TOWARDS RULE-GUIDED CLASSIFICATION FOR VOLUNTEERED GEOGRAPHIC INFORMATION

Ahmed Loai Ali[1,3], Falko Schmid[1,2], Zoe Falomir[1], Christian Freksa [1,2]

[1] Cognitive Systems Research Group, University of Bremen, Bremen, Germany
[2] SFB/TR 8 Spatial Cognition, University of Bremen, Bremen, Germany
[3] Information System Department, Faculty of Computers and Information, Assuit University, Assuit, Egypt
{loai, schmid, zfalomir, freksa}@informatik.uni-bremen.de

Commission II, WG II/4

KEY WORDS: Volunteered Geographic Information (VGI), Spatial Data Quality, Spatial Data Mining, Classification

ABSTRACT:

Crowd-sourcing, especially in form of Volunteered Geographic Information (VGI) significantly changed the way geographic data is collected and the products that are generated from them. In VGI projects, contributors' heterogeneity fosters rich data sources, however with problematic quality. In this paper, we tackle data quality from a *classification* perspective. Particularly in VGI, data classification presents some challenges: In some cases, the classification of entities depends on individual conceptualization about the environment. Whereas in other cases, a geographic feature itself might have ambiguous characteristics. These problems lead to inconsistent and inappropriate classifications. To face these challenges, we propose a guided classification approach. The approach employs data mining algorithms to develop a classifier, through investigating the geographic characteristics of target feature classes. The developed classifier acts to distinguish between related classes like *forest, meadow* and *park*. Then, the classifier could be used to guide the contributors during the classification process. The findings of an empirical study illustrate that the developed classifier correctly predict some classes. However, it still has a limited accuracy with other related classes.

1. INTRODUCTION

The advance of Web technologies (e.g. Web 2.0) and the increasing availability of hand-held location sensing devices (e.g. smart phones) empower the public to participate in mapping activities. Those activities, which were formerly conducted by mapping agencies and cartographers, now attract volunteers. Collaborative mapping is one form of *Volunteered Geographic Information* (VGI), when a group of volunteers acts to collect, share, maintain, and use information about geographic features (Goodchild, 2007). Among others, OpenStreetMap[1](OSM), Google Mapmakers[2] and Wikimapia[3] are examples of collaborative mapping projects which aim to produce a digital map of the world. During the last decade, VGI has played a significant role in the GIScience community. Various applications and services have been developed based on VGI data including – but not limited to – environmental monitoring, crisis management, urban planning, mapping services, etc.

Despite of the increasing dependency on VGI data, its questionable quality results – in some cases – in limited use (Elwood et al., 2012). Among other things, the lack of detailed information about data quality and the difficulty of applying traditional spatial quality measures for assessing the data are key reasons behind its questionable quality (Flanagin and Metzger, 2008, Elwood et al., 2012). Generally, multiple measures are used to describe the quality of spatial data from different perspectives such as completeness, positional accuracy, thematic accuracy, logical consistency, and lineage. In this paper, we tackle the quality from a classification perspective. Classification is one facet of data quality that influences thematic accuracy.

[1] www.openstreetmap.org
[2] www.google.com/mapmaker
[3] www.wikimapia.org

In most VGI projects, a large amount of data is contributed remotely by tracking satellite images. The contribution method itself poses a classification challenge: whether a piece of land covered by grass is classified as *park, garden, meadow*, or *grass*, if a water body classified as *pond* or *lake* – the classification answers to these questions mainly depend on contributors' perspectives and need some sense of locality. Moreover, some classes are semantically related (e.g., *park* or *garden*), while others have ambiguous characteristics (e.g., *grass*). Hence, in such cases an entity could be inappropriately classified resulting in problematic quality.

In this paper, we present an approach for rule-guided classification aiming to improve the quality of VGI data. The approach consists of two phases: *Learning* and *Guiding*. During the *Learning* phase, the task is to learn the unique geographic characteristics that distinguish between related classes. Learning mainly depends on topological investigation of classes. During learning, data mining algorithms are applied to extract the characteristics of specific classes in form of a set of predictive rules. Based on the extracted rules, a rule-based classifier is developed that guides the contributors, during the *Guiding* phase, towards the most appropriate classes.

In an empirical study, we investigate the classification of grass-covered land. We analyze the classes *forest, garden, grass, meadow, park*, and *wood*. The classification of these features represent a challenge: they are commonly covered by grass, however each class has unique characteristics. For example, the classes *park* and *garden* have entertainment characteristics, *forest* and *wood* are usually covered with trees or other woody vegetation, the class *meadow* has agriculture characteristics, etc. The findings indicate the feasibility of the approach; The developed classifier is able to precisely classify some of the target feature classes, while other classes still have poor classification accuracy.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A.-M. Olteanu-Raimond, C. de-Runz, and R. Devillers
doi:10.5194/isprsannals-II-3-W5-211-2015

The paper is organized as follows: Section 2 presents a literature review of VGI data quality. Section 3 gives insight into the main factors behind the heterogeneous classifications in VGI data. Section 4 presents the proposed approach and its phases. Section 5 presents an empirical study. The last section outlines the conclusions and the current state of the work.

