Subgroup Analytics and Interactive Assessment on Ubiquitous Data

Martin Atzmueller[1] and Juergen Mueller[1,2]

[1] University of Kassel, Knowledge and Data Engineering Group
[2] L3S Research Center
{atzmueller, mueller}@cs.uni-kassel.de

Abstract. This paper applies subgroup discovery for obtaining interesting descriptive patterns in ubiquitous data. Furthermore, we provide a novel graph-based analysis approach for assessing the relations between the obtained subgroup set, and for comparing subgroups according to their relations to other subgroups. We present and discuss first results utilizing real-world data, given by 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 their inter-relations.

This paper reports first results on analyzing ubiquitous data: objective (sensor) data and subjective perceptions. We apply subgroup discovery as a versatile method for descriptive data mining. We utilize the VIKAMINE [3] tool [6] for subgroup discovery and analytics in order to implement a semi-automatic pattern discovery process. For the analysis, we apply real-world data from the EveryAware project[4] – focusing on noise measurements. The individual data points include the measured noise in decibel (dB), associated subjective perceptions (feeling, disturbance, isolation, artificiality) and a set of tags for providing semantic context for the individual measurements. 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. We present first results analyzing subgroups patterns for hot-spots of low/high noise levels.

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

[3] http://www.vikamine.org
[4] www.everyaware.eu

2 Related Work

Ubiquitous data mining covers many subfields, including spatio-temporal data mining [15], mining sensor data or mining social media with geo-referenced data, c.f., [3]. Applications include destination recommenders, e.g., for tourist information systems [10], or geographical topic discovery [21]. Often established problem statements and methods have been transferred to this setting, for example, considering association rules [2]. Related approaches consider, for example, social image mining methods, cf., [17] for a survey. The concept of collecting information in ubiquitous systems, especially for crowd-sourced and citizen-driven applications is discussed in [18]. Basic issues of measuring noise pollution using mobile phones are presented in [19].

In contrast to the approaches discussed above, in this paper we focus on descriptive patterns. This allows for the flexible adaptation to the preferences of the users, since their interestingness can be flexibly tuned by altering the applied quality function and target concept. There are several variants of pattern mining techniques, e.g., frequent pattern mining [11], mining association rules [1], as well as subgroup discovery [13], which is the method applied in this work.

For analyzing a set of subgroups, these are typically clustered according to their similarity, e. g., [8], or based on their predictive power [14]. Other methods for pattern set refinement and selection, e. g., [16] 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 [12]. In the following, we focus on descriptive pattern mining methods. We apply subgroup discovery [13], 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, subgroup patterns, basics on graphs, and similarity measures.

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 $I \to \{0, 1\}$ 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 $sd = \{sel_1, \ldots, sel_l\}$, 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 subgroup description 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.

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). The result of a subgroup discovery task is the set of k subgroup descriptions $res_1, \ldots, res_k$ 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 $|ext(sd)|$ 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

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

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 (undirected) *graph* $G = (V, E)$ is an ordered pair, consisting of a finite set V containing the *vertices/nodes*, and a set E of *edges/connections* between the vertices. We freely use the term *network* as a synonym for graph. A *weighted* graph is a graph $G = (V, E)$ together with a function $\mathrm{w} : E \to \mathbb{R}^+$ that assigns a positive weight to each edge. 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., $\mathrm{s}(u) := \sum_{\{u,v\} \in E} \mathrm{w}(\{u, v\})$.

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., [20].We can measure vector similarity, for example, by the (normalized) Manhattan distance, defined as follows:

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

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

Another prominent measure 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:

$$\mathrm{sim}_{\cos}(\boldsymbol{v}_1, \boldsymbol{v}_2) := \cos \measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}\,. \tag{3}$$

4 Exploratory Subgroup Analytics

In the following, we first present the novel subgroup analytics approach using a graph-based representation for inspecting and assessing a set of subgroups. Next, we describe the utilized dataset. After that, we discuss first results of the analysis in our ubiquitous application context.

4.1 Overview

For subgroup analytics, we first obtain a set of the top-k subgroups for a specific target variable. Typically, an efficient subgroup discovery algorithm needs to be applied. In our experiments, we apply the SD-Map* [5] algorithm for efficient subgroup discovery, which is suitable for sparse tagging data, c.f., [7]. After that, the set of subgroups needs to be assessed and put into relation to each other.

