Knowledge-Based Mining of Exceptional Patterns in Logistics Data: Approaches and Experiences in an Industry 4.0 Context

Eric Sternberg[1] and Martin Atzmueller[2]

[1] University of Kassel (ITeG), Wilhelmshöher Allee 73, 34121 Kassel, Germany
est@cs.uni-kassel.de

[2] Tilburg University (CSAI), Warandelaan 2, 5037 AB Tilburg, The Netherlands
m.atzmuller@uvt.nl

Abstract. In the context of Industry 4.0 and smart production, industrial large-scale enterprise data is applied for enabling data-driven analysis and modeling methods. However, the majority of the currently applied approaches consider the data in isolated fashion such that data from different sources, e. g., from large data warehouses are only considered independently. Furthermore, connections and relations between those data, i. e., relating to semantic dependencies are typically not considered, while these would open up integrated semantic approaches for effective data mining methods. This paper tackles these issues and demonstrates approaches and experiences in the context of a real-world case study in the industrial logistics domain: We propose knowledge-based data analysis applying subgroup discovery for identifying exceptional patterns in a semantic approach using appropriately constructed knowledge graphs.

1 Introduction

In the industrial world of today, the amounts of data are increasing at a rapid pace, enabling large-scale business intelligence and data-driven decision support. Then, exploratory data mining provides a viable tool for obtaining relevant insights. Here, in particular methods for local pattern mining, e. g., subgroup discovery and exceptional model mining enable powerful approaches for detecting interesting, i. e., unexpected, anomalous and exceptional patterns with a broad range of applications for industrial data analytics.

Problem. However, so far the applied approaches consider the data in isolated fashion such that different data types ranging from unstructured to structured data, e. g., tabular and graph-data, are only considered independently. This requires an efficient and effective integrated approach for semantic modeling and data mining, which, however, has not been established at large-scale yet.

Objectives. In this paper, we exemplify an approach tackling these issues: We apply subgroup discovery for identifying exceptional patterns in the context of finding inventory differences. For that, a knowledge-based approach is presented, integrating large-scale data into a knowledge graph representation. We discuss experiences in the context of Industry 4.0 and smart production.

Preprint of: Eric Sternberg and Martin Atzmueller. Knowledge-Based Mining of Exceptional Patterns in Logistics Data: Approaches and Experiences in an Industry 4.0 Context. In: Foundations of Intelligent Systems - Proc. ISMIS 2018. Springer, 2018

In particular, we present results of a real-world case study in a productive logistic environment of a large scale industrial plant. One major goal for the client was to identify specific logistic processes that possibly lead to erroneous financial assessments, so-called inventory differences. The whole process was strongly assisted by domain experts, so another goal was to deliver reasonable explanations and transparency to them. One major issue was to represent the particular domain dependencies in an integrated knowledge graph and the construction of appropriate features to enable the synchronization between the experts domain understanding and the knowledge graph representation.

Contributions. Our contributions are summarized as follows:

1. We describe an integrated approach for large-scale industrial data analytics implemented by knowledge-based data mining utilizing knowledge graphs.
2. We demonstrate the application of local pattern mining in that context, focusing on subgroup discovery implemented in an intelligent system.
3. We present an anonymized real-world case study in the context of Industry 4.0 and smart production. Furthermore, we discuss insights and experiences in the application domain of production logistics.

The rest of the paper is structured as follows: Section 2 describes the industrial problem setting in more detail, before Section 3 provides related work on knowledge-based approaches and local pattern mining. Next, Section 4 summarizes the formal background on subgroup discovery. After that, Section 5 presents the applied approaches for knowledge graph construction and pattern mining in the industrial context, and discusses experiences in the context of a real-world industrial case study. Finally, Section 6 concludes with a summary and presents interesting directions for future work.