2. VGI DATA QUALITY

In VGI, particularly in collaborative mapping, contributors act as sensors to collect, update, and share information about geographic features. VGI employs the contributors' locality and their willingness to contribute in order to produce rich spatial data sources (Goodchild, 2007). However, the quality of the resulting data is heterogeneous. With increasing utilization of VGI in GIScience activities and applications, data quality becomes a concern of highest priority (Flanagin and Metzger, 2008, Elwood et al., 2012).

VGI data is evaluated either by comparison with authoritative data or by intrinsic analysis following crowd-sourcing, social, or geographic approaches (Goodchild and Li, 2012). (Girres and Touya, 2010, Haklay, 2010, Neis et al., 2011, Jackson et al., 2013) compare VGI data against authoritative data sources in France, UK, Germany, and USA, respectively. They emphasize the quality of VGI data particularly in urban areas. In (Hecht and Stephens, 2014), authors conclude that VGI data quality decreases with increased distance from urban areas. On the other side of research, (Bishr and Kuhn, 2007, Keßler et al., 2011, Neis et al., 2011, Mooney and Corcoran, 2012b, Barron et al., 2014) assess VGI data intrinsically. They asses VGI data by investigating the meta-data like contributors' mapping activities and reputation, entities' editing history, etc. Authors of (Neis et al., 2013) compare the development of contributors' communities in different cities around the world indicating the relation between the communities and data quality. The work in (Barron et al., 2014) presents 25 fitness-for-purpose measures to assess VGI data in specific uses.

Towards improving data quality, (Pourabdollah et al., 2013) conflate VGI data with authoritative data. In an attempt to improve the data quality at contribution time authors of (Vandecasteele and Devillers, 2013) provide an approach to guide contributors during the editing process aiming to improve the semantic data quality. Moreover, (Schmid et al., 2013) argue a task-specific interface approach toward acquiring higher data quality. In our previous work, we tackled the inconsistent classification problem in (Ali and Schmid, 2014) and proposed a learning-based approach to detect the problematic classification of VGI in (Ali et al., 2014).

Most of the research investigates quality measures like positional accuracy and completeness, while this paper tackles the thematic accuracy from a classification perspective. Moreover, assessment of VGI data through a comparison approach is no longer appropriate for the nature of VGI. As well as, more studies assess VGI data intrinsically following crowd-sourcing or social approaches, whereas we follow the geographic approach aiming to improve the data quality.

3. CLASSIFICATION CHALLENGES IN VGI

Classification ambiguity and vagueness in spatial data types are the fundamental sources beyond the problematic thematic accuracy of VGI (Fisher, 1999, Devillers et al., 2010). Particularly, the loose classification mechanisms and the absence of integrity checking mechanisms result in heterogeneous data classification. In most VGI projects, contributors are heterogeneous; they have

Figure 1: Appropriate classification as *park*

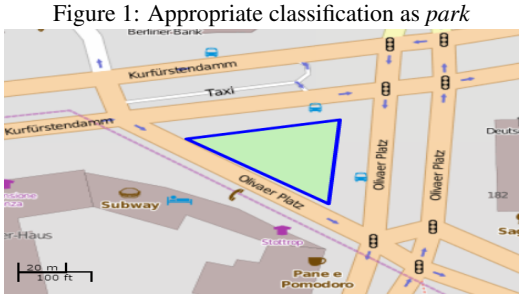

Figure 2: Inappropriate classification as *park*

diverse levels of knowledge about geography and cartography, and come from diverse cultures and educational backgrounds. On one hand, VGI harnesses the contributors' heterogeneity towards developing rich data sources and preserving the concept of locality; on the other hand, identical geographic features should be classified homogeneously as much as possible to support global applications (e.g. routing or map rendering). At the same time, there exist geographic entities that might appropriately belong to multiple classes (e.g. *park* or *garden*). However, if the characteristics and the geographic context of an entity is taken into account, this entity would be more appropriately belong to one class rather than to the other(s).

In this paper, we define *appropriate classification* as assigning a given entity a class which highly reflects its intrinsic and extrinsic characteristics and matches its geographic context. E.g., Figures 1 and 2 illustrate the terms of appropriate and inappropriate classifications, respectively. In Figure 1, the given entity contains some amusement facilities such as a playground, sport centers, and accessibility for walking. This entity is classified as *park*, which typically expresses the characteristics of the entity. Here, *park* represents the appropriate class of the entity. In contrast, in Figure 2 an entity represents a small piece of land covered by grass, located beside roads and roundabouts. The entity is classified as *park*, despite it being too far from being used for amusement or entertainment. Here, *park* is an inappropriate class and *grass* might be the appropriate class that truly reflects the characteristics of the entity. Hence, learning the intrinsic and extrinsic characteristics of a given geographic feature class is required towards guiding and recommending the contributors during the classification process.

