Exploratory pattern mining on social media using geo-references and social tagging information

Martin Atzmueller*

Knowledge and Data Engineering Group,
University of Kassel,
Wilhelmshöher Allee 73, 34121 Kassel, Germany
E-mail: atzmueller@cs.uni-kassel.de
*Corresponding author

Florian Lemmerich

Artificial Intelligence and Applied Computer Science Group,
University of Würzburg,
Am Hubland, 97074 Würzburg, Germany
E-mail: lemmerich@informatik.uni-wuerzburg.de

Abstract: This paper presents exploratory pattern mining techniques for describing communities of resources (e.g., images) and for characterising locations of interest. We utilise tagging information and collaborative geo-reference annotations for characterising resources locations by a set of descriptive patterns. The methods are embedded into an interactive approach for mining, browsing and visualising a set of patterns. As an exemplary use case, we focus on the social photo sharing application Flickr. Utilising publicly available real-world data from this platform, we provide a structural evaluation of the automatic approach as well as an exemplary case study for demonstrating the effectiveness and validity of the interactive approach.

Keywords: social web; social media; community detection; pattern mining; geospatial analysis; web science.

Reference to this paper should be made as follows: Atzmueller, M. and Lemmerich, F. (2013) 'Exploratory pattern mining on social media using geo-references and social tagging information', *Int. J. Web Science*, Vol. 2, Nos. 1/2, pp.80–112.

Biographical notes: Martin Atzmueller is a Senior Researcher at the University of Kassel. He studied computer science at the University of Texas at Austin (USA) and at the University of Würzburg (Germany) where he completed his MSc (diploma) in Computer Science. He earned his doctorate (PhD) from the University of Würzburg. His research areas include data mining, social computing, mining social media, web science, machine learning, and natural language processing.

Florian Lemmerich is a Researcher at the University of Würzburg. He studied computer science at the University of Würzburg (Germany), where he received his MSc diploma. Currently, he is completing his doctorate (PhD). His research focuses on data mining and machine learning, in especially pattern discovery, interestingness measures in data mining and the application of knowledge discovery techniques in education, medicine and web analysis.

Copyright © 2013 Inderscience Enterprises Ltd.

1 Introduction

The emergence of social networks, mobile systems and ubiquitous computing has created a number of novel location-aware services and applications. In order to analyse such social media environments, pattern mining provides convenient options, e.g., in order to identify interesting and relevant relations and resources. Especially in social media systems such as Twitter (http://www.twitter.com) or resource sharing systems like Flickr (http://www.flickr.com) – with geo-referenced and tagged data – the combination of interactive and automatic approaches enables powerful exploratory approaches.

In this paper, we present a two way perspective on exploring locations, tags, resources and their induced relations: First, we aim to describe sets of social media resources (e.g., photos) using location-information and tags, which are semantically related as well as focused on certain locations. Imagine, for example, browsing the map of Germany and taking an overview on the general Berlin/Brandenburg area in terms of tag descriptions. Second, we characterise given locations using tagging patterns and photos for interactive browsing. A user may click on a map to specify his point of interest, for example, and is then provided a set of tags that are specifically used for that region. We consider publicly available image data, e.g., from photo management and image sharing applications such as Flickr or Picasa (http://www.picasa.com). In our setting, each image is tagged by users with several freely chosen tags. Additionally, each picture is annotated with a geo-reference (latitude, longitude) that indicates, where the image was taken.

We propose an iterative two step approach for the exploration of locations and resources in social media: The first step uses pattern mining techniques (e.g., Atzmueller and Mitzlaff, 2011; Atzmueller and Lemmerich, 2009) to automatically generate a candidate set of potentially interesting descriptive tags. For a flexible characterisation of locations at different levels the search can be adapted by employing different location-based target measures for pattern mining. In the second step, a human explores this candidate set of patterns and introspects interesting patterns manually by browsing and viewing various visualisations. Based on the obtained results, pattern mining parameters can be adapted in an exploratory fashion. Additionally, background knowledge, e.g., on semantically equivalent tags, can be manually refined and included in the process.

In this way, we obtain an overview on the resources in terms of their location and describing tags. Furthermore, we can characterise different regions, areas or specific locations in terms of such descriptive information. The resulting patterns can be exploited by providing different visualisations and browsing options. Additionally, they can be filtered according to different interestingness criteria. We provide exploratory options for viewing the social media data on different abstraction levels according to the *information seeking mantra* by Shneiderman (1996): overview first (macroscopic view), browsing and zooming (mesoscopic analysis), and details on demand (microscopic focus). The presented approach is embedded into the comprehensive pattern mining and subgroup analytics environment VIKAMINE (http://www.vikamine.org), see Atzmueller and Lemmerich (2012), which was extended with a specialised user interface for handling, presenting and visualising geo-spatial information. Furthermore, VIKAMINE includes several editors for supporting the attribute construction and refinement steps that can be iteratively applied. From a scientific point of view, the tackled problem is interesting as it requires the combination of several distinct areas of research: pattern mining, knowledge

discovery in social media, community detection, mining (geo-)spatial data, visualisation, and interactive data mining.

The overall contributions of the paper can be summarised as follows:

1 We adapt and extend pattern mining techniques to the mining of combined geo-information and tagging information, focusing on two complementing perspectives:
   a first, we propose an approach to identify, describe and characterise closely related regions of resources in terms of descriptive information such as tags
   b second, we present a method for characterising specific points of interest.
2 We propose an incremental approach for including background knowledge about related tags and (semantic) tag similarity. This can be utilised to define tag hierarchies corresponding to topics.
3 In order to avoid a bias in the resource collection, we propose a weighting schema taking the individual user – resource contributions into account.
4 For interactive analysis, we provide a set of visualisations for exploration and inspection of the set of candidate patterns.
5 We demonstrate the impact and validity of the presented approach using publicly available data from the social photo sharing application *Flickr*.

The remainder of the paper is structured as follows: Section 2 discusses related work. After that, Section 3 summarises basics of descriptive pattern mining, and provides general notions of graphs and community mining measures. Section 4 describes the proposed exploratory mining approach. For demonstrating the effectiveness and validity of the presented approach, Section 5 features a structural evaluation and analysis of the automatic approach, and a case study of the exploratory techniques using publicly available data from Flickr. Finally, Section 6 concludes the paper with a summary and directions for future research.

2 Related work

This paper combines approaches from three distinct research areas, that is, pattern mining, mining (geo-)spatial data, and mining social media. There are several variants of pattern mining techniques (Novak et al., 2009), e.g., frequent pattern mining (Han et al., 2007), graph mining approaches (Horváth and Ramon, 2010; Horváth et al., 2006), mining association rules (Agrawal and Srikant, 1994; Lakhal and Stumme, 2005) and closed representations (Boley et al., 2007, 2010) as well as subgroup discovery (Klösgen, 1996; Wrobel, 1997; Atzmueller and Puppe, 2006). We extend common pattern mining approaches in two directions: first, we adapt community pattern mining to the handling of spatial resources, tags and network information. Second, we introduce different specialised target concept functions extending typical *k*-optimal pattern mining approaches.

Atzmueller and Mitzlaff (2011) considered the descriptive mining of user communities in order to identify common interests, e.g., for recommending or browsing indicators of interests and relevant information/tags. A first approach for the

characterisation and description of communities was introduced in Atzmueller et al. (2009), focusing on the description of spammers in the social bookmarking system BibSonomy. In contrast to the approaches mentioned above, in this paper we focus on exploratory pattern mining methods for describing communities, resources and locations. We consider a two way perspective on the respective relations: We describe interesting sets of resources (e.g., photos) using location-information and tags, but also apply tags, patterns and photos for describing locations, e.g., for interactive browsing, extending (Lemmerich and Atzmueller, 2011).

(Geo-)spatial data mining (Koperski et al., 1998) aims to extract new knowledge from spatial databases. This includes destination recommenders, e.g., for tourist information systems (Ceci et al., 2010), and for geographical topic discovery (Yin et al., 2011). Often established problem statements and methods have been transferred to this setting, for example, considering association rules (Appice et al., 2003). Similar to those methods, we incorporate geo-spatial elements for mining communities, and construct distance-based target concepts according to different intuitions. However, for the combination of pattern mining and geo-spatial data, we provide a set of visualisations and interactive browsing options for a semi-automatic mining approach.

Regarding mining social media, specifically 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 (cf. Liu, 2011). Sigurbjörnsson and van Zwol (2008) also analyse Flickr data and provide a characterisation 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 Lindstaedt et al. (2008) who analyse different aspects and contexts of the tag and image data. Abbasi et al. (2009) present a method to identify landmark photos using tags and social Flickr groups. They apply group information and statistical pre-processing of the tags for obtaining interesting landmark photos. In contrast to previously proposed techniques for related tasks, see for example, Kennedy and Naaman (2008), our approach does not require a separate clustering step. Instead, we focus on descriptive patterns in this paper. 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. In contrast to the above automatic approaches, we also present and extend different techniques for a semi-automatic interactive approach.

3 Preliminaries

In the following, we briefly introduce basic notions with respect to graphs and to descriptive pattern mining using subgroup discovery.

3.1 Graphs

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 $w : E \rightarrow \mathbb{R}^+$ that assigns a positive weight to each edge. We identify a *community* of nodes as a set of vertices $C \subseteq V$.

The *degree* d(u) of a node u in a network measures the number of connections it has to other nodes. In weighted graphs the *strength* s(u) is the sum of the weights of all edges containing u, i.e.,

$$\mathrm{s}(u) := \sum_{\{u,v\} \in E} \mathrm{w}\big(\{u, v\}\big).$$

The *adjacency matrix* of a graph is a matrix $A \in \mathbb{R}^{|V| \times |V|}$ such that $A_{u,v} = 1$ iff $\{u, v\} \in E$ for nodes $u, v \in V$. We identify a graph with its according adjacency matrix where appropriate.

For a given graph $G = (V, E)$ and a community $C \subseteq V$ we use the following notation: $n := |V|$, $m := |E|$, $n_C := |C|$, $m_C := |\{\{u, v\} \in E : u, v \in C\}|$ – the number of *intra-edges* of C.

3.2 Pattern mining

Next, we briefly summarise the pattern mining methods in the general context of descriptive pattern mining [also called supervised descriptive rule induction, see Novak et al. (2009)], subgroup discovery for continuous target concepts (Atzmueller and Lemmerich, 2009), and descriptive community mining (Atzmueller and Mitzlaff, 2011). Like subgroup discovery (Klösgen, 1996), descriptive pattern mining aims at identifying patterns, which are interesting with respect to a given target property of interest according to a specific quality (interestingness) measure. The top k patterns are then ranked according to the given quality measure. The main focus of the applied methods is thus the description of the data, that is, of certain communities or subgroups. In our context (see Section 4.2), the target property is either given by the quality of a community of resources, or specifically constructed using a provided location, i.e., a specific point of interest, landmark, or region, identified by geo-coordinates.