2 Industrial Problem Setting

In the field of logistics in industrial production, Industry 4.0 and smart production are important directions for implementing cost-effective measures. Since data is captured continuously during all the relevant processes, powerful data mining methods are required. The standard automation pyramid cf. [14, 15] on industrial processes depicts different systems corresponding to different levels of analysis. Data analytics is mainly performed on the upper levels – corresponding to the operations control level and the enterprise level. Here, one prominent case for data analytics is given by uncovering inventory differences. Basically, inventory differences cause deviations in the financial rating of the plants' current assets. In the past these differences where detected once a year and could reach deviations in the region of about EUR 100 million in our application domain, which corresponds to about one percent of the yearly turnover of a large plant. As a consequence a team of analysts from different departments permanently investigate these differences which decreases the deviations by a factor of 10. This leads to a trade-off between cost of human resources and the inventory difference because the desired state is to resolve as much as possible deviations using automated analysis, i. e., with a minimum commitment of human resources.

3 Related Work

Related work concerns both knowledge-intensive approaches as well as methods for local pattern mining. Domain knowledge is a natural resource for knowledge-intensive data mining methods, e. g., [8, 23, 20], for example in the context of prototype-based approaches [12] or knowledge graphs [21, 24]. However, in data mining, semantic knowledge is scarcely exploited so far. First approaches for integrating knowledge graphs, i. e., based on ontologies and a set of instance data has been proposed in the area of semantic data mining [23, 20]. [9] presents a mixed-initiative approach, for semantic feature engineering using a knowledge graph. In a semi-automatic process, the knowledge graph is engineered and refined. Finally, the engineered features are provided for data mining. A similar approach is applied in [5]. Here, data from heterogeneous data sources is integrated into a knowledge graph, which then provides the basis for data mining.

For data analytics, local pattern mining is a broadly applicable and powerful set of methods for exploratory data mining [7, 4, 3]. Common methods include those for association rule mining [1], subgroup discovery, e. g., [25, 3] and exceptional model mining [19, 3, 13]. Essentially, subgroup discovery is a flexible method for detecting relations between dependent (characterizing) variables and a dependent target concept, e. g., comparing the share or the mean of a nominal/numeric target variable in the subgroup vs. the share or mean in the total population, respectively [25, 3, 18]. The interestingness of a pattern can then be flexibly defined, e. g., by a significant deviation from a model that is derived from the total population. In the simplest case, (see the example above) a binary target variable is considered, where the share in a subgroup can be compared to the share in the dataset in order to detect (exceptional) deviations.

In contrast to the approaches discussed above, we focus on an integrated approach, exploiting knowledge-based semantic structures, i. e., sets of knowledge components connected to a local pattern mining method. Then, also the knowledge graph can be incrementally refined during the mining process.

4 Background: Subgroup discovery

Formally, a *database* $DB = (I, A)$ is given by a set of individuals I and a set of attributes A. For each attribute $a \in A$, a range $dom(a)$ of values is defined. An attribute/value assignment $a = v$, where $a \in A, v \in dom(a)$, is called a *feature*. We define the feature space V to be the (universal) set of all features. Intuitively, a *pattern* describes a *subgroup*, i. e., the subgroup consists of instances that are covered by the respective pattern. A subgroup *pattern* P is then defined as a conjunction $P = s_1 \wedge s_2 \wedge \ldots \wedge s_n$ of (extended) features $s_l \subseteq V$, which are then called selection expressions, where each s_l selects a subset of the range $dom(a)$ of an attribute $a \in A$. A *subgroup (extension)* $I_P := ext(P) := \{i \in I | P(i) = true\}$ is the set of all individuals which are covered by pattern P. The set of all possible subgroup descriptions, and thus the possible search space is then given by 2^{Σ}, i. e., all combinations of the patterns contained in Σ denoting the set of all possible selection expressions. A *quality function* $q\colon 2^{\Sigma} \to \mathbb{R}$ maps every pattern in the

search space to a real number that reflects the interestingness of a pattern (or the pattern cover, respectively). The result of a subgroup discovery task is the set of k subgroup descriptions $P_1, \ldots, P_k$ with the highest interestingness according to the selected quality function. For subgroup models there exist various quality functions, cf. e. g., [3] for a detailed discussion. A simple example compares the share p of a binary target in the subgroup pattern P to its share in the database p_0, weighted by the size of the subgroup n, i. e., $q(P) = (p - p_0) \cdot n$. For numeric targets, this can easily be adapted replacing the shares by the respective means.

