Advances in Exploratory Pattern Analytics on Ubiquitous Data and Social Media

Martin Atzmueller

University of Kassel
Research Center for Information System Design
Wilhelmshöher Allee 73, 34121 Kassel, Germany

atzmueller@cs.uni-kassel.de

Abstract. Exploratory analysis of ubiquitous data and social media includes resources created by humans as well as those generated by sensor devices. This paper reviews recent advances concerning according approaches and methods, and provides additional review and discussion. Specifically, we focus on exploratory pattern analytics implemented using subgroup discovery and exceptional model mining methods, and put these into context. We summarize recent work on description-oriented community detection, spatio-semantic analysis using local exceptionality detection, and class association rule mining for activity recognition. Furthermore, we discuss results and implications.

1 Introduction

In ubiquitous and social environments, a variety of heterogenous data is generated, e. g., by sensors and social media, cf. [3]. For obtaining first insights into the data, description-oriented *exploratory* data mining approaches can then be applied.

Subgroup discovery [2,8,50,87,88] is such an exploratory approach for discovering interesting subgroups – as an instance of *local pattern detection* [52,67,68]. The interestingness is usually defined by a certain property of interesting formalized by a quality function. In the simplest case, 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. More complex target concepts consider sets of target variables. In particular, *exceptional model mining* [8,55] focuses on more complex quality functions, considering complex *target models*, e. g., given by regression models or Bayesian networks with a deviating behavior for certain subgroups, cf. [37,38]. In the context of ubiquitous data and social media [3–5], interesting target concepts are given, e. g., by densely connected graph structures (communities) [17], exceptional spatio-semantic distributions [24], or class association rules [18].

This paper summarizes recent work on community detection, behavior characterization and spatio-temporal analysis using subgroup discovery and exceptional model mining. We start with the introduction of necessary foundational concepts in Section 2. After that, Section 3 provides a compact overview of recent scientific advances summarizing our recent work [4,6,7,17,21,24]. Furthermore, we describe exemplary results, and conclude with a discussion of implications and future directions in Section 4.

Preprint of: Martin Atzmueller (2016) Advances in Exploratory Pattern Analytics on Ubiquitous Data and Social Media. In: Stefan Michaelis, Nico Piatkowski, Marco Stolpe (eds.) Solving Large Scale Learning Tasks: Challenges and Algorithms. Festschrift in Honour of Prof. Dr. Katharina Morik. LNCS 9580, Springer, Berlin/Heidelberg (In Press).

2 Background

Below, we first introduce some basic notation. After that, we provide a brief summary of basic concepts with respect to subgroup discovery

2.1 Basic Notation

Formally, a *database* $D = (I, A)$ is given by a set of individuals I and a set of attributes A. A *selector* or *basic pattern* $sel_{a_i = v_j}$ is a Boolean function $I \rightarrow \{0, 1\}$ that is true if the value of attribute $a_i \in A$ is equal to v_j for the respective individual. The set of all basic patterns is denoted by S.

For a numeric attribute a_{num} selectors $sel_{a_{\mathit{num}} \in [min_j; max_j]}$ can be defined analogously for each interval $[min_j; max_j]$ in the domain of a_{num}. The Boolean function is then set to true if the value of attribute a_{num} is within the respective range.

2.2 Patterns and Subgroups

Basic elements used in subgroup discovery are patterns and subgroups. Intuitively, a *pattern* describes a *subgroup*, i. e., the subgroup consists of instances that are covered by the respective pattern. It is easy to see, that a pattern describes a fixed set of instances (subgroup), while a subgroup can also be described by a set of patterns, if there are different options for covering the subgroup' instances. In the following, we define these concepts more formally.

Definition 1. *A* subgroup description *or (complex)* pattern sd *is given by a set of basic patterns* $sd = \{sel_1, \ldots, sel_l\}$ *, where* $sel_i \in S$*, which is interpreted as a conjunction, i.e.,* $sd(I) = sel_1 \wedge \ldots \wedge sel_l$*, with* $length(sd) = l$.

Without loss of generality, we focus on a conjunctive pattern language using nominal attribute–value pairs as defined above in this paper; internal disjunctions can also be generated by appropriate attribute–value construction methods, if necessary. We call a pattern sd' a *superpattern* (or *refinement*) of a *subpattern* sd, iff $sd \subset sd'$.

Definition 2. *A* subgroup (extension)

$$sg_{sd} := ext(sd) := \{i \in I | sd(i) = true\}$$

is the set of all individuals which are covered by the pattern sd*.*

As search space for subgroup discovery the set of all possible patterns 2^S is used, that is, all combinations of the basic patterns contained in S. Then, appropriate efficient algorithms, e. g., [19, 27, 58] can be applied.

2.3 Interestingness of a Pattern

A large number of quality functions has been proposed in literature, cf.. [41] for estimating the interestingness of a pattern – selected according to the analysis task.

Definition 3. *A* quality function $q\colon 2^S \to \mathbb{R}$ *maps every pattern in the search space to a real number that reflects the interestingness of a pattern (or the extension of the pattern, respectively).*

Many quality functions for a single target concept (e. g., binary [8, 50] or numerical [8, 56]), trade-off the size $n = |ext(sd)|$ of a subgroup and the deviation $t_{sd} - t_0$, where t_{sd} is the average value of a given target concept in the subgroup identified by the pattern sd and t_0 the average value of the target concept in the general population. In the binary case, the averages relate to the *share* of the target concept. Thus, typical quality functions are of the form

$$q_a(sd) = n^a \cdot (t_{sd} - t_0),\ a \in [0; 1]\,. \tag{1}$$

For binary target concepts, this includes, for example, the *weighted relative accuracy* for the size parameter $a = 1$ or a simplified binomial function, for $a = 0.5$. *Multi-target concepts*, e. g., [24,51,88] that define a target concept captured by a set of variables can be defined similarly, e. g., by extending an univariate statistical test to the multivariate case, e. g., [24]: Then, the multivariate distributions of a subgroup and the general population are compared in order to identify interesting (and exceptional) patterns.

While a quality function provides a *ranking* of the discovered subgroup patterns, often also a statistical assessment of the patterns is useful in data exploration. Quality functions that directly apply a statistical test, for example, the Chi-Square quality function, e. g., [8] provide a p-Value for simple interpretation. However, the Chi-Square quality function estimates deviations in two directions. An alternative, which can also be directly mapped to a p-Value is given by the *adjusted residual* quality function q_r, since the values of q_r follow a large standard normal distribution, cf. [1]:

$$q_r = n(p - p_0) \cdot \frac{1}{\sqrt{np_0(1-p_0)(1-\frac{n}{N})}} \tag{2}$$

The result of top-k subgroup discovery is the set of the k patterns $sd_1, \ldots, sd_k$, where $sd_i \in 2^S$ with the highest interestingness according to the applied quality function. A subgroup discovery task can now be specified by the 5-tuple: (D, c, S, q, k), where c indicates the target concept; the search space 2^S is defined by set of basic patterns S. In addition, we can consider constraints with respect to the *complexity* of the patterns. We can restrict the length l of the descriptions to a certain maximal value, e. g., with length $l = 1$ we only consider subgroup descriptions containing one selector, etc.

For several quality functions *optimistic estimates* [8, 19, 43, 56] can be applied for determining upper quality bounds: Consider the search for the k best subgroups: If it can be proven, that no subset of the currently investigated hypothesis is interesting enough to be included in the result set of k subgroups, then we can skip the evaluation of any subsets of this hypothesis, but can still guarantee the optimality of the result. More formally, an optimistic estimate $\mathrm{oe}(q)$ of a quality function q is a function such that $p \subseteq p' \to \mathrm{oe}(q(p)) \geq q(p')$, i. e., such that no refinement p' of the pattern p can exceed the quality obtained by $\mathrm{oe}(q(p))$.

3 Methods

With the rise of ubiquitous and mobile devices, social software and social media, a wealth of user-generated data is being created covering the according interactions in the respective systems an environments. In the following, we focus on social media and ubiquitous data: We adopt an intuitive definition of social media, regarding it as online systems and services in the ubiquitous web, which create and provide social data generated by human interaction and communication, cf. [4, 7].

