Exploratory Subgroup Analytics on Ubiquitous Data

Martin Atzmueller[1](✉), Juergen Mueller[1], and Martin Becker[2]

[1] Knowledge and Data Engineering Group, University of Kassel, Kassel, Germany
{atzmueller,mueller}@cs.uni-kassel.de
[2] Data Mining and Information Retrieval Group,
University of Würzburg, Würzburg, Germany
becker@informatik.uni-wuerzburg.de

Abstract. This paper presents exploratory subgroup analytics on ubiquitous data: We propose subgroup discovery and assessment approaches for obtaining interesting descriptive patterns and provide a novel graph-based analysis approach for assessing the relations between the obtained subgroup set. This exploratory visualization approaches allows for the comparison of subgroups according to their relations to other subgroups and to include further parameters, e.g., geo-spatial distribution indicators. We present and discuss analysis results utilizing real-world data given by geo-tagged noise measurements with associated subjective perceptions and a set of tags describing the semantic context.

1 Introduction

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, and analyzing their inter-relations.

In this paper, we propose exploratory subgroup analytics on ubiquitous data. Subgroup discovery is a versatile method for descriptive data mining. We extend it in two analytical directions focusing on the applied quality functions and the relations between subgroups. First, we propose novel quality functions for estimating the quality of subgroups in the multivariate setting. In addition, we propose a novel graph-based approach for assessing sets of subgroups including multiple quality criteria. Specifically, we focus on the interrelation between sensor measurements, subjective perceptions, and descriptive tags. For assessing the relations between the result set of subgroups, we propose a novel graph-based analysis approach. This method is applied for visualizing subgroup relations, and can be utilized for comparing subgroups according to their relationships to other subgroups. The automatic discovery and visual analysis methods complement each other in exploratory fashion: The quality function used for ranking and estimating the quality of subgroups and the relationship function can be selected according to analysis goals. In addition, further subgroup parameters can be shown in the visualization using visual markers. Here, we specifically present an adapted technique for deriving characteristic indicators of the geo-spatial distribution of ubiquitous data in the context of subgroup analytics.

M. Atzmueller et al. (Eds.): MUSE/MSM 2013, LNAI 8940, pp. 1–20, 2015.
DOI: 10.1007/978-3-319-14723-9_1

Overall, we analyze real-world sensor data with associated semantic information and subjective measurements. For that, we utilize the VIKAMINE[1] tool [8] for subgroup discovery and analytics; it is complemented by methods of the *R* environment for statistical computing [34] in order to implement a semi-automatic pattern discovery process.[2]

For the analysis, we apply real-world data from the EveryAware project:[3] Our application context is given by the *WideNoise* smartphone app 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 for providing semantic context for the individual measurements. We present results analyzing subgroups patterns for hot-spots of low/high noise levels. Our results indicate, that there are indeed distinctive patterns in terms of descriptive tags. Furthermore, we analyze the characteristics of subgroups according to their geo-spatial distribution given by the covered set of noise measurements. In addition to investigating patterns that are characteristic for areas with low or high noise, we also analyze subgroups with respect to a distinctive perception profile – relating to subjective *perception patterns* – which we describe in terms of their assigned tags.

Our contribution can be summarized as follows:

1. We present an exploratory subgroup analytics approach covering the subgroup assessment and specifically the relations between subgroups. For this, we propose flexible quality functions and a set of relationship functions that are used to model dependencies and relations in a set of subgroups.
2. For the proposed automatic exploratory approach, we present a novel visualization method complementing the automatic methods. The presented visualization method allows to inspect subgroup relations and further influence parameters in context.
3. We provide an adapted technique for deriving geo-spatial distributional indicators in the context of subgroup analytics.
4. Finally, we present an analysis of ubiquitous data using subgroup discovery methods utilizing data from a real-world application. We describe a case study applying the proposed approaches and discuss the results in detail.

The remainder of the paper is organized as follows: Sect. 2 discusses related work. Then, Sect. 3 introduces necessary basic notions. Next, Sect. 4 proposes the novel approach for graph-based subgroup analytics. After that, Sect. 5 describes the applied dataset, presents the experiments, and discusses the results in our application setting. Finally, Sect. 6 concludes with a summary and presents interesting options for future work.

2 Related Work

Ubiquitous data mining covers many subfields, including spatio-temporal data mining [27], mining sensor data or mining social media with geo-referenced data,

[1] http://vikamine.org.

[2] http://rsubgroup.org.

[3] http://everyaware.eu.

c.f., [4]. Applications include destination recommenders, e.g., for tourist information systems [17], or geographical topic discovery [43]. Often established problem statements and methods have been transferred to this setting, for example, considering association rules [3]. Related approaches consider, for example, social image mining methods, cf., [33] for a survey.

In this area, specifically considering social image data, there have been several approaches, and the problem of generating representative tags for a given set of images is an active research topic. Reference [40] analyze Flickr data and provide a characterization on how users apply tags and which information is contained in the tag assignments. Their approach is embedded into a recommendation method for photo tagging, similar to [32] who analyze different aspects and contexts of the tag and image data. Reference [1] present a method to identify landmark photos using tags and social Flickr groups. They apply group information and statistical preprocessing of the tags for obtaining interesting landmark photos.

The concept of collecting information in ubiquitous systems, especially for crowd-sourced and citizen-driven applications is discussed in [36]. Basic issues of measuring noise pollution using mobile phones are presented in [38]. From a distributional point of view, the proposed approach seamlessly generalizes similar ones for analyzing event and place semantics using a user-specified quality function. Then we can capture techniques, e.g., for burst detection [24,35] or the analysis of peaks of temporal and spatial distributions [44]. Furthermore, such techniques can be incorporated in our exploratory analytics approach using visualization techniques.

In contrast to the approaches discussed above, this paper focuses on descriptive patterns. This allows for the flexible adaptation to the preferences of the users, since their interestingness can be flexibly tuned by selecting an appropriate quality function and target concept. There are several variants of pattern mining techniques, e.g., frequent pattern mining [20], mining association rules [2,28], and closed representations [16] as well as subgroup discovery [12,25,42], which is the method applied in this work. This work also extends existing subgroup discovery methods [9] for analyzing geo-tagged social media, especially in the direction of handling arbitrary sets of target properties and visual analytics. The latter methods are directly integrated into the subgroup analytics approach, for a holistic setting, similar to such visualizations for community mining [37].