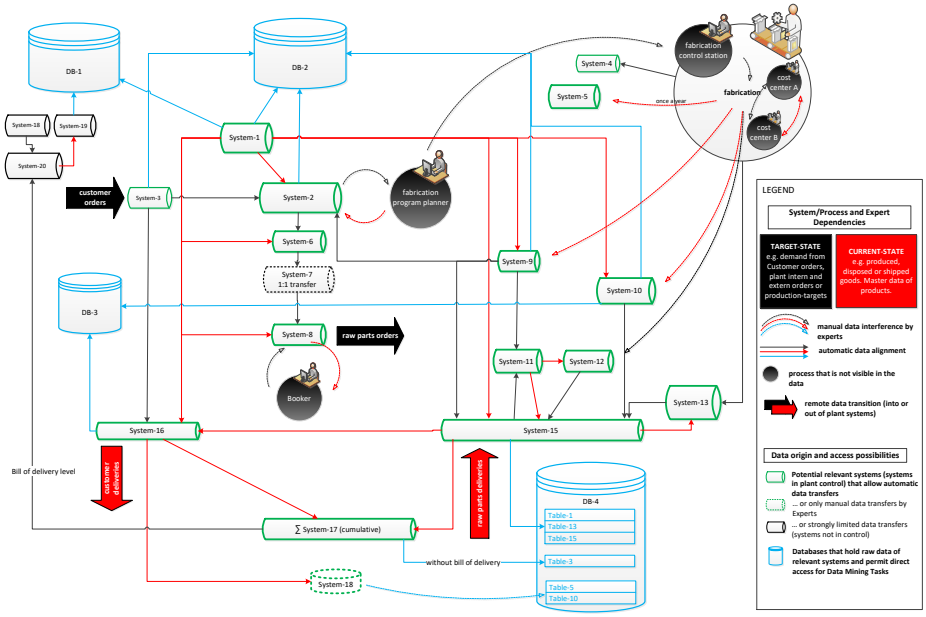

Fig. 1. Overview on the dependencies between systems and processes in our industrial context domain. The blue databases contain raw data accessible for data analytics.

5 Case Study

This section summarizes results of a real-world case study in industrial logistics. For the task of detecting inventory differences, the domain experts provided background knowledge, including exemplary cases leading to inventory differences in the past, relevant data sources for their analysis and how they detect the causes of those differences. A first step for data analysis was then given by business and data understanding, relating to the CRISP-DM model [11]. As a result, it was possible to identify the most relevant data sources contributing to the inventory difference problem. Figure 1 shows the resulting complex

structures. Here, different databases and processes were identified for data preprocessing and integration. This structure then was one of the main outcomes in order to transform the data into a knowledge graph. For data transformation, we implemented a generic framework to load and transform the given data, cf. Figure 2 for a schematic view, utilizing the VIKAMINE [6] system for subgroup discovery. For the knowledge transformation and modeling we applied Gephi [10] and GraphstreamLib [17]. Here, we distinguish between the *structure data* graph representing information and dependencies taken from the bills of materials, and the *accounting data* graph capturing material flow information within the production processes which are discussed in more detail below.