In this context, exploratory analytics provides the means to get insights into a number of exemplary analysis options, e. g., focusing on social behavior in mobile social networks. In the context of ubiquitous and social enviroments, exploratory data analysis is therefore a rather important approach, e. g., for getting first insights into the data: Here, subgroup discovery and exceptional model mining are prominent methods that can be configured and adapted to various analytical tasks. As outlined above, subgroup discovery [2,8,50,87,88] has been established as a general and broadly applicable technique for descriptive and exploratory data mining: It aims at identifying descriptions of subsets of a dataset that show an interesting behavior with respect to certain interestingness criteria, formalized by a quality function. Standard subgroup discovery approaches commonly focus on a *single* target concept as the property of interest [87], that can already be applied for common analytical questions like deviations of some parameters. Furthermore, since the quality function framework also enables *multi-target concepts*, e. g., [8, 24, 51, 88] these enable even more powerful approaches for data analytics.

Figure 1 shows an overview on methods adapted and extended to the specific analytical tasks in the context of social media and ubiquitous data. Below, we discuss these in more detail, summarizing our recent work [17, 18, 24].

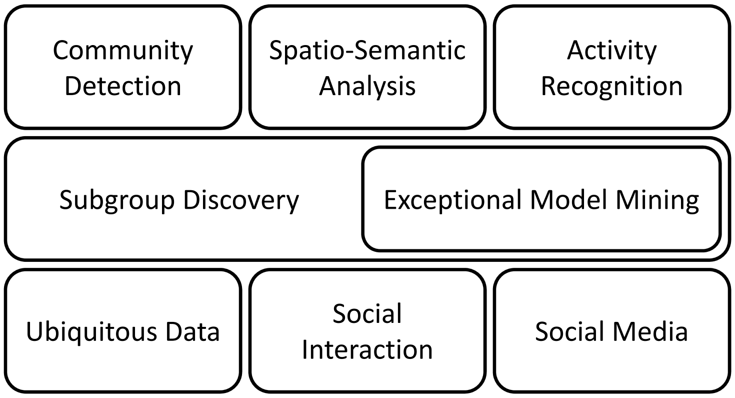


Fig. 1. Overview on the applied subgroup discovery and exceptional model mining approaches: We focus on the exploratory mining and analysis of social interaction in ubiquitous data and social media, tackling communities, human activities and behavior, and spatial-temporal characteristics, e. g., relating to events.

3.1 Description-Oriented Community Detection using Subgroup Discovery

Important inherent structures in ubiquitous and social environments are given by communities, cf. [72, 86]. Typically, these are seen as certain subsets of nodes of a graph with a dense structure. Classic community detection, e. g., [39] for a survey, just identifies subgroups of nodes with a dense structure, lacking an interpretable description. That is, no concise nor easily interpretable community description is provided.

In [17], we focus on *description-oriented community detection* using subgroup discovery. For providing both structurally valid and interpretable communities we utilize the graph structure as well as additional descriptive features of the graph's nodes. Using additional descriptive features of the nodes contained in the network, we approach the task of identifying communities as sets of nodes together with a *description*, i. e., a logical formula on the values of the nodes' descriptive features. Such a *community pattern* then provides an intuitive description of the community, e. g., by an easily interpretable conjunction of attribute-value pairs. Basically, we aim at identifying communities according to standard community quality measures, while providing characteristic descriptions at the same time.

As a simple example, we can consider a friendship graph common in online social systems. In the social bookmarking system BibSonomy[1] [35], for example, users can declare their friendship toward other users. This results in a directed graph, where nodes are denoted by users, and edges denote the friendship relations. Furthermore, in BibSonomy each user can tag resources like publications and web pages, i. e., assign a set of descriptive tags to certain resources. Then, the set of tags of a user can be considered as a description of that user's interests. Thus, the description-oriented community detection task in this context is to find user groups, where users are well connected given the friendship link structure, and also share a set of tags, as common features. Description-oriented community detection thus both needs to mine the graph-space and the description-space in an efficient way. In the following, we summarize the approach presented in [17], outlining the COMODO algorithm for fast description-oriented community detection, and present exemplary results.

Overview The COMODO algorithm for description-oriented community detection aims at discovering the top-k communities (described by community patterns) with respect to a number of standard community evaluation functions. The method is based on a generalized subgroup discovery approach [23, 57] adapted to attributed graph data, and also tackles typical problems that are not addressed by standard approaches for community detection such as pathological cases like small community sizes.

In [17] the approach is demonstrated on data sets from three social systems namely, i. e., from the social bookmarking systems BibSonomy and delicious[2], and from the social media platform last.fm[3]. However, the presented approach is not limited to such systems and can be applied to any kind of graph-structured data for which additional descriptive features (node labels) are available, e. g., certain activity in telephone networks, interactions in face-to-face contacts [16], and according edge-attributed graphs.

[1] http://www.bibsonomy.org

[2] http://www.delicious.com

[3] http://last.fm

Algorithm COMODO is a fast branch-and-bound algorithm utilizing optimistic estimates [43, 87] which are efficient to compute. This allows COMODO to prune the search space significantly, as we will see below.

As outlined above, COMODO utilizes both the graph structure, as well as descriptive information of the attributed graph, i. e., the label information of the nodes. This information is contained in two data structures: The graph structure is encoded in graph G while the attribute information is contained in database D describing the respective attribute values of each node. In a preprocessing step, we merge these data sources. Since the communities considered in our approach do not contain isolated nodes, we can describe them as sets of edges. We transform the data (of the given graph G and the database D containing the nodes' descriptive information) into a new data set focusing on the edges of the graph G: Each data record in the new data set represents an edge between two nodes. The attribute values of each such data record are the common attributes of the edge's two nodes. For a more detailed description, we refer to [17].

The FP-growth algorithm (cf. [44]) for mining association rules, the SD-Map* algorithm for fast exhaustive subgroup discovery [19], as well as quality functions operating on the graph structure form the basis of COMODO. COMODO utilizes an extended FP-tree structure, called the *community pattern tree* (CP-tree) to efficiently traverse the solution space. The tree is built in two scans of the graph data set and is then mined in a recursive divide-and-conquer manner, cf. [19, 57]. The CP-tree contains the frequent nodes in a header table, and links to all occurrences of the frequent basic patterns in the tree structure.

Algorithm 1 COMODO

```
procedure COMODO-Mine (cf. [17] for an extended description)
Input: Current community pattern tree $CPT$, pattern $\hat{p}$, priority queue top-k, int $k$ (max. number
    of patterns), int $maxLength$ (max. length of a pattern), int $\tau_n$ (min. community size)
$COM$ = new dictionary: $basicpattern \rightarrow pattern$
$minQ = minQuality(top\text{-}k)$
for all $b$ in $CPT$.getBasicPatterns do
    $p = createRefinement(\hat{p}, b)$
    $COM[b] = p$
    if $size(p, CPT) \geq \tau_n$ then
        if $quality(p, F) \geq minQ$ then
            $addToQueue(top\text{-}k, p)$
            $minQ = minQuality(top\text{-}k)$
if $length(\hat{p}) + 1 < maxLength$ then
    $refinements = sortBasicPatternsByOptimisticEstimateDescending(COM)$
    for all $b$ in $refinements$ do
        if $optimisticEstimate(COM[b]) \geq minQ$ then
            $CCPT = getConditionalCPT(b, CPT, minQ)$
            Call COMODO-Mine($CCPT$, $COM[b]$ , top-k)
```

The main algorithmic procedure of COMODO is shown in Algorithm 1. First, patterns containing only one basic pattern are mined. Then recursively, patterns conditioned on the occurrence of a (prefixed) complex pattern (as a set of basic patterns, chosen in the previous recursion step) are considered. For each following recursive step, a conditional CP-tree is constructed, given the conditional pattern base of a frequent basic pattern (CP-node). The conditional pattern base consists of all the prefix paths of such a CP-node, i.e., all the paths from the root node to the CP-node.

Given the conditional pattern base, a (smaller) CP-tree is generated: the *conditional CP-tree* of the respective CP-Node. If the conditional CP-tree just consists of one path, then the community descriptions can be generated by considering all the combinations of the nodes contained in the path. Otherwise, the new tree is subjected to the next recursion step. We refer to [44] for more details on CP-trees and FP-growth.