For analyzing a set of subgroups, these are typically clustered according to their similarity, e.g., [10], or based on their predictive power [26]. Other methods for pattern set refinement and selection, e.g., [29] focus on similarities on the instance and/or description level. In contrast to these approaches, the proposed approach for subgroup set analytics generalizes those methods. We provide a general approach for analyzing subgroup relations based on a freely configurable "relationship" function, embedded in a graph-based framework for the assessment of sets of subgroups.

3 Preliminaries

Data mining includes descriptive and predictive approaches [21]. In the following, we focus on descriptive pattern mining methods. We apply subgroup discovery [25], a broadly applicable data mining method which aims at identifying interesting patterns with respect to a given target property of interest according to a specific quality function. This section first introduces the necessary notions concerning the data representation, subgroups and patterns, basics on graphs, and similarity measures.

3.1 Patterns and Subgroups

Below, we summarize basic notions on patterns and subgroups. We define subgroups and their descriptions as well as interestingness measures that are necessary in the context of exploratory subgroup analytics on ubiquitous data. For a more general view and a detailed discussion, we refer to, e.g., [9,12].

Basic Definitions. Formally, a *database* $DB = (I, A)$ 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 $sel_{a_{num} \in [min_j; max_j]}$ 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.

A *subgroup description* or (complex) *pattern* sd is then 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. 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.

Interestingness of a Pattern. A *quality function* $q \colon 2^S \rightarrow \mathbb{R}$ 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).

The result of a subgroup discovery task is the set of k (subgroup) patterns $res_1, \ldots, res_k$, where $res_i \in 2^S$ with the highest interestingness according to the quality function. While a large number of quality functions has been proposed in literature, many quality measures 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.

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 $a = 0.5$.

An extension to a target concept defined by a set of variables can be defined similarly, by extending common statistical tests. For comparing multivariate means for a set of m numeric attributes T_M, with $m = |T_M|$, for example, we can make use of Hotelling's T-squared test [22], for the quality measure q_H:

$$q_H = \frac{n(n-m)}{m(n-1)}(\mu_P^{T_M} - \mu_0^{T_M})^\top {CV_P^{T_M}}^{-1}(\mu_P^{T_M} - \mu_0^{T_M})\,,$$

where $\mu_P^{T_M}$ is the vector of the model attribute means in the subgroup sg_{sd}, $CV_P^{T_M}$ is the covariance matrix, and $\mu_0^{T_M}$ is the vector of the target concept means in DB.

3.2 Distributional Geo-Spatial Subgroup Analysis

For analyzing the geo-spatial distribution of a subgroup, we adapt the approach presented in [44], which we briefly summarize below. We are basically interested in the geographical distribution of a subgroup, and in a characterization of this distribution using an adequate measure. Specifically, we model the geo-spatial distribution of a subgroup, and derive a *peakiness* measure of this distribution, indicating the overall shape of the distribution.

For modeling the distribution, first the geo-space of the world is split into b bins by r_{lat} degrees of latitude and r_{lon} degrees of longitude. Then, the latitude and longitude coordinates of the individuals $i \in sg_{sd}$ of the pattern sd can be mapped into those bins. We obtain a vector $v(sd) = (v_1, \ldots, v_b)$ of occurrence counts v_i of the subgroup pattern for each bin $i, i = 1 \ldots b$, i. e., how many distinct users have used the specific pattern sd there. We normalize this vector, by dividing each entry by the sum of the absolute values of the vector's entries (L1-Norm $\|\cdot\|_1$). Finally, we obtain the *peakiness* value $\phi(sd)$ of a pattern utilizing the vector $v = v(sd)$ by computing its second moment, as follows:

$$\phi(sd) = \frac{v \cdot v}{\|v\|_1^2} = \frac{1}{\|v\|_1^2}\sum_{i=1}^{b} v_i^2 \tag{2}$$

A high *peakiness* indicates rather characteristically peaky distributions, while a low value is observed for distributions close to a uniform distribution, cf. [44].

3.3 Graphs

An (undirected) *graph* $G = (V, E)$ is an ordered pair, consisting of a finite set V containing the *vertices* (also called *nodes*), and a set E of *edges* denoting the *connections* between the vertices. In the following, we freely use the term *network* as a synonym for the term graph. A *weighted* graph is a graph $G = (V, E)$ together with a function $\mathrm{w} : E \rightarrow \mathbb{R}^+$ that assigns a positive weight to each edge. The *density* of G is the fraction of possible edges that are actually present. The *degree* $\mathrm{d}(u)$ of a node u in a network measures the number of connections it has to other nodes. In weighted graphs the *strength* $\mathrm{s}(u)$ is the sum of the weights of all edges containing u, i. e., $\mathrm{s}(u) := \sum_{\{u,v\}\in E} \mathrm{w}(\{u, v\})$. A (weakly) connected component of G is a subset $U \subseteq V$, such that there exists an (undirected) path between every pair of nodes $\{u, v\}\,, u, v \in U$, i. e., that u and v are connected by a sequence of edges. For more details, we refer to standard literature, e.g., [18].

3.4 Similarity Measures

Given two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^X$, there are a variety of *similarity measures* for assessing the similarity between the contained values, e.g., [23,41] for a detailed overview. We can measure a *Manhattan similarity*, for example, by utilizing the (normalized) Manhattan distance, defined as follows:

$$\mathrm{sim}_{\mathrm{man}}(\boldsymbol{v}_1, \boldsymbol{v}_2) := 1 - \frac{\sum_{i=1}^{X} |\boldsymbol{v}_{1i} - \boldsymbol{v}_{2i}|}{X}\,, \tag{3}$$

where v_{ij} denotes the j-th component of vector v_i.

An alternative measure known from information retrieval is the cosine measure. The *cosine similarity* between two vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^X$ is then defined as:

$$\mathrm{sim}_{\cos}(\boldsymbol{v}_1, \boldsymbol{v}_2) := \cos\measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}\,. \tag{4}$$

4 Exploratory Subgroup Analytics

In the following, we first present our novel subgroup analytics approach using a graph-based representation for inspecting and assessing a set of subgroups. We describe a set of analysis methods for exploratory data analysis using subgroup discovery, and provide automatic and interactive techniques. In summary, we propose the following methods for exploratory subgroup analytics that are described in more detail below.

- Exploratory ubiquitous subgroup analytics of noise and perception patterns.
- Subgroup-based analysis of spatio-temporal peakiness.
- Visual exploration and assessment of subgroup relations.