The proposed approach especially focuses on this specific step: It considers a relation between subgroups such that their "connections" according to this relation can be modeled as a graph. More formally, given a certain criterion implemented by a relation function $rel : I \times I \to \mathbb{R}$ 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.

It is easy to see that – depending on the applied relationship function *rel*– this 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 Figure 5 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.

4.2 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 12, 2013.

WideNoise Plus allows the storage of noise measurements using ubiquitous mobile devices, and includes sensor data from the microphone given as noise level in dB and data from the location sensors (i.e., GPS-sensor, GSM- and WLAN-locating) represented as latitude and longitude coordinate as well as a timestamp. Furthermore, *WideNoise Plus* captures the user's perceptions about the recordings, expressed using the four slider 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. The data are stored and processed using the EveryAware platform [9], which is based on the UBICON framework [4].

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

- Objective: Level of noise (dB).
- Subjective perceptions about the environment:
  - "Feeling" (hate/love) encoded in the interval $[-5; 5]$, where -5 is most extreme for "hate" and 5 is most extreme for "love".
  - "Disturbance" (hectic/calm), encoded in the interval $[-5; 5]$, where -5 is most extreme for "hectic" and 5 is most extreme for "calm".
  - "Isolation" (alone/social), encoded in the interval $[-5; 5]$, where -5 is most extreme for "alone" and 5 is most extreme for "social".
  - "Artificiality" (man-made/nature), encoded in the interval $[-5; 5]$, 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.

The applied dataset contains 5,237 data records and 1,056 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.

Figures 1-4 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. As can be observed in Figure 1 and Figure 4, 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 high interest. Therefore, we present first analysis results of interesting patterns in the case study described below. We focus on the relation between semantics and perceptions as indicated by the different subjective perception values.

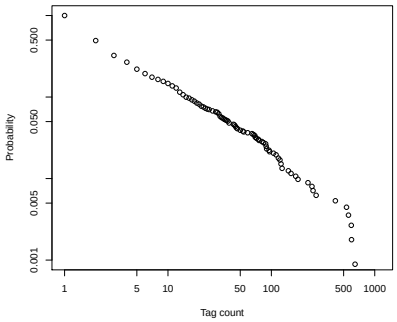

Fig. 1. Cumulated tag count distribution in the dataset. The y-axis provides the probability of observing a tag count larger than a certain threshold on the x-axis.

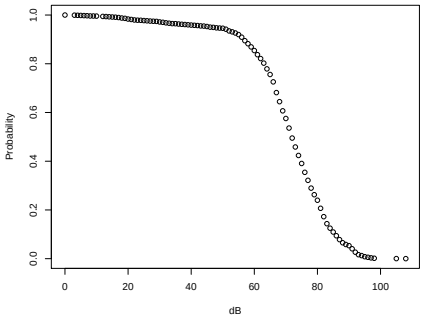

Fig. 2. Cumulated distribution of noise measurement (dB). The y-axis provides the probability for observing a measurement with a dB value larger than a certain threshold on the x-axis.

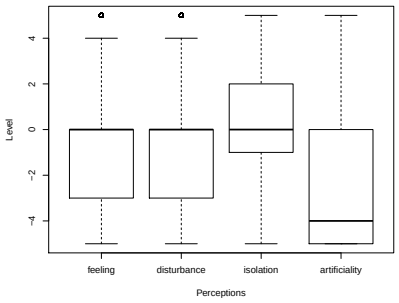

Fig. 3. Overview on the value distribution of the different perceptions.

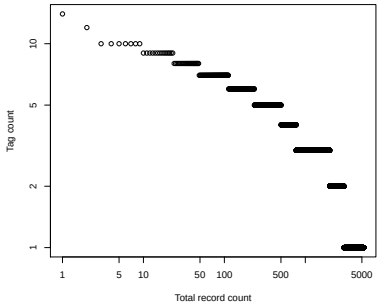

Fig. 4. Distribution of assigned tags per resource/data record.

4.3 Case Study: First Results and Discussion

In the following, we present first analysis results in the context of the *WideNoise Plus* data. According to the proposed approach, we applied subgroup discovery for the target variable *noise (dB)* focusing on subgroups with a large deviation comparing the mean of the target in the subgroup and the target in the whole database. We applied the simple binominal quality function, c.f., Section 3. Table 1 shows the resulting 20 patterns combining the two top-10 result sets.