As shown in the algorithm, we consider three options for pruning and sorting according to the current optimistic estimates:

1. **Sorting**: During the iteration on the currently active basic pattern queue when processing a (conditional) CP-tree, we can dynamically reorder the basic patterns that have not been evaluated so far by their optimistic estimate value. In this way, we evaluate the *more promising* basic patterns first. This heuristic can help to obtain and to propagate higher values for the pruning threshold early in the process, thus, helping to prune larger portions of the search space (line 11).
2. **Pruning**: If the optimistic estimate for the conditioning basic pattern is below the threshold given by the k best community pattern qualities (line 13), then we can omit a branch.
3. **Pruning**: When building a (conditional) community pattern tree, we can omit all the CP-nodes with an optimistic estimate below the mentioned quality threshold (line 14).

To efficiently compute the community evaluation functions together with their optimistic estimates COMODO stores additional information in the *community pattern nodes* (CP-nodes) of the CP-tree, depending on the used quality function. Each CP-node of the CP-tree captures information about the aggregated edge information concerning the database D and the respective graph. For each node, we store the following information:

- The basic pattern (selector) corresponding to the attribute value of the CP-node. This selector describes the community (given by a set of edges) covering the CP-node.
- The edge count of the (partial) community represented by the CP-node, i.e., the aggregated count of all edges that are accounted for by the CP-node and its basic pattern, respectively.
- The set of nodes that are connected by the set of edges of the CP-node, i. e., the nodes making up the respective subgroup.

Each edge data record also stores the contributing nodes and their degrees (in- and out-degree in the directed case). Then, as outlined in [17] we can compute standard quality functions efficiently, e. g., for the *Modularity* [71–73] or the *Segregation Index* [40].

Exemplary Evaluation Results In our evaluation, we focused on two aspects: The efficiency of the proposed optimistic estimates, and the validity of the obtained community patterns. In order to evaluate the efficiency, we count the number of search steps, i. e., community allocations that are considered by the COMODO algorithm. We compared the total number of search steps (no optimistic estimate pruning) to optimistic estimate pruning using different commmunity quality measures. Additionally, we measured the impact of using different minimal community size thresholds. Exemplary results are shown in Figures 2–3 for the BibSonomy click graph and the delicious friend graph, for $k = 10, 20, 50$ and minimal size thresholds $\tau_n = 10, 20$. We consider a number of standard community quality functions: The *segregation index* [40], the *inverse average ODF (out degree fraction)* [59], and the *modularity* [71].

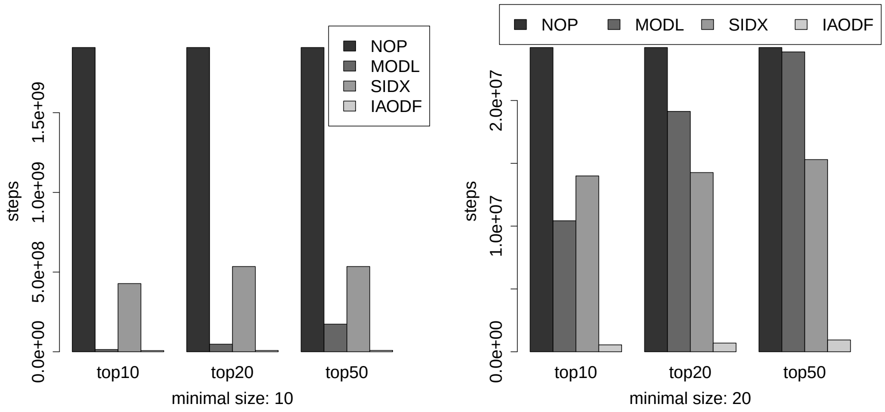


Fig. 2. Runtime performance of COMODO on BibSonomy click graph [17]: Search steps with no optimistic estimate pruning (*NOP*) vs. community quality functions with optimistic estimate pruning: MODL (Local Modularity), SIDX (Segregation Index) and IAODF (Inverse Average-ODF), for minimal size thresholds $\tau_n = 10, 20$.

The large, exponential search space can be exemplified, e. g., for the click graph with a total of about $2 \cdot 10^{10}$ search steps for a minimal community size threshold $\tau_n = 10$. The results demonstrate the effectiveness of the proposed descriptive mining approach applying the presented optimistic estimates. The implemented pruning scheme makes the approach scalable for larger data sets, especially when the local modularity quality function is chosen to assess the communities' quality. Concerning the validity of the patterns, we focused on structural properties of the patterns and the subgraphs induced by the respective comunity patterns. We applied the significance test described in [53] for testing the statistical significance of the density of a discovered subgraph. Furthermore, we compared COMODO to three baseline community detection algorithms [42,63,75], where COMODO shows a significantly better performance concerning validity and description length (for more details, we refer to [17]).

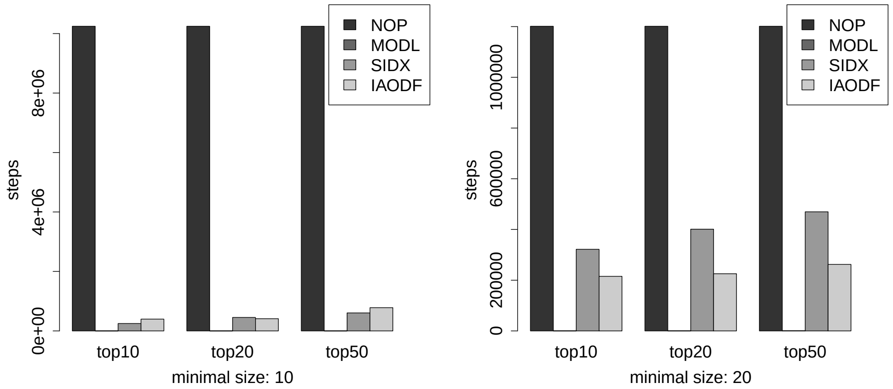


Fig. 3. Runtime performance of COMODO on the Delicious friend graph [17]: Search steps with no optimistic estimate pruning (*NOP*) vs. community quality functions with optimistic estimate pruning: MODL (Local Modularity), SIDX (Segregation Index) and IAODF (Inverse Average-ODF), for minimal size thresholds $\tau_n = 10, 20$.

Overall, the results of the structural evaluations indicate statistically valid and significant results. Also, these show that COMODO does not exhibit the typical problems and pathological cases such as small community sizes that are often encountered when using typical community mining methods. Furthermore, COMODO is able to detect communities that are typically captured by shorter descriptions leading to a lower description complexity, compared to the baselines, cf. [17].

3.2 Exceptional Model Mining for Spatio-Semantic Analysis

Ubiquitous data mining has many facets including descriptive approaches: These can help for obtaining a first overview on a dataset, for summarization, for uncovering a set of interesting patterns, analyzing their inter-relations [6, 24, 65, 66], and refinement [9]. Exploratory analysis on ubiquitous data needs to handle different heterogenous and complex data types, e. g., considering a combination of a dataset containing attributive and context information about certain data points with spatial and/or temporal information, cf. [46, 62, 79, 80]. Then, also semantic aspects concerning attributes, locations, and time can be considered.

In [6, 24], we present an adaptation of subgroup discovery using exceptional model mining formalizations on ubiquitous data – focusing the on spatio-semantic analysis in [24]: We consider subgroup discovery and assessment approaches for obtaining interesting descriptive patterns, cf. [28, 29, 32]. The proposed exploratory approach enables to obtain first insights into the spatio-semantic space. In the context of an environmental application, the presented approach provides for the detailed inspection and analysis of objective and subjective data and according measurements. Below, we sketch the approach presented in [24] and summarize illustrating results.

Overview The approach for exploratory subgroup analytics utilizes concepts of exceptional model mining in order to analyze complex target concepts on ubiquitous data. In particular, we focus on the interrelation between sensor measurements, subjective perceptions, and descriptive tags. Here, we propose a novel multi-target quality function for ranking the discovered subgroups, based on the Hotelling's T-squared test [45], see [24] for a detailed discussion.

