Describing Locations using Tags and Images: Explorative Pattern Mining in Social Media

Florian Lemmerich[1] and Martin Atzmueller[2]

[1] Artificial Intelligence and Applied Computer Science
University of Würzburg
97074 Würzburg, Germany,
lemmerich@informatik.uni-wuerzburg.de
[2] Knowledge and Data Engineering Group,
University of Kassel
atzmueller@cs.uni-kassel.de

Abstract. This paper presents an approach for explorative pattern mining in social media for describing image media based on tagging information and collaborative geo-reference annotations. We utilize pattern mining techniques for obtaining sets of tags that are specific for the specified point, landmark, or region of interest. Next, we show how these candidate patterns can be presented and visualized for interactive exploration using a combination of general pattern mining visualizations and views specialized on geo-referenced tagging data. We present a case study using publicly available data from the Flickr photo sharing platform.

1 Introduction

Given a specific location, it is often interesting to obtain representative and interesting descriptions for it, e.g., for planning touristic activities. In this paper, we present an approach for modeling location-based profiles of social image media by obtaining a set of relevant image descriptions (and their associated images) for a specific point of interest, landmark, or region, described by geo-coordinates provided by the user. We consider publicly available image data, e.g., from photo management and image sharing applications such as Flickr[3] or Picasa[4].

In our setting, each image is tagged by users with several freely chosen tags. Additionally, each picture is annotated with a geo-reference, that is the latitude and the longitude on earth surface where the image was taken. Based on this information, we try to explore the collaborative tagging behavior in order to identify interesting and representative tags for a specific location of interest. This can be either a point or a region, so that the method can be applied both for macroscopic (regional) and microscopic (local) analysis. Furthermore, by appropriate tuning and a fuzzified focus, also mesoscopic analyses combining both microscopic and macroscopic views can be implemented.

[3] http://www.flickr.com
[4] http://www.picasa.com

Since the problem of identifying *interesting* and representative descriptions of a location is to a certain degree subjective, one cannot expect to identify the best patterns in a completely automatic approach. On the other hand, considering datasets with thousands of tags, manual browsing is usually not an option.

Therefore, we propose a two step approach for tackling this problem: The first step uses pattern mining techniques, e.g., [1, 2] to automatically generate a candidate set of potentially interesting descriptive tags. For this task, we present three different options for constructing target concepts. In the second step, a human explores this candidate set of patterns and introspects interesting patterns manually. In a user-guided environment, explorative pattern mining can then be applied iteratively adapting the process steps according to the analysis goals. Additionally, background knowledge regarding the set of tags can be easily incorporated in a semi-automatic process, such that new attributes are generated from tag hierarchies that can be manually refined and included in the process. To further improve the results, we propose a simple but effective method for incorporating a weighting schema to avoid a bias towards very active users.

The presented approach is thus implemented in a semi-automatic way. In such contexts, typically advanced methods for the visualization and browsing of the respective tags sets are required according to the *Information Seeking Mantra* by Shneiderman [3]: *Overview*, *Zoom and Filter*, *Details on Demand*. We propose a set of techniques for exploring the statistics and spatial distribution of the candidate tags. These include visualizations adapted from statistics, from the area of pattern mining, and also domain specific views developed for spatial data. The presented approach is embedded into the comprehensive pattern mining and subgroup discovery environment VIKAMINE [4], which was extended with specialized plug-ins for handling and visualizing geo-spatial information.

From a scientific point of view, the tackled problem is interesting as it requires the combination of several distinct areas of research: Pattern mining, mining social media, mining (geo-)spatial data, visualization, knowledge acquisition and interactive data mining. Our contribution can be summarized as follows:

1. We adapt and extend pattern mining techniques to the mining of combined geo-information and tagging information.
2. To avoid bias towards users with very many resources, we propose a user weighting schema.
3. We show how background knowledge about similar tags can be included to define or refine topics consisting of multiple tags.
4. For the explorative mining approach we provide a set of visualizations.
5. The presented approach is demonstrated in a case study using publicly available data from Flickr with respect to two well-known locations in Germany.

The rest of the paper is structured as follows: Section 2 describes the candidate generation through pattern mining. After that, Section 3 introduces the interactive attribute construction and visualization techniques. Next, Section 4 features two real-world case studies using publicly available data from Flickr. Section 5 discusses related work. Finally, Section 6 concludes the paper with a summary and interesting directions for future research.