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* $sel_{a=a_j}$ is a Boolean function $I \rightarrow \{0, 1\}$ that is true, iff the value of attribute a is equal to a_j for the respective individual. The set of all basic patterns is denoted by S. A *subgroup description* or (complex) *pattern* $sd = \{sel_1,\ldots,sel_l\}$ is then given by a set of basic patterns, 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, since 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_s, iff $sd_s \subset sd$. A *subgroup* (*extension*) $sg_{sd} := ext(sd) := \{i \in I \mid sd(i) = true\}$ 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 in S.

A quality function $Q : 2^S \rightarrow \mathbb{R}$ maps every pattern in the search space to a real number that reflects the interestingness of a pattern (or the pattern's extension, 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 (cf. Geng and Hamilton, 2006), many quality measures trade-off the size $|ext(sd)|$ of a subgroup and the deviation $t - t_0$, where t is the average value of a given target concept in

the subgroup and t_0 the average value of the target concept in the general population. Thus, typical quality functions are of the form

$$q_a(sd) = \left|ext(sd)\right|^a \cdot \left(t - t_0\right),\ a \in [0;1]$$

For binary target concepts, this includes for example the *weighted relative accuracy* for the size parameter $a = 1$ or a simplified binomial function, for $a = 0.5$.

For descriptive community mining, there are special community quality functions: The concept of a *community* intuitively describes a group C of individuals out of a population such that members of C are strongly 'related' among each other but sparsely 'related' to individuals outside of C. This notion translates to vertex sets $C \subseteq V$ of a graph $G = (V, E)$. For descriptive community mining, we associate a description sd_C with C such that $ext(sd_C) = C$. A prominent measure to determine the amount of relatedness is given by the *modularity* MOD (Newman, 2004, 2006; Newman and Girvan, 2004) of a partitioning of a graph with k communities $C_1,\ldots,C_k \subseteq V$. It focuses on the number of edges *within* a community and compares that with the *expected* such number given a null-model (i.e., a corresponding random graph where the node degrees of G are preserved):

$$\mathrm{MOD} = \frac{1}{2m} \sum_{u,v \in V} \left(A_{u,v} - \frac{\mathrm{d}(u)\mathrm{d}(v)}{2m} \right) \delta\left(C(u), C(v)\right), \tag{1}$$

where $C(i)$ denotes for $i \in V$ the community to which node i belongs. $\delta(C(u), C(v))$ is the *Kronecker delta* symbol that equals 1 if $C(u) = C(v)$, and 0 otherwise. The *modularity contribution* of a single community C in a *local context* (sub-graph) can then be computed (Newman, 2006; Nicosia et al., 2009) as:

$$\mathrm{MODL}(C) = \frac{1}{2m} \sum_{u,v \in C} \left(A_{u,v} - \frac{\mathrm{d}(u)\mathrm{d}(v)}{2m} \right),$$

yielding

$$\mathrm{MODL}(C) = \frac{2m_C}{2m} - \sum_{u,v \in C} \frac{\mathrm{d}(u)\mathrm{d}(v)}{4m^2} = \frac{m_C}{m} - \sum_{u,v \in C} \frac{\mathrm{d}(u)\mathrm{d}(v)}{4m^2}.$$

For weighted graphs, the *modularity* measures introduced above can be adapted by accumulating the edges' weights instead of the edges. While the degree of a node is replaced by the node's strength, m, m_C and $\bar{m}_C$ have to be rewritten as follows:

$$m \coloneqq \sum_{\{u,v\} \in E} \mathrm{w}\left(\{u,v\}\right), \quad m_C \coloneqq \sum_{\substack{\{u,v\} \in E,\\ u,v \in C}} \mathrm{w}\left(\{u,v\}\right), \quad \bar{m}_C \coloneqq \sum_{\substack{\{u,v\} \in E,\\ \left|\{u,v\} \cap C\right| = 1}} \mathrm{w}\left(\{u,v\}\right).$$

3.3 Algorithms for descriptive pattern mining

In the following, we briefly summarise two state-of-the-art algorithms that we apply for descriptive pattern mining. Both are efficient algorithms based on branch-and-bound techniques using optimistic estimates for reducing the pattern search space.

For descriptive community mining, we apply the COMODO algorithm (Atzmueller and Mitzlaff, 2011). Using *extended frequent pattern trees* (Atzmueller and Mitzlaff, 2011), COMODO conducts an exhaustive search by traversing a representation of the solution space compiled into a *community pattern tree* (CP-tree). The CP-tree is a compact version of the database D, that also contains relevant information about the graph structure. Using this tree, the patterns can be efficiently computed using only the information contained in the tree. Alltogether, the algorithm requires only two passes through the generated graph dataset. For more details, we refer to Atzmueller and Mitzlaff (2011).

Additionally, we apply the SD-Map* algorithm (Atzmueller and Lemmerich, 2009) for location description in terms of tags. SD-Map* is based on the efficient FP-growth (Han et al., 2000) algorithm for mining frequent patterns. FP-growth applies a divide and conquer method, first mining frequent patterns containing one selector and then recursively mining patterns of size 1 conditioned on the occurrence of a (prefix) 1-selector. SD-Map* utilises the FP-tree structure built in one database pass to efficiently compute quality functions for all subgroups. Furthermore, SD-Map* applies pruning strategies by utilising optimistic estimates of subgroup qualities.

Both COMODO and SD-Map* can apply tight optimistic estimates for pruning the search space by orders of magnitude. This allows for an efficient (interactive) mining process, especially for the quality functions applied in the scope of this paper (cf. Atzmueller and Mitzlaff, 2011; Atzmueller and Lemmerich, 2009).

4 Exploratory pattern mining on social media

In the sections below, we present our approach for exploratory pattern mining on social media: In an interactive and iterative process, we first utilise pattern mining techniques to generate a candidate set of interesting patterns. These candidate patterns are then presented to the user, who can refine the obtained patterns, visualise the patterns and dependencies between these, and adapt parameters for candidate generation in a subsequent iteration. For generating candidate patterns we propose two methods. The first method is based on location-aware descriptive community mining (e.g., for browsing the Berlin/Brandenburg area on a the map of Germany), the second method focuses on identifying characterisations for pre-specified locations (e.g., when clicking on a currently unknown location in the vicinity of the city of Berlin). Thus, we tackle the location-image relation continuum from two opposite but complementing directions.

4.1 Location-aware descriptive community mining

In the following, we present an approach for identifying *characteristic* communities for sets of resources based on their descriptions, e.g., in terms of tags, and information about their geo-location. In this way, we mine a set of *descriptive* patterns for larger areas, which can also be restricted to certain regions of interest. For example, in an interactive browsing approach, the user could select the larger area of Berlin in Germany, for closer inspection of the resulting description, which are concentrated in that area.

For mining location-aware descriptive communities of resources, a community is intuitively defined as a set of nodes that has more and/or better links between its

members compared to the rest of the network. In an intuitive sense, community mining is thus concerned with the identification of dense (cohesive) subgroups (Wasserman and Faust, 1994). Hence, subgroups and communities are rather similar, and we will use the terms interchangeably. Intuitively, we aim at discovering semantically similar resources (photos) which are close together and thus describe certain points of interest well. In the following, we first provide an overview on the proposed approach, before we describe how to generate the used data representation, merging a graph structure with descriptive information.

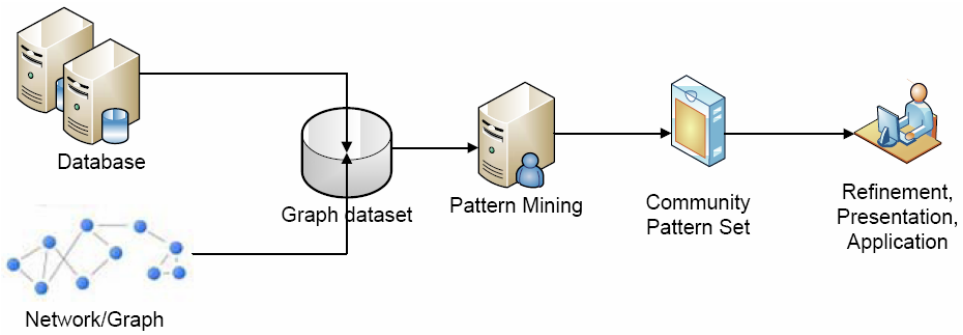
4.1.1 Descriptive community mining on location-resource data

For the location-aware community mining approach, we focus on location-resource data, e.g., photo resources, to which descriptive tags and geo-location information are assigned. Using this data, we can construct a graph G based on the similarity between the nodes and the location information connecting the different nodes (resources), e.g., photos that are taken in close proximity. Additionally, we enrich this graph using the descriptive information of the resources as described in the next section. The construction of the graph is performed according to the semantic similarity of the description of the nodes (resources); in our case, we consider the set of tags assigned to a photo for computing the similarity between photos. Additionally, we consider the *collective relevance* of a photo according to the number of views of the respective resource.

Overall, this approach is able to discover sets of communities described by sets of tags, respectively. For example, we could discover a community described by the tags *museum*, *daniel liebeskind*, and *architecture* corresponding to the location of the jewish museum in Berlin, Germany.

The nodes (resources) of these communities are given by photos which are semantically related and which are close together; for example, they could describe the jewish museum as a point of interest. The information contained in these communities provides then for interesting points, enabling for example exploratory browsing options. In this way, location and descriptive information can be presented at the same time. So, technically our goal is to discover the k best communities in a graph G, that can be described by the attributes of their nodes and that maximise a given community evaluation function. A bird's eye view of the approach is sketched in Figure 1.

Figure 1 Overview on the applied descriptive community mining approach (see online version for colours)



For the description of the communities, we require a database D containing a record for each graph node; in our Flickr example, a data record corresponds to a photo and contains the set of tags for this photo as well as the location (GPS coordinate) where the image was taken. Given a dataset of resources, tags and assigned geo-locations we can then create a location-resource graph using semantic similarity of resources and their distances as described below. The result of the mining process is a set of the k best community patterns characterising specific subsets of the resources, e.g., patterns given in terms of tags describing a set of photos. Intuitively, communities are densely connected sub-graphs. Therefore, we consider only node sets without isolated nodes as candidates for the communities. These top k patterns in the result set are selected according to the given evaluation function. For exploratory mining, introspection, and refinement, a pattern can always be mapped to its *extension*, i.e., a set of resources (photos). This enables a direct visualisation and browsing option, for example, by presenting the patterns together with their extensions being shown on a map.

4.1.2 Location – resource network construction

As outlined above, we need to pre-process the data in order to construct a consolidated data representation for capturing location – resource (photo) relations. For community mining, we aim to obtain a network annotated with descriptive information. For that, we first construct a graph G containing the resources as nodes of the graph.

While it is quite natural to represent the resources by nodes in the graph and to assign the resource properties, e.g., tags, to the respective nodes, there are different options for creating the edges between these nodes. An edge is created between two nodes, if the respective resources are *closely related*, according to their *semantic closeness*. In our case, we base this decision on the semantic similarity using the applied tagging information. Additionally, we assign a weight to an edge denoting the *locational closeness* of the respective resources: This weight is computed according to the distance between the respective locations, such that the closer the location the higher the weight.

