TheMa: An API for Mining Linked Datasets

Chrysostomos Tsoukalas, Dimitris Dervos
Department of Information Technology
A.T.E.I of Thessaloniki
57 400 Thessaloniki, P.O BOX 141, Greece
tsukalasCh@gmail.com, dad@it.teithe.gr

Jorge Martinez-Gil, Jose F. Aldana-Montes
Department of Computer Science
University of Malaga
29071 Malaga, Spain
jorgemar@acm.org, jfam@lcc.uma.es

Abstract—Linked Open Data is a paradigm for linking the data available on the Web in a structured format in order to make it accessible for computers and people. This leads to having more people and services publish their data on the web and as a result the graph that contains all this information is getting bigger. This paper proposes the usage of some network analysis algorithms on a Linked Dataset in order to extract useful information which in turn leads to a better understanding/interpretation of the data involved, plus comprises a first step in the direction of mining hidden information from the dataset.

Keywords: Linked Open Data, Graph Mining, Network Theory

INTRODUCTION

Linked Open Data (LOD) is an emerging paradigm for connecting the data available on the World Wide Web (WWW) in a well defined way so that it could be accessible for computers and people [1]. Nowadays, the amount of linked data available on the WWW is growing very quickly. The data reflect a wide range of resources like institutional data, user-generated content, and so on. With more and more sources publishing their content in the form of linked data, this amount of data is exploding, and therefore very difficult to process. Therefore, the development of new techniques and tools for facilitating this task comprises a challenging task for the research community.

More specifically, dealing with this huge amount of linked data content can be seen to comprise a challenge for many analysts. For example, exploiting implicit information from the linked datasets can lead to improved effectiveness of information retrieval operations. Also, the extraction of useful knowledge that is not explicit in its current form [3] can lead to important competitive advantages. To the best of our knowledge, only a few software tools for mining information from this kind of datasets have been reported in the literature, today [2].

In this paper, we present TheMa (derived from *Thessaloniki* and *Malaga*: author-city affiliations in the current project) , an Application Programming Interface (API) suitable for calculating useful information related to the Linked Datasets available on the WWW. This information includes the identification of the most prestigious nodes, the key predicates, the reachability of the nodes, and more statistics concerning the linked datasets. In this first version of the TheMa API, a simple, yet useful set of methods from the field of the network theory have been included. To the best of our knowledge, this API comprises one of the first reported attempts to provide a set of useful methods for analyzing and mining Linked Datasets available on the Web.

The remainder of this paper is organized as follows: Section 2 presents the state-of-the-art relating to data mining techniques for dealing with Linked Data. Section 3 describes our contribution, including the preliminary notions of the concepts involved, the design decisions and the development details taken into account when developing TheMa. In Section 4, we evaluate our implementation using a linked dataset extracted from DBPedia. In section 5, we summarize on the work presented and suggest future lines of research.

STATE-OF-THE-ART

The great amount of linked data in the form of RDF triples on the Web can be considered an important step forward in the direction of establishing a structured web which may allow not only to humans, but also computers to process and interpret data, information or knowledge available on the WWW. In fact, today, a number of software applications benefit from billions of triples available in repositories like DBpedia (www.dbpedia.org) [4]. At the same time, experts specializing in areas like finance, medicine or bioinformatics, demand the existence of more formal and expressive knowledge models for their data.

Therefore, an important challenge for linked data mining relates with the problem of mining structured datasets, where entities are linked in some way. Links among entities belonging to the same dataset may exhibit certain patterns, which can be useful for many mining tasks and they are

978-0-7695-4825-8/12 $26.00 © 2012 IEEE
DOI 10.1109/PCi.2012.58

usually hard to reveal using traditional statistical models over conventional databases.

Such type of problems have been traditionally studied by the link mining community who collectively label them as "data mining techniques that explicitly consider links when building predictive or descriptive models of the linked data" [6].

