Artificial Intelligence 239 (2016) 143–167

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Building knowledge maps of Web graphs

Valeria Fionda [a], Claudio Gutierrez [b], Giuseppe Pirrò [c,*]

[a] DeMaCS, University of Calabria, Italy
[b] DCC, University of Chile, Chile
[c] ICAR-CNR, Rende, CS, Italy

ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form 3 July 2016
Accepted 11 July 2016
Available online 18 July 2016

Keywords:
Maps of the web
Navigation
Web of data
Linked data
Semantic web

ABSTRACT

We research the problem of building knowledge maps of graph-like information. There exist well-consolidated cartographic principles and techniques for mapping physical landscapes. However, we live in the digital era and similarly to the Earth, the Web is simply too large and its interrelations too complex for anyone to grasp much of it through direct observation. Thus, the problem of applying cartographic principles also to digital landscapes is intriguing. We introduce a mathematical formalism that captures the general notion of map of a graph and enables its development and manipulation in a semi-automated way. We present an implementation of our formalism on the Web of Linked Data graph and discuss algorithms that efficiently generate and combine (via an algebra) regions and maps. We present the MaGe tool, implementing the map framework, and discuss examples of knowledge maps.

1. Introduction

The Web can be seen as a vast space of interconnected information that users commonly access via navigation enabled by browsers. However, the Web is simply too large and its interrelations too complex for anyone to grasp much only by direct observation. Consider the task of navigating a citation network by using, for instance, Google Scholar.[1] One typically starts from a seed paper. By clicking on the *cited by* link, one navigates towards papers that have cited the seed paper, selects those of interest (e.g., by bookmarking them) and then continues. After a while it is very hard to reconstruct the network of citations in terms of papers of interest and connections between them. Moreover, the whole process is manual. Having an automatic way of identifying the portion of the citation network of interest (i.e., papers and their connections) and then some form of abstract representation, where only salient papers (e.g., papers with certain keywords in the title) and links between them are represented, is an extremely useful support to the navigation.

To cope with the huge amount of interconnected information available on the Web, we take inspiration from cartography and introduce a framework to build *maps of the Web*. In the physical space, the process of map making can be summarized in two main steps, that is, *selection* and *abstraction* [1]. Selection enables one to focus only on the particular pieces of information that will serve the map's purpose; specifically, in this phase *the region* to be mapped is chosen. In our previous example about navigating a citation network, the region would consist in nodes (i.e., papers) and *cited by* links visited during the navigation. Abstraction is the fundamental property of a map, which states that a map is always smaller than the region

* Corresponding author.
E-mail addresses: fionda@mat.unical.it (V. Fionda), cgutierr@dcc.uchile.cl (C. Gutierrez), pirro@icar.cnr.it (G. Pirrò).

[1] http://scholar.google.com.

http://dx.doi.org/10.1016/j.artint.2016.07.003
0004-3702/

it portrays. Abstracting the region visited while navigating our citation network could be done by considering only nodes with certain properties (e.g., papers published in some specific conference) and links between them.

We see Web Maps as useful navigational cues and powerful ways of conveying complex information via abstract representations; in our context abstraction is used to remove unwanted details that are out of the scope of the map's subject. This gives Web Maps the role of *navigational charts* that help cope with the size of the Web (cyber) space and elude the "lost in the cyberspace" syndrome [2]. Thanks to Web Maps, users can explore complex digital territories, find routes toward destinations and discover previously unknown connections between knowledge items. Web Maps are also useful for analyzing information. For instance, the availability of a series of chronologically sequential maps enables complex map analysis (e.g., longitudinal analysis) for the detection and forecasting of trends in specific domains. This is useful, for instance, in the analysis of knowledge flows in scientific literature to show how the interlinking between disciplines is changing [3]. Another example are maps of social networks that can be analyzed to forecast friendship [4].

Recent progress in Web technologies and languages originating from the Semantic Web proposal *as well as* the availability at planetary scale of structured information in the Resource Description Framework (RDF) standard data format, open new opportunities *toward automating* the construction of Web Maps. On one hand, interlinks between data items, encoded in RDF predicates, carry a precise semantic meaning, thus allowing for precise characterization of the nature of reachability that *is crucial* in extracting regions of the Web. On the other hand, maps can be given an RDF representation and then be processed not only by humans, via visual interpretations, but also by machines, due to the machine-processable nature of RDF. This will foster the exchange, combination and reuse of maps. We believe that the availability of Web Maps can help human users cope with the complexity of Web Regions in the same way as geographic maps help users cope with the complexity of large physical regions.

Notions of maps for digital landscapes have been around for some time. Doemel [5] introduced the idea of Web Map as a means to capture user navigational activities. This kind of map is not a map in the cartographic sense as it misses the abstraction phase, which is the raison d'être of a map [1]. There are also many tools (partially) touching upon the problem of building maps of the Web. The most traditional and popular are bookmarks: a list of URLs specified by a user when navigating the Web. This idea has been enhanced to incorporate, for instance, social features (share, rank, tag) and/or annotations of different types of data (e.g., not only pages but also documents). Delicious,[2] Diigo,[3] and Google Bookmarks[4] are among the most popular bookmarking systems. Some systems go beyond simple bookmarks by enabling one to organize URLs to also highlight connections between *the two*. Results are grouped and presented in the form of a graph, which simulates the idea of a region of the Web. Examples of such systems are search engines like Tag Galaxy,[5] navigational history tools (e.g., [5,6]), visual HTML site maps (for users) and atlases of the Web (e.g., [2]). More recent approaches focus on providing visual representations of information in specific domains such as publications or news (e.g., [7,8]).

Existing approaches, discussed in Sections 2.1 and 6, do not comply with the idea of a map that we envision. First, even if they partially simulate the idea of capturing regions of the Web, they do not consider the abstraction of Web Regions to build maps. Second, they are designed for human visualization; hence their automatic processing, composition and reuse are not considered, which hinders the exchange, automatic combination and interpretation of maps. Third, they do not enable the declarative specification of the region (e.g., portion of interest of the Web) to be mapped thus hindering the automation of the process of creating maps. Fourth, they lack a formal mathematical model. They do not guarantee formal/provable (reachability) relations between the points in the map; formal notions of granularity and scope; and formal provable relations between the map and the region it represents. These limitations obstruct the generation of formal deductions from maps.

1.1. Contributions

In this paper we formally introduce the notion of Web Map and face several challenges toward this goal. First, given a user need or a conceptual notion, *provide a way to specify* a region of the Web that represents or encompasses it. Second, given a region of the Web, define what is a map of it and investigate its formal properties. Third, devise algorithms and compose the procedures efficiently. The contributions of this paper are:

- We provide a mathematical formalization of the notions of region and map of a graph; we discuss several types of maps, present algorithms for constructing such maps and study their complexity.
- We introduce an algebra for maps and study its formal properties.
- We discuss the problem of obtaining regions of the Web via graph navigational languages; to this aim we introduce a general navigational language to specify Web Regions.
- We showcase an implementation of the formal map framework and navigational language on the current Web of Linked Data via a tool called MaGe (Map Generator).[6]
- We discuss some examples of Web Maps with real data.

[2] http://delicious.com/.
[3] http://www.diigo.com/.
[4] http://www.google.com/bookmarks/.
[5] http://taggalaxy.de/.
[6] The tool is available at the Website http://mapsforweb.wordpress.com.