For a more convenient notation, we introduce the function $d(u, v)$ to compute the *distance* between the nodes u and v according to their assigned geo-location, for which the distance in km on the earth surface of two points $u = (lat_u, long_u)$ and $v = (lat_v, long_v)$ given latitudes and longitudes can be computed by:

$$d(u, v) = r_e \arccos\left(\sin\left(lat_u\right)\sin\left(lat_v\right) + \cos\left(lat_u\right)\cos\left(lat_v\right)\cos\left(long_v - long_u\right)\right),$$

where r_e is the earth radius in km.

For the derivation of an edge between two nodes u and v, we compute the *semantic closeness* using the descriptive information assigned to the resources, i.e., using the tagging information of the considered photos. There are a number of similarity (or distance) functions for computing semantic closeness for tagging data (Markines et al., 2009; Cattuto et al., 2008). In our case, we opted for a simple but easily interpretable measure. We selected the *jaccard coefficient* (Cattuto et al., 2008) which is defined as follows:

$$jaccard\left(T_u, T_v\right) = \frac{\left|T_u \cap T_v\right|}{\left|T_u \cup T_v\right|},$$

where $T_u \subseteq S$, $T_v \subseteq S$ are the sets of tags (basic patterns, as defined above), which are assigned to u, v respectively. So, if the semantic similarity between the nodes u and v is larger than a certain threshold τ_s, i.e., $jaccard(u, v) \geq \tau_s$, then we create an edge in the graph between u and v. Additionally, we construct weights for the potential edges according to different weighting strategies discussed below. If the weight of a potential edge is 0, then we skip this edge, that is, the edge is not created at all.

In order to support different analysis options and analysis *ranges*, i.e., macroscopic, mesoscopic or microscopic view, we consider different weighting options. This enables the exclusion of *outliers*, i.e., images that are not relevant for the regions of interest.

Continuous distance weight

For deriving a weight $w(\{u, v\})$ based on the 'raw' *continuous* distances between u and v, we can just derive it inversely to the distance, i.e.,

$$w(\{u, v\}) = \min\left(1, \frac{1}{d(u, v)}\right).$$

The advantage of this simple approach is the fact that it is parameter free – no parameter needs to be determined by the user. However, this yields also disadvantages: while it provides a broad *macroscopic* overview, it is rather unfocused. As we will see below, it can be applied for a macroscopic view, i.e., as a first overview on the set of interesting patterns.

Neighbourhood distance weight

For the *neighbourhood distance weight* we create an edge between two nodes, if the spatial *distance* between the resources denoted by the respective nodes is smaller than a certain threshold $d_{\max}$. This threshold can be specified by the user and allows a convenient tuning of the analysis results as discussed below. Basically, the value serves as a distance *cutoff*. Using the maximal distance threshold $d_{\max}$ we obtain a weight

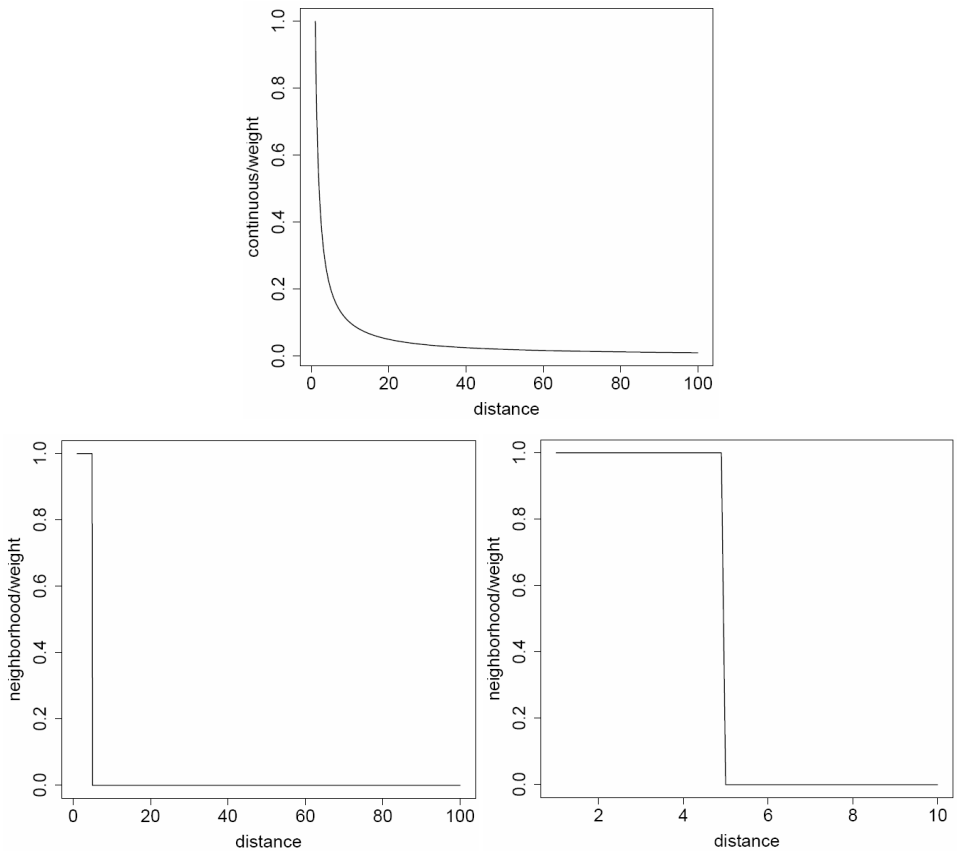
$$w(\{u, v\}) = \begin{cases} 1, & \text{if } d(u, v) < d_{\max} \\ 0, & \text{otherwise} \end{cases},$$

where $distance(u, v)$ computes the distance between the nodes u and v. It is easy to see that this approach is sensitive to the selection of the parameter $d_{\max}$. If $d_{\max}$ is too large then the analysis will not be too focused; if $d_{\max}$ is too small then we will consider only regions corresponding to very small components of the network.

Constructing the network dataset

Finally, using the given graph G and the database D containing the nodes' descriptive information, we build a new dataset focusing on the edges of the graph G: each data record in the new dataset represents an (undirected) edge between two nodes. The attribute values of each such data record are the common attributes of the edge's two nodes. The rationale behind storing only the common attributes is the observation, that an edge can only belong to a community described by a certain attribute value, if this respective attribute value is the same for both nodes of that edge.

Figure 2 The proposed weighting options *continuous distance weight* (*continuous*/*weight*) and *neighbourhood distance weight* (*neighbourhood*/*weight*) (with different scaling factors) with d_{max} = 5 km



In our Flickr example, we consider two photos r_1 and r_2 with tags t_1, t_2, and t_3 and t_1, t_3, and t_4 respectively. If r_1 is connected to r_2, then the transformed data would contain an edge $e = (r_1, r_2)$ with the tags t_1 and t_3 as description. The edge is then represented in the created dataset by a single data record, using the tagging information for the edge as attribute values.

Each such data record also stores the two nodes of the respective edge and their degrees in G to have them available during the evaluation of the quality function q. As described above, we can then directly calculate the given community quality measure.

We apply the efficient COMODO algorithm (Atzmueller and Mitzlaff, 2011) for descriptive community mining using the pattern mining techniques introduced in Section 3.2, especially focusing on the *modularity* quality function. Utilising this algorithm, the presented approach computes a set of communities representing sets of resources (images) that share a similar spatial distribution and are semantically close. In contrast to naive clustering approaches including only tags or distances, we thus enable a much more sophisticated approach: We include semantic information for determining if images are close enough in addition to their respective locations. In this way, we can easily exclude outliers that could affect the clustering results. Furthermore, we propose

flexible weighting options for the resources that can be tuned according to the analysis goals.

4.2 Location-based profile generation of social image media

In this section, we propose an approach for determining descriptions of certain locations by applying pattern mining. It is important to note, that we now focus on describing *specific* locations, instead of considering resource communities for *discovering* locations as above.

The most critical issue for formulating the location-based tag mining problem as a pattern mining task is how to construct a proper target concept capturing the locational interestingness. Therefore, we propose three approaches based on similar principles as discussed above concerning the weighting strategies. Based on the distance to the location of interest we aim at *minimising* a given quality function. Thus, smaller values for the target concept indicate proximity to the location of interest, i.e., better descriptions according to the characterisation task. We apply subgroup discovery for descriptive pattern mining utilising the tagging information. In contrast to clustering approaches that utilise only tagging or distance information, we are able to include both into the mining process: the approach below guarantees to identify the k-best patterns for a given interestingness measure – which are formulated using the distance to a certain point of interest.

4.2.1 Target concept construction

In the following, we propose three different approaches: using the *continuous* distance, a parametrised *neighbourhood* function, and a *fuzzified* neighbourhood function.

Continuous target distance

As the first approach, we could use the 'raw' *continuous* distance of an image to the point of interest as a numeric target property. As discussed above, given latitudes and longitudes the distance on the earth surface of any point $p = (lat_p, long_p)$ to the specified point of interest $c = (lat_c, long_c)$ can be computed by:

$$d(p) = r_e \arccos\Big(\sin\big(lat_p\big)\sin\big(lat_c\big) + \cos\big(lat_p\big)\cos\big(lat_c\big)\cos\big(long_c - long_p\big)\Big),$$

where r_e is the earth radius.

Using the *continuous target distance* as the numeric target concept, the task is to identify patterns, for which the average distance to the point of interest is relatively small. For example, the target concept for an interesting pattern could be described as: "pictures with this tag are on average 25 km from the specified point of interest, but the average distance for all pictures to the point of interest is 455 km".

The advantages of using the numeric target concept is that it is parameter-free and can be easily interpreted by humans. However, it is unable to find tags, which are specific to more than one location. For example, for the location of the Berlin olympic stadium the tag 'olympic' could be regarded as specific for this location. However, if considering other olympic stadiums (e.g., in Munich) the average distance to Berlin is quite large for the tag 'olympic'. Therefore, using the continuous weighting option, images in both locations tagged with 'olympic' would be included but the image for the olympic stadium

in Munich would be assigned a smaller weight and therefore decrease the interestingness of the subgroup.

Neighbourhood target distance

In order to address a better customisability according to the requirements of the user, we propose a second distance function based on the concept of a *close neighbourhood*: The neighbourhood distance requires a maximum distance $d_{\max}$ to the location of interest. Then, the target concept is given by:

$$neighbour(p) = \begin{cases} 0, & \text{if } d(p) \leq d_{\max} \\ 1, & \text{otherwise} \end{cases}$$

Tags are then considered as interesting, if they occur relatively more often in the neighbourhood than in the total population. For example, the target concept for an interesting pattern in this case could be described as: "while only 1% of all pictures are in the neighbourhood of the specified point of interest, 33% for pictures with tag x are in this neighbourhood". The disadvantage of this approach is however, that it is strongly dependent on the chosen parameter $d_{\max}$. If this parameter is too large, then the pattern mining step will not return tags specific for the point of interest, but for the surrounding region. On the other hand, if $d_{\max}$ is too small, then the number of instances in the respective area is very low and thus can easily be influenced by noise.

Fuzzified target distance

The third approach considers a fuzzy variant of the second approach: Instead of a single distance $d_{\max}$ we define a minimum distance $d_{l\max}$ and a maximum distance $d_{u\max}$ for our neighbourhood. Images with a distance smaller than $d_{l\max}$ are then assigned to the neighbourhood completely but only partially for distances between $d_{l\max}$ and $d_{u\max}$.