Commonly addressed link mining tasks include object ranking, group detection, collective classification, link prediction and sub graph discovery [5]. Therefore, mining Linked Data can be useful in a number of analogous situations, some of which are explained below.

A. Use Cases

The problem of mining linked data on the Web is becoming relevant as more and more information is made available online. Some of the most popular mining tasks focus on

- *The identification of customer networks [9].* Mining Linked Datasets can help provide a better understanding one has on a dataset. This can be very useful from the point of view of the organizations who want to cluster people on the basis of a given common profile.
- *The identification of crime or fraud networks [10].* Mining Linked Datasets can help experts who want to identify possible fraud scenarios by discovering fraud indicators and connections between nodes. Obviously, it is supposed that criminals are not going to publish their data on the WWW, but for example, institutional data on public funds spending are usually published and many misuses can be discovered using computer algorithms.

These are only a few examples, but we are confident that over time, users and practitioners are likely to propose more areas of application for this type of software.

CONTRIBUTION

Firstly, we are going to outline the formal aspects of our model. Next, we are going to explain the design and evaluation of our API.

A Linked Dataset is a set of triples LD = (S, P, O) where

- S is a concept which is called subject
- O is a concept or a literal data which is called object
- P is an ordered pair which includes S and O. It is called predicate

A predicate p = (S, O) is always directed from S to O; O is also called the head and S is called the tail of the predicate; O is said to comprise a direct successor of S, and S is said to comprise a direct predecessor of O. If a path leads from S to O, then O is said to be a successor of S and reachable from S, and S is said to be a predecessor of O.

On the other hand, a Linked Dataset (LD) is called symmetric if, for every predicated in LD, the corresponding inverted predicated also belongs to LD.

For the rest of this section we are going to explain the design and the development details of TheMa.

A. Design of TheMa

We have found inspiration in network theory in order to design our API. The reason is that network theory provides the foundations concerning the study of graphs as a representation of relations between discrete objects; this is direct correspondence with the proposed data model.

In this first version of our API, we have decided to implement only basic operations like: the computation of the prestige measure for each one node, the discovery of bridges, and the computation of the reachability for a given node. All these methods are overloaded, thus, they are offered under different versions so that users can select the most appropriate function for each application. We are going to explain the details for these operations now.

1) Prestige: The prestige of node in a given dataset relates to the reputation or importance that a node has in a dataset. To represent this measure, we count the links that converge on, and those that originate from this node. The larger the number of links that converge on or originate from the node, the more prestigious the node is. Two types of node prestige measures are calculated: the out-going prestige and in-coming prestige. The two concepts are defined more formally as follows:

Definition: *The input prestige of a node n is the number of predicates terminating at n (Figure 1).*

Definition: *The output prestige of a node n is the number of predicates beginning at n (Figure 2).*

Out-going prestige measures the links in which this node is a subject pointing to other nodes. A very prestigious node is one that appears as a subject in many triples in the dataset. One may claim that the node in question is the source to a considerable amount of information in the dataset. Consequently, using a prestigious node as a starting point in order to retrieve information present in the dataset is likely to lead to a more reliable result.

In-going prestige measures the links that points to a specific node. Equivalently, the number of links the node in question comprises the object of. A node with a high in-going prestige value tends to be an important node for the dataset because it represents useful information for most of the nodes in the dataset. This in turn implies that following the links to high in-going prestige nodes in the graph, one is able to efficiently discover information originally hidden in the dataset.

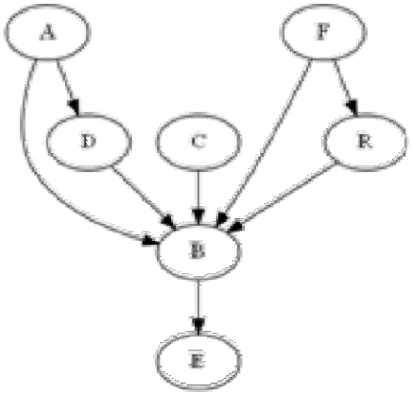

Fig. 1. The input prestige for the node B is 4.

