 AND rank_1≥0.00027 AND funding_1≥9.0 AND 25201020_1=0 AND 403010_1=0	→ comp_2=cluster7	[0, 1, 0, 0, 1, 1, 7, 0]	6 : 11 : 6 : 6 : 11 : 11 : 44 : 6	0.112	5
8	funding_1≥2.0 AND investors_1≤11.0 AND funding_1≥7.0 AND 45205010_1=0	→ comp_2=cluster3	[1, 0, 4, 0, 0, 1, 1, 0]	13 : 7 : 33 : 7 : 7 : 13 : 13 : 7	0.102	4
9	investors_1≤7.0 AND 25_1≠0 AND investors_1≥3.0	→ comp_2=cluster1	[5, 2, 0, 0, 0, 4, 2, 0]	29 : 14 : 5 : 5 : 5 : 24 : 14 : 5	0.095	3
10	rank_1≥0.00097 AND foundationYear_1≥2005.0	→ comp_2=cluster2	[0, 3, 0, 0, 0, 1, 0, 0]	8 : 33 : 8 : 8 : 8 : 17 : 8 : 8	0.18	2
11	20_1≠0	→ comp_2=cluster8	[0, 0, 0, 0, 0, 0, 0, 2]	10 : 10 : 10 : 10 : 10 : 10 : 10 : 30	0.222	1
12	funding_1≥36.0	→ comp_2=cluster6	[0, 0, 0, 0, 0, 2, 0, 0]	10 : 10 : 10 : 10 : 10 : 30 : 10 : 10	0.25	1
13	rank_1≥0.00055	→ comp_2=cluster4	[0, 0, 0, 1, 0, 0, 0, 0]	11 : 11 : 11 : 22 : 11 : 11 : 11 : 11	0.25	1
14	TRUE	→ comp_2=cluster5	[0, 0, 0, 0, 1, 0, 0, 0]	11 : 11 : 11 : 11 : 22 : 11 : 11 : 11	0	

Figure 10 Rules for filtering the investment opportunities from the associated cluster and selection of top-N investment recommendations, generated by CN2

	IF conditions	THEN class	Distribution	Probabilities [%]	Quality	Length
0	countryStr_2=1835841 AND 451020_2=0 AND foundationYear_2≤2006.0	→ rule1=rule11	[119, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 13, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 7, 0, 0, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	51 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 6 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 3 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 3 : 0 : 0 : 0 : 0 : 0 : 0 : 2 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0	0.511	3
1	15_2≠0 AND rank_2≥0.00068	→ rule1=rule11	[56, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	38 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 5 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1	0.383	2
2	countryStr_2=3017382 AND country_match≠0 AND 452030_2=0 AND 2020_2=0 AND 201040_2=0 AND 402010_2=0 AND 4040_2=0 AND 452020_2=0 AND 45103020_2=0 AND employeeArray_2≠51-100 AND 50101020_2=0 AND ?_2=0 AND 351020_2=0 AND 303020_2=0 AND 35101010_2=0	→ rule1=rule47	[1, 0, 0, 0, 0, 9, 10, 11, 12, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 14, 0, 0, 0, 0, 8, 87, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 3, 9, 0, 0, 0, 29, 0, 0, 1, 0, 0, 0, 0, 0, 0, 6, 0, 0, 17, 0, 0, 3, 10, 0, 0, 0, 0, 0]	1 : 0 : 0 : 0 : 0 : 3 : 3 : 4 : 4 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 2 : 0 : 0 : 0 : 0 : 0 : 5 : 0 : 0 : 0 : 0 : 3 : 27 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 2 : 0 : 0 : 1 : 3 : 0 : 0 : 0 : 9 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 2 : 0 : 0 : 5 : 0 : 0 : 1 : 3 : 0 : 0 : 0 : 0 : 0	0.268	15
3	countryStr_2=2635167 AND ind_match≠0 AND 25_2=0 AND employeeArray_2≠101-250 AND 3510_2=0 AND employeeArray_2≠501-1000 AND 45203010_2=0 AND 452050_2=0	→ rule1=rule44	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 45, 0, 2, 3, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 20, 0, 0, 0, 16, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 11, 0, 0, 0, 0, 0, 0]	1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 25 : 1 : 2 : 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 11 : 1 : 1 : 1 : 9 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 6 : 1 : 1 : 1 : 1 : 1 : 1	0.246	8
4	funding_2≤1.57 AND ind_match≠0 AND countryStr_2=6252001 AND funding_2≥0.13 AND rank_2≥0.00031 AND 25_2=0 AND funding_2≤1.53 AND 352010_2=0 AND 2540_2=0	→ rule1=rule43	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 74, 0, 0, 0, 0, 0, 0, 0, 13, 0, 0, 0, 0, 0, 0, 0, 61, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 30 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 6 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 25 : 0 : 0 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 5 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0	0.305	9
5	countryStr_2=1861060 AND city_match≠0 AND foundationYear_2≥2008.0 AND investors_2≤9.0 AND 202010_2=0 AND 452050_2=0	→ rule1=rule62	[0, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 16, 0, 10, 0, 7, 14, 0, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 51, 0, 0, 0, 0, 0, 0, 0, 11, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	0 : 0 : 0 : 0 : 0 : 0 : 0 : 5 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 1 : 8 : 0 : 5 : 0 : 4 : 7 : 0 : 5 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 24 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 5 : 0 : 1 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0 : 0	0.236	6
6	451030_2≠0 AND city_match≠0 AND countryStr_2≠1814991 AND countryStr_2≠1269750 AND countryStr_2≠3017382 AND rank_2≤0.00098 AND foundationYear_2≤2016.0 AND countryStr_2≠6251999 AND countryStr_2≠1861060 AND 25401020_2=0 AND 351030_2=0 AND 45103030_2=0 AND 501010_2=0	→ rule1=rule44	[0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 7, 44, 0, 1, 1, 1, 0, 4, 0, 0, 1, 1, 0, 0, 0, 0, 3, 0, 0, 0, 2, 5, 10, 2, 1, 0, 0, 0, 4, 0, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 0, 0]	1 : 1 : 1 : 1 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 4 : 24 : 1 : 1 : 1 : 1 : 1 : 3 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 2 : 3 : 6 : 2 : 1 : 1 : 1 : 1 : 3 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 1 : 1 : 1	0.243	13