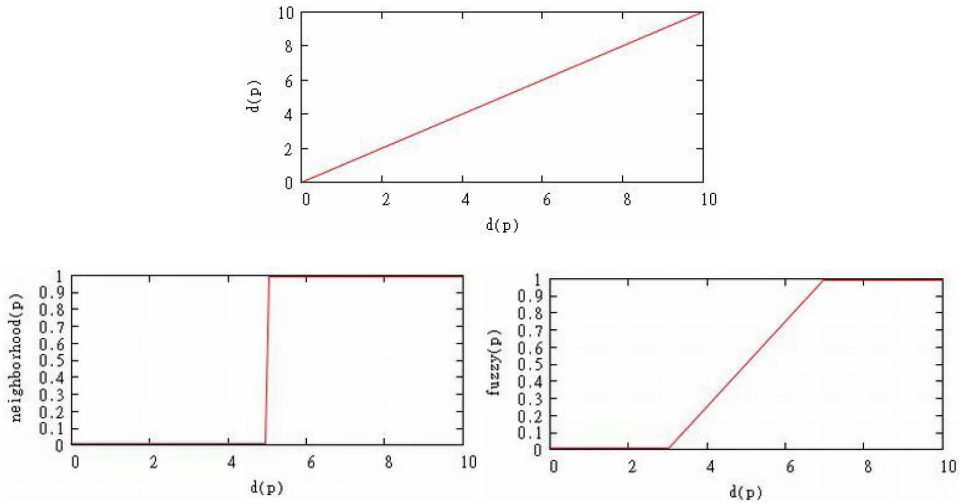
For the transition region between $d_{l\max}$ and $d_{u\max}$ any strictly monotone function could be used. In this paper, we concentrate on the most simple variant, that is, a linear function (possible alternatives include sigmoid-functions like the generalised logistic curve):

$$fuzzy(p) = \begin{cases} 0, & \text{if } d(p) \leq d_{l\max} \\ \dfrac{d(p) - d_{l\max}}{d_{u\max} - d_{l\max}}, & \text{if } d(p) > d_{l\max} \text{ and } d(p) < d_{u\max} \\ 1, & \text{otherwise} \end{cases}$$

In this way, we require the selection of two parameters; however, using such soft boundaries the results are less sensible to slight variations of the chosen parameters. Thus, we achieve a smooth transition between instances within or outside the chosen neighbourhood. Additionally, the selection can often be conveniently supported, e.g., by using a map visualisation.

Figure 3 depicts the described options: the fuzzy function can be regarded as a compromise between the other two function. It combines the steps for the neighbourhood function with a linear part that reflects the common distance function.

Figure 3 The three proposed distance functions $d(p)$, $neighbour(p)$ with a threshold of $dist_{max} = 5$ and $fuzzy(p)$ with thresholds $d_{-} = 3$ and $d_{+} = 7$ as a function over $d(p)$ (see online version for colours)



Note: It can be observed, that d(p) is (obviously) linear, $neighbour(p)$ is a step function, and $fuzzy(p)$ combines both properties in different sections.

4.2.2 Avoiding user bias: user-resource weighting

So far, in the approaches described above, all images are treated as equally important. However, due to the common *power law* distribution between users and resources (images) in social media systems, only a few but very active users contribute a substantial part of the data. Since images from a specific user tend to be concentrated on certain locations and users also often apply a specific vocabulary, this can induce a bias towards the vocabulary of these active users. As an extreme example, consider a single 'power user', who shared hundreds of pictures of a specific event at one location and tags all photos of this event with a unique term. This term could then be considered as very important for that location, although the tag is not commonly used by the overall user base.

One possibility to solve this issue could be to utilise an interestingness measure that also incorporates the user count. That is, one could extend the standard quality function given above by adding a term, that reflects the number of different users that own a picture in the evaluated subgroup. Such an extended quality function could be defined as $q_a(sd) = | ext(sd) |^a \cdot (t - t_0) \cdot | u(sd) |$, where $| u(sd) |$ is the user count for images in the respective subgroup. Unfortunately, such interestingness measures are not supported by efficient exhaustive algorithms for subgroup discovery, e.g., SD-Map* (Atzmueller and Lemmerich, 2009) or BSD (Lemmerich et al., 2010). On the other hand, more basic algorithms, for example exhaustive depth-first search without a specialised data structure scale not very well for the problem setting of this paper, with thousands of tags as descriptions and possibly millions of instances.

Therefore, we propose to apply a slightly different approach to reduce user bias in our application. We assume that a single picture might be overall less important, if a user

shared a large amount of images. This is implemented by applying an instance weight for each resource, that is, for each image in our application. Thus, when generating statistics for a subgroup, the overall count and the target value, which is added, if the respective image i is part of this subgroup, is multiplied by the corresponding weight $w(i)$. The weight is smaller, if more pictures are contributed by the owner of the image. For our experiments, we utilised the weighting function

$$w(i) = \frac{1}{\sqrt{\left(\left|\{ j \mid j \text{ is contributed by the user that contributed } i \}\right|\right)}}.$$

Instance weighting is supported by SD-Map* as well as many other important subgroup discovery algorithms, since it is also applied in pattern set mining approaches such as weighted covering (cf. Lavrac et al., 2004).

4.3 Interactive exploration

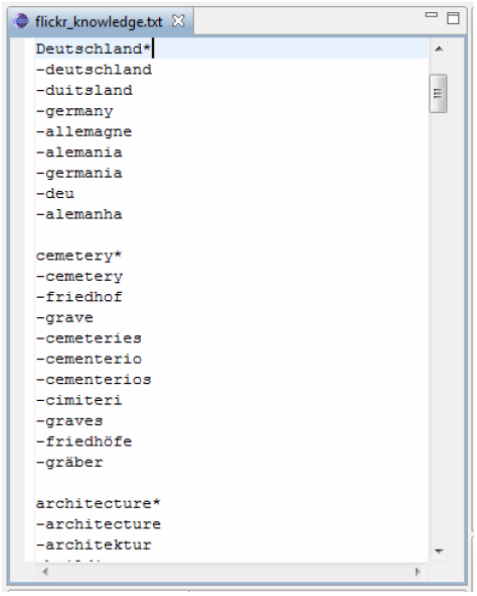
The result of the two methods outlined above, the location-guided descriptive community mining approach, and the location-aware profile generation technique results in a set of patterns, i.e., communities according to interesting (spatially-oriented) topics and descriptions of specified locations. These *candidate patterns* then need to be assessed by the user. Background knowledge for organising and refining the descriptive information is usually helpful (cf. Atzmueller et al., 2005). In the following, we first describe the options for including background knowledge for semi-automatic attribute construction. After that, we describe the different visualisation options.

4.3.1 Semi-automatic attribute construction

Since tags can be freely chosen by the users, often different tags are used for the same concept or (semantically) similar concepts. For an improved analysis such tags should be combined in a single new *meta-tag* or *topic*. To provide such knowledge to the system we propose to apply a semi-automatic approach: In a pre-processing step, we apply an automatic technique, e.g., a LDA-based approach [*latent dirichlet allocation* (Blei et al., 2003)], for generating topic proposals. In this way, we efficiently build interpretable tag clusters, i.e., for obtaining descriptive topic sets. The LDA method itself builds topics capturing semantically similar tags and thus helps to inhibit the problem of synonyms, semantic hierarchies, etc.

In a subsequent refinement step, the set of proposed topics is tuned and refined by the user. The refinement can be performed by editing a text document using dashtrees (Reutelshoefer et al., 2010) as a simple intuitive syntax for defining taxonomic structures, see Figure 4 for an example. For each parent node in the tree a new attribute (topic, *meta tag*) is constructed in the system, that is set to *true* for a single instance, iff at least one of the attributes identified by a child node is *true* in this instance. In this way, hierarchical relations can be effectively modelled.

Figure 4 Editor for specifying background knowledge (tag hierarchies) in textual form, implemented in the VIKAMINE system (http://www.vikamine.org) (see online version for colours)



Notes: The tag hierarchies can be generated, e.g., by LDA-based approaches, and can then be refined by the user manually. For example, the new attribute *cemetery** is constructed that is true, iff the respective image has been tagged by any of the tags beyond (*cemetery*, *friedhof*, *grave*, *cemeteries*, *cementerios*, *cimiteri*, *graves*, *friedhöfe*, *gräber*).

4.3.2 Visualisation

The proposed approaches for location-aware mining are formulated as pattern mining tasks. While such tasks can generate candidate patterns, often only manual inspection by human experts can reveal the most informative patterns. In many cases, the interestingness of images and locations is subjective and dependent on prior knowledge.

As a simple example, if you knowingly choose a point of interest in the city of Berlin, the information, that the tag '*berlin*' is often used there, will not add much knowledge. However, if a point is chosen arbitrarily on the map without any information about the location, then the information that this tag is used frequently in that area is supposedly rather interesting. Therefore, we consider possibilities to interactively explore, analyse and visualise the candidate patterns. We consider three kinds of visualisations:

1 For the exploration of location-resource relations specialised visualisation methods can be exploited that focus on the spatial information. These are especially relevant for browsing and inspecting patterns for a region or specific points of interest.

a The *tag-resource map* visualises the spatial distribution of tags on a dragable and zoomable map. Figure 5 shows an example. Each picture for a specific pattern is represented by a marker on the map. Since for one pattern easily several thousand pictures could apply, we recommend to limit the number of displayed markers. In our case study (see Section 5) we chose a sample of at most 1,000 markers. In a variant of this visualisation also the distribution of sets of tags can be displayed on a single map in order to compare their distributions, see Figure 5. Furthermore, this view also allows for the characterisation of selected areas and regions by browsing interesting tag sets identified by mining interesting communities. Overall, this view allows for a quick and intuitive overview on which tag is used for images at which location.

b The *exemplification view* displays sample images for the currently displayed tag. The view can be filtered with respect to a set of pattern. This is especially important, since pattern exemplification has shown to be essential for many applications (e.g., Atzmueller and Puppe, 2008) and can be applied for characterising both subgroups and communities. Using this view, the overall application can be used to not only browse and explore the used tags with respect to their geo-spatial distribution, but also allows for interactive browsing of the images itself. Since there are possibly too many pictures described by a set of tags to be displayed at once, we propose to select the shown images also with respect to their popularity, i.e., the number of views of the images.