3.1 Ambiguous Classification

As a case study, this paper addresses the classification of grass-covered land. A piece of land covered by grass could be classified as *garden, grass, park, meadow*, or even *forest* or *wood*. These classes represent a sample among other potential classes (e.g. *recreation ground, scrubs*). Our previous study in (Ali et al., 2014) demonstrates how contributors are unlikely to agree between themselves on a certain class for a given set of entities. The participants of the study typically reflect the nature of VGI contributors: diversity of age, gender, culture, education, and geographic knowledge. The findings indicate the following: (1) the

difficulties of classifying such of these entities; (2) the massive need for multiple classes for some entities; and (3) the demand for rule-guided classification. During remote classification, it is difficult, even for experts, to recognize the intrinsic properties of an entity to assign the most appropriate class. Thus recommendations and guides are both required particularly for non-expert contributors, which represent the majority in VGI projects.

We utilized OSM data, as a common example of VGI projects. In OSM, the classification is done by means of tags in form of $key = value$, where the key represents a classification perspective and the $value$ represents a class of that perspective. For example, tag $leisure = park$ the key $leisure$ is associated with the set of entities that are used for entertainment purposes, while $park$ represents one class between others like $garden$, $pitch$, $recreation$, etc. There are no restrictions on the number of tags that are associated with an entity; each entity could be related to no tags or several tags with arbitrary combinations of tags (Mooney and Corcoran, 2012a). The flexibility of contribution mechanisms itself leads to problematic classifications. At the same time, OSM provides only recommendations of tags based on discussions between mappers communities. However, most contributors do not spend enough time to check the given recommendations. Moreover, particularly for non-experts, some recommendations might be conceptually misinterpreted (e.g. *wood* or *forest* and *landuse* or *landcover*).

4. RULE-GUIDED CLASSIFICATION APPROACH TO IMPROVE CLASSIFICATION QUALITY

The proposed approach to improve the quality of classification exclusively depends on VGI data. We aim to develop a classification system able to guide the contributors during the classification process. Through guiding we aim to obtain data of consistent/homogeneous classification. Figure 3 illustrates the proposed approach, which consists of two phases: *Learning phase* (see Section 4.1) and *Guiding phase* (see Section 4.2).

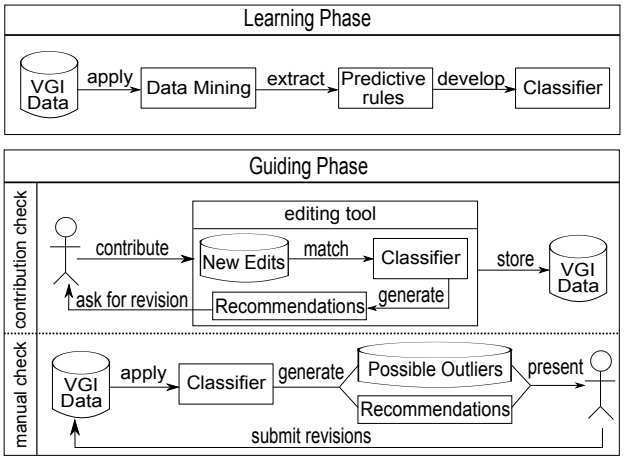

Figure 3: Guided classification approach

4.1 Learning Phase

The objective of the *Learning phase* is mining VGI data to extract a set of predictive rules. The rules describe the geographic characteristics of specific feature classes. The extracted characteristics have the form of:

$$head \leftarrow body \tag{1}$$

where the body describes the characteristics of an entity and the head points to the recommended (predicted) class. The combination of rules would be able to describe a specific feature class. Afterwards, the extracted rules are organized into a rule-based classifier, which consequently would be able to predict the most appropriate class for a given set of characteristics of an entity. The proposed approach maintains the locality principle. We assume that at country level a certain geographic feature should have the same characteristics; learning the characteristics of a specific feature in China and applying the developed classifier in Germany may not make sense. During the learning process, we depend on topological investigations to understand the geographic context of target feature classes.

4.1.1 Topological Investigation Based on the first law of geography (Tobler, 1970): "Everything is related to everything else, but near things are more related than distant things", we investigate the topological relations between pairs of entities in order to understand the geographic context of specific classes of entities. In short, this is to find the frequent relations between entities that uniquely distinguish each class. For example, *park* typically contains playgrounds, pathways, etc., whereas *grass* and *meadow* contain less infrastructure; also *park* is located within or near residential areas, whereas *meadow* is typically located near farms and rural areas, etc.