2 Location-based Profile Generation and Interactive Exploration of Social Image Media

The problem of generating representative tags for a given set of images is an active research topic, see [5]. In contrast to previously proposed techniques, cf. [6], our approach does not require a separate clustering step. Furthermore, we also include interactive exploration into our overall discovery process: The approach starts by obtaining a candidate set of patterns from an automated pattern mining task. However, since it is difficult to extract exactly the most interesting patterns automatically, we propose an interactive and iterative approach: Candidate sets are presented to the user, who can refine the obtained patterns, visualize the patterns and their dependencies, add further knowledge, or adapt parameters for a refined search iteratively.

2.1 Background on Pattern Mining

Since the number of used tags in a large dataset usually is huge, it is rather useful to provide the user with a targeted set of interesting candidates for interactive exploration. For this task, we utilize the data mining method of pattern mining, specifically subgroup discovery [1, 2, 7, 8]. This allows us to identify not only interesting single tags efficiently, but also combinations of tags, which are used unusually more frequently together in a given area of interest.

Subgroup discovery aims at identifying interesting patterns with respect to a given target property of interest according to a specific interestingness measure. In our context, the target property is constructed using a user-provided location, i.e., a specific point of interest, landmark, or region, identified by geo-coordinates.

Pattern mining is thus applied for identifying relations between the (dependent) target concept and a set of explaining (independent) variables. In the proposed approach, these variables are given by (sets of) tags that are as specific as possible for the target location. The top patterns are then ranked according to the given interestingness measure.

Formally, a database $D = (I, A)$ is given by a set of individuals I (pictures) and a set of attributes A (i.e., tags). A *selector* or *basic pattern* $sel_{a=a_j}$ is a boolean function $I \to \{0, 1\}$ that is true, iff the value of attribute a is a_j for this individual. A (complex) *pattern* or *subgroup description* $sd = \{sel_1, \ldots, sel_d\}$ is then given by a set of basic patterns, which is interpreted as a conjunction, i.e., $sd(I) = sel_1 \wedge \ldots \wedge sel_d$. We call a pattern sd_s a generalization of its specialization sd_g, iff $sd_g \subset sd_s$. A subgroup (extension) sg is then given by the set of individuals $sg = ext(sd) := \{i \in I | sd(i) = true\}$ which are covered by the subgroup description sd.

A subgroup discovery task can now be specified by a 5-tuple (D, T, S, Q, k). The target concept $T\colon I \to \mathbb{R}$ specifies the property of interest. It is a function, that maps each instance in the dataset to a target value t. It can be binary (e.g., the instance/picture belongs to a neighborhood or not), but can use arbitrary target values (e.g, the distance of an instance to a certain point in space). The

search space 2^S is defined by a set of basic patterns S. Given the dataset D and target concept t, the 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. Finally, the integer k gives the number of returned patterns of this task. Thus, 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. Each of these descriptions could be reformulated as a rule $res_i \to t$.

While a huge amount of quality functions has been proposed in literature, cf. [9], the most popular interesting measures trade-off the size $|ext(sd)|$ of a subgroup and the deviation $t - t_0$, where t is the average value of the target concept in the subgroup and t_0 the average value of the target in the general population. Please note, that for binary t the average value of t reflects the likelihood of t in the respective set. Thus, the most used quality functions are of the form

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

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

2.2 Target Concept Construction

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. In this paper we propose and discuss the effects of three different approaches: Using the *raw* distance, a parametrized neighborhood function, and a "fuzzified" neighborhood function.

First, we could use the raw distance of an image to the point of interest as a numeric target property. 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:

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

where r_e is the earth radius.

Using this as the numeric target concept, the task is to identify patterns, for which the average distance to the point of interest is very small. For example, the target concept for an interesting pattern could be described as: "Pictures with this tag are on average 25km 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, while for the location of the Berlin olympic stadium the tag "olympic" could be a regarded as specific. However, if considering other olympic stadiums (e.g., in Munich) the average distance for the tag "olympic" is quite large. Therefore, we define a second function: The neighborhood distance requires a maximum distance d_{max} to the location of interest. Then, the target concept is given by:

$$neighbor(p) = \begin{cases} 0, \text{ if } d(p) < dist_{max} \\ 1, \text{ else} \end{cases}$$

Tags are then considered as interesting, if they occur relatively more often in the neighborhood 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 neighborhood of the specified point of interest, 33% for pictures with tag x are in this neighborhood." The downside 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 influenced by noise.