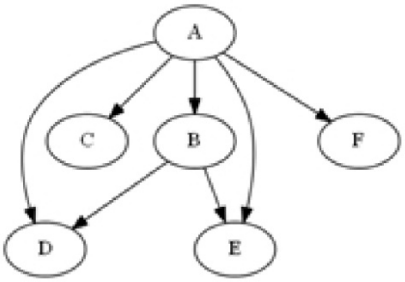

Fig. 2. The output prestige for node A is 5.

2) Bridges: A bridge of a given dataset is a predicate connecting two nodes (the subject node and the object node) which once deleted, there exist no alternative path from the given subject node to the object node in question. Equivalently, the removal of a bridge from a dataset results into the splitting of the latter into two datasets. To identify the cases whereby a given predicate comprises a bridge from those that it does not (because this predicate can appear several times in a dataset), we form a key from the whole triple in which this predicate appears and connects the two nodes that would otherwise be disconnected. More formally,

Definition: *A bridge is a predicate whose removal disconnects a Linked Dataset. (For example, a dataset with the form of a tree is made entirely of bridges). A disconnected Linked Dataset is a set of predicates whose removal increases the number of components.* Figure 3 presents an example of a bridge.

Statements involving predicates that are bridges comprise weak points for the dataset, because the latter becomes disconnected once these statements are removed, resulting in information being lost.

By collecting all the bridges in the dataset one can create a critical path inside the graph and use it in order to evaluate the importance of the result obtained from another procedure or use it to calculate the importance of selected connections.

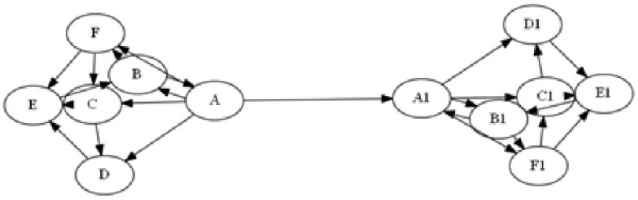

Fig. 3. The connection from node A to node A1 is a bridge in the dataset.

3) Reachability: Mean reachability is defined to be the number of nodes that can be reached in the whole dataset, using a given node as a starting point. More formally,

Definition: *Reachability is a measure that counts the number of nodes that can reached from a specific node.*

Combining these measures, one can obtain useful information about the connections in the dataset, and on the effectiveness of the alternative navigation routes in the corresponding graph. Moreover, by processing the results from the Bridge/Statements percentage method, one establishes a better view on the dataset and its cohesion.

B. Development of TheMa API

The algorithms behind the four measures presented were implemented in the Jena supporting software framework. Jena is a framework for building semantic applications. It has turned out to be very useful in our case because it provides a programmatic environment for RDF, RDFS and OWL, SPARQL (http://jena.apache.org).

1) Prestige: As commented earlier, the two methods are overloaded and they can be used to calculate the in-going/out-going prestige for the whole dataset or to create a new smaller (more specific) dataset by choosing the predicates of interest . The result of both methods is a Map data structure. As the key in the Map we have the name of the node and as a value the count of the prestige of the node.

2) Bridges: We have devised two versions of the method that calculates/identifies the bridges: one that calculates the bridges in the whole dataset and a second that calculates the bridges only between resources. Literals are not included in the results. The two versions are overloaded and they can be used on more specific models by selecting the predicates of interest. However,the latter may result in some information being lost. Because of this, the results obtained may not directly relate to the real life situation considered.

3) Bridge/Statements percentage: The measure reflects the percentage of statements that comprise bridges. It represents useful information in relation to the cohesion of the dataset. Applying the classic graph technique on cohesion in directed graphs is not useful because statement removal implies information. In linked data, the links that interconnect nodes represent information instances. In this

respect, the algorithm that calculates the percentage of statements that act as bridges in the graph comprises a better approach, and a useful statistical measure. A high value of the bridge/statements percentage implies a loosely connected dataset, one that can easily become disconnected. When the percentage is low, most of our nodes interconnected to each other more than once. This in turn implies a strongly connected dataset, one that is hard to split it and (consequently) lose information.