We employ the 9-Intersection Model (9IM) (Egenhofer, 1995) to investigate the topological relations between pairs of entities. As shown in Figure 4, 9IM describes the topological relations between pairs of entities as: *disjoint, meet, overlap, covers, covered By, contains, inside*, and *equal*. Basically, geographic features are represented by means of point, line, and polygon data elements. In this work, the target classes are usually represented by polygon. Thus, we consider all possible topological relations between polygon and other data elements; *polygon-point, polygon-line and polygon-polygon.*

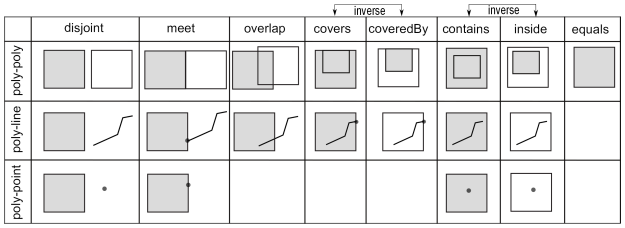

Figure 4: The 8 topological relations of the 9-Intersection Model

At Figure 4, assume that the gray entities represents the target entities. We consider *disjoint, meet, overlap, contains*, and *covers* relations. Regarding *disjoint* relation we analyze entities within distance of 10 meters far from target entities. Particularly, the *disjoint* relation gives insight about the external geographic context, while the others represent the relations resulting from the intersections of the interiors and boundaries of entities. We neglect the *inside, covers*, and *equals* relations for two reasons: (a) *inside* and *covers* are inverse relations of *contains* and *covered By*, respectively; and (b) the *equal* relation rarely occurs and does not add useful information for analysis.

4.1.2 Data Mining Process The topological analysis aims to find the frequent patterns (topological relations) involved between target classes and other geographic features, e.g. *park contains* playground, sport center, etc. We consider each combination of *key* and *value* as a new feature type. E.g. *leisure = playground* and *leisure = sport* are two different geographic features. We encode them as `leisure_playground` and `leisure_sport` respectively and relate each new feature with a unique identifier (ID) in an indexed file. The analysis includes the common map

features that are suggested by the OSM project on its Wiki page[4]. Due to the free contribution mechanism of the OSM project, the analysis results in more than 1,000 unique features, after filtering.

The mining process works to extract atomic rules in form of rule (1), which is translated into:

$$Class(X, C) \leftarrow R(X, F) \quad (2)$$

where X represents a target entity, C is the predicted class and $C \in \{$*park, meadow*, etc.$\}$, R is one of the topological relations where $R \in \{contains, meet, etc.\}$ and F represents the set of frequent features that is mostly involved in a relation R with entities of class C.

To extract such rules, we apply the Apriori algorithm (Agrawal et al., 1994). The Apriori algorithm is one of the common data mining algorithms that were initially developed to extract frequent item sets and to learn association rules from a transactional database (Witten and Frank, 2005). In this work, we particularly use a class association rule mining task, when rules have a predefined class (e.g. *park*) as their consequences (left side at rules (1) and (2)). Extracting interested rules among a large number of possibilities requires setting up some constraint parameters. Support (*supp*) and confidence (*conf*) are two common constraints that used to define the thresholds for extracting and evaluating the interesting rules, as follows: [*where* `l=leisure_playground` and `15=highway_footway`]

support is used to filter the interesting patterns. It is defined as the percentage of entities that hold the body description. e.g., $supp\,(contains(X, [1, 15]))$ = 20%, means 20% of the entire entities contains playground and footways features.

confidence is used to evaluate the extracted rules. It is equal to the percentage of entities that hold the body description and consequently the head. e.g., $conf\,(Class(X, park) \leftarrow contains(X, [1, 15]))$ = 80%, implies 80% of the entities hold the rule body is associated with class *park*.

4.1.3 Classifier development The main idea of association rule mining has adapted to solve other problems such as classification problem resulting in associative classification mining filed. *Associative Classification* (AC) is one branch of data mining that combines two mining tasks, associating rule mining and classification, to build a classifier based on a set of predictive association rule (Thabtah, 2007). Generally, developing a classifier based on a set of predictive rules consists of 4 steps:

Step 1 Find all interesting class association rules from a data set;

Step 2 Based on a *confidence* threshold, filter the extracted rules into a set of predictive association rules;

Step 3 Encode the rules into a classifier; then

Step 4 Evaluate the classifier on a test data set.

In geographic contexts, usually everything is possible (e.g., a building may be located in a desert, a highway crosses a residential area or a public park, etc.). Besides, in VGI projects there exist unlimited unique features (See section 4.1.2). Thus, we set the *support* threshold to 1% and consider patterns which occur with a frequency of more than 1% as frequent. During the learning process, we are mining to extract atomic rules per topological relation per class.

[4] `http://wiki.openstreetmap.org/wiki/Map_Features`

The extracted rules represent the output of Step 1. In the spatial context and due to the uncertainty of spatial data, the rules themselves represent a challenge at Steps 2 and 3. The aim at Step 2 is to organize the extracted rules into a set of predictive association rules for developing the classifier in Step 3. Hence, the difficulties come from the following points: (a) Step 1 results in rules of identical bodies associated with different heads (classes); (b) during Step 2, the higher the confidence threshold for filtering the interesting rules, the more possibility to dismiss useful information; (c) due to ambiguous classification (See section 3.1), an entity could plausibly belong to more than one class; and (d) due to geographic context, an entity could match with several atomic rules associated with different head (classes). In summary: How should we classify? By the majority of rules or by rules of higher confidence? In this paper, an *appropriate classification* is that which truly reflects the characteristics of an entity.