4.1 Exploratory Ubiquitous Subgroup Analytics

As a first step in our subgroup analytics approach, we obtain a set of the top-k subgroups for a specific target variable. Depending on the analytical questions, different quality functions can be applied, as sketched above, such that relatively simple deviations of a target variable from the "overall trend" observed in the general population can be analyzed. Furthermore, more complex exceptionality criteria, e.g., corresponding to more complex models such as captured by a set of target variables can be investigated.

In an exhaustive subgroup discovery approach, all combinations of possible selectors (e.g., tags) are analyzed for discovering interesting patterns. Since this search space is exponential in the number of selectors, typically an efficient subgroup discovery algorithm needs to be applied, e.g., the BSD [31], SD-Map [12], or the SD-Map* [7] algorithm. For large ubiquitous data with sparse distributions, the SD-Map* algorithm can often be successfully applied, cf. [9].

The result of the subgroup discovery step is then given by the top-k subgroups that need to be assessed and put into relation to each other. This is typically performed semi-automatically, based on automatic methods guided by user interaction. Then, the result set of subgroups can be inspected and validated. This can be supported by background knowledge [11,14], statistical approaches [8] and interactive techniques, e.g., [10] that are applied according to the *Information Seeking Mantra* by Shneiderman [39]: Overview first (macroscopic view), browsing and zooming (mesoscopic analysis), and details on demand (microscopic focus).

4.2 Subgroup-Based Analysis of Spatio-Temporal Peakiness

Basic subgroup discovery described in the previous section provides a first view on interesting patterns and data characteristics for selected target concepts. However, ubiquitous data typically also often contains temporal and specifically geo-spatial information. The latter is especially interesting for identifying, e.g., interesting places and locations.

In order to analyze the data further into this direction, we apply the technique described in Sect. 3 concerning the analysis of peakiness of subgroup patterns. *Peaky* subgroups are then those which are relatively specific for a certain *location* considering the geo-spatial information. Essentially, this information can be computed in addition to the subgroup information, i. e., a peakiness value is assigned to each subgroup that we collected using subgroup discovery.

For computing the peakiness value, we utilize the parameters proposed in [44] and divide the space into a one-by-one degree grid, and use the latitude and longitude values associated with our ubiquitous data measurements. Then, for each subset of the database covered by a specific subgroup, we can determine the peakiness value. These can then be added to the subgroup information and furthermore visualized in several ways as presented in the case study below.

4.3 Visual Exploration and Assessment of Subgroup Relations

After subgroups and their parameters, e.g., their peakiness values, have been determined, we apply an interative step for the visual exploration and assessment of a set of subgroups. Using a given relationship function, we consider specific *relations* between subgroups. Then, their "connections" according to this relation can be modeled as a graph.

More formally, given a certain criterion implemented by a relation function $rel : I \times I \rightarrow \mathbb{R}$ we obtain a value estimating the relationship between pairs of subgroups, identified by their respective subgroup descriptions. Possible relations include, for example, geographic distance, or semantic criteria. In our application setting, we focus on the latter, since we will use the given perceptions for noise measurements as semantic proxies for subgroup relatedness.

For assessing our result set of subgroups R, we obtain the *rel*-value for each pair of subgroups (u, v). After that, we construct a *subgroup assessment graph* G_R for R: The nodes of G_R are given by the subgroups contained in R. The edges between node pairs (u, v) are constructed according to the respective $rel(u, v)$ value: If the respective value between the subgroup pair is zero, then the edge is dropped; otherwise, an edge weighted by $rel(u, v)$ is added to the graph. In addition to the connections denoted by edges in the graph, we can furthermore visualize certain parameters such as properties of the nodes by colors and/or size of the nodes, as well as the weights of edges by different edge styles such as thickness or line types. In the graph, we can directly visualize the connectedness of a node using the degree information, as well as the peakiness in order to directly highlight highly interesting subgroups.

It is easy to see that, depending on the applied relationship function *rel*, the graph construction process can result in a fully connected graph which is hard to interpret. Therefore, a refinement of this process utilizes a certain threshold τ_{rel} which is used for pruning edges in the graph. If the relation "strength" $rel(u, v)$ between a subgroup pair (u, v) is below the threshold, i. e., $rel(u, v) < \tau_{rel}$ then we do not consider the edge between u and v, such that the edge is dropped. By carefully selecting a suitable threshold τ_{rel} the resulting subgroup network can then be easily inspected and assessed.

Typically, the situation becomes interesting when the graph is split into different components corresponding to certain clusters of subgroups. We will discuss examples of constructed networks below. For selecting a suitable threshold, a *threshold-component* visualization can be applied, see Fig. 10 for an example. This visualization plots the number of connected components of the graph depending on the applied threshold. Then, the "steps" within the plot can indicate interesting thresholds that can be interactively inspected. A related visualization plots the used threshold against the graph density for obtaining a first impression of the ranges of suitable threshold selections, cf. Fig. 11.

5 *WideNoise Plus* Case Study

In the following, we first describe the applied dataset. Then, we discuss a basic statistical analysis of the main parameters concerning the noise levels and tag distribution. After that, we describe a case study applying the presented techniques and discuss our results in context.

5.1 Applied Dataset

In this paper, we utilize data from the EveryAware project, 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. It includes sensor data from the microphone given as noise level in dB(A), the location from the GPS-, GSM-, and WLAN-sensor represented as latitude and longitude coordinate, as well as a timestamp. Furthermore, the user can enter his perceptions about the measurement, expressed using the four sliders for feeling (love to hate), disturbance (calm to hectic), isolation (alone to social), and artificiality (nature to man-made). In addition, tags can be assigned to the recording. We collected data from all around the world using iOS and Android devices. The largest user group is located around Heathrow Airport London where the residents map the noise pollution caused by the airport to profile and monitor their environmental situation.

The data are stored and processed using the EveryAware backend [15], which is based on the UBICON software platform [5,6].

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. In our analysis, we utilize the following objective and subjective information for each measurement:

- Objective: Level of noise (dB).
- Subjective perceptions about the environment, encoded in the interval $[-5; 5]$:
  - “Feeling” (hate/love) where -5 is most extreme for “hate” and 5 is most extreme for “love”.
  - “Disturbance” (hectic/calm) where -5 is most extreme for “hectic” and 5 is most extreme for “calm”.
  - “Isolation” (alone/social) where -5 is most extreme for “alone” and 5 is most extreme for “social”.
  - “Artificiality” (man-made/nature) where -5 is most extreme for “man-made” and 5 is most extreme for “nature”.
- Tags, e.g., “noisy”, “indoor”, or “calm”, providing the semantic context of the specific measurement.