Therefore, the third considered approach is to "fuzzify" the second approach: Instead of a single distance d_{max} we define a minimum distance d_{lmax} and a maximum distance d_{umax} for our neighborhood. Images with a distance smaller than d_{lmax} are counted fully to the neighborhood but only partially for distances between d_{lmax} and d_{umax}. For the transition region between d_{lmax} and d_{umax} any strictly monotone function could be used. In this paper, we concentrate on the most simple variant, that is, a linear function. Alternatives could be sigmoid-functions like the generalized logistic curve.

$$fuzzy(p) = \begin{cases} 0, & \text{if } d(p) < d_{lmax} \\ \frac{d(p)-d_{lmax}}{d_{umax}-d_{lmax}}, & \text{if } d(p) > d_{lmax} \text{ and} \\ & \quad d(p) < d_{umax} \\ 1, & \text{otherwise} \end{cases}$$

In doing so, we require one more parameter to chose, 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 neighborhood.

Figure 1 depicts the described options: The fuzzy function can be regarded as a compromise between the other two function. It combines the steps for the neighborhood function with a linear part that reflects the common distance function.

2.3 Avoiding User Bias: User–Resource Weighting

In the previously described process for candidate generation 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

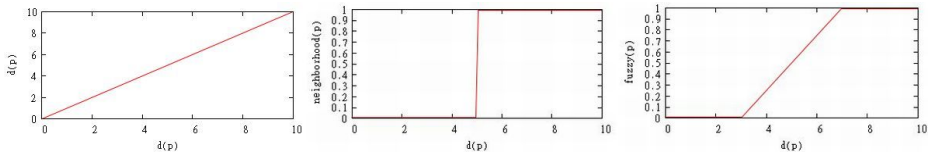

Fig. 1. The three proposed distance functions $d(p)$, $neighbor(p)$ with a threshold of $dist_{max} = 5$ and $fuzzy(p)$ with thresholds $d_- = 3$ and $d_+ = 7$ as a function over $d(p)$. It can be observed, that $d(p)$ is (obviously) linear, $neighbor(p)$ is a step function, and $fuzzy(p)$ combines both properties in different sections.

photos of this event with a unique name. Given the approach presented above this name is then considered as very important for that location, although the tag is not commonly used by the user base.

One possibility to solve this issue could be to utilize 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 [10] or BSD [11]. On the other hand, more basic algorithms, for example exhaustive depth-first search without a specialized data structure scale not very well for the problem setting of this paper, with thousands of tags as descriptions and possibly millions of instances in an interactive setting.

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 computing statistics of a subgroup the overall count and the target value, which is added if the respective image is part of a 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 utilized a weighting function of

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

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 [7].

3 Interactive Exploration

In the following, we first describe the options for including background knowledge for semi-automatic attribute construction. After that, we describe the different visualization options.

3.1 Semi-Automatic Attribute Construction

In social environments similar semantics are often expressed using diverse sets of tags, e.g., due to different languages. For an improved analysis it can be helpful to combine multiple tags into *topics* (meta-tags), that is, sets of semantically related attributes. The attribute hierarchy editor shown in Figure 2 allows an easy but fine-grained specification of topics by editing a text document using dash-trees [12] as a simple intuitive syntax: A tree structure can easily be defined by adding "-" characters at the start of the respective lines, see Figure 2. The root of the tree defines the topic name, the tree children declares included tags for this topic. For each topic a new attribute 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. The hierarchies are directly specified in VIKAMINE and propagated to the applied dataset.

In addition to providing the knowledge purely manually, we can also apply a semi-automatic approach. This is implemented, e.g., using LDA-based approaches (*latent dirichlet allocation* [13]). LDA provides for a convenient data preprocessing option. Following the semi-automatic approach, we apply it for generating topic proposals, which then are tuned interactively. The LDA method itself builds topics capturing semantically similar tags and thus helps to inhibit the problem of synonyms, semantic hierarchies, etc. After that, the set of proposed topics can then be tuned and refined by the user. In this way, we efficiently build interpretable tag clusters, i.e., for obtaining descriptive topic sets.