2 For a detailed exploration of the mined profiles and their descriptions given by tag sets, we can utilise various established techniques for interactive pattern mining and subgroup analytics (cf. Novak et al., 2009; Atzmueller and Lemmerich, 2012; Atzmueller and Puppe, 2005). These user interfaces include for example:

a The *zoomtable* allows for interactive browsing of tag distributions considering a currently selected pattern. For numeric targets, it shows the distribution of tags concerning the currently active pattern. For the binary 'neighbour' target concept, it shows more details within the zoom bars, e.g., showing the most interesting factors (tags) for the current pattern and target concept. Clicking on a non-selected tag in the zoomtable adds this tag to the currently selected combination of tags, clicking on an already selected tag removes it from this collections. Thus, the zoomtable allows for interactive exploration of tag combinations. For a more detailed description of this visualisation, we refer to Atzmueller and Puppe (2005). Figure 6 shows an exemplary view on a set of tags.

b The *nt-plot* compares the size and target concept characteristics of many different patterns, see Figure 8 for an example. In this ROC-space related plot (e.g., Flach, 2010), each pattern is represented by a single point in two dimensional space. The position on the x-axis denotes the size of the subgroup, that is, the number of pictures covered by the respective tags. The position on the y-axis describes the value of the target concept for the respective pattern. Thus, a pattern with a high frequency that is not specific for the target location is displayed on the lower right corner of the plot, while a very specific tag, which was not frequently used is displayed on the upper left corner. This visualisation

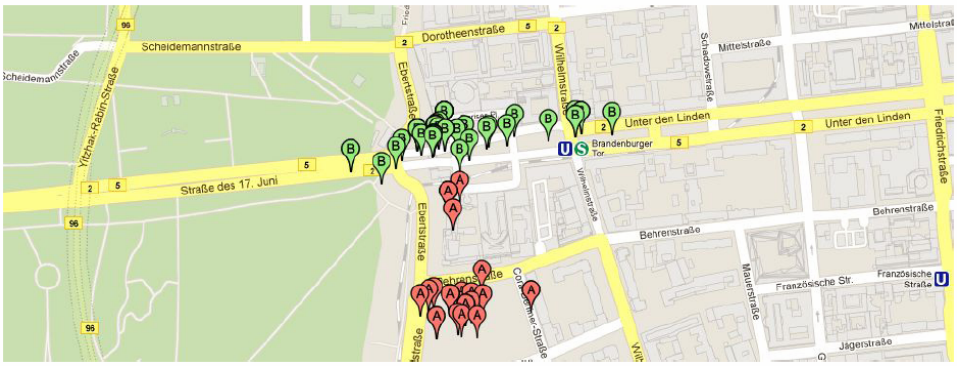
is especially suited to compare the statistical properties for a large amount of patterns.

c The *specialisation graph* is used to show the dependencies between tags (cf. Klösgen and Lauer, 2002). In this graph, each pattern is visualised by a node represented by a two-part bar. The total length of these bars represents the number of cases covered by this pattern, while the ratio between the two parts of the bar represent the value/share of the target concept within the extension of the pattern. Generalisation relations between patterns are depicted by directed edges from more general to more specific patterns. For example, the patterns *fluss* and (*fluss* $\wedge$ *elbe*) are connected by an edge pointing at the latter pattern. Figure 7 shows an example of an specialisation graph.

3 Furthermore, we can apply 'low-level' visualisations for the tag sets and patterns that are mainly used for introspection of candidate patterns, providing a very specific level of detail. Typical visualisations include the contingency table, pie charts, and box plots. A short recent overview on those visualisation techniques including a discussion of usefulness, correctness and intuitiveness is provided in Novak et al. (2009). An especially important visualisation of this category proved to be a distance histogram, cf. Figure 9 for an example. This histogram shows on the x-axis the distances $d(p)$ from the location of interest and on the y-axis the number of images with the specified tag(s) at that distance.

In an iterative approach, the user obtains new insights on the data and can then enter this knowledge, e.g., on different tags describing the same concept, into the system, see Section 4.3.1 or adapt the automatic candidate generation accordingly. The proposed features were implemented as a plugin for the interactive pattern mining and subgroup analytics environment VIKAMINE. For incorporating the traditional plots the VIKAMINE R-Plugin was used as a bridge to the R (http:/www.r-project.org) language for statistical computing.

Figure 5 Example tag-resource visualisation (see online version for colours)

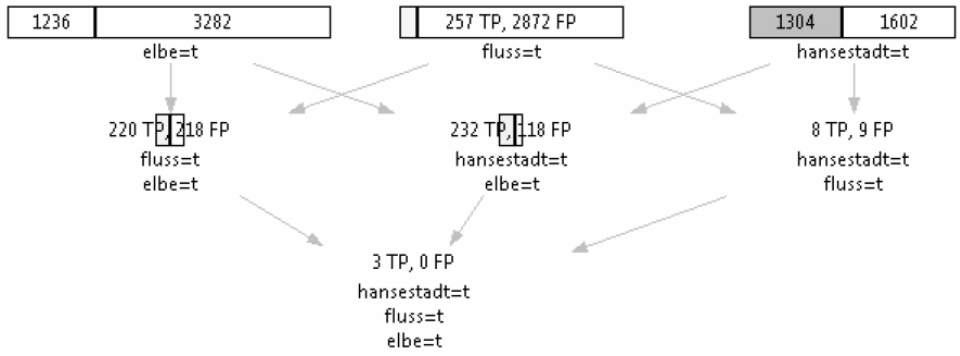


Note: Pictures with tag 'holocaust' are marked with an red 'A', while pictures for the tag 'brandenburgertor' are marked with a green 'B'.

Figure 6 The zoomtable displaying a set of exemplary tags (see online version for colours)

Notes: The rows show the distributions of the individual tags, i.e., a 't' if the tag occurs in the dataset. Green markings show that adding the tag to the current combination of tags will increase the share of images that are geographically near the selected target location.

Figure 7 An exemplary specialisation graph showing the generalisations of the pattern *hansestadt* ∧ *fluss* ∧ *elbe*

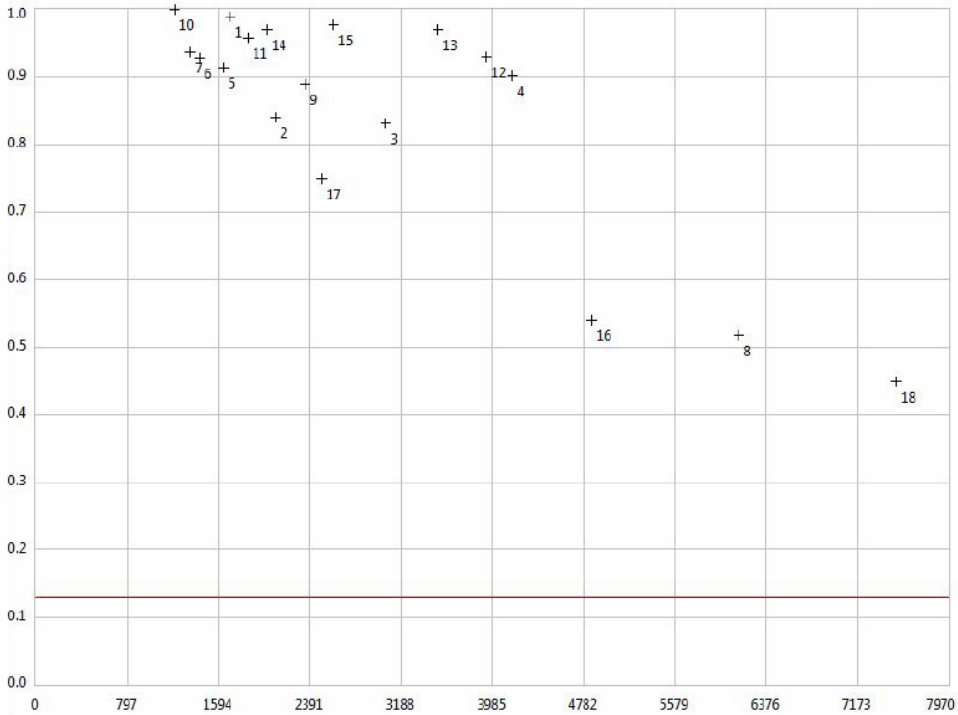


5 Case study: mining Flickr

In the following, we describe an exemplary application of the presented techniques using publicly available data from the resource sharing system Flickr. We collected those images that were taken in 2010 and have a geo-tag within Germany resulting in about 1.1 million images. In the crawling process, we ensured that the crawled images had been available for at least half a year – in order to have a chance for a high view count for each of the images.

Below, we start with a structural analysis and evaluation of the automatic method for mining resource communities, comparing the presented approach to state-of-the-art algorithms. After that, we focus on two application scenarios of the complete exploratory approach: We first describe how to discover interesting locations using descriptive community mining. Second, we show how to characterise locations. For both approaches, we provide an assessment using exemplary results in the respective case study context. The proposed approach includes iterative and incremental steps in order to incorporate subjective views of the user which are fundamental for supporting a final evaluation of the mined results by the user. For the dataset we considered all tags that were used at least 100 times. This resulted in about 11,000 tags.

Figure 8 An exemplary nt-plot for the location Brandenburgertor, for tags with a maximum distance of 5 km (see online version for colours)



Notes: Tags that were used more often are shown on the right side of the diagram, for example, ‘streetart’ (16), ‘graffiti’ (8), or ‘urban’ (18). Tags that are very specific for the given target concept, that is, they appear almost only within a 5 km area of the Berlin Brandenburger Tor, are displayed at the top of the diagram. For example, tags such as ‘heinrichböllstiftung’ (10), ‘alexanderplatz’ (1), or ‘potsdamerplatz’ (14) are very specific (and interesting) for the specified location.

5.1 Structural analysis

For a first analysis, we consider structural properties of the communities discovered by the automatic approach. These communities are candidate patterns in the interactive approach – as hypotheses. Therefore, a solid basis provided by the automatic methods provides the foundation of the whole process.

For the analysis, we constructed different networks using different minimal view count thresholds (τ_{count}) for selecting relevant resources, and different minimal semantic similarity thresholds (τ_{sim}) for constructing the network. We applied the weighting strategies discussed above. Table 1 and 2 depict the properties of the networks including the average node degrees, number of nodes, number of edges, diameter, density and cluster coefficient of the respective network.

Table 1 Different networks for the continuous distance weighting option, with different minimal viewcount and similarity thresholds

τ_{count}	τ_{sim}	*avg(deg)*	*#nodes*	*#edges*	*d*	*density*	*C*
1,000	0.3	57.42	2,160	62,009	18	0.030	0.840
1,000	0.5	56.95	1,482	42,203	6	0.040	0.804
1,000	0.8	43.53	769	16,739	6	0.060	0.580
500	0.3	68.78	6,858	235,835	27	0.010	0.751
500	0.5	47.64	4,929	117,403	22	0.010	0.770
500	0.8	35.52	2,309	41,005	7	0.015	0.534
200	0.3	89.17	27,419	1,222,457	26	0.003	0.772
200	0.5	63.48	20,846	661,653	32	0.003	0.802
200	0.8	36.75	10,959	201,393	12	0.003	0.520

Note: The table shows the average node degree (avg(deg)), the number of nodes (#nodes), the number of edges (#edges), the diameter (d), the density of the graph (density) and its cluster coefficient (C).

For the different networks, we applied the COMODO algorithm as described above searching for 100 best patterns (k = 100). In order to assess the structural validity of the proposed approach, we compared our approach to prominent approaches for detecting overlapping communities. We considered the MOSES and the COPRA algorithms, see McDaid and Hurley (2010) and Gregory (2009) respectively, as a reference. In these experiments, we required a minimal community size of at least ten nodes. Since MOSES and COPRA do not accept a minimum size as input, we applied a post-processing step for the their discovered communities and filtered all communities below that size. Additionally, for the COMODO algorithm we applied a minimal improvement filter (cf. Bayardo et al., 2000), for the community patterns, and pruned all specialisations for which the absolute difference to the quality of their parent patterns was smaller than 0.01.