Figures 1, 2, 3 and 4 show the value distributions of the different perception values as histograms.

5.2 Statistical Analysis

In this section, we perform some basic statistical analysis of the observed distributions as well as initial experiments on correlating the subjective and objective data. As we will see, we observe typical phenomena in the domain of tagging data, while the correlations are expressed on a medium level; this directly leads

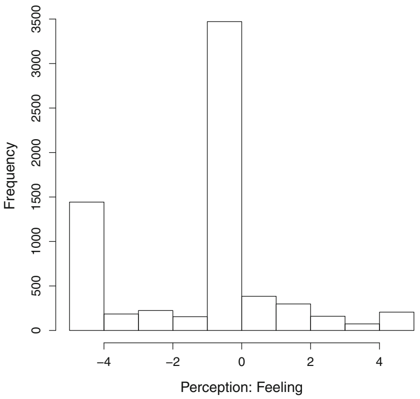


Fig. 1. Histogram for the subjective perception ‘feeling’.

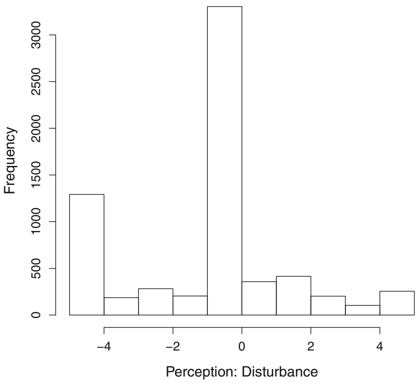


Fig. 2. Histogram for the subjective perception ‘disturbance’.

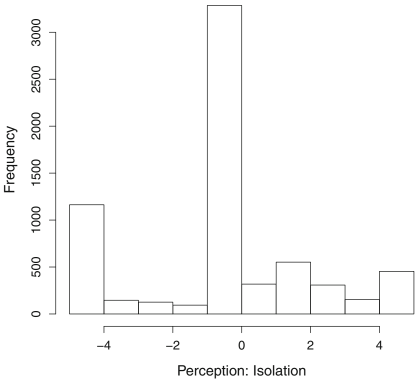


Fig. 3. Histogram for the subjective perception ‘isolation’.

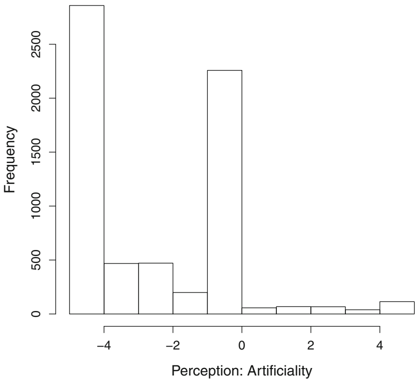


Fig. 4. Histogram for the subjective perception ‘artificiality’.

to the advanced techniques using our subgroup discovery that we describe in the subsequent sections, where we analyze the relation between objective and subjective data given patterns of tagging data in more detail.

Figures 5, 6, 7 and 8 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 6 shows the distribution of the collected dB values, with a mean of 67.42 dB.

In Fig. 7 we observe a typical heavy-tailed distributions of the tag assignments. Also, as can be observed in Figs. 5 and 8, 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.

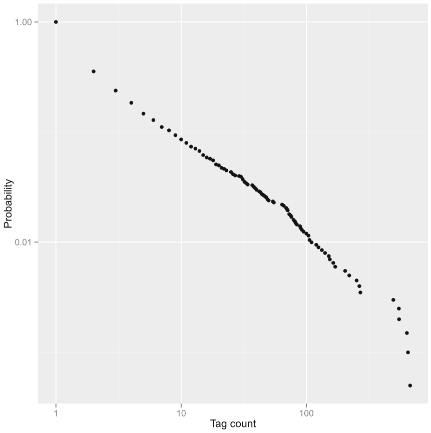


Fig. 5. Cumulated tag count distribution in the dataset. The y-axis provides the probability of observing a tag count larger than a certain threshold on the x-axis.

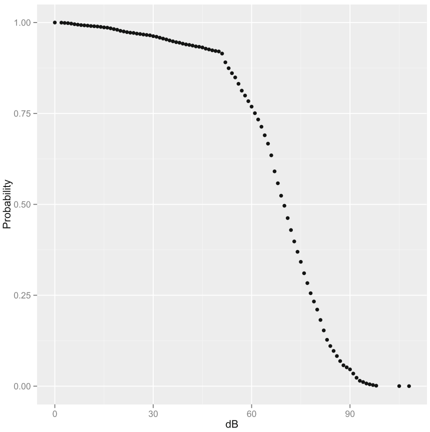


Fig. 6. Cumulated distribution of noise measurement (dB). The y-axis provides the probability for observing a measurement with a dB value larger than a certain threshold on the x-axis.

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

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.

5.3 Exploratory Subgroup Analytics: Results and Discussion

In the following, we present exploratory subgroup discovery results in the context of the *WideNoise Plus* data. First, we focus on the analysis of the (subjective) perceptions as target concepts, for which we identify both highly deviating and “conforming” patterns, i. e., those that are close to the means of the perceptions observed in the complete dataset. After that, we analyze characteristic patterns for high and low noise levels. Finally, we present a combined analysis, also including the geo-spatial distribution of specific subgroups.

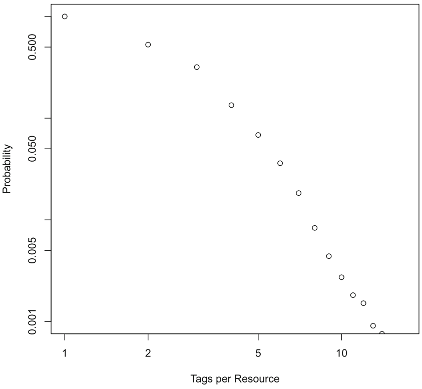


Fig. 7. Cumulated tag per record distribution in the dataset. The y-axis provides the probability of observing a tag per record count larger than a certain threshold on the x-axis.