4.2 Guiding Phase

During the *Guiding phase*, the aim is to enhance the classification quality of VGI by applying the developed classifier. The proposed approach presents two different ways of guiding: First, contribution checking, when the classifier is implemented in an editing tool. At contribution time, the tool informs the contributor about the potential problem, based on the classifier. The editor provides the contributor with recommendations. Thereafter, the contributor considers the guidance provided and responds with correction (if required). Second, manual checking, when the classifier is applied directly on an existing data set. The classifier points out entities with problematic classification, which don't match any of the predictive rules. The classifier generates the problematic entities combined with some recommendations. Afterwards, both are presented for assessment and correction (if required). Through both ways, the guiding could indirectly enrich the data source, when the contributors add more information to satisfy the recommended class.

5. EMPIRICAL STUDY

To evaluate the approach, we perform an empirical study. This study checks the ability of the developed classifier to distinguish between similar classes. During the study, we use the OSM data set of Germany dated December, 2013. Reasons behind selecting Germany for the study are the following: (1) active mappers communities; (2) no authoritative bulks are imported to data, so it still reflects the voluntary nature; and (3) several studies conclude the higher quality of OSM data in Germany relative to other places (Zielstra and Zipf, 2010, Ludwig et al., 2011, Neis et al., 2013). We extract all entities, that are represented by polygons and classified as *forest, garden, grass, meadow, park*, or *wood*. The entities are extracted from the 10 most densest cities at Germanyto ensure active mappers communities and acceptable level of quality. The cities are: *Berlin, Bremen, Cologne, Dortmund, Dusseldorf, Essen, Frankfurt, Hamburg, Munich*, and *Stuttgart*. The data set consists of 3,724 *forest*, 3,030 *garden*, 7,336 *grass*, 4,277 *meadow*, 4,445 *park*, and 1454 *wood* entities. We processed each entity individually by analyzing the topological relations between pairs of entities within its geographic context. Each entity is described by a set of topological relations with other surrounded features and assigned to a specific class.

5.1 Learning Process

During the learning process, we apply the Apriori algorithm to investigate the frequent topological relations describing each class. We consider *support* threshold of 1% to find the interesting patterns. Each topological relation is processed individually with a

given class producing a set of predictive rules of the class. The rules represent the output of Step 1 (see Section 4.1.3).

We extract 9,193 rules; 4,100 describe *forest*, 215 describe *garden*, 745 describe *grass*, 506 describe *meadow*, 2,938 describe *park*, and 689 describe *wood*.

5.2 Classification Hypothesis

As mentioned previously, the rules resulting from the learning process represent a challenge for developing the classifier. To overcome the mentioned difficulties, we do the following:

- **Pruning**: Redundant rules are removed based on the rules' *conf* threshold. The rules with identical bodies are integrated into one rule assigned to the head (class) of higher *conf*.
- **Filtering using the *confidence* threshold**: The classification is done once by considering the entire rule set and once by considering rules with $conf \geq 50\%$.
- **Grading 1st and 2nd recommendations**: During the classification process, we consider the 1st and 2nd recommended classes given by the predictive rules.
- **Classification assumptions**: Due to an unbalanced number of rules describing each class, depending on the majority of rules assigned to a specific class might be biased. Thus, we consider only rules with maximum *conf* per class to define 1st and 2nd potential classes.

During the classification process, each entity is matched with the predictive rules. For example, Figure 5 shows an entity[5] with $osm\_id = 25422214$. At writing time, the entity has 28 editing versions and is tagged with `leisure=park` and `name=Revierpark Wischlingen`. It matches 401 rules: 232 *park*, 132 *forest*, 25 *grass*, 8 *meadow*, 2 *wood*, and 2 *garden*. Table 1 shows some of the matched rules with this entity:

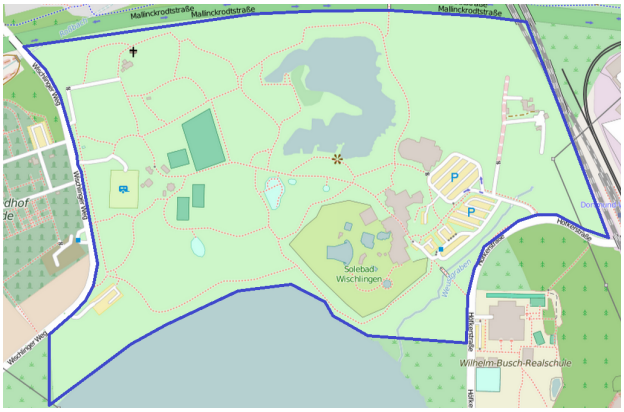

Figure 5: An entity with osm_id = 25422214

Rule	conf
$Class(X, "park") \leftarrow contains(X, [1, 15, 27, 89])$	94%
$Class(X, "park") \leftarrow contains(X, [1, 15, 21, 22])$	83%
$Class(X, "park") \leftarrow meet(X, [6, 15])$	70%
where 1=`leisure_playground`, 6=`highway_residential`, 15=`highway_footway`, 21=`sport_soccer`, 22=`leisure_pitch`, 27=`building_yes` and 89=`nature_water`	