In the table, we can identify several distinctive tags for noisy environments, for example, *craft*, *aircraft*, *plane*, *heathrow AND plane* which relate to Heathrow noise monitoring, c.f., [4] for more details. These results confirm the basic analysis in [4]. For more quiet environments, we can also observe typical patterns, e. g., focusing on the tags *indoor*, *background* and *work*, and combinations. Some

Table 1. Patterns: 1-10 - target: large mean noise (dB); 11-20 - target: small mean noise (dB); Overall mean (population): 70.12 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	deg (t=0.9)	deg (t=0.95)
1	craft	67	92.10	-3.06	-3.21	3.21	-4.61	4	1
2	air	72	89.72	-3.07	-3.10	2.97	-4.57	4	1
3	arriva	252	78.64	-0.02	-0.01	0.01	0.00	9	8
4	plane	415	76.26	-3.47	-2.61	-0.59	-3.75	5	0
5	heathrow AND plane	31	87.81	-4.61	-4.48	-0.32	-4.65	3	2
6	runway	107	79.62	-3.78	-3.45	-1.45	-3.94	3	2
7	runway AND plane	92	79.92	-3.75	-3.67	-1.38	-3.78	3	2
8	aeroporto	13	94.08	-5.00	0.00	0.00	0.00	6	1
9	ciampino	16	91.13	-4.06	0.00	0.00	0.00	10	2
10	departure	14	92.50	-0.71	0.57	-0.29	-1.36	11	8
11	home	124	45.58	1.10	1.31	-0.96	-0.99	9	7
12	bosco	17	35.35	3.29	3.53	-1.65	1.88	0	0
13	indoor	111	56.69	0.81	0.71	-0.17	-1.29	9	8
14	office	172	59.78	0.10	0.68	-0.35	-1.68	11	9
15	borgo	12	31.33	3.00	3.25	-1.00	1.67	0	0
16	background	35	48.06	0.40	2.11	-2.46	-0.97	10	9
17	work	74	55.76	-0.49	0.19	-0.35	-1.86	11	5
18	indoor AND background	22	44.32	0.55	1.91	-2.14	-0.73	10	8
19	kassel	96	58.67	-0.17	0.64	0.17	-1.41	10	9
20	work AND background	23	47.43	0.61	1.74	-2.00	-0.74	10	8

further interesting subgroups are described by the tags *bosco* (forest) and *borgo* (village). These also show a quit distinct perception profile, shown in the respective columns of Table 1. 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 *borgo* and *bosco* are quite isolated.

In order to analyze subgroup relations with respect to the perceptions, we apply the Manhattan similarity as defined in Section 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 distance, we consider the overall "closeness" of the vectors; alternatively, the cosine similarity would focus on similar perception "profiles", i. e., uniformly expressed perceptions.

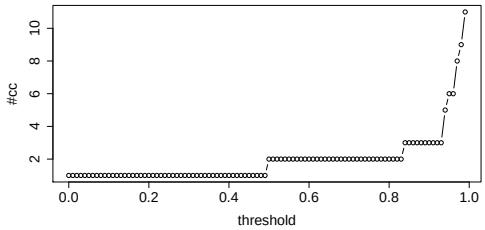

Fig. 5. Thresholded connected component plot based on a minimal *rel* value.

For determining appropriate thresholds τ_{rel}, Figure 5 shows a threshold vs. connected component plot using the Manhattan similarity defined above. Then, appropriate thresholds can be selected by the analyst. As can be observed in Figures 6-7 the respective networks for thresholds 0.90 and 0.95 show a distinct structure. Starting with the lowest threshold $\tau_{rel} = 0.90$ we can already observe the special structure of patterns 12 and 15. At this level, the remaining graph stays connected. With threshold $\tau_{rel} = 0.95$, several clusters emerge – the "Heathrow cluster" (5, 6, 7), 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 (3, 8, 9, 10), which are rather unexpected and therefore quite interesting for subsequent analysis. The connecting subgroup patterns can then be simply extracted by tracing the connections in the graph.

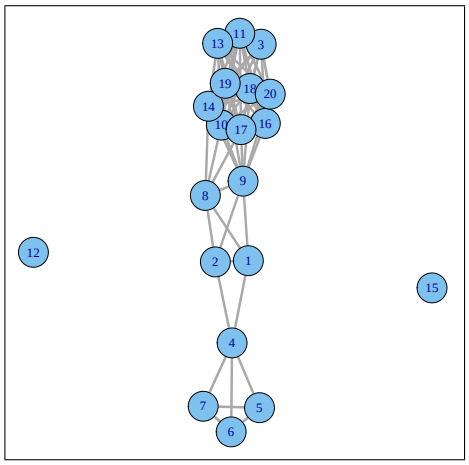

Fig. 6. Assessment graph: $\tau_{rel} = 0.90$.