Table 2 Different networks for the neighbourhood distance weighting option, with a distance threshold of 1 km and different minimal viewcount and similarity thresholds

τ_{count}	τ_{sim}	*avg(deg)*	*#nodes*	*#edges*	*d*	*density*	*C*
1,000	0.3	26.43	1,015	13,413	9	0.030	0.03
1,000	0.5	30.73	728	11,186	4	0.040	0.027
1,000	0.8	29.42	359	5,280	4	0.080	0.005
500	0.3	30.27	2,987	45,207	9	0.010	0.140
500	0.5	24.92	1,948	24,271	7	0.010	0.180
500	0.8	22.36	923	10,317	4	0.020	0.130
200	0.3	30.94	14,106	218,187	26	0.003	0.390
200	0.5	29.92	9,829	147,055	11	0.003	0.350
200	0.8	18.44	4,959	45,721	9	0.004	0.285

Note: The table shows the average node degree (avg(deg)), the number of nodes (#nodes), the number of edges (#edges), the diameter (d), the density of the graph (density) and its cluster coefficient (C).

Tables 3 to 6 show exemplary results for the networks in Tables 1 to 2. In addition to the number of the discovered communities, we include the respective sizes for a first overview of the properties of the communities (and induced sub-graphs). However, for a

comprehensive assessment, they need to be inspected with some insight (e.g., Schaeffer, 2007). Therefore, we also evaluated the obtained communities using the significance test described in Koyuturk et al. (2007) for testing the statistical significance of the density of the sub-graph induced by a community against a corresponding null-model.

In our experiments we observed that COMODO tends to return substantially larger communities in comparison to the other algorithms. Additionally, the communities described by COMODO are always statistically significant. In contrast, for the Moses and Copra algorithms up to 60% of the discovered communities do not pass a significance test of the required significance level of $\alpha < 0.01$, cf. Tables 3 to 6. Furthermore, in these experiments the *p*-values obtained from the COMODO results are usually by far stricter than those by the other algorithms and much stricter than required. In particular, for none of the communities discovered by COMODO the p-value exceeded 10^{-10}. This is especially important, since community mining – as pattern mining in general – suffers from the multiple comparison problem (see Holm, 1979).

Table 3 Comparison of different community detection algorithms on the continuous distance networks for a minimal viewcount of 1,000

τ_{sim}	*Comodo*			*Copra*			*Moses*		
	n	*Size*	*PS*	*n*	*Size*	*PS*	*n*	*Size*	*PS*
0.3	100	126.8 ± 74.0	100%	37	37.1 ± 37.2	68%	44	26.4 ± 32.3	68%
0.5	91	119.4 ± 71.3	100%	21	31.2 ± 43.6	76%	25	31.2 ± 43.8	56%
0.8	72	129.9 ± 62.4	100%	6	41.5 ± 67.0	67%	11	31.0 ± 34.9	82%

Note: The table includes the semantic similarity threshold τ_{sim}, the number of communities (n), the mean sizes, and the share (PS) of statistically significant communities according to a p-value of at least 0.01.

Table 4 Comparison of different community detection algorithms on the continuous distance networks for a minimal viewcount of 500

τ_{sim}	*Comodo*			*Copra*			*Moses*		
	n	*Size*	*PS*	*n*	*Size*	*PS*	*n*	*Size*	*PS*
0.3	58	158.5 ± 76.1	100%	133	32.0 ± 51.0	54%	178	29.1 ± 39.9	46%
0.5	61	150.0 ± 78.7	100%	74	29.4 ± 40.0	46%	101	27.5 ± 28.4	57%
0.8	58	133.0 ± 9.3	100%	22	34.4 ± 41.3	73%	37	27.8 ± 29.5	56%

Note: The table includes the semantic similarity threshold τ_{sim}, the number of communities (n), the mean sizes, and the share (PS) of statistically significant communities according to a p-value of at least 0.01.

Table 5 Comparison of different community detection algorithms on a network with a neighbourhood distance of 1 km for a minimal viewcount of 500

τ_{sim}	*Comodo*			*Copra*			*Moses*		
	n	*Size*	*PS*	*n*	*Size*	*PS*	*n*	*Size*	*PS*
0.3	73	180.0 ± 83.7	100%	33	17.3 ± 9.4	48%	37	21.4 ± 17.8	46%
0.5	56	190.0 ± 85.2	100%	13	20.1 ± 13.0	46%	19	21.2 ± 16.6	63%
0.8	58	167.10 ± 60.35	100%	6	18.2 ± 5.7	67%	5	28.8 ± 24.5	80%

Note: The table includes the semantic similarity threshold τ_{sim}, the number of communities (n), the mean sizes, and the share (PS) of statistically significant communities according to a p-value of at least 0.01.

Table 6 Comparison of different community detection algorithms on a network with a neighbourhood distance of 1 km for a minimal viewcount of 200

τ_{sim}	*Comodo*			*Copra*			*Moses*		
	n	*Size*	*PS*	*n*	*Size*	*PS*	*n*	*Size*	*PS*
0.3	100	434.0 ± 194.8	100%	199	21.3 ± 15.6	40%	224	23.2 ± 27.8	44%
0.5	100	435.3 ± 173.9	100%	104	20.1 ± 16.6	48%	129	24.1 ± 32.7	50%
0.8	100	450.5 ± 262.6	100%	40	21.0 ± 13.9	55%	68	20.8 ± 17.9	44%

Note: The table includes the semantic similarity threshold τ_{sim}, the number of communities (n), the mean sizes, and the share (PS) of statistically significant communities according to a p-value of at least 0.01.

5.2 Explorative application scenarios

In the following, we focus on exemplary application scenarios of the presented exploratory pattern mining techniques. In an iterative approach, we first describe how to discover interesting locations using descriptive community mining. Second, we show how to characterise locations. In the examples below, we experimented with different parameters and thresholds. These always need to be refined by the user in an interactive approach in order to include all of the subjective interestingness criteria of the user. As we will see below, the parameters and thresholds can be quite intuitively adapted, from general to specific, or vice versa.

For the collected tagging data, we applied data cleaning and pre-processing methods, e.g., stemming and LDA for synonym identification as outlined above. In order to identify equivalent tags and combine them within the system we used our semi-automatic attribute construction technique. To do so, first a latent dirichlet allocation was performed on the dataset to obtain a set of 100 candidate topics. The results were manually evaluated and transformed in a dash-tree format, see Section 4.3.1. The input format was then used to construct new meta-tags (topics) that are treated like regular tags. Additionally, the tags that were used to build these meta-tags were excluded from candidate generation.

The automatically constructed tags were of mixed quality: For a few topics the describing tags could be almost directly used as equivalent tags. For example, one resulting topic of the LDA was given by the tags: *cemetery*, *friedhof*, *grave*, *cimetičre*, *cemeteries*, *cementerio*, *friedhöfe*, *cementerios*, *cemitério*, *cimiteri*, *cimetičres*, *cemitérios* and *graves*. The majority of the topics included several tags that can be considered as equivalent, but include other tags as well, for example: *architecture*, *building*, *architektur*, *church*, *dom*, *cathedral*, *germany*, *tower*, *gebäude*, *window*, glass. Some of these tags can be used to construct a new meta-tag by manual refinement, e.g., *architecture*, *building* and *architektur*, however the tags *germany* or *glass* should not be used for this purpose. The last group of topics consisted of rather loosely related tags, for example: *winter*, *thuringia*, *snow*, *town*, *tree*, *village*, *sky*. These topics were considered inappropriate for the purpose of constructing expressive attributes.

In summary, LDA provided for a very good starting point to find equivalent tags. However, applying only the automatic method was far from a quality level that enabled us to use the results directly to construct clear meaningful and comprehensible combined

tags. The text-based format in our mining environment proved to be easy to use and well-fit for this purpose.

5.2.1 Discovering interesting locations: descriptive community mining

For discovering interesting locations using descriptive community mining on networks of related photos, we applied different weighting options discussed in Section 4.1. We considered the 1.1 million images for discovering photo communities, with different minimal semantic similarity thresholds. For identifying prominent images, we restricted the analysis to photos with a view count of at least 100.

In our case study, we focused on the larger Berlin/Brandenburg area, e.g., using a *tag-resource* view, cf. Figure 5. That is, we discover patterns describing interesting pictures (resources) that are densely connected and semantically similar according to their assigned tags, focusing on the resources in the larger Berlin/Brandenburg area. In this way, we obtain subgroups of images occurring in this target region.

Table 7 shows the results of the continuous weighting option for the region of interest, using a tag similarity threshold $\tau_s = 0.3$. It is evident, that this results in very general relations (and communities), e.g., focusing on sports, athletics, cars, and architecture related images; several track and field athletics and athletic championships took place in this area in 2010. Thus, the continuous weighting option gives a very *broad view* on the relations and can be used for a first browing and overview on selected regions – for a macroscopic view. Compared to the microscopic results shown below, it is easy to see that the macroscopic option contains a diverse set of topics and can be used for a first overview and browsing.

Table 7 Top community patterns with the continuous distance weighting option selected from the greater Berlin/Brandenburg area for a minimal semantic similarity threshold of $\tau_s = 0.3$

Description	*Community size*	*Quality*
athletics	246	0.24
leichtathletik	231	0.22
sport	347	0.14
sport $\wedge$ athletics	176	0.13
car	483	0.09
arquitectura	470	0.05
coche	129	0.05
vintage $\wedge$ car	122	0.05
design	294	0.05
exotic	238	0.05

If we increase the similarity threshold a little $\tau_s = 0.8$, we observe that the patterns tend to concentrate on more *specific* topics, while overall they still show a broad view on the data, cf. Table 8.

Table 8 Top community patterns with the continuous distance weighting option selected from the greater Berlin/Brandenburg area for a minimal semantic similarity threshold of $\tau_s = 0.8$

Description	*Community size*	*Quality*
sport	265	0.24
athletics	159	0.22
arquitectura	366	0.16
vidrio	210	0.15
museo	157	0.14
fachada ∧ arquitectura	152	0.13
patio ∧ vidrio	152	0.13
daniellibeskind	145	0.12
kreuzberg	145	0.12
daniellibeskind ∧ irregular	132	0.12

For a *microscopic view*, the neighbourhood distance weighting approach provides more focused results as shown in Table 9. It is easy to see that this approach provides for interesting *topic* communities, for example, regarding architecture, museums, sports, or specific districts of Berlin (Kreuzberg).

If we compare Table 7, Table 8 and Table 9, then we observe, that the microscopic results (Table 9) contain much longer descriptions and more level of detail. This is actually what we expected, since longer descriptions allow for much more detailed information – in a microscopic view.