Table 1: Matched rules for the entity with osm_id = 25422214

[5] `http://www.openstreetmap.org/way/25422214`

Classification method	Accuracy		
		corrected classified ent.	%
max(*conf*) per class	1st	14418	**60**
	1st or 2nd	18333	**75**
max(*conf*) per class *where* $conf \geq 50\%$	1st	12165	50
	1st or 2nd	13487	55
	not match any rule	6276	25

Table 2: General accuracy of the proposed classifier

Regarding rules *conf*, the top 50 rules have *conf* range form 94% to 83% and all of them have $head$ of $Class(X, park)$. While considering 1st and 2nd recommended classes requires looking into the maximum *conf* per class. E.g., the same entity matches with *park*, *forest*, *grass*, *meadow*, *garden*, and *wood* classes with descending *conf* of 94%, 65%, 54%, 48%, 45% and 20% respectively. Hence, the given entity could belong to *park* (1st prediction) or *forest* (2nd prediction) classes rather than any other potential classes.

5.3 Results and Discussions

We depend on the accuracy (*acc*) measure to evaluate the results, where accuracy represents the percentage of corrected classified entities. Table 2 demonstrates results from applying different classification hypotheses. Due to the classification ambiguity 1st and 2nd recommended classes are considered.

First, we take into account the entire set of extracted rules. The classification is based on rules with the maximum *conf* per class; when classes of rules with the 1st and 2nd maximum *conf* are assigned to 1st and 2nd recommended classes, respectively. Considering only the 1st recommended class, the classifier correctly classified 60% of entities. Whereas 75% of entities are correctly classified considering 1st or 2nd recommendations.

Second, we repeat the previous process, considering only rules with $conf \geq 50\%$. As Table 2 indicates, besides lower classification accuracies, a large number of entities does not match any rule. The clarification of that is the filtered rules are not able to cover all cases and do not give enough descriptions for the classes; some useful information might be hidden behind rules with low *conf*. E.g., $Class(X, park) \leftarrow meet(X, [highway\_footway])$ has *conf* of 38%. However this rule exactly exists in OSM Wiki recommendations[6]. To remove redundant rules, we do the pruning process (see Section 5.2). The 9,193 rules are reduced to 5,826 rules, while the accuracies are remaining mostly the same.

Due to the unbalanced distribution of classes, depending on the overall accuracy might be biased. Thus, looking into more details of the accuracy per class is important as well. Table 3 gives insight into the classification accuracy per class. According to Table 3, *park*, *grass*, and *garden* have higher classification accuracies (80-94 %, 72-87%, and 71-81%, respectively), whereas *forest*, *meadow* have moderate accuracies with 38-67% and 42-58%. While *wood* has a noticeably lower accuracy. The entities of *park, grass*, and *garden* match the 1st recommendations within 70% to 80%. They also match 1st or 2nd recommendations with higher accuracies between 81% to 94%. In contrast, entities of *forest, meadow*, and *wood* match even 1st or 2nd recommendations by an average of 46%.

The lower classification accuracy of the class *wood* might result from the limited number of entities in the training data set (1454).

[6] `http://wiki.openstreetmap.org/wiki/Tag:leisure\%3Dpark`

Class	Accuracy		
		corrected classified ent.	%
forest (3724)	1st	1447	38
	1st or 2nd	2501	67
garden (3030)	1st	2167	**71**
	1st or 2nd	2472	**81**
grass (7336)	1st	5355	**72**
	1st or 2nd	6424	**87**
meadow (4277)	1st	1826	42
	1st or 2nd	2499	58
park (4445)	1st	3516	**80**
	1st or 2nd	4216	**94**
wood (1454)	1st	107	7
	1st or 2nd	221	15

Table 3: Proposed classifier accuracies per class

Thus, the generated rules have lower *conf*; about 96% of the predictive rules of *wood* class has *conf* <50%. Besides, at OSM Wiki[7] different tagging approaches are presented for *forest* and *wood*. The same regarding *meadow* classes as 90% of the extracted rules has *conf* < 50%. In contrast, 21%, 15%, and 11% of rules describing the classes*garden, grass*, and *park*, respectively have *conf* ≥ 75%. The various classification accuracies might return to dealing with VGI data itself; some features might be better mapped than others. Moreover, the training data set is not free of incorrect classified entities. We assumed the correctness of a large partition of data.

