Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 566 – 571

Procedia Computer Science

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.

Peer-review under responsibility of the Program Chairs

Keywords: VGI, volunteered geographic information, reputation model, trust model, VGI data model, 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

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.088

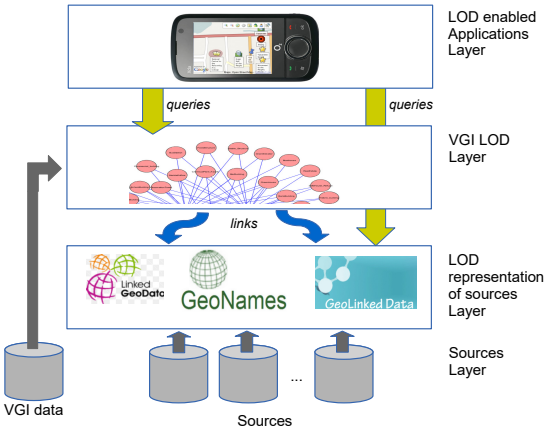

Fig. 1. Positioning of the Corrections LOD Layer.

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

attributes and their values of an object and can change during the object lifespan (e.g., the address has changed). For an object state, a user may produce a VGI version and another user may correct (e.g., proposing a change to an attribute value) or complete (e.g., adding one or more pairs attribute-value) thus originating another version. A user can also give a positive or a negative feedback to corrections and completions in order to express her agreement or disagreement. In our framework, we consider corrections and completions of both VGI data and authoritative geospatial data exposed as linked data. The collection of corrections and completions is stored as Linked Open Data (LOD). According to a conceptual perspective, this data can be considered as an additional *VGI LOD layer* in the architecture of geospatial linked data, which is set between the layer of the linked data sources and the applications one. In this way, geospatial applications can use and query the additional layer and the original sources in a joint way as depicted in Fig. 1. Basic requirements for the framework are that the maintained data collection has to be complementary to the existing sources and compliant to the LOD practices: (i) the *VGI LOD layer* is published as LOD as the original data; this means that, each VGI version includes as well the original URI(s) which can be resolved to access the original description of the geospatial object in the corresponding source; (ii) the *VGI LOD layer* constitutes a mediation layer built over the original sources and linked to them; (iii) Applications and users can browse this layer by using either specific applications or linked data querying tools (e.g., RDF data browsers and SPARQL/GeoSPARQL query languages).

3. The Personalized PageRank algorithm

In the Google PageRank algorithm the relevance of a web page B is recursively determined by the total number of links pointing at it[3]. Where each of these links has a weight proportional to the relevance of the web page A, source of the link, divided by the number of the links outgoing from A. The resulting relevance is also called *rank* of the page. Linked pages define a directed graph where pages are represented as nodes and links as edges. In the original formalization of PageRank the rank measures the authority of a web page and is meant as probability of visiting a the page by a web surfer. The rank of a page is calculated as:

$$PR(p_i) = \frac{1-\alpha}{|N|} + \alpha \cdot \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{1}$$

where $PR(p_j)$ is the page p_j rank, N is the number of pages, α a factor usually set to 0.85, called damping factor, $M(p_i)$ the number of pages with links to p_i and $L(p_j)$ is the number of outgoing links from page p_j. Initially, all pages are assigned with the same probability, that is, $PR(p_i) = 1/N$. The approach actually used to compute the recursive Equation (1) is the following. First, a $N \times N$ transition matrix P is defined with elements $p_{ij} = 1/L(p_j)$ if there is a link from page j to page i, otherwise $p_{ij} = 0$. Then the matrix P is modified to deal with rows of 0s associated with *dangling* pages, i.e., pages with no outgoing links. These rows are replaced with vectors $(1/n)e^T$, where e^T is a vector of length n with 1s elements and n is the order of the matrix. The resulting matrix is denoted as $\overline{P}$. In order to compute the $PR(p_i)$ values it is required a further manipulation of the matrix (to make it *irreducible*). This consists of computing a matrix $\overline{\overline{P}} = \alpha\overline{P} + (1-\alpha)ee^T/n$ with α the damping factor above mentioned. Finally, $PR(p_i)$ is found as i-th component of the eigenvalue π^*, where π^* is the solution of the system:

$$\pi\overline{\overline{P}} = \pi \tag{2}$$

This solution can be obtained by applying simple and effective iterative procedure[3].