For example for the company "Even Financial, Inc." was generated as top 5 investment alternative recommendations of 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.

The experimental results support the main objective of the investment recommendation system to diversify the investment portfolio.

6. CONCLUSION AND FURTHER WORK

In this paper, we presented an approach to creating a system for an automated suggestion of investment alternatives based on a limited number of features. The process of building the system consisted of roughly three steps. First, we used statistical analysis to identify pairs of investment alternatives in our Knowledge Graph which contains over 7.5 million companies and 1.5 million financial transactions. Then, we clustered companies into several types based on the known features. Finally, we tested a number of different algorithms for producing criteria that identify further pairs of potential investment alternatives. In this last step, we demonstrated that a CN2 induction algorithm is a suitable tool for generating these criteria. Based on some early-stage feedback from domain experts, the lists of investment candidates produced by the final system contain some promising leads although it very much serves as pre-selection steps followed by in-depth research and analysis by human experts.

The next stage of our work is to develop an evaluation strategy that takes into account the subjective and non-binary nature of the investment alternatives being suggested. Then we will engage domain experts to evaluate the performance of this automated approach through a more rigorous metric. This will reveal the strengths and weaknesses of the algorithm as well as give an easily comparable score of its performance which will allow us to use it as a baseline in the evaluation of future iterations and alternative algorithms.

Company descriptions provide a clear opportunity for further development from a feature engineering perspective. This would require the use of state-of-the-art NLP systems that perform advanced text processing, most likely in the form of deep neural networks that detect subtle semantic similarities between company descriptions. The advantage of these algorithms is that they reduce the incredibly nuanced texts comparison task to a series of numerical values that can be simply put into the algorithms in the following steps we've examined.

References

1. Clark, P. and Niblett, T., 1989. The CN2 induction algorithm. Machine learning, 3(4), pp.261-283.
2. Cohen, W.W., 1995. Fast effective rule induction. In Machine learning proceedings 1995 (pp. 115-123). Morgan Kaufmann.
3. Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231).
4. Grahne, G. and Zhu, J., 2003, May. High performance mining of maximal frequent itemsets. In 6th International Workshop on High Performance Data Mining (Vol. 16, p. 34).

5. Han, J., Pei, J., Yin, Y. and Mao, R., 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery, 8(1), pp.53-87.
6. Huang, W., Nakamori, Y. and Wang, S.Y., 2005. Forecasting stock market movement direction with support vector machine. Computers & operations research, 32(10), pp.2513-2522.
7. Musto, C., Semeraro, G., Lops, P., De Gemmis, M. and Lekkas, G., 2015. Personalized finance advisory through case-based recommender systems and diversification strategies. Decision Support Systems, 77, pp.100-111.
8. Paranjape-Voditel, P. and Deshpande, U., 2013. A stock market portfolio recommender system based on association rule mining. Applied Soft Computing, 13(2), pp.1055-1063.
9. Quah, T.S., 2006. Improving returns on stock investment through neural network selection. In Artificial Neural Networks in Finance and Manufacturing (pp. 152-164). IGI Global.
10. Shiue, W., Li, S.T. and Chen, K.J., 2008. A frame knowledge system for managing financial decision knowledge. Expert Systems with Applications, 35(3), pp.1068-1079.
11. Tan, P.N., Kumar, V. and Srivastava, J., 2000, September. Indirect association: Mining higher order dependencies in data. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 632-637). Springer, Berlin, Heidelberg.
12. Tseng, C.C., 2004, March. 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.
13. Yingsaeree, C., Nuti, G. and Treleaven, P., 2010. Computational finance. Computer, 43(12), pp.36-43.
14. Zibriczky12, D., 2016. 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.

ABOUT THE AUTHORS

Svetla Boytcheva, Ph.D., Sirma AI trading as Ontotext, Sofia, Bulgaria & Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria, svetla.boytcheva@ontotext.com

Andrey Tagarev, Sirma AI trading as Ontotext, Sofia, Bulgaria, andrey.tagarev@ontotext.com