Table 9 Top community patterns with the neighbourhood weighting option ($d_{\max} = 1$ km) selected from the greater Berling/Brandenburg area, using a minimal semantic similarity threshold $\tau_s = 0.8$

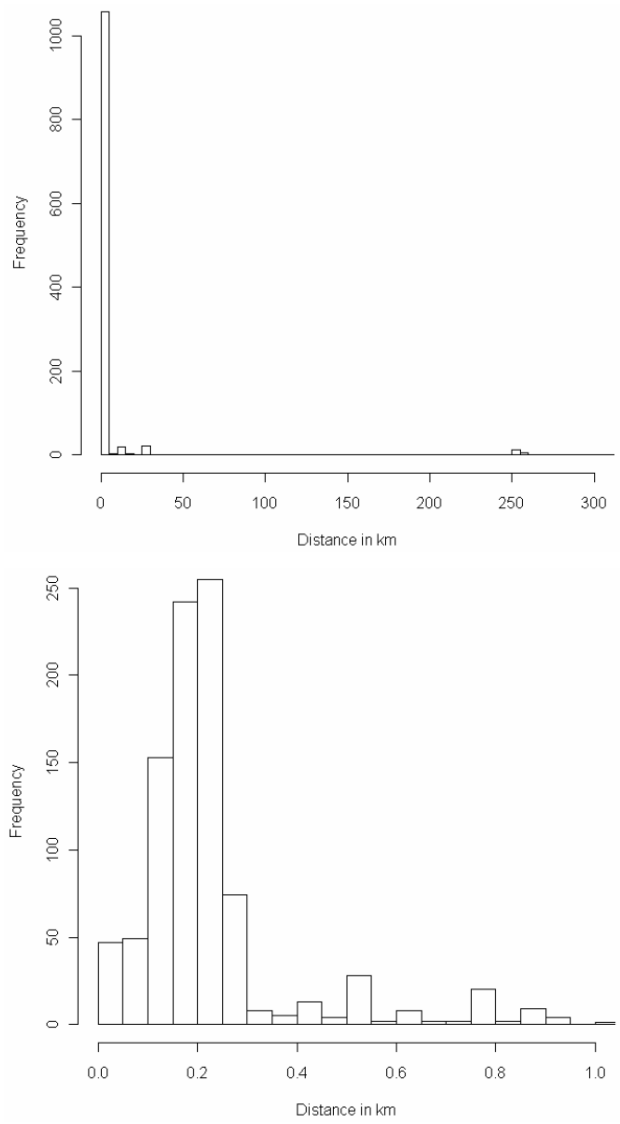
Description	*Community size*	*Quality*
arquitectura	363	0.24
vidrio ∧ arquitectura	210	0.23
fachada ∧ arquitectura	152	0.22
kreuzberg	181	0.21
holocaust ∧ daniellibeskind ∧ kreuzberg	127	0.21
fachada ∧ patio ∧ vidrio ∧ museo	125	0.21
kreuzberg ∧ arquitectura ∧ museo	125	0.21
historic ∧ museo ∧ patio	125	0.21
potsdamerplatz ∧ arquitectura	108	0.11
leichtathletik ∧ athletics ∧ sport	98	0.11

5.2.2 Characterising locations: profile generation using social media

In the following, we characterise locations by identifying tag combinations which are interesting for the specified location. In our first example we consider the city centre of Berlin, more precisely, the location of the Brandenburger Tor. The expected tags were,

for example, *brandenburgertor*, *reichstag*, *holocaustmemorial* (since this memorial is nearby). Of course, also the tag *berlin* is to be expected. An exemplary tag-map for the tag *brandenburgertor* is shown in Figure 5. Figure 9 shows the distance distribution of this tag to the actual location.

Figure 9 Histogram showing the distances of pictures with the tag 'brandenburgertor' to the actual location



Notes: It can be seen in the left histogram that the tag is very specific, since the vast majority of pictures with this tag is within a 5 km range of the location. The histogram on the right side shows the distance distribution up to 1 km in detail. It can be observed that most pictures are taken at a distance of about 200 m to the sight.

First we investigated, which candidate tags were returned by an automatic search using the different proposed target concept options. The results are shown in Tables 10 to 14. For pattern mining, we applied the proposed quality function with $a = 0.5$.

Table 10 shows, that the results include several tags, which are not very specific for the location of interest, but for another nearby location, for example the tags *potsdam* or *leipzig* for cities close to Berlin. This can be explained by the fact, that these tags are quite popular and the average distance for pictures with this tag is relatively low in comparison to the total population even if pictures do not correspond to the location of interest itself, but for a nearby location. Since the use of the distance function $d(p)$ does not allow for parametrisation, it is difficult to adapt the search, such that those tags are excluded.

Tables 11 to 13 show the *neighbour* function with different distance thresholds d_{max}, from 0.1 km to 5 km. It is important to note that we show the *neighbour share* in the result tables, i.e., the share of pictures within the range of interest defined by the maximal distance d_{max}. The results for this target concept are strongly dependent on this threshold. For a very small value of $d_{max} = 0.1$ km the results seem to be strongly influenced by some kind of noise, since the number of pictures in this neighbourhood is relatively small. For example it includes the tags *metro*, *gleis* (translated: 'rail track') or *verkehrsmittel* (translated 'means of transport'). While these tags should occur more often in urban areas, they are by no means the most representative tags for the area around the Brandenburger Tor.

Table 10 Brandenburger Tor: the top patterns (max. description size 1) for the common mean target distance function, cf. Section 4.2.1

Tag	*Size*	*Mean target distance (km)*
berlin	113,977	10.48
potsdam	5,533	26.83
brandenburg	5,911	47.33
leipzig	10,794	147.87
kreuzberg	3,935	14.11
leute	4,547	53.37
berlinmitte	3,054	4.76

Table 11 Brandenburger Tor: the top patterns (description size 1) for the target distance function *neighbour*, cf. Section 4.2.1, with $d_{max} = 0.1$ km

Description	*Subgroup size*	*Neighbour share*
wachsfigur	322	0.99
madametussauds	177	0.853
verkehrsmittel	163	0.313
metro	469	0.277
berlinunderground	158	0.247
brandenburgertor	1,136	0.085
gleis	375	0.085

Notes: The *neighbour share* indicates the share of pictures within the range of interest defined by the maximal distance.

Table 12 Brandenburger Tor: the top patterns (description size 1) for the target distance function *neighbour*, cf. Section 4.2.1, with d_{max} = 1 km

Description	*Subgroup size*	*Neighbour share*
reichstag	2,604	0.829
heinrichböllstiftung	1,211	0.988
brandenburgertor	1,136	0.816
sonycenter	803	0.923
gendarmenmarkt	696	0.885
potsdamer	577	0.88
panoramapunkt	271	1

Notes: The *neighbour share* indicates the share of pictures within the range of interest defined by the maximal distance.

Table 13 Brandenburger Tor: the top patterns (description size 1) for the target distance function *neighbour*, cf. Section 4.2.1, and a threshold d_{max} = 5 km

Description	*Subgroup size*	*Neighbour share*	*User count*
kreuzberg	3,933	0.961	405
mitte	3,507	0.972	404
reichstag	2,604	0.976	680
potsdamerplatz	2,017	0.97	375
karnevalderkulturen	1851	0.958	36
alexanderplatz	1,699	0.989	546
heinrichböllstiftung	1,211	1	3

Notes: The last column shows the overall count of users that used this description. The *neighbour share* indicates the share of pictures within the range of interest defined by the maximal distance.

In contrast, the parameter d_{max} = 1 km yields results that do meet our expectations. The resulting tags reflects the most important sites in that area according to travel guides, including *reichstag*, *brandenburgertor*, *potsdamerplatz* and *sonycenter*. We consider these tags as the most interesting and representative for this given location. However, we do not assume that this parameter will lead to the best result in all circumstances. For example, in more rural areas, where more landscape pictures with a larger distances to depicted objects are taken, we expect that a larger value of d_{max} might be needed.

As shown in Table 13, for a parameter of d_{max} = 5 km the results are tags, which are specific for Berlin as a whole, but not necessarily for the area around the Brandenburger Tor. The results include tags like *kreuzberg* or *alexanderplatz*, which describe other areas in Berlin.

Table 14 exemplifies the fuzzified distance function ranging from 1 km to 5 km as lower and upper thresholds. The results indicate, that this function is less sensitive to the parameter choices. Therefore, selecting the parameter is less difficult; distances like 1–5 km as in the presented example can be applied for a microscopic to a mesoscopic perspective. The collected results form a nice compromise between the results of the respective *neighbour* functions with the different thresholds discussed above (see Tables 11–13 for reference).

Table 14 Brandenburger Tor: the top 7 patterns (description size 1) for the 'fuzzified' target distance function, cf. Section 4.2.1, ranging from 1 km to 5 km

Description	*Subgroup size*	*Mean target distance*
reichstag	2,604	0.05
potsdamerplatz	2,017	0.05
berlinmitte	3,053	0.30
heinrichböllstiftung	1,211	0.01
brandenburgertor	1,136	0.10
alexanderplatz	1,699	0.28
sonycenter	803	0.02

5.2.3 Including instance weighting

Taking a closer look at the results of Table 13 most of the resulting tags provide a good description of the larger area of Berlin. However, there are a few exceptions: *karnevalderkulturen* describes a seasonal well known, but not indicative event in Berlin. *heinrichböllstiftung* is a political foundation, for which the headquarters are located in Berlin. While both tags are certainly associated with Berlin, one would not expect them to be as important or typical for Berlin as other descriptions. The occurrence of these tags can be explained by a few 'power users' that extensively used these tags for many images.

To show this effect, we added an additional column to Table 13, which computes the overall count of users that used that description. For example the tag *heinrichböllstiftung* was applied for 1,211 images, but only by three different users.

To avoid such results in the candidate generation, we apply an instance (resource) weighting as described in Section 2.3. The results are presented in Table 15 and show a more focused tag presentation. Thus, we consider the attribute weighting as appropriate to reduce bias towards the vocabulary of only a few but very active users, as shown in the example.

Table 15 Brandenburger Tor: top patterns (description size 1) using instance weighting for the target distance function *neighbour*, cf. Section 4.2.1, and a threshold $d_{\max} = 5$ km

Description	*Weighted subgroup size*	*Weighted neighbour share*	*User count*
reichstag	431.9	0.972	680
mitte	366.3	0.97	404
kreuzberg	371	0.96	405
alexanderplatz	275.6	0.982	546
berlinwall	237.8	0.945	275
potsdamerplatz	196.4	0.963	375
brandenburgertor	139.4	0.931	332

Notes: The last column shows the overall count of users that used this description.

6 Conclusions

In this paper, we presented exploratory pattern mining techniques for describing social media based on tagging information and collaborative geo-reference annotations. We proposed methods for obtaining sets of tags that describe interesting communities of resources (e.g., images), and discover interesting tag descriptions for locations of interest. For assessing both interesting resources and tag descriptions, we considered semantic closeness and geographical distance. Additionally, we provided an exploratory approach for mining, browsing and visualising a set of candidate patterns. This enables several options including selectable analysis-specific interestingness measures and semiautomatic feature construction techniques. In an interactive process, the results can then be visualised, introspected and refined. For demonstrating the applicability and effectiveness, we presented a case study using real-world data from the photo sharing application Flickr.