3.2 Visualization

In our approach, the problem of identifying tags specific for a region is formulated as a pattern mining task. While this task can generate candidate patterns, often only manual inspection by human experts can reveal the most informative patterns. This is especially the case, when considering that the interestingness is often 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, analyze and visualize the candidate tags and tag combinations as essential for effective knowledge discovery in our setting. We consider three kinds of visualizations:

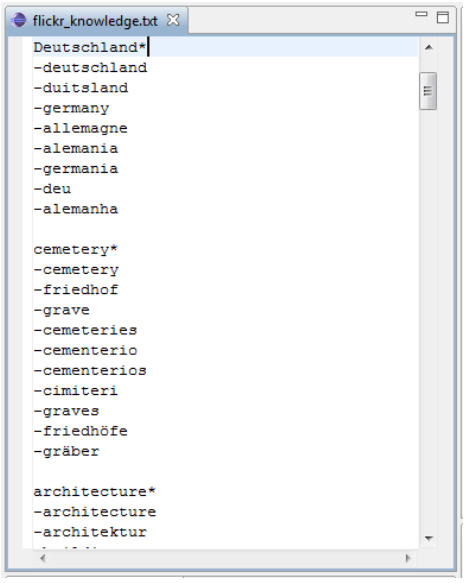

Fig. 2. Editor for specifying background knowledge (tag hierarchies) in textual form. The tag hierarchies can be generated, e.g., by LDA-based approaches, and can be refined in a semi-automatic step. In this example for instance the new attribute *cemetery** is constructed that is true, iff the respective image has been tagged by any of the tags beyond (*cemetery, friedhof, grave, cemeteries, cementerios, cimiteri, graves, friedhöfe, gräber*).

1. Traditional visualizations are mainly used for introspection of candidate patterns. Typical visualizations include the contingency table, pie charts, and box plots. An especially important visualization of this category proved to be a distance histogram. 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.
2. For an interactive exploration of the mined profiles and the tag sets and comparative visualization we can utilize various established visualizations for interactive subgroup mining, cf. [4]. These user interfaces include for example:
   (a) The *Zoomtable* which is used to browse over on the refinements of the currently selected pattern. For numeric targets, it includes the distribution of tags concerning the currently active pattern. For the binary 'neighbor' target concept, it shows more details within the zoom bars, cf. [4], e.g., showing the most interesting factors (tags) for the current pattern and target concept.
   (b) The *nt-Plot* compares the size and target concept characteristics of many different pattern. In this ROC-space related plot, e.g., [4], 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.

(c) The *Specialization Graph* is used to observe the dependencies between tag combinations, cf. [14]. In this graph, each pattern is visualized by a node in the graph. Each node is 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. Generalization relations between patterns are depicted by directed edges from more general to more specific patterns. For example, the patterns $arts$ and $arts \wedge night$ are connected by an edge pointing at the latter patterns.

For a more specific exploration of the location-based profiles of social image media advanced visualization methods can furthermore be exploited:

(a) The *Distance Attribute Map* is a view, that allows for the interactive creation of distance attributes ($d(p)$, $neighbor(p)$ and $fuzzy(p)$) by selecting a point p on a dragable and zoomable map. Future improvements could incorporate online search function, e.g., by using the Google Places API.

(b) The *Tag Map* visualizes the spatial distribution of tags on a dragable and zoomable map. 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 4) we chose a sample of at most 1000 markers. In a variant of this visualization also the distribution of sets of tags can be displayed on a single map in order to compare their distributions. An exemplary zoomed-in Tag-Map for the tags *brandenburgertor* and *holocaust* (for the memorial) is shown in Figure 3. Figure 4 shows the distance distribution of the tag to the actual location.

(c) The *Exemplification View* displays sample images for the currently displayed tag. This is especially important, since pattern exemplification has shown to be essential for many applications, e.g., [15]. 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 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, if this information is available.

The interactive exploration also can utilize background knowledge concerning the provided tags, which is entered either in a textual or graphical form.

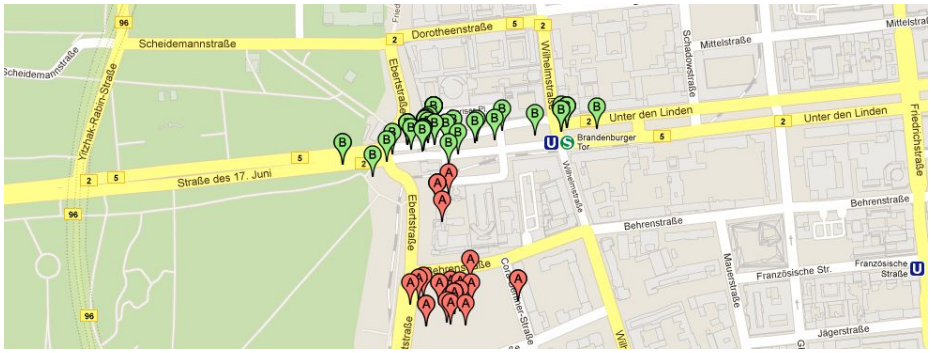

Fig. 3. Example comparative Tag-Map visualization from the case study (zoomed in): Pictures with tag "holocaust" are marked with an red "A", while pictures for the tag "brandenburgertor" are marked with a green "B"

The proposed features were implemented as a plugin for the interactive subgroup discovery environment VIKAMINE[5]. For incorporating the traditional plots the VIKAMINE R-Plugin was used as a bridge to the R[6] language for statistical computing.