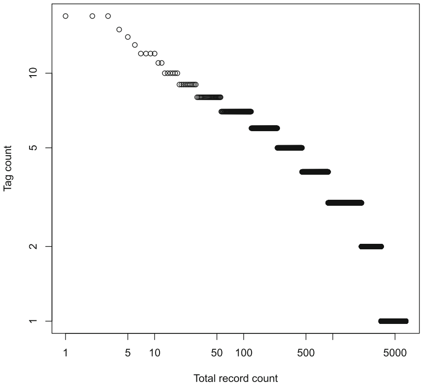


Fig. 8. Distribution of assigned tags per resource/data record.

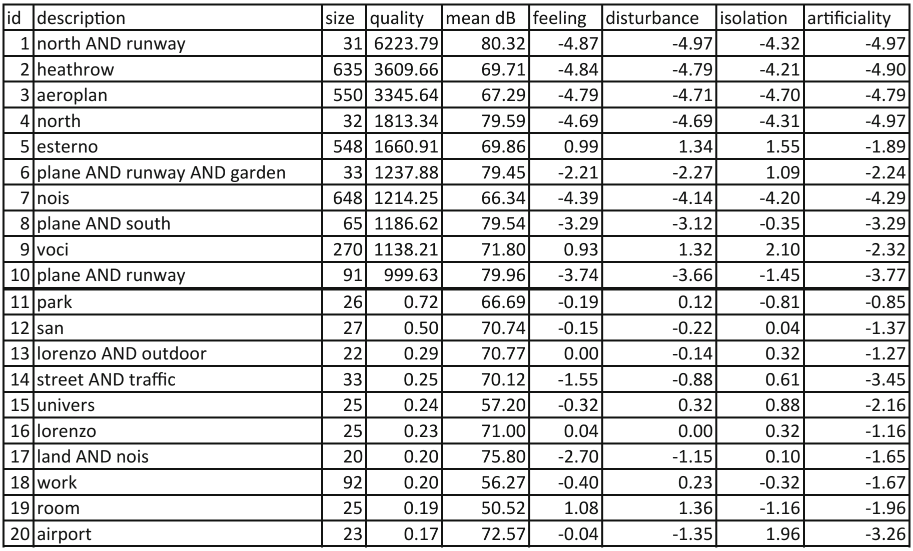
Analysis of Perception Patterns. For analyzing the characteristics of the subjective data given by the perception values assigned to the individual measurements we applied the multi-target quality function q_H (based on the Hotelling's T-squared test); in that way, we measured, which patterns show a perception profile (given by the means of the individual perceptions) that is exceptionally different from the overall picture of the perceptions (means 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 (q'_H).

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* – residents close to Heathrow obviously tend to measure noise 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*.

Table 2. Perception patterns: rows 1–10 - deviating patterns; rows 11–20 - 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

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. While the analysis of the perceptions provides some initial insights on subjective and objective data, again these results motivate our proposed approach for analyzing subgroups and their relations modeled by arbitrary parameters in more detail. This will be discussed in the next section, where we provide an integrated approach for assessing noise and perceptions patterns and their inter-relations.

Analysis of Noise (dB) Patterns. For identifying characteristic noise patterns, we applied subgroup discovery for the target variable *noise (dB)* focusing on subgroups both with a large deviation comparing the mean of the target in the subgroup and the target in the complete database. It is easy to see that increasing deviations (above the global mean) indicate *noisy* environments, while decreasing deviations (below the global mean) indicate more *quiet* situations. For analysis, we applied the simple binominal quality function, c.f., Sect. 3 and discovered the top-20 patterns for *noisy* and *quiet* environments, respectively.

Table 3 shows 40 patterns combining the two top-20 result sets. The patterns in rows $1-20$ denote the top-20 patterns for the target concept "high dB Value" whereas the subgroup patterns in rows $21 - 40$ denote the top-20 patterns for the target concept "low dB Value".

Table 3. Patterns: rows 1–20 - target: large mean noise (dB); rows 21–40 - target: small mean noise (dB); Overall mean (population): 67.42 dB. The last two columns include the node degree in the subgroup assessment graph, for $\tau_{rel} = 0.90$ and $\tau_{rel} = 0.95$.

id	description	size	mean dB	feeling	disturbance	isolation	artificiality	peakiness	deg (t=0.90)	deg (t=0.95)
1	craft	66	92.14	-3.03	-3.18	3.18	-4.61	1.00	9	2
2	air	75	88.43	-3.03	-3.09	2.73	-4.49	0.85	9	2
3	arriva	252	78.64	-0.02	-0.01	0.01	0.00	0.99	23	19
4	plane	495	73.88	-3.40	-2.25	-0.53	-3.38	0.97	8	5
5	aircraft	154	77.44	-1.18	-0.31	-0.71	-2.81	0.85	10	4
6	garden	674	72.19	-0.32	-0.17	-0.20	-4.39	0.90	10	4
7	plane AND runway	91	79.96	-3.74	-3.66	-1.45	-3.77	0.98	7	2
8	runway	100	79.12	-3.70	-3.62	-1.60	-3.88	0.98	7	2
9	heathrow AND plane	33	86.79	-4.33	-4.21	-0.30	-4.36	0.94	8	2
10	aeroporto	15	94.00	-5.00	0.00	0.00	0.00	1.00	10	4
11	ciampino	18	91.39	-4.17	0.00	0.00	0.00	1.00	17	5
12	plane AND south	65	79.54	-3.29	-3.12	-0.35	-3.29	0.97	8	3
13	runway AND south	65	79.54	-3.29	-3.12	-0.35	-3.29	0.97	8	3
14	south	66	79.41	-3.24	-3.08	-0.35	-3.32	0.97	8	3
15	plane AND street	46	80.83	-4.85	-3.87	-2.35	-4.87	1.00	4	1
16	ferri	20	86.95	0.05	0.05	0.10	-0.10	0.59	23	19
17	runway AND street	40	80.68	-4.83	-4.58	-2.83	-4.85	1.00	2	1
18	rout	120	75.07	-0.04	-0.02	0.03	-0.03	1.00	23	19
19	bus	170	73.83	-0.93	-1.24	0.76	-2.16	0.17	23	6
20	eva	71	77.11	0.00	0.00	0.00	-0.07	0.89	23	19
21	home	164	43.71	1.15	1.31	-0.99	-0.94	0.23	22	19
22	background	79	49.84	0.15	1.61	-1.08	-0.49	0.50	23	19
23	indoor	151	54.78	0.60	0.56	-0.08	-1.19	0.21	23	19
24	bosco	14	33.36	3.21	3.36	-1.71	1.93	1.00	0	0
25	night	32	45.53	1.78	2.38	-1.59	0.00	0.09	20	1
26	offic	204	58.77	0.08	0.73	-0.46	-1.69	0.22	27	20
27	fan AND music	14	37.50	0.21	0.29	0.00	-0.29	0.87	23	19
28	background AND nois	37	49.14	0.32	1.95	-2.30	-1.05	1.00	27	20
29	fan	23	44.35	-0.26	0.22	-1.30	-1.48	0.39	25	6
30	general	42	50.45	0.00	1.31	0.00	0.00	1.00	21	15
31	fan AND indoor	15	39.40	0.13	0.33	-0.07	-0.33	0.76	23	19
32	work	92	56.27	-0.40	0.23	-0.32	-1.67	0.33	26	11
33	morn	30	47.93	0.43	1.80	-1.93	-1.20	0.60	24	20
34	background AND indoor	23	45.35	0.43	1.74	-2.00	-0.91	1.00	24	19
35	indoor AND nois	23	45.35	0.43	1.74	-2.00	-0.91	1.00	24	19
36	background AND work	23	47.43	0.61	1.74	-2.00	-0.74	1.00	24	20
37	nois AND work	23	47.43	0.61	1.74	-2.00	-0.74	1.00	24	20
38	kassel	110	58.35	-0.05	0.54	0.12	-1.27	0.91	24	20
39	room	25	50.52	1.08	1.36	-1.16	-1.96	0.07	24	20
40	indoor AND music	20	48.65	0.60	0.15	0.50	-0.70	0.55	22	19

