ISPRS Journal of Photogrammetry and Remote Sensing 127 (2017) 3–15

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Rule-guided human classification of Volunteered Geographic Information ☆

Ahmed Loai Ali [a,b,*,1], Zoe Falomir [a], Falko Schmid [a], Christian Freksa [a]

[a] Bremen Spatial Cognition Center, University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
[b] Information System Department, Faculty of Computers and Information, Assuit University, Assuit, Egypt

ARTICLE INFO

Article history:
Received 26 January 2016
Received in revised form 27 April 2016
Accepted 8 June 2016
Available online 22 June 2016

Keywords:
Volunteered Geographic Information (VGI)
Spatial data quality
Spatial data mining
Classification
Topology
Qualitative spatial reasoning

ABSTRACT

During the last decade, web technologies and location sensing devices have evolved generating a form of crowdsourcing known as Volunteered Geographic Information (VGI). VGI acted as a platform of spatial data collection, in particular, when a group of public participants are involved in collaborative mapping activities: they work together to collect, share, and use information about geographic features. VGI exploits participants' local knowledge to produce rich data sources. However, the resulting data inherits problematic data classification. In VGI projects, the challenges of data classification are due to the following: (i) data is likely prone to subjective classification, (ii) remote contributions and flexible contribution mechanisms in most projects, and (iii) the uncertainty of spatial data and non-strict definitions of geographic features. These factors lead to various forms of problematic classification: inconsistent, incomplete, and imprecise data classification. This research addresses classification appropriateness. Whether the classification of an entity is appropriate or inappropriate is related to quantitative and/or qualitative observations. Small differences between observations may be not recognizable particularly for non-expert participants. Hence, in this paper, the problem is tackled by developing a rule-guided classification approach. This approach exploits data mining techniques of Association Classification (AC) to extract descriptive (qualitative) rules of specific geographic features. The rules are extracted based on the investigation of qualitative topological relations between target features and their context. Afterwards, the extracted rules are used to develop a recommendation system able to guide participants to the most appropriate classification. The approach proposes two scenarios to guide participants towards enhancing the quality of data classification. An empirical study is conducted to investigate the classification of grass-related features like *forest*, *garden*, *park*, and *meadow*. The findings of this study indicate the feasibility of the proposed approach.

1. Introduction

The advanced technologies of Web 2.0, geo-tagging, geo-referencing, Global Navigation Satellite System (GNSS), and broadband communication enable the public to generate spatial content known as User Generated Geographic Content (UGGC) (Goodchild, 2008). They empower ordinary citizens to participate in mapping activities producing geo-spatial content, such activities were formerly conducted by mapping agencies and professional organizations. This trend results in evolving a form of crowdsourcing data known as Volunteered Geographic Information (VGI) (Goodchild, 2007). In this research, we are concerned with the form of VGI, when a group of participants collaboratively work to collect, share, update, and use information about geographic features. Among others, OpenStreetMap[2] (OSM), Google Map Maker[3] and Wikimapia[4] are examples of collaborative mapping projects

☆ This paper is an extended version of the research titled *Towards Rule-Guided Classification for Volunteered Geographic Information* published in the International Symposium of Spatial Data Quality (ISSDQ-2015).

* Corresponding author at: Bremen Spatial Cognition Research Center (BSCC), Universität Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen Germany.

E-mail addresses: loai@informatik.uni-bremen.de (A.L. Ali), zfalomir@informatik.uni-bremen.de (Z. Falomir), schmid@informatik.uni-bremen.de (F. Schmid), freksa@uni-bremen.de (C. Freksa).

[1] Ahmed Loai Ali is sponsored by a DAAD scholarship as a member of the Information System Department, Faculty of Computers and Information, Assuit University, Assuit, Egypt.

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

http://dx.doi.org/10.1016/j.isprsjprs.2016.06.003
0924-2716/

which aim to produce a digital map of the world. During the last decade, VGI played a significant role in the GIScience community. Various applications and services have been developed based on VGI data sources including – but not limited to – environmental monitoring (Gouveia and Fonseca, 2008), crisis management (Roche et al., 2013), urban planning (Foth et al., 2009; Song and Sun, 2010), land use mapping (Arsanjani et al., 2015), and mapping provision (Haklay and Weber, 2008). Moreover, VGI acted as a means of geographic data collection and as a complementary component of Spatial Data Infrastructure (SDI) (McDougall, 2009).

However, the dramatic increase of VGI triggers questions about the resulting data quality (Flanagin and Metzger, 2008; Elwood et al., 2012). Among other things, the lack of detailed information about data quality and the difficulty of applying the conventional spatial quality measures 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. However, this paper addresses the quality from the perspective of data classification.