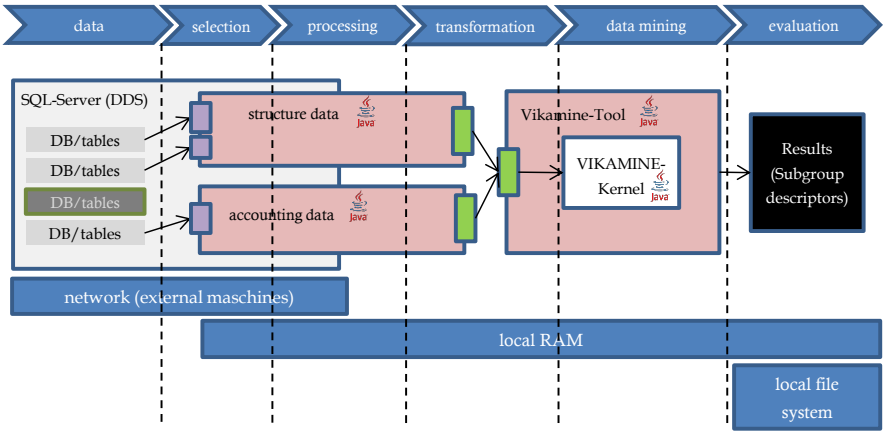

Fig. 2. Schematic view on the data pipeline. The top shows the CRISP steps, the bottom the utilized infrastructure. Gray depicts the present data sources (relates to blue in Figure 1). Red depicts implementation done for the project where the structure and accounting data boxes essentially depict the knowledge graph components.

5.1 Knowledge Graph Construction

As outlined above, two logistic data sources emerged in our case study holding essential information for data analytics. The first related to information on logistic bill of material (BOM). This data is stored in several distributed relational databases, and describes the composition of basic parts up into an end-product in a hierarchical way. Utilizing those assembly dependencies we developed a parser for constructing a graph structure using the GraphStream framework [17], resulting in the *structure data* graph. An (anonymized) visualization of the (complete) generated graph is shown in Figure 3. Each dependency is represented by a directed edge from the basic material to the (complex) processed material, the node size corresponds to its degree.

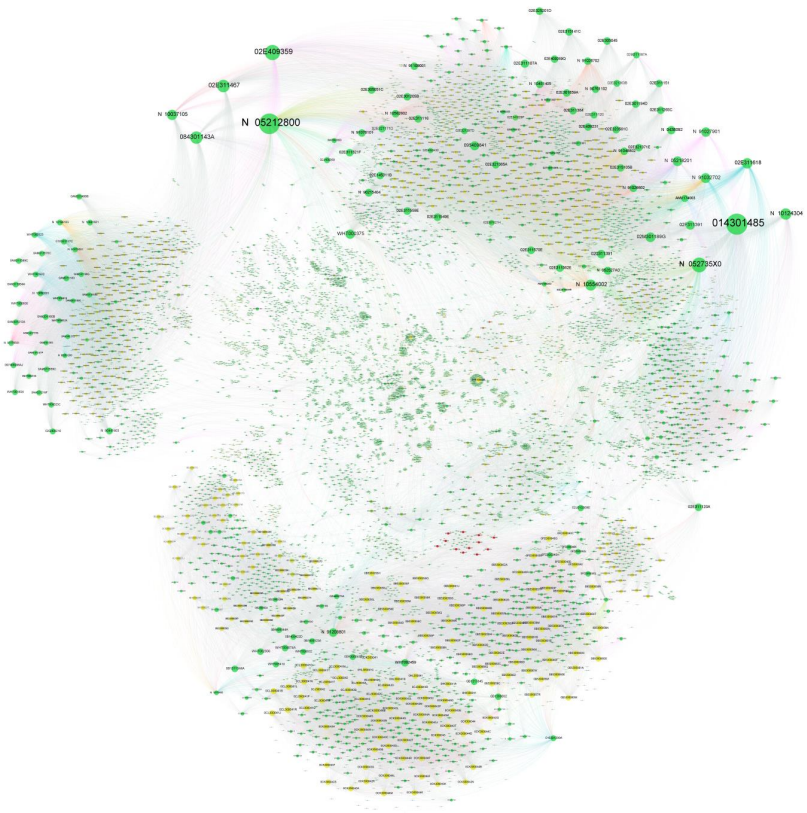

Fig. 3. Visualization of the bill of material (*structure data* graph). Each node with a size according to its degree represents a material. The node color indicates the nodes price where inexpensive (basic) materials are green and more expensive material go over yellow to red (normally end-products are most expensive).