Our application context is given by the *WideNoise Plus* smartphone application for measuring environmental noise. The individual data points include the measured noise in decibel (dB), associated subjective perceptions (feeling, disturbance, isolation, and artificiality) and a set of tags (free text) for providing an extended semantic context for the individual measurements. For the practical implementation, we utilize the VIKAMINE[4] tool [20] for subgroup discovery and analytics; it is complemented by methods of the *R* environment for statistical computing [77] in order to implement a semi-automatic pattern discovery process[5] based on automatic discovery and visual analysis methods.

Dataset – *WideNoise Plus* For the analysis, we utilize real-world data from the EveryAware project[6], specifically, on collectively organized noise measurements collected using the *WideNoise Plus* application between December 14, 2011 and June 6, 2014. *WideNoise Plus* allows the collection of noise measurements using smartphones, including noise level (dB) measured using the microphone, location (latitude/longitude), as well as a timestamp when the measurement was taken. In addition, when taking a measurement the user can add subjective information about the context perceptions, encoded in the interval $[-5;5]$ for *feeling*: [hate;love], *disturbance*: [hectic;calm], *isolation*: [alone;social], *artificiality*: [man-made;nature]. Furthermore, the user can assign tags to the measurement for additional descriptive information, e. g., "noisy", "indoor", or "calm", providing the semantic context of the specific measurement. The data are stored and processed by the backend based on the UBICON platform [13, 14].[7]

The applied dataset contains 6,600 data records and 2,009 distinct tags: The available tagging information was cleaned such that only tags with a length of at least three characters were considered. Only data records with valid tag assignments were included. Furthermore, we applied stemming and split multi-word tags into distinct single word tags.

Exemplary Analysis Results In our experiments, we initially performed some basic statistical analysis of the observed distributions as well as experiments on correlating the subjective and objective data. Doing that, we observed typical phenomena in the domain of tagging data, while the correlations are expressed on a medium level. This directly motivated the development and application of the proposed advanced techniques using our subgroup analytics approach. This allows us to focus on the relation between objective and subjective data given patterns of tagging data in more detail.

[4] `http://vikamine.org`

[5] `http://rsubgroup.org`

[6] `http://www.everyaware.eu`

[7] `http://www.ubicon.eu`

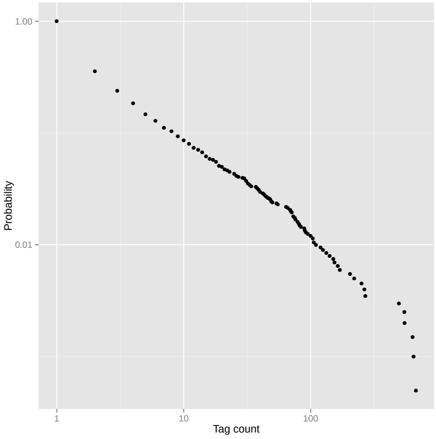


Fig. 4. Cumulated tag count distribution in the dataset. The y-axis provides the probability of observing a tag count larger than a certain threshold on the x-axis, cf. [24].

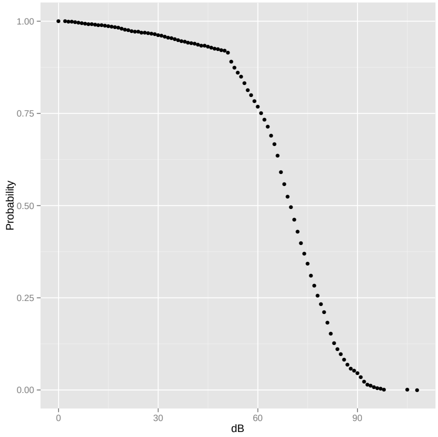


Fig. 5. Cumulated distribution of noise measurement (dB). The y-axis provides the probability for observing a measurement with a dB value larger than a certain threshold on the x-axis, cf. [24].

Figures 4-7 provide basic statistics about the tag count and measured noise distributions, as well as the value distributions of the perceptions and the number of tags assigned to a measurement. Figure 5 shows the distribution of the collected dB values, with a mean of 67.42 dB.

In Figure 6, we observe a typical heavy-tailed distributions of the tag assignments. Also, as can be observed in Figure 4 and Figure 7, the tag assignment data is rather sparse, especially concerning larger sets of assigned tags. However, it already allows to draw some conclusions on the tagging semantics and perceptions. In this context, the relation between (subjective) perceptions and (objective) noise measurements is of special interest. Table 1 shows the results of analyzing the correlation between the subjective and objective data. As shown in the table, we observe the expected trend that higher noise values correlate with the subjective “hate”, “hectic” or “man-made” situations. While the individual correlation values demonstrate only medium correlations, they are nevertheless statistically significant.

Table 1. Correlation analysis between subjective (perceptions) and objective (dB) measurements; all values are statistically significant ($p < 0.01$).

	Feeling	Disturbance	Isolation	Artificiality
dB	-0.27	-0.32	-0.32	0.19

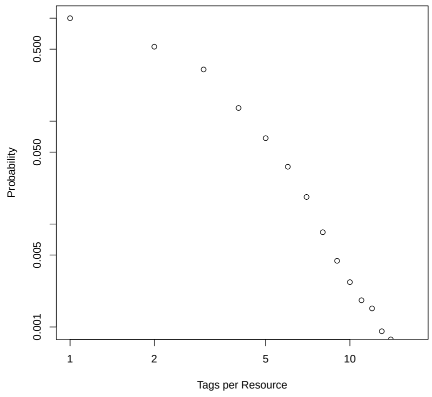


Fig. 6. Cumulated tag per record distribution in the dataset. The y-axis provides the probability of observing a tag per record count larger than a certain threshold on the x-axis, cf. [24].

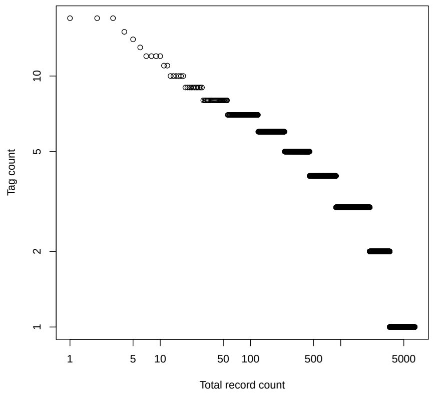
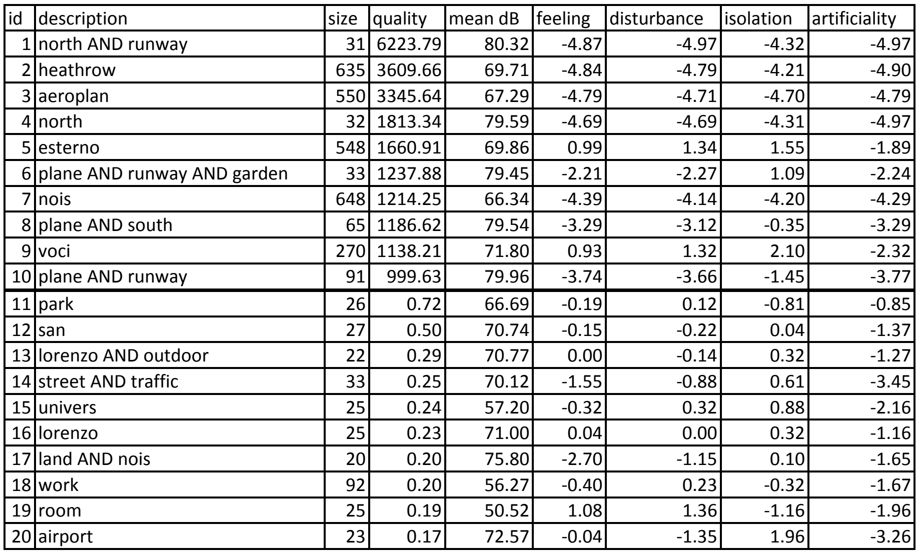


Fig. 7. Distribution of assigned tags per resource/data record, cf. [24].