In the table, we can identify several distinctive tags for noisy environments, for example, *north AND runway*, *heathrow*, and *aeroplan*, which relate to Heathrow noise monitoring, c.f., [5] for more details. These results confirm the results of the basic analysis in [5]. For more quiet environments, we can also observe typical patterns, e.g., focusing on the tags *park*, *lorenzo AND outdoor*, and *room*, and combinations. Some further interesting subgroups are described by the tags *bosco* (forest) and *night*. These also show a quit distinct perception profile, shown in the respective columns of Table 3. This can also be observed in the last two columns of the table indicating the degree in the subgroup assessment graph (see below): The subgroups described by *bosco* and *night* are quite isolated.

When considering the *peakiness* of the subgroups, we observe that most of the patterns are relatively specific for certain locations, since they exhibit rather high peakiness values, e.g., *craft*, *plane*, *aeroporto* etc. These correspond to relatively

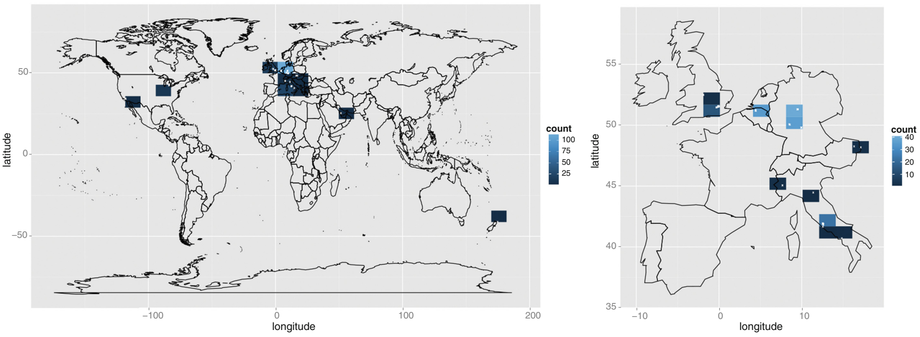


Fig. 9. Exemplary peakiness visualization for the pattern *bus* (worldwide, Europe).

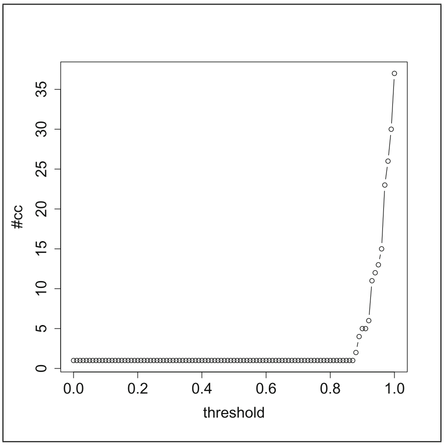


Fig. 10. Thresholded connected component plot based on a minimal *rel* value.

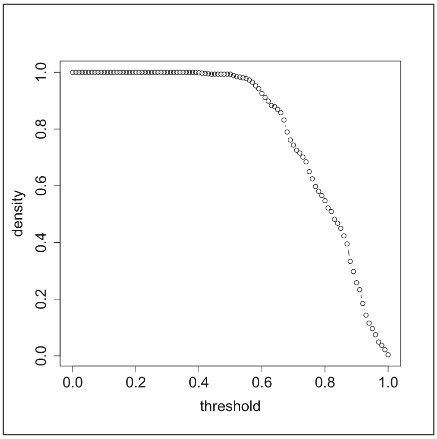


Fig. 11. Thresholded density plot based on a minimal *rel* value.

specific terms indicating certain locations. In contrast, there are some tags for which we estimate rather low peakiness values, for example, *bus*, *indoor*, *night*, *offic(e)*, *work*, *room*. It is easy to see, that those tags correspond to more *general* terms that are not so specific for certain locations. Therefore, these results are first a first indication for the relevance of the peakiness indicator with respect to identifying specific locations as described by specific tags. Figure 9 shows an example of visualizing the pattern *bus* on a worldwide view and a map of Europe, respectively. The subfigures show the location of the individual measurements, while the counts observed in different regions are color-coded: Lighter blue-areas indicate locations with higher counts of measurements. The figures are an example for a pattern with a smaller peakiness value, since the measurements are distributed more widely on a worldwide scale. In contrast, a pattern such as *kassel* tends to be more focused arund the kassel area (top-right light-blue area in the right plot of Fig. 9).

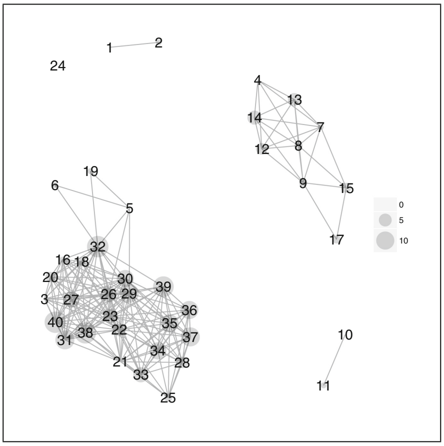


Fig. 12. Assessment graph: $\tau_{rel} = 0.90$; the size of a node depicts its degree.

