International Workshop on Geospatial Big Data - Trends, Applications, and Challenges (GBD - TAC)

A PageRank-based reputation model for VGI data

Carlo Lodigiani[a], Michele Melchiori[a,*]

[a]*Dept. of Information Engineering - University of Brescia, via Branze, 38, Brescia, 25123, Italy*

Abstract

Quality of data is one of the key issues in the domain of Volunteered geographic information (VGI). To this purpose, in literature VGI data has been sometime compared with authoritative geospatial data. Evaluation of single contributions to VGI databases is more relevant for some applications and typically relies on evaluating reputation of contributors and using it as proxy measures for data quality. In this paper, we present a novel approach for reputation evaluation that is based on the well known PageRank algorithm for Web pages. We use a simple model for describing different versions of a geospatial entity in terms of corrections and completions. Authors, VGI contributions and their mutual relationships are modeled as nodes of a graph. In order to evaluate reputation of authors and contributions in the graph we propose an algorithm that is based on the personalized version of PageRank.

1. Introduction

Models for evaluating quality of data have been been proposed and widely discussed in the context of authoritative geospatial data sources [15]. In the domain of Volunteered geographic information (VGI) this is considered one of the key and most studied issues [1,7,14]. Various works have compared VGI datasets with authoritative geospatial ones on the same spatial region in order to provide a general evaluations of the average VGI quality according to various metrics (e.g.,[7]). However, quality estimation of single VGI contributions, like the description and location of a point of interest, may be more relevant for typical end users which operate focusing on a single or few geospatial objects. As well as, using traditional geospatial datasources for evaluating VGI data by comparison is not always viable or convenient due to licensing fees and access restrictions. Therefore, in literature (e.g.,[5]) metrics have been defined for measuring trust or reputation level of authors of VGI contributions based on feedbacks and these metrics have been used as proxy measures for quality of VGI contributions. Alternatively, reputation of users is inferred by their activities of feature editing and how these activities have been treated later on by other users [10]. Reputation evaluation of users and descriptions of VGI objects have been also performed based on a model explicitly including production time and history of versions of the geospatial object [16]. As well as, measuring the reputation of users, when they produce ratings

* Corresponding author. Tel.: +39-030-3715845, Fax +39-030-380014.
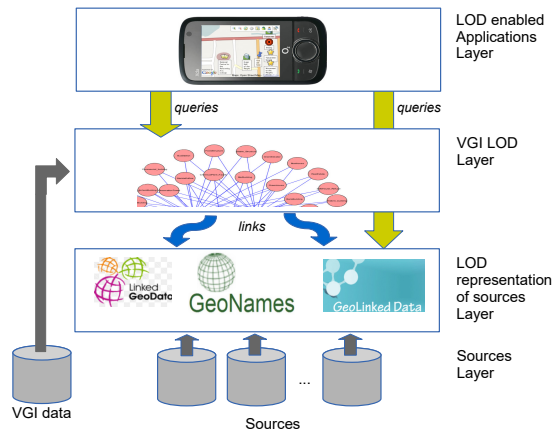 *E-mail address:* michele.melchiori@unibs.it

doi:10.1016/j.procs.2016.09.088

Fig. 1. Positioning of the Corrections LOD Layer.

or when we need to weight their experience, is relevant in other domains, e.g., Web services[4]. Additional motivations for dealing with VGI quality in a different way w.r.t. the approaches for quality of authoritative geospatial datasets, are heterogeneity, contributors' behavior, environment and their mutual interactions. For example, heterogeneity of coverage occurs since contributions are not usually distributed in a uniform way on the map. This phenomena is well discussed in literature and known as generating "cupcakes" of information, with zones well detailed and other ones that are not covered or are covered partially[2]. Other heterogeneities are possibly due to non uniform motivations of users. This brings to the production of datasets detailed and structured in a way that reflects the users' personal interests. In literature, the outcome of uneven production is known as "patchworks of geographic information"[8].

Paper contribution and positioning. In this paper, we propose a novel approach for estimating reputation of VGI data and authors based on the well known PageRank algorithm for ranking Web pages. Some other works deal with the estimation of reputation in VGI. For example, Bishr and Khun[5] describe a reputation model based on coherence between volounter's reports on drinkability of wells in developing countries. The drinkability status has only two possible values: good and bad. Time is explicitly included in the model. In fact, trustworthiness in reports on wells is reduced proportionally to the passed time. Our model considers more general VGI scenarios allowing for complex descriptions for objects. Another approach is given in Zhao et al.[16] that bases the trustworthiness on contributor's reputation and on analyzing editing sequences of VGI versions, similarly to Keßler et al.[11] and D'Antonio et al.[6]. Each version is created by a contributor and describes the current status of a geospatial object. Level of trust for a specific version depends on: (i) contributor's reputation; (ii) distance between this version and the previous one (smaller is better) for the same object; (iii) level of trust in the previous version. This approach looks actually inspired by ideas in D'Antonio et al.[6] but comes with a complete data model. As per our model, these Authors distinguish between implicit and explicit assessment of contributors. Ours is a light model (e.g., it does not include in the current version, effects of information obsolescence) and focuses on potentialities and properties of using the PageRank algorithm in VGI that, as far as we know, is something not before discussed in literature. Moreover in our approach we explicitly consider the issue of defending the reputation assessment from undesired manipulations performed by users.

Our paper is organized as follows. A simple model for describing versions of the same geospatial object in terms of corrections and completions is given in Sect. 2, in the context of a linked data application scenario. Authors, VGI contributions and their mutual relationships are then modeled as nodes of a graph. In order to evaluate reputation we propose in Sect. 4 an adaptation of the personalized version of PageRank described in Sect. 3 to our problem. Finally, we provide some experimentation based on a preliminary implementation of the approach in Sect. 5.

2. A linked data model for VGI

In this section, we briefly present in a linked data application scenario the reference framework[9] for our proposal. According to the model in[16] a VGI version describes the state of a geospatial object. A state makes reference to