For a detailed analysis, we first focused on subgroup patterns for hot-spots of low or high noise levels, i. e., on patterns that are characteristic for areas with low or high noise. We were able to identify several characteristic tags for noisy environments, for example, *north AND runway*, *heathrow*, and *aeroplan*, which relate to Heathrow noise monitoring case study, cf. [13] for more details. For more quiet environments, we also observed typical patterns, e. g., focusing on the tags *park*, *forest*, *outdoor*, and *room*, and combinations of these. Due to the limited space, we refer to [24] for more details on this analysis. We also extended the analysis in exploratory fashion by providing a semi-automatic approach for inspecting the geo-spatial characteristics of the discovered patterns by assessing their *geo-spatial* distribution in terms of its *peakiness* [90].

In the following, we focus on the discovery of subgroups with respect to a distinctive perception profile – relating to subjective *perception patterns* – which we describe in terms of their assigned tags. For analyzing the characteristics of the subjective data given by the perception values assigned to the individual measurements we applied the multi-target quality function q_H (based on the Hotelling's T-squared test), cf. [24]. This function allows us to detect exceptional subgroups, i. e., patterns that show a perception profile (given by the means of the individual perceptions) that is exceptionally different from the overall picture of the perceptions (respectively, their means estimated on the complete dataset). In addition, we also analyzed, which patterns show a rather "conforming" behavior to the overall mean values. For that, we applied the quality function $q'_H = \frac{1}{q_H}$. Using the reciprocal of q_H we could then identify patterns for which their deviation was quite small, i. e., close to the general trend in the complete dataset. Table 2 presents the obtained results, where the rows 1-10 in the table denote deviating patterns (q_H), while rows 11-20 show conforming patterns.

Table 2. Exemplary perception patterns [24]: rows 1-10 show deviating patterns, while rows 11-20 show conforming patterns. Overall means (perceptions): feeling=-0.83, disturbance=-0.64, isolation=-0.19, artificiality=-2.33. The table shows the size of the subgroups, their quality according to the applied quality function, the mean of the measured dB values, and the means of the individual perceptions.



id	description	size	quality	mean dB	feeling	disturbance	isolation	artificiality
1	north AND runway	31	6223.79	80.32	-4.87	-4.97	-4.32	-4.97
2	heathrow	635	3609.66	69.71	-4.84	-4.79	-4.21	-4.90
3	aeroplan	550	3345.64	67.29	-4.79	-4.71	-4.70	-4.79
4	north	32	1813.34	79.59	-4.69	-4.69	-4.31	-4.97
5	esterno	548	1660.91	69.86	0.99	1.34	1.55	-1.89
6	plane AND runway AND garden	33	1237.88	79.45	-2.21	-2.27	1.09	-2.24
7	nois	648	1214.25	66.34	-4.39	-4.14	-4.20	-4.29
8	plane AND south	65	1186.62	79.54	-3.29	-3.12	-0.35	-3.29
9	voci	270	1138.21	71.80	0.93	1.32	2.10	-2.32
10	plane AND runway	91	999.63	79.96	-3.74	-3.66	-1.45	-3.77
11	park	26	0.72	66.69	-0.19	0.12	-0.81	-0.85
12	san	27	0.50	70.74	-0.15	-0.22	0.04	-1.37
13	lorenzo AND outdoor	22	0.29	70.77	0.00	-0.14	0.32	-1.27
14	street AND traffic	33	0.25	70.12	-1.55	-0.88	0.61	-3.45
15	univers	25	0.24	57.20	-0.32	0.32	0.88	-2.16
16	lorenzo	25	0.23	71.00	0.04	0.00	0.32	-1.16
17	land AND nois	20	0.20	75.80	-2.70	-1.15	0.10	-1.65
18	work	92	0.20	56.27	-0.40	0.23	-0.32	-1.67
19	room	25	0.19	50.52	1.08	1.36	-1.16	-1.96
20	airport	23	0.17	72.57	-0.04	-1.35	1.96	-3.26

For comparison, the overall means of the perceptions are given by: feeling=-0.83, disturbance=-0.64, isolation=-0.19, artificiality=-2.33. As we can observe in the table, the deviating patterns tend to correspond to more *noisy* patterns; the majority of the patterns shows a dB value above the mean in the complete dataset (67.42 dB). Furthermore, most of the patterns relate to the Heathrow case study, e. g., *north AND runway*, *plane AND south*; an interesting pattern is given by *plane AND runway AND garden* – people living close to Heathrow obviously tend to measure noise often in their garden. For the *conforming* patterns we mostly observe patterns with a mean dB close to the general mean. However, interestingly there are some patterns that show an increased mean and also “unexpected” patterns, e. g., *street AND traffic* or *airport*.

Overall, these results confirm the trends that we observed in the statistical analysis above indicating a medium correlation of the perceptions with the noise patterns. However, combinations of descriptive tags, and the contributions of individual perceptions is only provided using advanced techniques, like the proposed subgroup discovery approach using a complex multi-target concept for the detection of local exceptional patterns. While the initial statistical analysis of the perceptions provides some initial insights on subjective and objective data, again these results motivate our proposed approach as a flexible and powerful tool for the analysis of subgroups and their relations in this spatio-semantic context. Further steps then include appropriate visualization and introspection techniques, e. g., [2, 8, 25, 28].

3.3 Class Association Rule Mining using Subgroup Discovery

With more and more ubiquitous devices, sensor data capturing human activities is becoming a universal data source for the analysis of human behavioral patterns. In particular, *activity recognition* has become a prominent research field with many successful methods for the classification of human activities. However, often the learned models are either "black-box" models such as neural networks, or are rather complex, e. g., in the case of random forests or large decision trees. In this context, we propose exploratory pattern analytics for constructing rule-based models in order to aid interpretation by humans, supported using appropriate quality and complexity measures [11,12].

Below, we summarize a novel approach for *class association rule mining* [60,61, 84,89] presented in [18]. We propose an *adaptive framework* for mining such rules using *subgroup discovery*, and demonstrate the effectiveness of our approach using real-world activity data collected using mobile phone sensors. We summarize the proposed approach and algorithmic framework, before we provide exemplary results of an evaluation using real world activity data obtained by mobile phone sensors. The effectiveness of the approach is demonstrated by a comparison with typical *descriptive* models, i. e., using a rule-based (*Ripper* [36]) and a decision tree classifier (*C4.5* [76]) as a baseline.

Overview Associative classification approaches integrate association rule mining and classification strategies. Basically, class association rules are special association rules with a fixed class attribute in the rule consequent. In order to mine such rules, we apply subgroup discovery. In the case of class association rules, the respective class can be defined as the target concept (i. e., the rule head) of the subgroups. Then, subgroup discovery can be adapted as a rule generator for class association rule mining.

In summary, in [18] we adapt subgroup discovery to class association rule mining, and embed it into an adaptive approach for obtaining a rule set that aims to target a simple rule base with an adequate level of predictive power, i. e., combining simplicity and accuracy. We utilize standard methods of rule selection and evaluation, that can be integrated into our framework: Liu et al. [61], for example, propose the *CBA* algorithm, which includes association rule mining and subsequent rule selection. It applies a covering strategy, selecting rules one by one, minimizing the total error. In addition to the rule mining and selection techniques, there are several strategies for the final decision of how to combine rules for classification ("voting" of the matching rules), e. g., [82].

Algorithmic Framework For our adaptive framework, we distinguish the *learning phase* that constructs the model, and the *classification phase* that applies the model.

Model Construction For the construction of the model, we apply the steps described in Algorithm 2. Basically, CARMA starts with discovering class association rules for each class c contained in the dataset. Using subgroup discovery, we collect a set of class association rules for the specific class, considering a maximal length of the concerned patterns. After that, we apply a boolean *ruleset assessment* function a in order to check, if the quality of the ruleset is good enough. If the outcome of this test is positive, we continue with the next class. Otherwise, we increase the maximal $length$ of a rule (up to a certain user-definable threshold $\mathcal{T}_l$). After the final set of all class association rules for all classes has been determined, we apply the *rule selection function* r in order to obtain a set of class association rules that optimizes predictive power on the trainingset. That is, the rule selection function aims to estimate classification error and should select the rules according to coverage and accuracy of the rules on the trainingset.