4 Case Study: Flickr

We show the effectiveness of our approach in two case studies. These application scenarios utilize 1.1 million images collected from Flickr. We selected those that were taken in 2010 and are geotagged with a location in Germany.

For the collected tagging data, we applied data cleaning and preprocessing methods, e.g., stemming. We considered all tags that were used at least 100 times. This resulted in about 11,000 tags. In the case studies we show how the combination of automated pattern mining, visualization and specialized views for geo-referenced tagging data enables the identification of tag combinations which are interesting for the specified location. For pattern mining, we applied the proposed quality function with $a = 0.5$.

For our case studies, we present results for two example locations: The famous Brandenburger Tor in Berlin and the Hamburg harbor area. The goal was to enable the identification of tags, which are representative especially for this region, for people without knowledge of the respective location.

4.1 Example 1: Berlin, Brandenburger Tor

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

[5] `www.vikamine.org`
[6] `http:/www.r-project.org`

course, also the tag *berlin* is to be expected. As an example, Figure 4 shows the distance distribution of the tag *brandenburgertor* to the actual location.

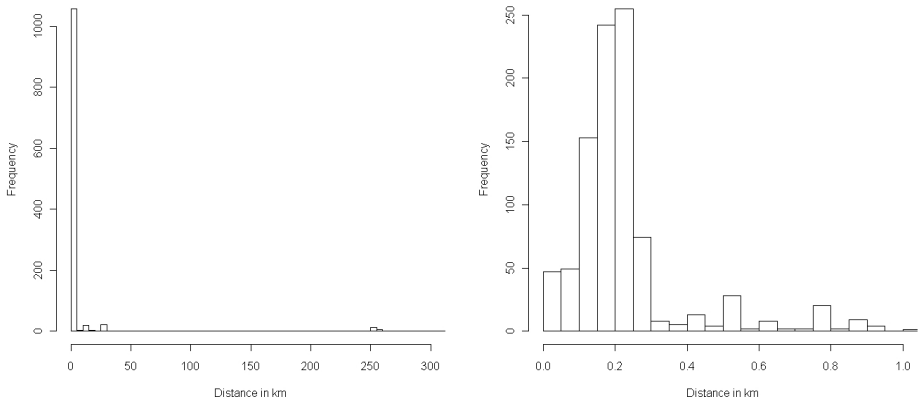

Fig. 4. Histogram showing the distances of pictures with the tag "brandenburgertor" to the actual location. It can be seen in the left histogram that the tag is very specific, since the vast majority of pictures with this tag is within a 5km range of the location. The histogram on the right side shows the distance distribution up to 1km in detail. It can be observed that most pictures are taken at a distance of about 200m to the sight.

Target Concept Options First we investigated, which candidate tags were returned by an automatic search using the different proposed target concept options. The results are shown in the Tables 1-5.

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

Tag	Subgroup Size	Mean Target Distance (km)
berlin	113977	10.48
potsdam	5533	26.83
brandenburg	5911	47.33
charlottenburg	4738	10.90
art	24067	211.28
leipzig	10794	147.87
kreuzberg	3935	14.11
nachbarn	3691	6.16
leute	4547	53.37
strassen	6899	126.83
berlinmitte	3054	4.76

Table 1 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 parametrization, it is difficult to adapt the search, such that those tags are excluded.

Tables 2-4 show the *neighbor* function with different distance thresholds d_{max}, from 0.1 km to 5 km. 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 that neighborhood 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. 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 4, for a parameter of $d_{max} = 5\ km$ the results show to be tags, which are specific for Berlin as a whole, but not necessarily for the area around the Brandenburger Tor. The results include tags like *tiergarten*, *kreuzberg* or *alexanderplatz* which describe other areas in Berlin.

Finally, Table 5 shows the fuzzified distance function, ranging from 1km to 5km 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 since, e.g., distances like 1-5km 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 *neighbor* functions.