The personalized version. In our framework we rely on a version of the PageRank algorithm that is known as Personalized PageRank[12]. The difference between the standard algorithm and the personalized one is the presence of a vector, called personalization v^T, that modifies the irreducible matrix $\overline{\overline{P}} = \alpha\overline{P} + (1-\alpha)ee^T/n$ of PageRank. The matrix ee^T/n is called teleportation matrix and intuitively it permits to set for each page A the probability that the web surfer dumps the current page and without following any link she chooses arbitrarily to visit page A. In the standard version, each page is assigned with the same probability to be visited. In the Personalized version we want that the probabilities that a user teleports any web page are not uniform so the teleportation matrix is redefined as ev^T. The i-th element of v^T is the probability to teleport to node i and higher this probability, higher will be the rank of the page

Table 1. User's actions and impact on the personalization vector v^T.

Type	Subject	Object	Author of the object	Result
Correction	U_j	I_{ik}	U_i	• Decrease the element associated with V_i in v^T • Decrease the element associated with I_{ik} in v^T
Completion	U_j	I_{ik}	U_i	• Increase the element associated with V_i in v^T • Increase the element associated with I_{ik} in v^T
Feedback +	U_j	I_{ik}	U_i	• Increase the element associated with V_i in v^T • Increase the element associated with I_{ik} in v^T
Feedback -	U_j	I_{ik}	U_i	• Decrease the element associated with V_i in v^T • Decrease the element associated with I_{ik} in v^T

i. Therefore, v can be used to influence the rank of a page, increasing or decreasing it. This approach has been used in web engines to personalize search results based on user requirements (e.g., ranking better pages about sport for users that expressed this prefence) or to penalize that pages which try to increase illegally their rank (also known as, link farms technique).

4. PageRank-based reputation evaluation

In order to apply the PageRank technique to the VGI data model presented in Section 2, we describe VGI data as a graph and evaluate each node using an adapted version of Personalized PageRank algorithm. Stated in other terms, we use PageRank to assign a reputation value to VGI contributors and versions of data based on the computed rank. We consider four different types of nodes representing: user, VGI versions, correction and completion. When a new user registers, she is assigned with a default value in the personalization vector and she receives a new rank after a first iteration of the algorithm. Moreover, a new node is added to the graph in order to represent the new user. When a user generates a new VGI version, this action creates a new node linked from the user node. Each VGI version can be completed or corrected more times. Conceptually, a completion is a set of new pairs $< attribute, value >$ added to an existing VGI version (e.g., adding <phone2,+390303333020> to the version). A correction consists of a new value for an existing pair $< attribute, value >$ (e.g., the pair <opening, Monday-Friday> corrects <opening,Monday-Saturday>). Completion nodes are connected to the VGI version (the link is oriented from the VGI version to the completion node) and to the author (the link is oriented from the author to the completion node). Correction nodes are connected to either a VGI version or a completion node and to the author. For example, in Fig. 2, nodes 1,3,4,5,6 represent users. Node 2 is a VGI version created by user 1. Node 7 is a completion produced by user 3 to this VGI version (node 2). Note that corrections of corrections are not included in the current version of the model. We distinguish between explicit and implicit feedbacks. A user can give an explicit positive or a negative feedback to a VGI version, correction, completion. This is not represented in the graph organization but it impacts on the ranking process as explained in the following. Moreover, a correction is also considered as an implicit negative feedback, since it reveals something wrong about the corrected information, and a completion as an implicit positive feedback, since it confirms the correctness of the information by completing it.

Operationally, our approach is divided into two steps. In the first step, the personalization vector v^T is modified according to users' actions. As we said, changes in v^T set the precondition for changing ranks of nodes. For example, if a version receives a correction, both the rank of information and the rank of user which generated it are decreased since this is an implicit negative feedback. Table 2 represents all the admitted user'a actions and their effects on v^T. Every change (increase or decrease) in v^T depends on the rank of the subject performing the action. Obviously, every change in v^T is such that the sum of all its elements remains equal to one by normalizing the vector.