Algorithm 2 CARMA: Framework for Adaptive Class Association Rule Mining [18]

```
Input: Database $D$, set of classes $C$, parameter $k$ specifying the cardinality of top-$k$ pattern set,
    parameter $\mathcal{T}_l$ denoting the maximal possible length of a subgroup pattern, quality function $q$,
    ruleset assessment function $a$, rule selection function $r$.
Patterns $P = \emptyset$
for all $c \in C$ do
    Current length threshold $length = 1$
    while true do
        Obtain candidate patterns $P^*$ by $SubgroupDiscovery(D, c, S, q, k)$
        if Current candidate patterns are good enough, i. e., $a(P^*) = true$ then
            $P = P \cup P^*$
            break
        else if $length > \mathcal{T}_l$ then
            break
        else
            $length = length + 1$
Add a default pattern (rule) for the most frequent class to $P$
Apply rule selection function: $P = r(P)$
return P
```

Classification In the classification phase, we apply the rules contained in a model. For aggregating the predictions of the matching rules, we apply a specific *rule combination* strategy, cf. [82]. Examples include *unweighted voting* (majority vote according to the matching rules for the respective class), *weighted voting* (including weights for the matching rules), or *best rule* (classification according to the matching rule with the highest confidence).

Summary In contrast to existing approaches, the CARMA framework is based on subgroup discovery for class association rule mining. This allows for selection of a suitable quality function for generating the rules, in constrast to (simple) confidence/support-based approaches. Then, e. g., significance criteria can be easily integrated. Furthermore, CARMA applies an adaptive strategy for balancing rule complexity (size) with predictive accuracy by applying a ruleset assessment function, in addition to the rule selection function. The framework itself does not enforce a specific strategy, but leaves this decision to a specific configuration. In our implementation in [18], for example, we follow the rule selection strategy of CBA; the ruleset assessment is done by a median-based ranking of the according confidences of the rules, i. e., estimated by the respective shares of the class contained in the subgroup covered by the respective rule. Here, we test if the median of the rules' confidences is above a certain threshold $\tau_c = 0.5$.

Exemplary Evaluation Results In [18] we compared an instantiation of the CARMA framework against two baselines: The Ripper algorithm [36] as a rule-based learner, and the C4.5 algorithm [76] for learning decision trees. For the subgroup discovery step in the CARMA framework, we apply the BSD algorithm [58], utilizing the *adjusted residual* quality function, cf. Section 2, which directly maps to significance criteria. Furthermore, we apply an adaptation of the CBA algorithm [61] for the rule selection function, We opted for interpretable patterns with a maximal length of 7 conditions, and set the respective threshold $\mathcal{T}_l = 7$ accordingly. In the evaluation, we used three different *TopK* values: 100, 200 and 500. For the rule combination strategy, we experimented with four strategies: taking the best rule according to confidence and Laplace value, the unweighted voting strategy, and the weighted voting (Laplace) method, cf. [18, 82] for a detailed discussion. All experiments were performed using 10-fold cross-validation on an activity dataset with 27 activities (classes) and 116 features, cf. [18] for details.

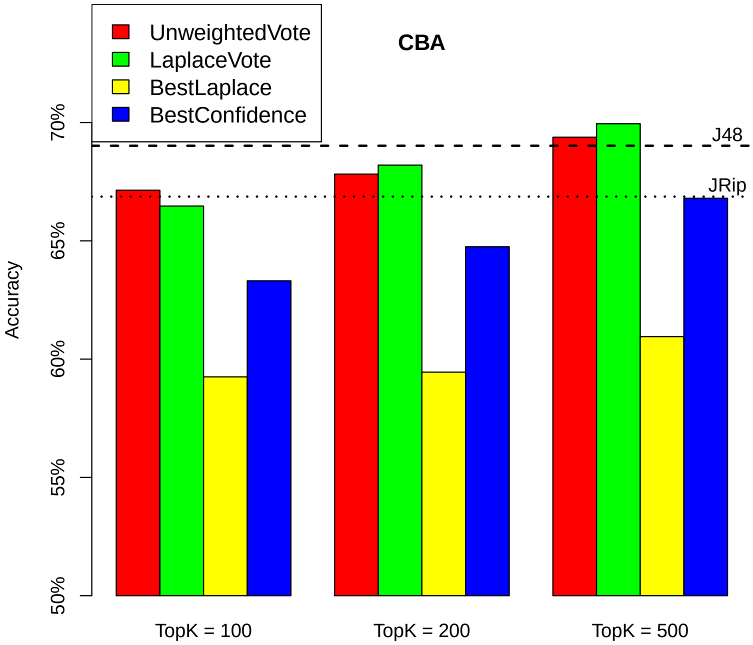


Fig. 8. Comparison of the accuracy of CARMA using the standard CBA method for rule selection, with different rule combination strategies to the baselines, cf. [18].

Figure 8 shows the accuracy of CARMA using these parametrizations. Overall, it is easy to see that the proposed approach is able to outperform the baselines in accuracy. Furthermore, it outperformed both as well in complexity, since it always had a significantly lower average complexity regarding the average number of conditions in a rule. For the baselines, C4.5 showed a better performance than Ripper, however, with a more complex model (1394 rules) that were also more complex themselves; Ripper had a slightly lower accuracy but a signicantly lower number of rules and average rule length. The proposed CARMA approach outperforms both concerning the combination of accuracy and simplicity. Considering the voting functions, we observe that the functions (unweighted voting, and weighted Laplace) always outperforms the rest. In our experiments, using larger values of k indicates a higher accuracy – here also the compexity (in the number of rules) can be tuned. We observe a slight trade-off between accuracy and complexity. Basically, the parameter k seems to have an influence on the complexity, while the remaining instantiations do not seem to have a strong influence.

In summary, the proposed framework always provides a more compact model than the baseline algorithms. In our experiments, it is at least in the same range or even better than the baselines. Considering the best parameter instantiation, the proposed approach is able to outperform both baselines concerning the accuracy and always provides a more compact model concerning rule complexity, cf. [18] for more details.

4 Conclusions and Outlook

Subgroup discovery and exceptional model mining provide powerful and comprehensive methods for knowledge discovery and exploratory analyis. In this paper, we summarized recent advances concerning according approaches and methods in the context of ubiquitous data and social media. Specifically, we focused on exploratory pattern analytics implemented using subgroup discovery and exceptional model mining methods, summarizing recent work on description-oriented community detection, spatio-semantic analysis using local exceptionality detection, and class association rule mining using subgroup discovery. The methods were embedded into evaluations and case studies demonstrating their theoretical as well as practical impact and implications.

Interesting future directions include the adaptation and extension of knowledge-intensive approaches, e. g., [22, 26, 30, 31, 54, 69, 85]. This also concerns the incorporation of multiple relations, e. g., in the form of *partitioning knowledge* [10], or making use of multiplex and multi-modal networks [47, 64, 70, 78, 83], for modeling complex relations on ubiquitous data and social media and the analysis of emerging semantics [15, 65, 66]. Furthermore, the extended analysis of sequential data can be applied in both spatio-temporal dimensions [74], also concerning dynamics in the spatio-temporal space, e. g., for an extended temporal modeling of ubiquitous relations [48, 49]: Here, possible methods for extension and adaptation include temporal pattern mining for event detection [34], or temporal subgroup analytics [81], especially considering sophisticated exceptional model classes in that area. In addition, for including dynamics of spatial and temporal properties, for example, Markov chain approaches can be extended towards exceptional model mining, e. g., for modeling and analyzing sequential hypotheses and trails [33] in order to detect exceptional sequential transition patterns.