5.4 Evaluation

Due to the unavailability of an authoritative data for these types of features, we adopt two ways for the evaluation process. First, we visually investigate the results to check the recommendations given by the proposed approach. Figure 6 and 7 illustrate examples of problematically classified entities and the recommended classifications. In Figure 6, the entity classified as *grass*, whereas the recommended classification is *park*; it contains sport areas, footways, etc. and is adjacent to a forest area, thus the appropriate classification might be *park*. While in Figure 7, the entity is wrongly classified as *park* and the recommendation given is *grass*; it contains nothing and is located between roundabouts. The findings indicate that applying the proposed classifier and following the given recommendations might potentially result in an improved classification quality.

Second, we depend on the intrinsic properties (e.g. tags, version, mapper, etc.) and extract a data set for the validation process.

[7] `http://wiki.openstreetmap.org/wiki/Tag:natural\%3Dwood`

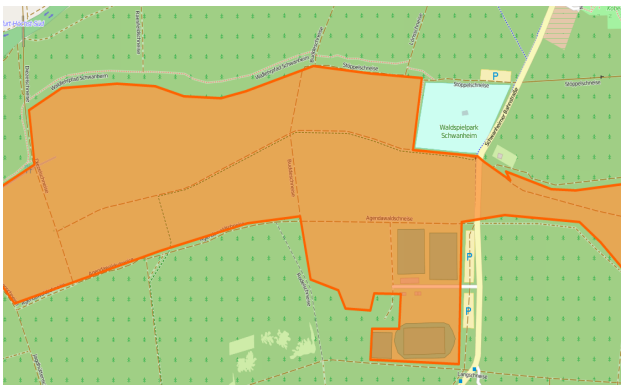

Figure 6: An entity problematic classified as *grass*, 1st recommendation *park*

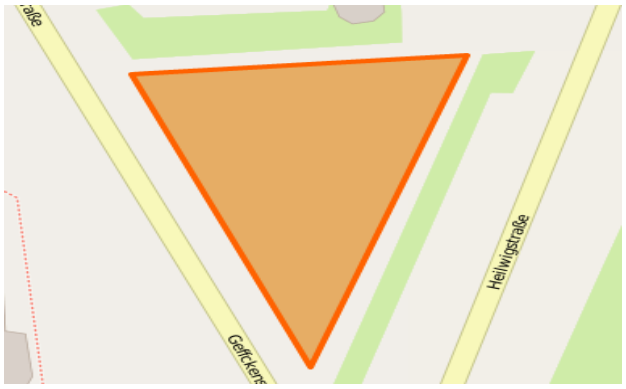

Figure 7: An entity problematic classified as *park*, 1st recommendation *grass*

For example, the proposed class appears to to correctly describe and classify *park* entities. Hence, we extract all entities that have names like *park* and are tagged by *leisure=park* as a validation data set. The extraction done from the entire Germany data set resulted in 1,856 *park* entities. We applied the developed classifier on the extracted entities. The results show that 87% of entities are correctly classified by the 1st recommendation; 95% of the entities are correctly classified by the 1st or 2nd recommendation. The validation reflects the classifier's is efficiency in distinguishing a specific class based on learning its intrinsic and extrinsic properties. Hence, applying the classifier on the entire *park* entities of Germany would point out inappropriately classified entities. The problematic classification might be relevant to incomplete mapping of an area or incorrect editing attitude of a contributor, which could be enhanced by applying the classifier at contribution time.

6. CONCLUSIONS

The increasing utilization of VGI for GIScience research results in a demand of higher data quality. Contributors' diversities result in rich data sources, however with questionable quality. In this research, we are concerned with *classification* as a facet of data quality. Definitely, identical geographic features should be homogeneously classified to support global applications. As a case study, we tackled the classification of grass-covered land, where a piece of land covered by grass could be classified as *park, garden, forest*, etc. Classifications of these features are difficult and provide multiple challenges.

This paper presents an approach for rule-guided classification. The approach harnesses the availability of VGI data to learn the characteristics of specific feature classes. The proposed approach has two phases: *Learning* and *Guiding* phases. During the *Learning* phase, we depend exclusively on the investigation of topological relations to understand the geographic context of target classes. Data mining algorithms are applied resulting in a set of predictive rules, which describe the intrinsic and extrinsic characteristics of target classes. The rules are then organized into a classifier. Whereas during the *Guiding* phase, the developed classifier could be applied in different ways guiding and recommending the contributors towards appropriate classification.

An empirical study covering the *Learning* phase was conducted. The results indicate the feasibility of learning from VGI data. The classifier we developed is able to predict classes like *park, grass*, and *garden* with higher accuracy. While the approach still has limited accuracy with other classes like *meadow, wood*, and *forest*. Further investigations are required to evaluate the generated

rules. In future work, we will focus on implementing the *Guiding* phase and measure the classification improvements based on the provided recommendations. We plan to study the OSM ontology (e.g., OSMonto (Codescu et al., 2011)) to determine whether the semantic distance between the ontological concepts could solve the ambiguity between similar classes.

ACKNOWLEDGMENT

This work is partially funded by the German Academic Exchange Service (DAAD), the German Research Foundation (DFG) through the Transregional Collaborative Research Center Spatial Cognition SFB/TR 8, and the European Commission Marie Curie project COGNITIVE-AMI. We would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

Agrawal, R., Srikant, R. et al., 1994. Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB, Vol. 1215, pp. 487–499.