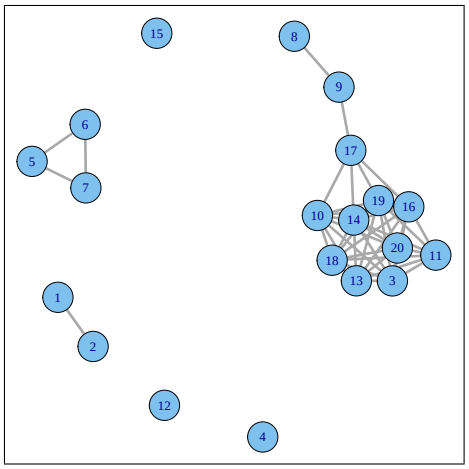

Fig. 7. Assessment graph: $\tau_{rel} = 0.95$.

5 Conclusions

In this paper, we presented exploratory subgroup analytics for obtaining interesting descriptive patterns in ubiquitous data. Specifically, we provided a novel graph-based analysis approach for assessing the relations between sets of subgroups. Using real-world data from a ubiquitous application we presented and discussed first analysis results. The analyzed noise measurements and associated subjective perceptions described by a set of tags confirmed the semantic context and provided interesting patterns towards a more comprehensive analysis.

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 interrelations between perceptions, tags, and sentiments based on the tagging data. These can then also be applied, for example, for enhanced event detection, recommendations, or community mining, in combination with spatio-temporal patterns.

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 by the EU project EveryAware.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th Int. Conf. Very Large Data Bases. pp. 487–499. Morgan Kaufmann (1994)
2. Appice, A., Ceci, M., Lanza, A., Lisi, F., Malerba, D.: Discovery of Spatial Association Rules in Geo-Referenced Census Data: A Relational Mining Approach. Intelligent Data Analysis 7(6), 541–566 (2003)
3. Atzmueller, M.: Mining Social Media: Key Players, Sentiments, and Communities. WIREs: Data Mining and Knowledge Discovery 1069 (2012)
4. 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 CPSCom (2012)
5. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. ISMIS 2009. LNCS (2009)
6. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: ECML/PKDD 2012. Springer, Berlin (2012)
7. Atzmueller, M., Lemmerich, F.: Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. IJWS 2(1/2) (2013)
8. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science) 11(11), 1752–1765 (2005)
9. Becker, M., Mueller, J., Hotho, A., Stumme, G.: A Generic Platform for Ubiquitous and Subjective Data. In: Proc. 1st Intl. Workshop on Pervasive Urban Crowdsensing Architecture and Applications, PUCAA 2013 (2013)
10. Ceci, M., Appice, A., Malerba, D.: Time-Slice Density Estimation for Semantic-Based Tourist Destination Suggestion. In: Proc. ECAI 2010. pp. 1107–1108. IOS Press, Amsterdam, The Netherlands, The Netherlands (2010)
11. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent Pattern Mining: Current Status and Future Directions. Data Mining and Knowledge Discovery 15, 55–86 (2007)
12. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco, USA (2006)
13. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI (1996)
14. Knobbe, A., Fürnkranz, J., Cremilleux, B., Scholz, M.: From Local Patterns to Global Models: The LeGo Approach to Data Mining. In: Proc. ECML/PKDD'08 LeGO Workshop (2008)
15. Koperski, K., Han, J., Adhikary, J.: Mining Knowledge in Geographical Data. Communications of the ACM 26 (1998)
16. van Leeuwen, M., Knobbe, A.J.: Diverse Subgroup Set Discovery. Data Min. Knowl. Discov. 25(2), 208–242 (2012)
17. Liu, Z.: A Survey on Social Image Mining. Intelligent Computing and Information Science pp. 662–667 (2011)
18. Richter, K.F., Winter, S.: Citizens as Database: Conscious Ubiquity in Data Collection. In: Advances in Spatial and Temporal Databases, Lecture Notes in Computer Science, vol. 6849, pp. 445–448. Springer, Berlin (2011)
19. 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: Proc. Intl. Conf. on Networked Sensing Systems (INSS),. pp. 1 –8 (2009)
20. 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)
21. 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)