In the second step, the PageRank algorithm is applied to the graph using as input the personalization vector v^T updated by the first step. This has the effect of actually changing the rank of the graph nodes in order to reflect a possible new value of reputation for users, VGI versions, correction and completions . As said before, a user can also give a feedback to a versions, a correction or a completion and we permit a user to give at most one feedback per each node.

Some properties of the node ranking are the following:

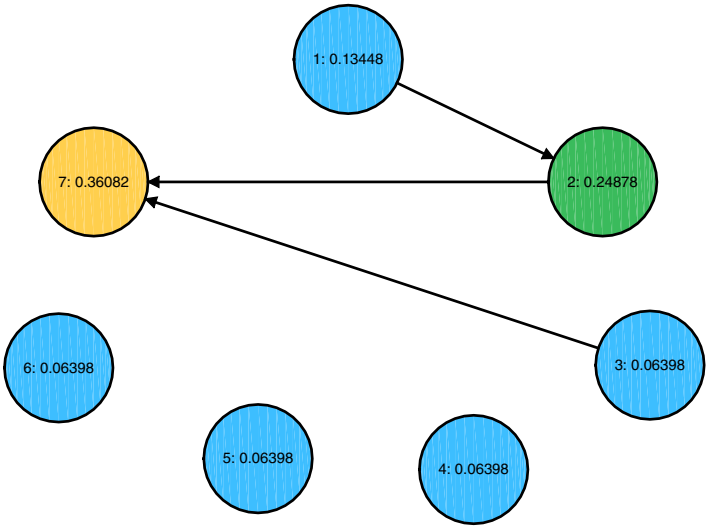

Fig. 2. Initial graph configuration.

- The initial rank of a VGI version is proportional to the rank of the author. So better the author's rank, more reliable is considered the version. As well as, initial rank of a VGI version is inversely proportional to the number of versions produced by an author.
- Rank of a version is changed based on received feedbacks (explicit or implicit). The impact of a feedback on the version rank is proportional to the feedback author's rank. So better the author's ranking, more powerful the feedback effect.

5. Preliminary experimentation

In this section we provide some preliminary experimentation of the approach based on a Matlab implementation. In particular, we consider how the approach performs when we have a mix of cooperative and non cooperative users on a simple case study. Cooperative users give a positive feedback to VGI versions that we assume correct and non cooperative ones assume a vandalistic attitude giving a negative feedback to the same version in order to damage its reputation [13].

We consider a version A that is truthful and has been confirmed by a number x (in this case we use $x = 5$) of users through positive feedback or completion. Initial situation in represented as graph in Fig. 2, where inside nodes we report the node identificator followed by the node rank.

We analyzed the behavior of the version rank as the number of non cooperative users increases. Fig. 3(a) shows how the ranks of a version (represented by node 2 in the graph) and of the author (node 1) as the the data evolve. In Fig. 3(b) the authors' rank variation is represented as percentage. We observe that it decreases by 30% when the number of non cooperative users equals the number of the cooperative ones, that is 5.

We analyzed also the case of two users trying to give each other positive feedbacks for all the information they have produced in order to raise artificially their ranks. Our results show that this type of manipulation is not very effective for its authors because the redistribution of ranks implemented in our approach bounds the rise of the ranks. Indeed, even if the sum of ranks of the users may increase (e.g., by 30%) the sum of the ranks of all the versions, correction and completions they have produced decreases by a value such that the sum of all values gives 1.

6. Conclusions

In this paper we presented a first version of a model for evaluating reputation of VGI users and data. Users and VGI contributions are modeled as nodes and authoring relationships between them as direct edges. A approach based