We improved the graph step by step from subgroup discovery in cooperation with the domain knowledge of the experts by additional data sources (e. g., using material prices), in order to represent the relevant dependencies and parts-information according to the experts' judgment, cf. Section 5.3. For that, we formulated different dependencies/expectations as target concepts for subgroup discovery and automatically investigated specific patterns that either occurred very rarely or dominated the observed dependencies in the data, applying the quality function described above.

For a better assessment of potentially problematic dependencies indicated by the patterns, we also implemented a visual explorer for the structure data graph, see Figure 4: This screenshot shows top-level materials (final products), where each material is represented by one node. The user is able to search specific materials (or groups) and to drill down on dependencies via a right-click.

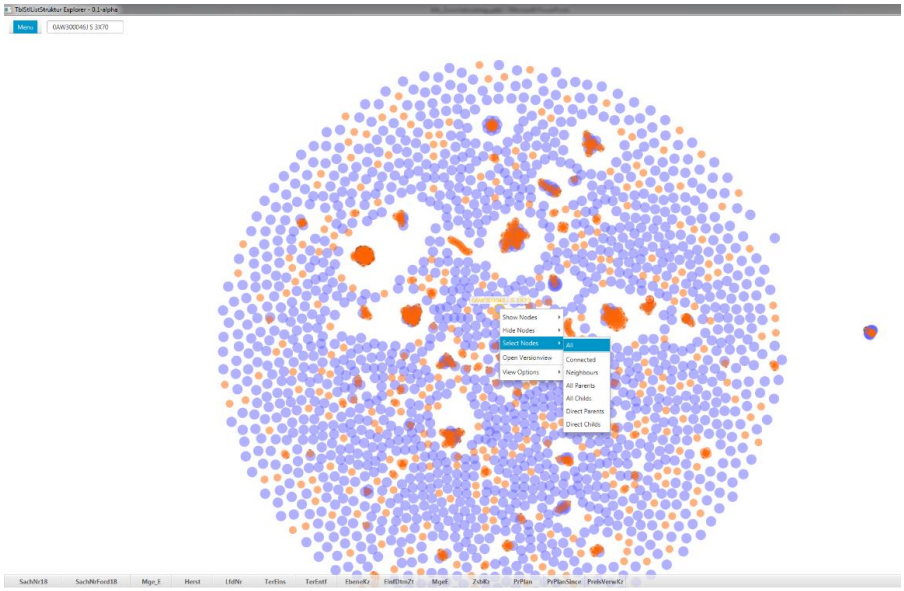

Fig. 4. Screenshot of the interactive BOM explorer. It allows experts to identifi problematic parts and their assembly dependencies in an intuitive and exploratory way. The graph is layouted by a dynamic force layout available from the utilized libary [17] and shows only end-products (end/root nodes in the *structure data* graph).

Regarding the inventory differences it also became apparent that additional data sources with information about the logistic material movements were essential to the experts' analysis. We identified a special system that collected booking information from several systems and created daily summaries, e. g., regarding essential attributes of a booking record like sender, material, amount, time, receiver and booking reason. That information basically describes a material flow network (for more information on logistics and logistic material flow networks we refer to, e. g., [22]). Therefore, analogously to the *structure data* graph, we built the *accounting data* graph corresponding to material flow information (see Figure 2). Both graphs exhibit very significant dependencies which is used in pattern analysis by experts when investigating inventory differences.

5.2 Investigating Expected Relations using Local Pattern Mining

An essential problem to the experts was to check the relevant bookings of a material and all its dependent materials. Because of the complexity and large amounts of the data such manual analysis is always only performed partially, also there are lots of domain specific cases to consider. Therefore, we modeled a key performance indicator (KPI) as a feature for data analytics for each material-node in the *structure data* graph, called the *relbook* feature.