Including Instance Weighting Taking a closer look at the results of Table 4 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 for to Table 4, which notes the overall count of users that used that description. For example the tag *hein-*

richböllstiftung was applied for 1211 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 6. The table shows, for example that the tags *heinrichböllstiftung* and *karnevalderkulturen* have disappeared and are replaced by more broadly used descriptions of Berlin attractions such as *fernsehturm* (translated: television tower) or *memorial* (for the previously mentioned holocaust memorial). 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.

Attribute Construction As can been seen from this example (Table 4), the automatic candidate generation tends to return semantically equivalent or very closely related tags in the results, i.e. translations of tags into other languages, for example *berlin, berlino and berlijn*. Such results fill slots in the result set of the candidate generation, suppress further interesting and make the results more difficult to comprehend. Additionally, one wants to perform the next step of the analysis— the interactive exploration — for these descriptions at once. In order to identify such equivalent tags and combine them within the system we used our semi-automatic attribute construction technique. To do so, first a latent dirichlet allocation is 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 3.1. The input format was then used to construct new combined 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. The automatic method (LDA) proposed suitable sets of tags which could be manually refined. Depending on the amount of total tags this requires a certain amount of manual work. Accordingly, the decision, which tags can be considered semantically

equivalent is also subjective to a certain degree. Nonetheless, this only emphasizes the need of a simple interactive environment that enables also system users without a data mining background to combine attributes as they see them fit. This technique of attribute construction also enables the user to investigate self-constructed topics by interactive exploration by just creating a meta tag with certain selected keywords.

Table 2. Brandenburger Tor: top patterns (description size 1) for the target concept function *neighbor*, with $d_{max} = 0.1$ km.

Tag	Subgroup Size	Target Share
wachsfigur	322	0.99
madametussauds	177	0.853
celebrity	345	0.435
verkehrsmittel	163	0.313
metro	469	0.277
berlinunderground	158	0.247
kitty	185	0.227
brandenburgertor	1136	0.085
u55	114	0.263
ubahn	4295	0.034
unterdenlinden	573	0.075
gleis	375	0.085
bahnsteig	551	0.058

4.2 Example 2: Hamburg Harbor - "Landungsbrücken"

The second example considers the Hamburg harbor, especially the famous "Landungsbrücken". For this location, Figure 6 shows the distribution of several interesting tags in the zoomtable.

For the Hamburg example, we also show complex patterns, i.e., combinations of tags, in the result tables. Table 7 shows the results of applying the standard mean distance target concept, while Table 8 shows the results of the fuzzified target concept, ranging from 1km to 5km (lower, upper parameters).

It is easy to see, that these results support the findings for the Berlin example: The fuzzified approach is more robust and concentrates on the important tags well, while the standard approach is suitable on a very macroscopic scale. It includes tags that are specific for the region, e.g., *schleswigholstein* or relatively close cities such as *Lingen* and *Hannover*.

5 Related Work

This paper combines approaches from three distinct research areas, that is, pattern mining, mining (geo-)spatial data, and mining social media. First, in con-

Table 3. Brandenburger Tor: top patterns (description size 1) for the target concept function *neighbor*, with $d_{max} = 1$ km.

Tag	Subgroup Size	Target Share
berlin	113977	0.225
reichstag	2604	0.829
potsdamerplatz	2017	0.797
heinrichböllstiftung	1211	0.988
berlino	4162	0.461
brandenburgertor	1136	0.816
sonycenter	803	0.923
gendarmenmarkt	696	0.885
potsdamer	577	0.88
bundestag	1096	0.611
brandenburggate	643	0.776
brandenburger	401	0.913
friedrichstrasse	558	0.735
unterdenlinden	573	0.705
panoramapunkt	271	1
holocaustmemorial	301	0.93

Table 4. Brandenburger Tor: top patterns (description size 1) for the target concept function *neighbor* and a threshold $d_{max} = 5$ km. The last column shows the overall count of users that used this description.

Tag	Subgroup Size	Target Share	Users
berlin	113977	0.745	5703
kreuzberg	3933	0.961	405
berlino	4162	0.915	392
mitte	3507	0.972	404
reichstag	2604	0.976	680
berlinmitte	3053	0.832	96
potsdamerplatz	2017	0.97	375
hauptstadt	2350	0.892	106
karnevalderkulturen	1851	0.958	36
alexanderplatz	1699	0.989	546
berlijn	2094	0,844	120
berlinwall	1635	0.914	275
graffiti	6136	0.525	838
tiergarten	2497	0.749	287
berlín	1431	0.931	119
heinrichböllstiftung	1211	1	3

trast to the common pattern mining approaches, we introduce different target concept (functions), extending the traditional definition of target concepts.