For future work, we aim to consider richer location descriptions as well as further descriptive data besides tags, e.g., social friendship links in the photo sharing application, or other link data from social networks. Also, the integration of information extraction techniques (e.g., Atzmueller et al., 2011) seems promising, in order to add information from the textual descriptions of the images. Furthermore, we plan to include more semantics concerning the tags, such that a greater detail of relations between the tags can be implemented in the pre-processing, the mining, and the presentation.

Acknowledgements

This work has partially 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

Abbasi, R., Chernov, S., Nejdl, W., Paiu, R. and Staab, S. (2009) 'Exploiting Flickr tags and groups for finding landmark photos', *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR '09*, Berlin, Heidelberg, Springer-Verlag, pp.654–661.

Agrawal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rules', in Bocca, J.B., Jarke, M. and Zaniolo, C. (Eds.): *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, Morgan Kaufmann, pp.487–499, ISBN 1-55860-153-8.

Appice, A., Ceci, M., Lanza, A., Lisi, F.A. and Malerba, D. (2003) 'Discovery of spatial association rules in geo-referenced census data: a relational mining approach', *Intelligent Data Analysis*, Vol. 7, No. 6, pp.541–566.

Atzmueller, M. and Lemmerich, F. (2009) 'Fast subgroup discovery for continuous target concepts', *Proc. 18th International Symposium on Methodologies for Intelligent Systems (ISMIS 2009)*, LNCS.

Atzmueller, M. and Lemmerich, F. (2012) 'VIKAMINE – open-source subgroup discovery, pattern mining, and analytics', *Proc. ECML/PKDD 2012: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Berlin, Springer Verlag.

Atzmueller, M. and Mitzlaff, F. (2011) 'Efficient descriptive community mining', *Proc. 24th International FLAIRS Conference*, AAAI Press, pp.459–464.

Atzmueller, M. and Puppe, F. (2005) 'Semi-automatic visual subgroup mining using VIKAMINE', *Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining*, Vol. 11, No. 11, pp.1752–1765.

Atzmueller, M. and Puppe, F. (2006) 'SD-map – a fast algorithm for exhaustive subgroup discovery', *Proc. 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, No. 4213 in LNAI, Berlin, Springer Verlag, pp.6–17.

Atzmueller, M. and Puppe, F. (2008) 'A case-based approach for characterization and analysis of subgroup patterns', *Journal of Applied Intelligence*, Vol. 28, No. 3, pp.210–221.

Atzmueller, M., Beer, S. and Puppe, F. (2011) 'Data Mining, validation and collaborative knowledge capture', in Brüggemann, S. and d'Amato, C. (Eds.): *Collaboration and the Semantic Web: Social Networks, Knowledge Networks and Knowledge Resources*, IGI Global.

Atzmueller, M., Lemmerich, F., Krause, B. and Hotho, A. (2009) 'Who are the spammers? Understandable local patterns for concept description', *Proc. 7th Conference on Computer Methods and Systems*.

Atzmueller, M., Puppe, F. and Buscher, H-P. (2005) 'Exploiting background knowledge for knowledge-intensive subgroup discovery', *Proc. 19th Intl. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland, pp.647–652.

Bayardo, R., Agrawal, R. and Gunopulos, D. (2000) 'Constraint-based rule mining in large, dense databases', *Data Mining and Knowledge Discovery*, Vol. 4, Nos. 2–3, pp.217–240, ISSN 1384-5810. 10.1023/A:1009895914772.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent Dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, No. 3/1/2003, pp.993–1022.

Boley, M., Horváth, T., Poigné, A. and Wrobel, S. (2007) 'Efficient closed pattern mining in strongly accessible set systems (extended abstract)', *PKDD*, pp.382–389.

Boley, M., Horváth, T., Poigné, A. and Wrobel, S. (2010) 'Listing closed sets of strongly accessible set systems with applications to data mining', *Theoretical Computer Science*, Vol. 411, No. 3, pp.691–700.

Cattuto, C., Benz, D., Hotho, A. and Stumme, G. (2008) 'Semantic grounding of tag relatedness in social bookmarking systems', in Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W. and Thirunarayan, K. (Eds.): *The Semantic Web – ISWC 2008, Proc. Intl. Semantic Web Conference 2008*, Vol. 5318 of LNAI, Heidelberg, Springer, pp.615–631.

Ceci, M., Appice, A. and Malerba, D. (2010) 'Time-slice density estimation for semantic-based tourist destination suggestion', *Proc. 19th European Conference on Artificial Intelligence (ECAI 2010)*, Amsterdam, The Netherlands, IOS Press, ISBN 978-1-60750-605-8, pp.1107–1108.

Flach, P.A. (2010) 'ROC analysis', in Sammut, C. and Webb, G.I. (Eds.): *Encyclopedia of Machine Learning*, Springer, ISBN 978-0-387-30768-8, pp.869–875.

Geng, L. and Hamilton, H.J. (2006) 'Interestingness measures for data mining: a survey', *ACM Computing Surveys*, Vol. 38, No. 3, pp.1–32.

Gregory, S. (2009) 'Finding overlapping communities using disjoint community detection algorithms', in Fortunato, S., Mangioni, G., Menezes, R. and Nicosia, V. (Eds.): *Complex Networks*, Vol. 207 of Studies in Computational Intelligence, Springer, Berlin, pp.47–61.

Han, J., Cheng, H., Xin, D. and Yan, X. (2007) 'Frequent pattern mining: current status and future directions', *Data Mining and Knowledge Discovery*, Vol. 15, No. 1, pp.55–86, ISSN 1384-5810.

Han, J., Pei, J. and Yin, Y. (2000) 'Mining frequent patterns without candidate generation', *2000 ACM SIGMOD Intl. Conf. on Management of Data*, ACM Press, ISBN 1-58113-218-2, pp.1–12.

Holm, S. (1979) 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics*, Vol. 6, pp.65–70.

Horváth, T. and Ramon, J. (2010) 'Efficient frequent connected subgraph mining in graphs of bounded tree-width', *Theor. Comput. Sci.*, June, Vol. 411, Nos. 31–33, pp.2784–2797, ISSN 0304-3975.

Horváth, T., Ramon, J. and Wrobel, S. (2006) 'Frequent subgraph mining in outerplanar graphs', *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, New York, NY, USA, ACM, ISBN 1-59593-339-5, pp.197–206.

Kennedy, L.S. and Naaman, M. (2008) 'Generating diverse and representative image search results for landmarks', *Proceeding of the 17th International Conference on World Wide Web*, ACM, pp.297–306.

Klösgen, W. (1996) 'Explora: a multipattern and multistrategy discovery assistant', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.): *Advances in Knowledge Discovery and Data Mining*, pp.249–271, AAAI Press, Palo Alto, CA, USA.

Klösgen, W. and Lauer, S.R.W. (2002) *Handbook of Data Mining and Knowledge Discovery*, Chapter 20.1: Visualization of Data Mining Results, Oxford University Press, New York.

Koperski, K., Han, J. and Adhikary, J. (1998) 'Mining knowledge in geographical data', *Communications of the ACM*, Vol. 26, No. 1, pp.65–74.

Koyuturk, M., Szpankowski, W. and Grama, A. (2007) 'Assessing significance of connectivity and conservation in protein interaction networks', *Journal of Computational Biology*, Vol. 14, No. 6, pp.747–764.

Lakhal, L. and Stumme, G. (2005) 'Efficient mining of association rules based on formal concept analysis', in Ganter, B. et al. (Eds.): *Formal Concept Analysis, Foundations, and Applications*, pp.180–195, ISBN 3-540-27891-5.

Lavrac, N., Kavsek, B., Flach, P. and Todorovski, L. (2004) 'Subgroup discovery with CN2-SD', *Journal of Machine Learning Research*, February, Vol. 5, pp.153–188.

Lemmerich, F. and Atzmueller, M. (2011) 'Modeling location-based profiles of social image media using explorative pattern mining', *Proc. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (Passat)* and *2011 IEEE Third International Conference on Social Computing (Socialcom)*, Boston, MA, USA, IEEE Computer Society.

Lemmerich, F., Rohlfs, M. and Atzmueller, M. (2010) 'Fast discovery of relevant subgroup patterns', *Proc. 23rd FLAIRS Conference*.

Lindstaedt, S., Pammer, V., Mörzinger, R., Kern, R., Mülner, H. and Wagner, C. (2008) 'Recommending tags for pictures based on text, visual content and user context', *Proc. 3rd International Conference on Internet and Web Applications and Services*, pp.506–511, Washington, DC, USA, IEEE Computer Society.

Liu, Z. (2011) 'A survey on social image mining', *Intelligent Computing and Information Science*, Vol. 134, pp.662–667.

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A. and Stumme, G. (2009) 'Evaluating similarity measures for emergent semantics of social tagging', *18th Int'l. WWW Conference*, pp.641–641.

McDaid, A. and Hurley, N. (2010) 'Detecting highly overlapping communities with model-based overlapping seed expansion', *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, ASONAM '10*, Washington, DC, USA, IEEE Computer Society, ISBN 978-0-7695-4138-9, pp.112–119.

Newman, M.E. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, Vol. 69, No. 2, pp.1–15, 026113.

Newman, M.E.J. (2004) 'Detecting community structure in networks', *Europ. Physical J.*, Vol. 38, No. 2, pp.321–330.

Newman, M.E.J. (2006) 'Modularity and community structure in networks', *Proceedings of the National Academy of Sciences*, Vol. 103, No. 23, pp.8577–8582, doi: 10.1073/pnas.0601602103.

Nicosia, V., Mangioni, G., Carchiolo, V. and Malgeri, M. (2009) 'Extending the definition of modularity to directed graphs with overlapping communities', *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 3, No. 03024, pp.1–22, P03024.

Novak, P.K., Lavrač, N. and Webb, G.I. (2009) 'Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining', *J. Mach. Learn. Res.*, February, Vol. 10, pp.377–403, ISSN 1532-4435.

Reutelshoefer, J., Baumeister, J. and Puppe, F. (2010) 'Towards meta-engineering for semantic wikis', *5th Workshop on Semantic Wikis: Linking Data and People (SemWiki2010)*.

Schaeffer, S.E. (2007) 'Graph clustering', *Computer Science Review*, Vol. 1, No. 1, pp.27–64,. ISSN 1574-0137.

Shneiderman, B. (1996) 'The eyes have it: a task by data type taxonomy for information visualizations', *Proc. IEEE Symposium on Visual Languages*, Boulder, Colorado, pp.336–343.

Sigurbjörnsson, B. and van Zwol, R. (2008) 'Flickr tag recommendation based on collective knowledge', *Proceeding of the 17th International Conference on World Wide Web, WWW '08*, New York, NY, USA, ACM, pp.327–336.

Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*, No. 8 in Structural Analysis in the Social Sciences, 1st ed., Cambridge University Press, Cambridge, UK, ISBN 9780521387071.

Wrobel, S. (1997) 'An algorithm for multi-relational discovery of subgroups', *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, Berlin, Springer Verlag, pp.78–87.

Yin, Z., Cao, L., Han, J., Zhai, C. and Huang, T. (2011) 'Geographical topic discovery and comparison', *Proc. 20th International Conference on World Wide Web, WWW '11*, New York, NY, USA, ACM, ISBN 978-1-4503-0632-4, pp.247–256.