The *relbook* KPI operates on the two associated graphs that represent the bill of material (structure of parts) and the material flow described by bookings (movement of parts). The feature was designed with the experts expectation of correct interpretation of bookings and set as a target concept for subgroup discovery: *relbook* is essentially given by a calculation rule that outputs a real number between -1 and 1. The calculation is done on both the *structure data* graph and the *accounting data* graph and utilizes domain specific knowledge. The KPI basically traces the construction dependencies in the *structure data* graph from a basic material (a screw, an amount of aluminum etc.) up to all final products (a gearing or an exhaust system) containing that basic material. For each visited material-node the particular information from the *accounting data* graph and the corresponding effective booking amounts are determined.

Basically, the experts expected, that if the calculation rule (and the data) is correct, the respective amounts included from concepts of the graphs will add up and there are no deviations (yielding a *relbook* KPI around 0). Thus, if the KPI diverges from zero this indicates a potential problem in the data.

Overall, we discovered two major reasons for deviations from the experts' expectations. The first is indicated by a positive *relbook* KPI generated by an accumulation in storage of specific materials, i. e., those that find no longer or not yet application in active products. On the other hand a negative deviation from the KPI was mainly observed in small groups of bookings. The reasons here where mainly special cost centers and storages with exotic processes e. g., temporary outsourcing of material for extern handling or a special type of (final) waste-booking of old material without actual use in production.

5.3 Identifying Anomalous Subgroup Patterns

Further interesting findings often showed significant deviations from the experts expectations. For detecting anomalous patterns, for example, we examined subgroup targets for the different working-shifts and found descriptions which identify that there are strong associations between specific types of logistic bookings and the current working shift. As an example, Table 1 shows two significant patterns found using the target variable that identifies bookings done in the normal shift (*SchichtKz=01*). The dataset properties (*defined individuals*) show that there are about 8.18 million bookings in the population, also nearly 45% of all bookings are done in the normal shift. The descriptor #2 *BstArt=NIK* states that the normal shift produces 82% of all waste-bookings which was interesting to the experts because they expected a homogeneous distribution over all four shifts. As a result it was uncovered that only particular personal is allowed to perform return or waste bookings – only available in specific shifts. Furthermore it was uncovered that the shifts handling in the booking system is different to the real shifts execution. Therefore the missing 18% of waste-bookings where found in the following system shift which still falls in the normal execution shift.

All results were discovered evaluating patterns from tables like Table 1, where often interesting patterns occured that lead to further investigation such as further expert interviews, individual data inspection, or further drill-down on the

Table 1. Population properties and two exemplary subgroup patterns for bookings done in the normal shift. Pattern #2 uncovers the strong dependence between this shift and waste bookings.

Target (nominal)	**SchichtKz=01**

population properties

defined individuals	undefined individuals	target share in population
8175938	0	44.9%

subgroup descriptors (mid-size groups)

#	quality	group size	target share	TP rate	coverage
1	**EmpfKz=V AND ZwAnkTag** $]-\infty; 44,5[$ **AND AnzBwgTag** $]-\infty; 26,5[$				
	238.15	1506843	64,3%	26,4%	18,4%
2	**BstArt=NIK**				
	214.75	334828	82%	7,5%	4,1%

patterns. Using different targets, for example, concerning the kind of material application and the fluctuation of material-prices we discovered technical problems in an import process for data. In Figure 1, this process would be represented by a blue arrow between a relevant system (green) and a database (blue). The problem was noticed because there were patterns that described bookings with empty storage groups, which was against the experts' expectation. Subgroup descriptions discovered for the target "component price fluctuation" also revealed fragmentary price data for a larger group of parts. In later iterations another systematic problem was uncovered where cost center IDs of physically the same cost center where not equal in different but dependent logistic systems.