Table 5. Brandenburger Tor: top patterns (description size 1) for the 'fuzzified' target concept distance function ranging from 1 km to 5 km.

Tag	Subgroup Size	Mean Target Share
berlin	113977	0.46
reichstag	2604	0.05
potsdamerplatz	2017	0.05
mitte	3507	0.42
berlinmitte	3053	0.30
heinrichböllstiftung	1211	0.01
hauptstadt	2350	0.34
brandenburgertor	1136	0.10
alexanderplatz	1699	0.28
city	18246	0.76
tiergarten	2497	0.42
platz	2171	0.4
touristen	2815	0.47
nachbarn	3691	0.55
sonycenter	803	0.02

Table 6. Brandenburger Tor: top patterns (description size 1) using instance weighting for the target concept function *neighbor* and a threshold $d_{max} = 5$ km.The last column shows the overall count of users that used this description.

Tag	Subgroup Size	Target Share	Users
berlin	13790.6	0,804	5703
berlino	806.2	0,916	392
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
berlijn	291.7	0,85	120
fernsehturm	310.8	0,794	725
berlín	224.9	0,908	119
potsdamerplatz	196.4	0,963	375
wall	548.6	0,597	959
memorial	287.7	0,721	488
eastsidegallery	155.6	0,922	156
graffiti	661.6	0,506	838
brandenburgertor	139.4	0,931	332

Next, (geo-)spatial data mining [16] aims to extract new knowledge from spatial databases. In this context, often established problem statements and methods have been transfered to the geo-spatial setting, for example, considering

Fig. 5. An exemplary nt-plot for the location Brandenburgertor, for tags with a maximum distance of 5km. Tags that were used more often are shown on the right side of the diagramm, for example, "streetart" (16), "graffiti" (8), or "urban" (18). Tags that are very specific for the given target concept, that is, within a 5km area of the Berlin Brandenburger Tor, are displayed at the top of the diagramm. For example, the tag "urban" (18) was used relatively often, but it is not specific for the specified location of interest. However, tags such as "heinrichböllstiftung" (10), "alexanderplatz" (1), or "potsdamerplatz" (14) are very specific (and interesting) for the specified location.

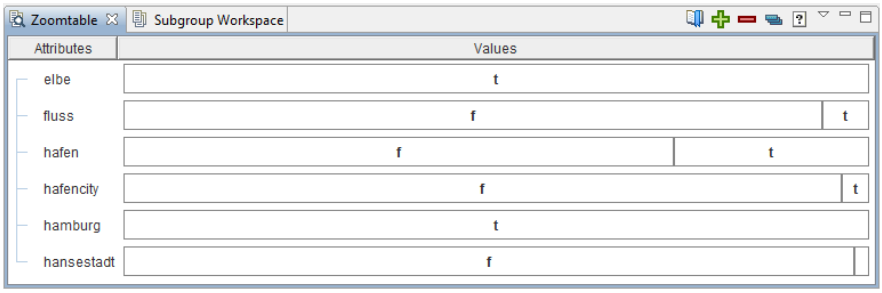

Fig. 6. The zoomtable showing some tags from the Hamburg Harbor

association rules [17]. We incorporate geo-spatial elements constructing distance-based target concepts according to different intuitions. Also, for the combination of pattern mining and geo-spatial data, we provide a set of visualizations 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, see e.g. [5]. Sigurbjörnsson and van Zwol also analyze Flickr data and provide a characterization of how users

Table 7. Hamburg Harbor: The top patterns (max. description size 2) for the mean distance target concept.

Tag	Subgroup Size	Mean Target Distance (km)
hamburg	29448	9.60
niedersachsen	34672	170.05
berlin	116979	258.34
schleswigholstein	9068	96.75
2010 AND hamburg	5255	7.81
oldenburg	10023	126.02
berlin AND germany	43280	256.95
ostsee	9565	154.41
hannover	8052	138.62
bremen	5656	99.06
lingen	14004	210.85
lingen AND germany	13909	210.82

Table 8. Hamburg Harbor: The top patterns (max. description size 2) for the 'fuzzified' target concept distance function ranging from 1 km to 5 km.