References

1. Agresti, A.: An Introduction to Categorical Data Analysis. Wiley-Blackwell (2007)
2. Atzmueller, M.: Knowledge-Intensive Subgroup Mining – Techniques for Automatic and Interactive Discovery, Dissertations in Artificial Intelligence-Infix (Diski), vol. 307. IOS Press (March 2007)
3. Atzmueller, M.: Mining Social Media. Informatik Spektrum 35(2), 132 – 135 (2012)
4. Atzmueller, M.: Mining Social Media: Key Players, Sentiments, and Communities. WIREs: Data Mining and Knowledge Discovery 1069 (2012)
5. Atzmueller, M.: Onto Collective Intelligence in Social Media: Exemplary Applications and Perspectives. In: Proc. 3rd International Workshop on Modeling Social Media (MSM 2012), Hypertext 2012. ACM Press, New York, NY, USA (2012)
6. Atzmueller, M.: Data Mining on Social Interaction Networks. Journal of Data Mining and Digital Humanities 1 (June 2014)
7. Atzmueller, M.: Social Behavior in Mobile Social Networks: Characterizing Links, Roles and Communities. In: Chin, A., Zhang, D. (eds.) Mobile Social Networking: An Innovative Approach, pp. 65–78. Computat. Social Sciences, Springer, Heidelberg, Germany (2014)
8. Atzmueller, M.: Subgroup Discovery – Advanced Review. WIREs: Data Mining and Knowledge Discovery 5(1), 35–49 (2015)
9. Atzmueller, M., Baumeister, J., Hemsing, A., Richter, E.J., Puppe, F.: Subgroup Mining for Interactive Knowledge Refinement. In: Proc. 10th Conference on Artificial Intelligence in Medicine (AIME 05). pp. 453–462. LNAI 3581, Springer, Heidelberg, Germany (2005)
10. Atzmueller, M., Baumeister, J., Puppe, F.: Evaluation of two Strategies for Case-Based Diagnosis handling Multiple Faults. In: Proc. 2nd Conf. Professional Knowledge Management (WM2003). Luzern, Switzerland (2003)
11. Atzmueller, M., Baumeister, J., Puppe, F.: Quality Measures and Semi-Automatic Mining of Diagnostic Rule Bases. In: Proc. 15th International Conference on Applications of Declarative Programming and Knowledge Management (INAP 2004). pp. 65–78. No. 3392 in LNAI, Springer, Heidelberg, Germany (2005)
12. Atzmueller, M., Baumeister, J., Puppe, F.: Semi-Automatic Learning of Simple Diagnostic Scores Utilizing Complexity Measures. Artificial Intelligence in Medicine. Special Issue on Intelligent Data Analysis in Medicine 37(1), 19–30 (2006)
13. Atzmueller, M., Becker, M., Doerfel, S., Kibanov, M., Hotho, A., Macek, B.E., Mitzlaff, F., Mueller, J., Scholz, C., Stumme, G.: Ubicon: Observing Social and Physical Activities. In: Proc. IEEE International Conference on Cyber, Physical and Social Computing (CPSCom). pp. 317–324. IEEE Computer Society, Washington, DC, USA (2012)
14. Atzmueller, M., Becker, M., Kibanov, M., Scholz, C., Doerfel, S., Hotho, A., Macek, B.E., Mitzlaff, F., Mueller, J., Stumme, G.: Ubicon and its Applications for Ubiquitous Social Computing. New Review of Hypermedia and Multimedia 20(1), 53–77 (2014)
15. Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Towards Mining Semantic Maturity in Social Bookmarking Systems. In: Proc. Workshop on Social Data on the Web, 10th International Semantic Web Conference (ISWC. Bonn, Germany (2011)
16. Atzmueller, M., Doerfel, S., Hotho, A., Mitzlaff, F., Stumme, G.: Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In: Modeling and Mining Ubiquitous Social Media, LNAI, vol. 7472. Springer, Heidelberg, Germany (2012)
17. Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Information Sciences 329, 965–984 (2016)
18. Atzmueller, M., Kibanov, M., Hayat, N., Trojahn, M., Kroll, D.: Adaptive Class Association Rule Mining for Human Activity Recognition. In: Proc. 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE), ECML/PKDD. Porto, Portugal (2015)

19. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. International Symposium on Methodologies for Intelligent Systems. LNCS, vol. 5722, pp. 1–15. Springer, Heidelberg, Germany (2009)
20. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer, Heidelberg, Germany (2012)
21. Atzmueller, M., Lemmerich, F.: Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. International Journal of Web Science 2(1/2) (2013)
22. Atzmueller, M., Lemmerich, F., Reutelshoefer, J., Puppe, F.: Wiki-Enabled Semantic Data Mining - Task Design, Evaluation and Refinement. In: Proc. 2nd International Workshop on Design, Evaluation and Refinement of Intelligent Systems (DERIS2009). CEUR-WS, vol. 545. Krakow, Poland (2009)
23. Atzmueller, M., Mitzlaff, F.: Efficient Descriptive Community Mining. In: Proc. 24th International FLAIRS Conference. pp. 459 – 464. AAAI Press, Palo Alto, CA, USA (2011)
24. Atzmueller, M., Mueller, J., Becker, M.: Mining, Modeling and Recommending 'Things' in Social Media, chap. Exploratory Subgroup Analytics on Ubiquitous Data. No. 8940 in LNAI, Springer, Heidelberg, Germany (2015)
25. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science 11(11), 1752–1765 (2005)
26. Atzmueller, M., Puppe, F.: A Methodological View on Knowledge-Intensive Subgroup Discovery. In: Proc. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2006). pp. 318–325. No. 4248 in LNAI, Springer, Heidelberg, Germany (2006)
27. Atzmueller, M., Puppe, F.: SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). pp. 6–17. Springer, Heidelberg, Germany (2006)
28. Atzmueller, M., Puppe, F.: A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence 28(3), 210–221 (2008)
29. Atzmueller, M., Puppe, F.: Semi-Automatic Refinement and Assessment of Subgroup Patterns. In: Proc. 21th International Florida Artificial Intelligence Research Society Conference. pp. 518–523. AAAI Press, Palo Alto, CA, USA (2008)
30. Atzmueller, M., Puppe, F., Buscher, H.P.: Towards Knowledge-Intensive Subgroup Discovery. In: Proc. LWA 2004, Workshop KDML, Germany. pp. 117–123 (2004)
31. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI). pp. 647–652. Edinburgh, Scotland (2005)
32. Atzmueller, M., Puppe, F., Buscher, H.P.: Profiling Examiners using Intelligent Subgroup Mining. In: Proc. 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005). pp. 46–51. Aberdeen, Scotland (2005)
33. Atzmueller, M., Schmidt, A., Kibanov, M.: DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In: Proc. WWW 2016 (Companion). IW3C2 / ACM (2016)
34. Batal, I., Fradkin, D., Harrison, J., Moerchen, F., Hauskrecht, M.: Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. In: Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 280–288. KDD '12, ACM, New York, NY, USA (2012)
35. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The Social Bookmark and Publication Management System BibSonomy. VLDB 19, 849 – 875 (2010)

36. Cohen, W.W.: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning. pp. 115–123. Morgan Kaufmann (1995)
37. Duivesteijn, W., Knobbe, A., Feelders, A., van Leeuwen, M.: Subgroup Discovery Meets Bayesian Networks–An Exceptional Model Mining Approach. In: Proc. International Conference on Data Mining (ICDM). pp. 158–167. IEEE, Washington, DC, USA (2010)
38. Duivesteijn, W., Feelders, A., Knobbe, A.J.: Different Slopes for Different Folks: Mining for Exceptional Regression Models with Cook's Distance. In: Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 868–876. ACM, New York, NY, USA (2012)
39. Fortunato, S.: Community Detection in Graphs. Physics Reports 486(3-5), 75 – 174 (2010)
40. Freeman, L.: Segregation In Social Networks. Sociological Methods & Research 6(4), 411 (1978)
41. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3) (2006)
42. Gregory, S.: Finding Overlapping Communities in Networks by Label Propagation . New J. Phys. (12) (2010)
43. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight Optimistic Estimates for Fast Subgroup Discovery. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. LNCS, vol. 5211, pp. 440–456. Springer, Heidelberg, Germany (2008)
44. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns Without Candidate Generation. In: Chen, W., Naughton, J., Bernstein, P.A. (eds.) Proc. SIGMOD. pp. 1–12. ACM Press (05 2000)
45. Hotelling, H.: The Generalization of Student's Ratio. Ann. Math. Statist. 2(3), 360–378 (1931)
46. Hotho, A., Ulslev Pedersen, R., Wurst, M.: Ubiquitous Data. In: Ubiquitous Knowledge Discovery, pp. 61–74. No. 6202 in Lecture Notes in Computer Science, Springer (2010)
47. Kibanov, M., Atzmueller, M., Illig, J., Scholz, C., Barrat, A., Cattuto, C., Stumme, G.: Is Web Content a Good Proxy for Real-Life Interaction? A Case Study Considering Online and Offline Interactions of Computer Scientists. In: Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE Press, Boston, MA, USA (2015)
48. Kibanov, M., Atzmueller, M., Scholz, C., Stumme, G.: On the Evolution of Contacts and Communities in Networks of Face-to-Face Proximity. In: Proc. IEEE International Conference on Cyber, Physical and Social Computing (CPSCom). IEEE Computer Society, Boston, MA, USA (2013)
49. Kibanov, M., Atzmueller, M., Scholz, C., Stumme, G.: Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions. Science China Information Sciences 57 (March 2014)
50. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI Press (1996)
51. Klösgen, W.: Handbook of Data Mining and Knowledge Discovery, chap. 16.3: Subgroup Discovery. Oxford University Press, New York (2002)
52. Knobbe, A.J., Cremilleux, B., Fürnkranz, J., Scholz, M.: From Local Patterns to Global Models: The LeGo Approach to Data Mining. In: From Local Patterns to Global Models: Proceedings of the ECML/PKDD-08 Workshop (LeGo-08). pp. 1 – 16 (2008)
53. Koyuturk, M., Szpankowski, W., Grama, A.: Assessing Significance of Connectivity and Conservation in Protein Interaction Networks. Journal of Computational Biology 14(6), 747–764 (2007)

54. Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I., Kralj Novak, P.: Using Ontologies in Semantic Data Mining with SEGS and g-SEGS. In: Proceedings of the 14th International Conference on Discovery Science (DS) (2011)
55. Leman, D., Feelders, A., Knobbe, A.: Exceptional Model Mining. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 5212, pp. 1–16. Springer (2008)
56. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery (2015), http://dx.doi.org/10.1007/s10618-015-0436-8
57. Lemmerich, F., Becker, M., Atzmueller, M.: Generic Pattern Trees for Exhaustive Exceptional Model Mining. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer, Heidelberg, Germany (2012)
58. Lemmerich, F., Rohlfs, M., Atzmueller, M.: Fast Discovery of Relevant Subgroup Patterns. In: Proc. Intl. FLAIRS Conference. pp. 428–433. AAAI Press, Palo Alto, CA, USA (2010)
59. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. CoRR abs/0810.1355 (2008)
60. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: Cercone, N., Lin, T.Y., Wu, X. (eds.) Proc. International Conference on Data Mining (ICDM). pp. 369–376. IEEE Computer Society (2001)
61. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 80–86. AAAI Press (August 1998)
62. May, M., Berendt, B., Cornuéjols, A., Gama, J., Giannotti, F., Hotho, A., Malerba, D., Menesalvas, E., Morik, K., Pedersen, R., et al.: Research Challenges in Ubiquitous Knowledge Discovery. Next Generation of Data Mining pp. 131–150 (2008)
63. McDaid, A., Hurley, N.: Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining. pp. 112–119. ASONAM '10, IEEE Computer Society, Washington, DC, USA (2010)
64. Mitzlaff, F., Atzmueller, M., Benz, D., Hotho, A., Stumme, G.: Community Assessment using Evidence Networks. In: Analysis of Social Media and Ubiquitous Data. LNAI, vol. 6904 (2011)
65. Mitzlaff, F., Atzmueller, M., Hotho, A., Stumme, G.: The Social Distributional Hypothesis. Journal of Social Network Analysis and Mining 4(216) (2014)
66. Mitzlaff, F., Atzmueller, M., Stumme, G., Hotho, A.: Semantics of User Interaction in Social Media. In: Ghoshal, G., Poncela-Casasnovas, J., Tolksdorf, R. (eds.) Complex Networks IV, Studies in Computational Intelligence, vol. 476. Springer, Heidelberg, Germany (2013)
67. Morik, K.: Detecting Interesting Instances. In: Hand, D., Adams, N., Bolton, R. (eds.) Pattern Detection and Discovery, Lecture Notes in Computer Science, vol. 2447, pp. 13–23. Springer Berlin Heidelberg (2002)
68. Morik, K., Boulicaut, J., Siebes, A. (eds.): Local Pattern Detection, International Seminar, Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers, Lecture Notes in Computer Science, vol. 3539. Springer (2005)
69. Morik, K., Potamias, G., Moustakis, V., Charissis, G.: Knowledgeable Learning using MOBAL: A Medical Case Study. Applied Artificial Intelligence 8(4), 579–592 (1994)
70. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community Structure in Time-dependent, Multiscale, and Multiplex Networks. Science 328(5980), 876–878 (2010)
71. Newman, M.E., Girvan, M.: Finding and Evaluating Community Structure in Networks. Phys Rev E Stat Nonlin Soft Matter Phys 69(2), 1–15 (2004)

72. Newman, M.E.J.: Detecting Community Structure in Networks. Europ Physical J 38 (2004)
73. Newman, M.E.J.: Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
74. Piatkowski, N., Lee, S., Morik, K.: Spatio-temporal Random Fields: Compressible Representation and Distributed Estimation. Machine Learning 93(1), 115–139 (2013)
75. Pool, S., Bonchi, F., van Leeuwen, M.: Description-driven Community Detection. Transactions on Intelligent Systems and Technology 5(2) (2014)
76. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
77. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009), `http://www.R-project.org`, ISBN 3-900051-07-0
78. Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C., Stumme, G.: New Insights and Methods For Predicting Face-To-Face Contacts. In: Kiciman, E., Ellison, N.B., Hogan, B., Resnick, P., Soboroff, I. (eds.) Proc. International AAAI Conference on Weblogs and Social Media. AAAI Press, Palo Alto, CA, USA (2013)
79. Scholz, C., Atzmueller, M., Stumme, G.: Unsupervised and Hybrid Approaches for On-Line RFID Localization with Mixed Context Knowledge. In: Proc. 21st Intl. Symposium on Methodologies for Intelligent Systems. Springer, Heidelberg, Germany (2014)
80. Scholz, C., Doerfel, S., Atzmueller, M., Hotho, A., Stumme, G.: Resource-Aware On-Line RFID Localization Using Proximity Data. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 129–144. Springer, Heidelberg, Germany (2011)
81. Sáez, C., Rodrigues, P., Gama, J., Robles, M., García-Gómez, J.: Probabilistic Change Detection and Visualization Methods for the Assessment of Temporal Stability in Biomedical Data Quality. Data Mining and Knowledge Discovery pp. 1–26 (2014)
82. Sulzmann, J.N., Fürnkranz, J.: A Comparison of Techniques for Selecting and Combining Class Association Rules. In: Baumeister, J., Atzmueller, M. (eds.) Proc. LWA. Technical Report, vol. 448, pp. 87–93. Department of Computer Science, University of Würzburg, Germany (2008)
83. Symeonidis, P., Perentis, C.: Link Prediction in Multi-Modal Social Networks. In: Machine Learning and Knowledge Discovery in Databases, pp. 147–162. Springer (2014)
84. Thabtah, F.: A Review of Associative Classification Mining. Knowl. Eng. Rev. 22(1), 37–65 (Mar 2007)
85. Vavpetič, A., Lavrač, N.: Semantic Subgroup Discovery Systems and Workflows in the SDM-Toolkit. The Computer Journal 56(3), 304–320 (2013)
86. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. No. 8 in Structural Analysis in the Social Sciences, Cambridge University Press, 1 edn. (1994)
87. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery. pp. 78–87. Springer, Heidelberg, Germany (1997)
88. Wrobel, S., Morik, K., Joachims, T.: Maschinelles Lernen und Data Mining. Handbuch der Künstlichen Intelligenz 3, 517–597 (2000)
89. Yin, X., Han, J.: CPAR: Classification based on Predictive Association Rules. In: Barbará, D., Kamath, C. (eds.) Proc. SIAM International Conference on Data Mining (SDM). pp. 331–335. SIAM (2003)
90. Zhang, H., Korayem, M., You, E., Crandall, D.J.: Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities. In: Proc. International Conference on Web Search and Data Mining (WSDM). pp. 33–42. ACM, New York, NY, USA (2012)