6 Conclusions

In this paper, we presented approaches for knowledge-based mining of exceptional patterns in logistics data and discussed experiences in the context of a real-world case study. In particular, we focused on modeling background knowledge in the form of knowledge graphs, and we applied subgroup discovery for identifying exceptional patterns. Overall, the process and the results were very well accepted by the domain experts, which especially favored explainability and transparency of the mined patterns. During iterative sessions, interesting and useful patterns were identified for enabling the automatic monitoring of inventory differences. For future work, we aim to explore network patterns for refinement of knowledge graphs, extending feature engineering methods, e. g., [9]. Here, multiplex approaches, e. g., [2, 16] and pattern-based anomaly detection are further interesting directions.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. VLDB. pp. 487–499. Morgan Kaufmann (1994)
2. Atzmueller, M.: Data Mining on Social Interaction Networks. JDMDH **1** (2014)
3. Atzmueller, M.: Subgroup Discovery. WIREs DMKD **5**(1), 35–49 (2015)
4. Atzmueller, M., Baumeister, J., Puppe, F.: Introspective Subgroup Analysis for Interactive Knowledge Refinement. In: Proc. FLAIRS. pp. 402–407. AAAI (2006)
5. Atzmueller, M., Kloepper, B., Mawla, H.A., Jäschke, B., Hollender, M., Graube, M., Arnu, D., Schmidt, A., Heinze, S., Schorer, L., Kroll, A., Stumme, G., Urbas, L.: Big Data Analytics for Proactive Industrial Decision Support: Approaches & First Experiences in the Context of the FEE Project. atp edition **58**(9) (2016)
6. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: Proc. ECML/PKDD. Springer (2012)
7. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science **11**(11), 1752–1765 (2005)
8. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. IJCAI. pp. 647–652 (2005)
9. Atzmueller, M., Sternberg, E.: Mixed-Initiative Feature Engineering Using Knowledge Graphs. In: Proc. K-CAP. ACM (2017)
10. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An Open Source Software for Exploring and Manipulating Networks (2009)
11. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0. CRISP-DM consortium (2000)
12. Duch, W., Grudzinski, K.: Prototype Based Rules - A New Way to Understand the Data. In: Proc. IJCNN. vol. 3, pp. 1858–1863. IEEE (2001)
13. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. Data Min. Knowl. Disc. **30**(1), 47–98 (2016)
14. Givehchi, O., Trsek, H., Jasperneite, J.: Cloud Computing for Industrial Automation Systems - A Comprehensive Overview. In: Proc. EFTA. pp. 1–4. IEEE (2013)
15. Hollender, M.: Collaborative Process Automation Systems. ISA (2010)
16. Kanawati, R.: Multiplex Network Mining: A Brief Survey. IEEE Intelligent Informatics Bulletin **16**(1), 24–27 (2015)
17. Laboratoire d'Informatique, du Traitement de l'Information et des Systmes (LITIS): Graphstream project, http://graphstream-project.org
18. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. DMKD **30**, 711–762 (2016)
19. Lemmerich, F., Becker, M., Atzmueller, M.: Generic Pattern Trees for Exhaustive Exceptional Model Mining. In: Proc. ECML/PKDD. Springer (2012)
20. Rauch, J., Simunek, M.: Learning Association Rules from Data through Domain Knowledge and Automation. In: Proc. RuleML. pp. 266–280. Springer (2014)
21. Ristoski, P., Paulheim, H.: Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey. Web Semantics **36**, 1–22 (2016)
22. Rushton, A., Croucher, P., Baker, P.: The Handbook of Logistics and Distribution Management: Understanding the Supply Chain. Kogan Page Publishers (2014)
23. Vavpetic, A., Podpecan, V., Lavrac, N.: Semantic Subgroup Explanations. J. Intell. Inf. Syst. **42**(2), 233–254 (2014)
24. Wilcke, X., Bloem, P., de Boer, V.: The Knowledge Graph as the Default Data Model for Learning on Heterogeneous Knowledge. Data Science pp. 1–19 (2017)
25. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. PKDD. pp. 78–87. Springer (1997)