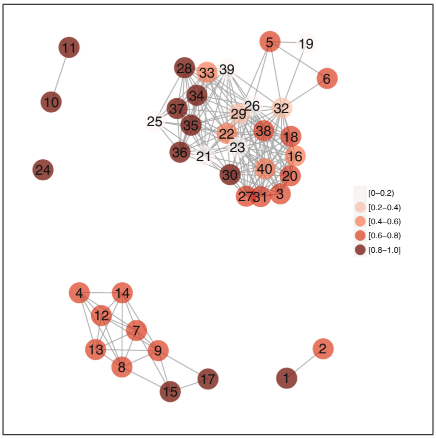


Fig. 13. Assessment graph: $\tau_{rel} = 0.95$ with color-coded peakiness values.

Graph-Based Exploration of Subgroup Relations. In order to analyze subgroup relations with respect to the perceptions, we apply the Manhattan similarity as defined in Sect. 3 as our assessment relation *rel*. We measure the similarity using the averaged perception vectors of the respective subgroup patterns, with normalized values in the interval $[0; 1]$. Using the Manhattan similarity, we consider the overall "closeness" of the vectors; alternatively, the cosine similarity would focus on similar perception "profiles", i. e., uniformly expressed perceptions.

For determining appropriate thresholds τ_{rel}, Fig. 10 shows a threshold vs. connected component plot, constructed using the given similarity measure for estimating the relations between the subgroup patterns. Figure 11 shows the according thresholded density plot. Then, appropriate thresholds can be selected by the analyst. Figures 12 and 13 show the graphs for a threshold $\tau_{rel} = 0.9$, for which the degrees of the nodes are visualized by the size of the individual nodes, and the peakiness is color-coded, respectively.

As can be observed in Figs. 12 and 14 the respective networks for thresholds 0.90 and 0.95 show a distinct structure. Starting with $\tau_{rel} = 0.90$ the networks start to break up into distinct components, for $\tau_{rel} = 0.95$ the number of component increases significantly. For the lowest threshold $\tau_{rel} = 0.90$ we can already observe the special structure of pattern 24, one larger, and two smaller clusters. With threshold $\tau_{rel} = 0.95$, several more clusters emerge – the "Heathrow clusters" (7, 8), (15, 17) as well as the large cluster covering most of the *lower noise* patterns. However, this cluster also contains some patterns from the *higher noise* patterns (5, 6), which are rather unexpected and therefore quite interesting for subsequent analysis of the assigned perceptions. The connecting subgroup patterns can then be simply extracted by tracing the connections in the graph.

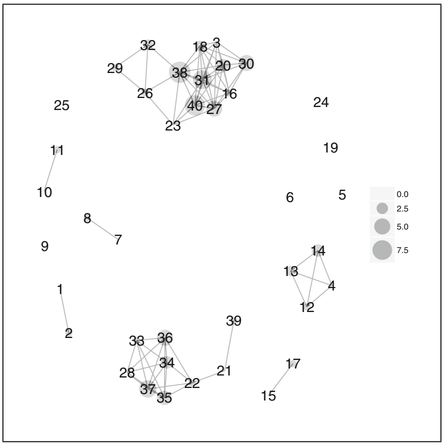


Fig. 14. Assessment graph: $\tau_{rel} = 0.95$; the size of a node depicts its degree.

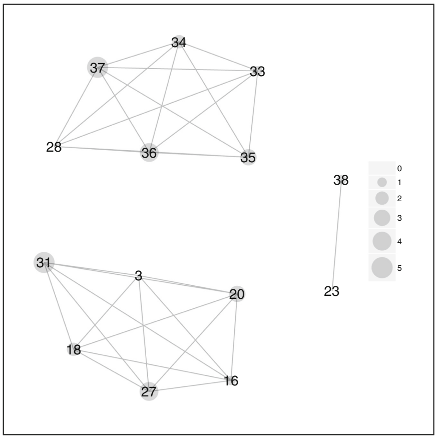


Fig. 15. Assessment graph: $\tau_{rel} = 0.97$ zoomed in on the large cluster in Fig. 14; the size of a node depicts its degree.

Using the *Information Seeking Mantra*, cf. [39], we can zoom in and analyze details on demand as described above. Figure 15 shows an example, focusing on the high-degree nodes (degree ≥ 5) of the large cluster discussed above (Fig. 14), for a threshold $\tau_{rel} = 0.97$. The cluster dissolves into distinct components and the strongly connected (and partially overlapping subgroups) remain, for example, the patterns $34, 35, 36$.

6 Conclusions

In this paper, we presented exploratory subgroup analytics for obtaining interesting descriptive patterns in ubiquitous data. The presented approach includes semi-automatic techniques for comprehensive analysis of target concepts ranging from single variables to multi-target analysis and geo-spatial patterns. Specifically, we provided a novel graph-based analysis approach for assessing the relations between sets of subgroups including additional properties such as connectivity of the patterns or their peakiness corresponding to geo-spatial locations. Using data from a ubiquitous application we presented the proposed approach and discussed analysis results of a real-world case study. The analyzed noise measurements and associated subjective perceptions described by a set of tags confirmed the semantic context and provided interesting patterns with respect to the analysis of subjective and objective data.

For future work, we aim to extend the approach to diverse relationship and similarity measures. Furthermore, we plan to investigate multi-relational representations, i. e., multi-graphs capturing a set of relationships for assessing a set of subgroups. A further direction for analysis concerns the analyis of interrelations between perceptions, tags, and sentiments based on the tagging data, e.g., extending case-based approaches, e.g., [13] and methods for local exceptionality detection, e.g., [30]. These can then also be applied, for example, for enhanced event detection, recommendations, or community mining.

Acknowledgements. This work has been supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University, and parts of this research was funded by the European Union in the 7th Framework programme EveryAware project (FET-Open).