Tag	Subgroup Size	Mean Target Share
hamburg	29448	0,89
deutschland AND hamburg	6127	0.80
hafen AND hamburg	2163	0.69
hansestadt AND hamburg	1376	0.60
deutschland AND hansestadt	1676	0.68
elbe AND hamburg	1786	0.70
schiff AND hamburg	996	0.58
hafen AND elbe	656	0.52
hansestadt	2906	0.81
ship AND hamburg	882	0.63

apply tags and which information is contained in the tag assignments [18]. Their approach is embedded into a recommendation method for photo tagging, similar to [19] who analyze different aspects and contexts of the tag and image data. Abbasi et al. present a method to identify landmark photos using tags and social Flickr groups [20]. They apply group information and statistical preprocessing of the tags for obtaining interesting landmark photos.

In contrast to previously proposed techniques, e.g., [6], our approach does not require a separate clustering step. Furthermore, we focus on descriptive patterns consisting of tags that are interesting for a specific location; the interestingness can also be flexibly scaled by tuning the applied quality function. In contrast to the above automatic approaches, we also present and extend different visualizations for a semi-automatic interactive approach, integrating the user.

6 Conclusions

In this paper, we have presented an approach for obtaining location-based profiles for social image media using explorative pattern mining techniques. Candidate sets of tags, which are specific for the target location are mined automatically by an adapted pattern mining search step and can be refined subsequently. The approach enables several options including selectable analysis-specific interestingness measures and semi-automatic feature construction techniques. In an interactive process, the results can then be visualized, introspected and refined. For demonstrating the applicability and effectiveness, we presented a case study using real-world data from the photo sharing application Flickr considering two well-known locations in Germany.

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, see for example [21], 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 preprocessing, the mining, and the presentation.

Acknowledgment

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

1. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97), Berlin, Springer Verlag (1997) 78–87
2. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. 18th International Symposium on Methodologies for Intelligent Systems (ISMIS 2009). LNCS (2009)
3. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proc. IEEE Symposium on Visual Languages, Boulder, Colorado (1996) 336–343
4. Atzmueller, M., Puppe, F.: Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining **11**(11) (2005) 1752–1765
5. Liu, Z.: A Survey on Social Image Mining. Intelligent Computing and Information Science (2011) 662–667
6. Kennedy, L., Naaman, M.: Generating Diverse and Representative Image Search Results for Landmarks. In: Proceeding of the 17th international conference on World Wide Web, ACM (2008) 297–306

7. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. Journal of Machine Learning Research **5** (2004) 153–188
8. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th Intl. Joint Conf. on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland (2005) 647–652
9. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys **38**(3) (2006)
10. Atzmueller, M., Puppe, F.: SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Number 4213 in LNAI, Berlin, Springer Verlag (2006) 6–17
11. Lemmerich, F., Rohlfs, M., Atzmueller, M.: Fast discovery of relevant subgroup patterns. In: Proc. 23rd FLAIRS Conference. (2010)
12. Reutelshoefer, J., Baumeister, J., Puppe, F.: Towards Meta-Engineering for Semantic Wikis. In: 5th Workshop on Semantic Wikis: Linking Data and People (SemWiki2010). (2010)
13. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (2003) 993–1022
14. Klösgen, W., Lauer, S.R.W.: 20.1: Visualization of Data Mining Results. In: Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York (2002)
15. Atzmueller, M., Puppe, F.: A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence **28**(3) (2008) 210–221
16. Koperski, K., Han, J., Adhikary, J.: Mining Knowledge in Geographical Data. Communications of the ACM **26** (1998)
17. 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) (2003) 541–566
18. Sigurbjörnsson, B., van Zwol, R.: Flickr Tag Recommendation based on Collective Knowledge. In: Proceeding of the 17th International Conference on World Wide Web. WWW '08, New York, NY, USA, ACM (2008) 327–336
19. Lindstaedt, S., Pammer, V., Mörzinger, R., Kern, R., Mülner, H., Wagner, C.: Recommending Tags for Pictures Based on Text, Visual Content and User Context. In: Proc. 3rd International Conference on Internet and Web Applications and Services, Washington, DC, USA, IEEE Computer Society (2008) 506–511
20. Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., Staab, S.: Exploiting Flickr Tags and Groups for Finding Landmark Photos. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval. ECIR '09, Berlin, Heidelberg, Springer-Verlag (2009) 654–661
21. Atzmueller, M., Beer, S., Puppe, F.: Data Mining, Validation and Collaborative Knowledge Capture. In Brüggemann, S., d'Amato, C., eds.: Collaboration and the Semantic Web: Social Networks, Knowledge Networks and Knowledge Resources. IGI Global (2011)