In a VGI context, the classification of data faces various challenges. On one hand, a large amount of data is contributed by arm-chair participants based on their local knowledge. This remote contribution method results in imprecise classification. On the other hand, human observations generate subjective data classification. Whether a water body is classified in VGI as *pond* or *lake*, depends on the participant's perceptions. In contrast, in the professional field, a strictly defined classification model is developed by experts in advance, and then data is classified according to measures and observations in comparison to the predefined model. Hence, remote contributions and subjective perceptions, among other reasons, produce problematic data classification, and consequently, difficulties for data integration and utilization.

For example, Fig. 1 shows one of the common interfaces (iD editor) of OSM project, where participants can edit geographic features using the appropriate geometric representation (point, line, or polygon) by tracking over satellite images provided by Bing.[5] Afterwards, they describe (classify) the sketched entity using tags (see Section 3.2). Whether this piece of land covered by grass – in the middle of Fig. 1 – is classified as *park, garden, meadow*, or generally *grass*, is not strictly defined. The human-centered classification generates multiple acceptable class labels with higher or lower degrees of appropriateness. The given entity can be recognized by a participant as *park*, even if it has been classified by others as *garden* or *forest*. The most appropriate classification of an entity is related to qualitative and/or quantitative observations. Small difference in observations might lead to different classification. These differences might be not recognizable by voluntary participants. Hence, this paper presents a rule-guided classification approach to tackle the classification problems of VGI.

The proposed approach exploits the dramatic increase of VGI towards enhancing data classification. It consists of two phases: *Learning* and *Guiding* phases. During the *Learning* phase, the task is to learn the qualitative characteristics that distinguish among similar classes. This task exclusively investigates qualitative topological characteristics of specific classes. The extracted characteristics are formulated into descriptive qualitative rules able to guide the participants towards the most appropriate classification. The *Guiding* phase presents two scenarios for applying the generated guidance and recommendations.

To validate the proposed approach, an empirical study has been conducted addressing the classification of grass-related features. Classes of *forest, garden, grass, meadow, park*, and *wood* are selected for the study. The classification of these features represents a challenge; they are commonly covered by grass, although each class has its unique features. For example, the classes *park* and *garden* have entertainment characteristics, *forest* and *wood* are usually covered with trees or other woody vegetation, *meadow* has agriculture characteristics, etc. The findings indicate the feasibility of the proposed approach. Specifically, the developed system is able to unambiguously classify some of the target classes, while other classes still have poor classification accuracy.

This paper is organized as follows. Section 2 presents a review of VGI data quality. Section 3 gives insight into the fundamental reasons behind the problematic classification of VGI. Section 4 presents an overview of the qualitative spatial reasoning field, which provides intuitive and well-defined semantics from spatial quantitative data. Section 5 presents the proposed approach and its phases and Section 6 describes the empirical study carried out. Section 7 envisions the application of the presented approach in emerging GIS trends. The last section concludes the findings and points to future work.

2. Issues of VGI data quality

In VGI, humans are the fundamental source of data. Particularly in collaborative mapping projects, participants record their observations by collecting, updating, and sharing information about geographic features. VGI employs participants' local knowledge and their willingness to contribute in order to produce rich spatial data sources (Goodchild, 2007). But the quality of the resulting data is questionable. With increasing utilization of VGI in GIScience research, data quality becomes a concern of highest priority (Flanagin and Metzger, 2008; Elwood et al., 2012). Various methods to assess data intrinsically or extrinsically can be found in the literature (Section 2.1), also methodologies/ approaches to improve data quality (Section 2.2), whereas there is only a limited number of research that addresses data classification problems (Section 2.3).

2.1. Extrinsic and intrinsic data assessment

Generally, VGI is evaluated by following either extrinsic or intrinsic procedures. In the extrinsic procedure, with the availability of ground-truth data, the VGI data set is compared with a comparable ground-truth data source. Girres and Touya (2010), Haklay (2010), Neis et al. (2011), and Jackson et al. (2013) compared OSM data against ground-truth data sources in France, UK, Germany, and USA, respectively. They emphasized the quality of VGI data particularly in urban areas. In Hecht and Stephens (2014), authors found that VGI data quality decreases with increased distance from urban areas.

In the intrinsic procedure, comparable data sources are not available. The data is assessed by analyzing its intrinsic properties like participants' mapping activities, data development, and participants' reputation. Goodchild and Li (2012) presented three dimensions that could be followed to ensure VGI quality intrinsically: the crowdsourcing, social, and geographic dimensions. Bishr and Kuhn (2007), Keßler et al. (2011), Neis et al. (2011), Mooney and Corcoran (2012b), and Barron et al. (2014) assessed VGI data intrinsically. They analyzed meta-data of VGI like contributors' mapping activities and reputation, editing history of entities, Neis et al. (2013) compared the development of contributors' communities in different cities around the world indicating the relation between the communities and data quality. Moreover, the nature of VGI results in new intrinsic measures of data quality like fitness of use and conceptual quality. Barron et al. (2014) developed 25

[5] www.bing.com/maps.