References

1. Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., Staab, S.: Exploiting flickr tags and groups for finding landmark photos. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 654–661. Springer, Heidelberg (2009)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of VLDB, pp. 487–499. Morgan Kaufmann (1994)
3. Appice, A., Ceci, M., Lanza, A., Lisi, F., Malerba, D.: Discovery of spatial association rules in geo-referenced census data: a relational mining approach. Intell. Data Anal. **7**(6), 541–566 (2003)
4. Atzmueller, M.: Mining social media: key players, sentiments, and communities. WIREs: Data Min. Knowl. Disc. **2**(5), 411–419 (2012)
5. 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: Proceedings of IEEE International Conference on Cyber, Physical and Social Computing, pp. 317–324. IEEE Computer Society, Washington, DC, USA (2012)
6. 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. N. Rev. Hypermedia Multimedia **20**(1), 53–77 (2014)
7. Atzmueller, M., Lemmerich, F.: Fast subgroup discovery for continuous target concepts. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 35–44. Springer, Heidelberg (2009)
8. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. LNCS, pp. 842–845. Springer, Berlin (2012)
9. Atzmueller, M., Lemmerich, F.: Exploratory pattern mining on social media using geo-references and social tagging information. Int. J. Web Sci. (IJWS), **1/2**(2) (2013)
10. Atzmueller, M., Puppe, F.: Semi-automatic visual subgroup mining using VIKAMINE. Journal of Universal Computer Science **11**(11), 1752–1765 (2005)
11. Atzmüller, M., Puppe, F.: A methodological view on knowledge-intensive subgroup discovery. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 318–325. Springer, Heidelberg (2006)
12. Atzmüller, M., Puppe, F.: SD-Map – A fast algorithm for exhaustive subgroup discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 6–17. Springer, Heidelberg (2006)

13. Atzmueller, M., Puppe, F.: A case-based approach for characterization and analysis of subgroup patterns. J. Appl. Intell. **28**(3), 210–221 (2008)
14. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting background knowledge for knowledge-intensive subgroup discovery. In: Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI-05), pp. 647–652. Edinburgh, Scotland (2005)
15. Becker, M., Mueller, J., Hotho, A., Stumme, G.: A generic platform for ubiquitous and subjective data. In: Proceedings of 1st International Workshop on Pervasive Urban Crowdsensing Architecture and Applications, PUCAA 2013 (2013)
16. Boley, M., Horváth, T., Poigné, A., Wrobel, S.: Listing closed sets of strongly accessible set systems with applications to data mining. Theor. Comput. Sci. **411**(3), 691–700 (2010)
17. Ceci, M., Appice, A., Malerba, D.: Time-slice density estimation for semantic-based tourist destination suggestion. In: Proceedings of ECAI 2010, pp. 1107–1108. IOS Press, Amsterdam, The Netherlands, The Netherlands (2010)
18. Diestel, R.: Graph Theory. Springer, Berlin (2006)
19. Ganter, B., Stumme, G., Wille, R. (eds.): Formal Concept Analysis, Foundations and Applications. Lecture Notes in Computer Science. Springer, Berlin (2005)
20. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. Data Min. Knowl. Disc. **15**, 55–86 (2007)
21. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
22. Hotelling, H.: The generalization of student's ratio. Ann. Math. Statist. **2**(3), 360–378 (1931)
23. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3), 264–323 (1999)
24. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of KDD, pp. 91–101. ACM, New York, NY, USA (2002)
25. Klösgen, W.: Advances in knowledge discovery and data mining. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Explora: A Multipattern and Multistrategy Discovery Assistant, pp. 249–271. AAAI, California (1996)
26. Knobbe, A., Fürnkranz, J., Cremilleux, B., Scholz, M.: From local patterns to global models: the lego approach to data mining. In: Proceedings of ECML/PKDD'08 LeGO Workshop (2008)
27. Koperski, K., Han, J., Adhikary, J.: Mining knowledge in geographical data. Commun. ACM **26**, 65–74 (1998)
28. Lakhal, L., Stumme, G.: Efficient mining of association rules based on formal concept analysis. In: Ganter, B., Stumme, G., Wille, R. (eds.) Formal Concept Analysis. LNCS (LNAI), vol. 3626, pp. 180–195. Springer, Heidelberg (2005)
29. van Leeuwen, M., Knobbe, A.J.: Diverse subgroup set discovery. Data Min. Knowl. Discov. **25**(2), 208–242 (2012)
30. Lemmerich, F., Becker, M., Atzmueller, M.: Generic pattern trees for exhaustive exceptional model mining. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2008, Part II. LNCS, vol. 5212, pp. 277–292. Springer, Heidelberg (2008)
31. Lemmerich, F., Rohlfs, M., Atzmueller, M.: Fast discovery of relevant subgroup patterns. In: Proceedings of 23rd International FLAIRS Conference, pp. 428–433. AAAI Press, Palo Alto, CA, USA (2010)
32. Lindstaedt, S., Pammer, V., Mörzinger, R., Kern, R., Mülner, H., Wagner, C.: Recommending tags for pictures based on text, visual content and user context. In: Proceedings of 3rd International Conference on Internet and Web Applications and Services, pp. 506–511. IEEE Computer Society, Washington, DC, USA (2008)

33. Liu, Z.: A survey on social image mining. Intell. Comput. Inf. Sci. **134**, 662–667 (2011)
34. 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
35. Rattenbury, T., Naaman, M.: Methods for extracting place semantics from flickr tags. ACM Trans. Web **3**(1), 1:1–1:30 (2009)
36. Richter, K.-F., Winter, S.: Citizens as database: conscious ubiquity in data collection. In: Pfoser, D., Tao, Y., Mouratidis, K., Nascimento, M.A., Mokbel, M., Shekhar, S., Huang, Y. (eds.) SSTD 2011. LNCS, vol. 6849, pp. 445–448. Springer, Heidelberg (2011)
37. Roitman, H., Raviv, A., Hummel, S., Erera, S., Konopniki, D.: Microcosm: visual discovery, exploration and analysis of social communities. In: Proceedings of IUI, pp. 5–8. ACM, New York, NY, USA (2014)
38. Santini, S., Ostermaier, B., Adelmann, R.: On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments. In: Proceedings of International Conference on Networked Sensing Systems (INSS), pp. 1–8 (2009)
39. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of IEEE Symposium on Visual Languages, pp. 336–343. Boulder, Colorado (1996)
40. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th International Conference on World Wide Web, pp. 327–336. WWW ’08, ACM, New York, NY, USA (2008)
41. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: AAAI WS AI for Web Search, pp. 58–64. Austin, TX, USA (2000)
42. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97), pp. 78–87. Springer, Berlin (1997)
43. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In: WWW 2011, pp. 247–256. ACM, New York, NY, USA (2011)
44. Zhang, H., Korayem, M., You, E., Crandall, D.J.: Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In: Proceedings of International Conference on Web Search and Data Mining, pp. 33–42. ACM, New York, NY, USA (2012)