Ali, A. L. and Schmid, F., 2014. Data quality assurance for Volunteered Geographic Information. In: Proc. of the 8th International Conf. on Geographic Information Science, Springer International Publishing Switzerland, Vienna, Austria, pp. 126–141.

Ali, A. L., Schmid, F., Al-Salman, R. and Kauppinen, T., 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information. In: Proc. of the 22nd ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems, ACM, Dallas, TX, pp. 143–152.

Barron, C., Neis, P. and Zipf, A., 2014. A comprehensive framework for intrinsic OpenStreetMap quality analysis. Transactions in GIS 18, pp. 877 – 895.

Bishr, M. and Kuhn, W., 2007. Geospatial information bottom-up: A matter of trust and semantics. In: The European information society, Springer, pp. 365–387.

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T. and Rau, R., 2011. Osmonto – an ontology of openstreetmap tags. In: M. Schmidt and G. Gartner (eds), Proceedings of the SOTM-EU 2011 : 1st State of the Map - Europe Conference, pp. 55 – 65.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P. and Shi, W., 2010. Thirty years of research on spatial data quality: achievements, failures, and opportunities. Transactions in GIS 14(4), pp. 387–400.

Egenhofer, M. J., 1995. On the equivalence of topological relations. International Journal of Geographical Information Systems 9, pp. 133–152.

Elwood, S., Goodchild, M. F. and Sui, D. Z., 2012. Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice. Annals of the Association of American Geographers 102(3), pp. 571–590.

Fisher, P. F., 1999. Models of uncertainty in spatial data. Geographical information systems 1, pp. 191–205.

Flanagin, A. J. and Metzger, M. J., 2008. The credibility of Volunteered Geographic Information. GeoJournal 72(3-4), pp. 137–148.

Girres, J.-F. and Touya, G., 2010. Quality assessment of the french OpenStreetMap dataset. Transactions in GIS 14(4), pp. 435–459.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), pp. 211–221.

Goodchild, M. F. and Li, L., 2012. Assuring the quality of Volunteered Geographic Information. Spatial statistics 1, pp. 110–120.

Haklay, M., 2010. How good is Volunteered Geographic Information? a comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning. B, Planning & design 37(4), pp. 682.

Hecht, B. and Stephens, M., 2014. A tale of cities: Urban biases in Volunteered Geographic Information. In: Proceeding of the 8th International Conference on Weblogs and Social Media (ICWSM), Michigan, USA.

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A. and Stefanidis, A., 2013. Assessing completeness and spatial error of features in volunteered geographic information. ISPRS International Journal of Geo-Information 2(2), pp. 507–530.

Keßler, C., Trame, J. and Kauppinen, T., 2011. Tracking editing processes in Volunteered Geographic Information: The case of OpenStreetMap. In: Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory, Vol. 12.

Ludwig, I., Voss, A. and Krause-Traudes, M., 2011. A comparison of the street networks of Navteq and OSM in Germany. In: Advancing Geoinformation Science for a Changing World, Springer, pp. 65–84.

Mooney, P. and Corcoran, P., 2012a. The annotation process in OpenStreetMap. Transactions in GIS 16(4), pp. 561–579.

Mooney, P. and Corcoran, P., 2012b. Characteristics of heavily edited objects in openstreetmap. Future Internet 4(1), pp. 285–305.

Neis, P., Zielstra, D. and Zipf, A., 2011. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4(1), pp. 1–21.

Neis, P., Zielstra, D. and Zipf, A., 2013. Comparison of Volunteered Geographic Information data contributions and community development for selected world regions. Future Internet 5(2), pp. 282–300.

Pourabdollah, A., Morley, J., Feldman, S. and Jackson, M., 2013. Towards an authoritative OpenStreetMap: conflating osm and os opendata national maps road network. ISPRS International Journal of Geo-Information 2(3), pp. 704–728.

Schmid, F., Frommberger, L., Cai, C. and Dylla, F., 2013. Lowering the barrier: How the What-You-See-Is-What-You-Map paradigm enables people to contribute volunteered geographic information. In: Proc. of the 4th Annual Symposium on Computing for Development, ACM, Cape Town, South Africa, pp. 8–18.

Thabtah, F., 2007. A review of associative classification mining. The Knowledge Engineering Review 22(01), pp. 37–65.

Tobler, W. R., 1970. A computer movie simulating urban growth in the detroit region. Economic geography 46, pp. 234–240.

Vandecasteele, A. and Devillers, R., 2013. Improving volunteered geographic data quality using semantic similarity measurements. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1(1), pp. 143–148.

Witten, I. H. and Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn, Morgan Kaufmann, San Francisco.

Zielstra, D. and Zipf, A., 2010. Quantitative studies on the data quality of OpenStreetMap in Germany. In: Proc. of the 6th International Conf. on Geographic Information Science, GIScience, Zurich, Switzerland, pp. 20–26.