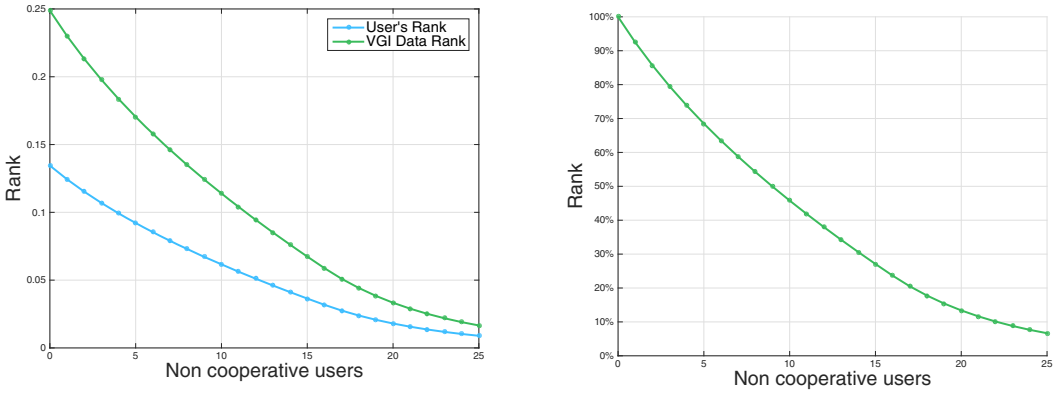

Fig. 3. Resistance to non cooperative users: (a) absolute ranks. (b) relative user's rank.

on an adaptation of the personalized PageRank algorithm, which is used for ranking web pages, is applied to the graph in order to compute reputations associated with nodes. The proposal is part of a wider framework we developed for dealing with VGI Linked Open Data. The model has to be further studied and improved. For example, currently it does not distinguish between corrections to errors and corrections describing updates required by the evolution of geographic objects. In this sense, an update action should have an effect on reputation different from a completion. We also plan to include the effect of aging on VGI data and on feedbacks as well as studying the possibility of integrating evidence from the editing history of a version as in[11,6,16]. Finally, a further possible refinement concerns dealing with some complexity indexes (e.g., density in the working area or type of objects classes) in order to evaluate the user reputation also proportionally to her ability to deal with complex situations.

References

1. Jamal Jokar Arsanjani, Peter Mooney, Alexander Zipf, and Anne Schauss. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. In *OpenStreetMap in GIScience*, pages 37–58. Springer, 2015.
2. D. Bégin, R. Devillers, and S. Roche. Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours. In *Proceedings of the 8th international symposium on spatial data quality ISSDQ*, pages 149–154, 2013.
3. Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.
4. Devis Bianchini, Valeria De Antonellis, and Michele Melchiori. An approach for service selection based on developers ranking. In *Web Services (ICWS), 2016 IEEE International Conference on*, pages 704–707, June 2016.
5. Mohamed Bishr and Werner Kuhn. Trust and reputation models for quality assessment of human sensor observations. In *Spatial Information Theory*, pages 53–73. Springer, 2013.
6. Fausto DÁntonio, Paolo Fogliaroni, and Tomi Kauppinen. Vgi edit history reveals data trustworthiness and user reputation. In *17th AGILE International Conference on Geographic Information Science (Short Paper)*, 2014.
7. Helen Dorn, Tobias Törnros, and Alexander Zipf. Quality Evaluation of VGI Using Authoritative Data - A Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information*, 4(3):1657–1671, 2015.
8. Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
9. Roula Karam and Michele Melchiori. Improving geo-spatial linked data with the wisdom of the crowds. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 68–74, New York, NY, USA, 2013. ACM.
10. Carsten Keßler and René Theodore Anton de Groot. Trust as a proxy measure for the quality of volunteered geographic information in the case of openstreetmap. In *Geographic information science at the heart of Europe*, pages 21–37. Springer, 2013.
11. Carsten Keßler, Johannes Trame, and Tomi Kauppinen. Tracking editing processes in volunteered geographic information: The case of openstreetmap. *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, 12, 2011.
12. Amy N Langville and Carl D Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.
13. Pascal Neis, Marcus Goetz, and Alexander Zipf. Towards automatic vandalism detection in openstreetmap. *ISPRS International Journal of Geo-Information*, 1(3):315–332, 2012.
14. Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Muki Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, May 2016.
15. Howard Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.
16. Yijiang Zhao, Xiaoguang Zhou, Guangqiang Li, and Hanfa Xing. A Spatio-Temporal VGI Model Considering Trust-Related Information. *ISPRS International Journal of Geo-Information*, 5(2):10, 2016.