4) Reachability/Mean Reachability: Reachability is a measure that counts the number of nodes that can be reached from a specific node. Mean reachability is the average reachability value across the entire dataset. By combining the two measures one can obtain useful information about the connections in the dataset and on how one can navigate through it. Moreover, if the above are combined/considered in parallel with the aforementioned Bridge/Statements percentage method in parallel, one establishes a clearer view on the dataset and its cohesion.

RESULTS

In order to evaluate our approach, we created a linked dataset on South American countries. The dataset was extracted from DBpedia. The TheMa API was tested against this dataset, giving the results summarized below.

Input prestige:

- The most prestigious node was found to be #Argentina with a value of 2001. That means that the node #Argentina appeared as an object in the dataset's statements for 2001 times, the highest input prestige value across the dataset.

Output prestige:

- The most prestigious node was found to be #Suriname with a value of 182. This meant that the node #Suriname appeared to be the subject in 182 statements, achieving the highest output prestige value across the entire dataset.

Bridges/Statements Percentage

- The dataset consisted of 16038 statements, 12418 of which represented bridges. Consequently, the bridges/statements percentage value was calculated to be 0.774.
- From the 12418 bridges, 11735 were found to be the instances of bridges pointing to resources and not to literal values. Thus, the percentage of statements-bridges that did not point to literals was calculated to be 0.732.

Reachability

- Using #Brazil as the start node, the total number of nodes that could be reached was 8. Thus, the reachability of the #Brazil node was measured to be equal to 8.

Mean reachability

- The mean reachability of the dataset was calculated to be equal to 257.159

CONCLUSION

We report on a new API involving algorithms used for extracting implicit information from Linked Datasets. The API incorporates basic network analysis algorithms applied to graphs representing relations between discrete objects, and includes methods for identifying the most prestigious nodes, bridges, as well as for calculating useful statistics, like reachability and mean reachability. The results indicate that this set of algorithms can be useful in extracting implicit information from datasets in the Web of Data.

In the future stages of our research, we intend to extend the TheMa API in the direction of calculating a richer set of measures. Measure like the number of cycles in the dataset the closure of a node in a given dataset, etc.. The main idea is to not only support basic network theory operations, but also more involved statistics. For example, we wish extend the API to calculate path similarity, as well as to identify/predict missing links.

Lastly, we intend to devise a set of benchmarking linked datasets to be used for comparing the TheMa API to other, analogous, API's that will be proposed by other researchers in the near future.

ACKNOWLEDGMENTS

This work has been funded by the Spanish Ministry of Innovation and Science project ICARIA: *From Semantic Web to Systems Biology*, Project Code: TIN2008-04844, and by the Regional Government of Andalucía Pilot Project for Training and Developing Applied Systems Biology Technology, Project Code: P07-TIC-02978.

REFERENCES

[1] C. Bizer (2009), The Emerging Web of Linked Data, IEEE Intelligent Systems, Vol. 24(5), pp. 87-92.

[2] V. Narasimha Pavan Kappara, R. Ichise, O.P. Vyas (2011), LiDDM: A Data Mining System for Linked Data, http://ceur-ws.org, Vol. 813

[3] S. Auer, J. Lehmann (2010), Creating knowledge out of interlinked data, Semantic Web, Vol. 1, pp. 97-104.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann (2009), DBpedia - A crystallization point for the Web of Data, J. Web Sem., Vol. 7(3), pp. 154-165.

[5] D. Chakrabarti, C. Faloutsos (2006), Graph Mining: Laws, Generators, and Algorithms, ACM Comput. Surv., Vol. 38(1).

[6] L. Getoor (2003), Link mining: A new data mining challenge, SIGKDD Exploration, Vol. 5.

[7] H. White and K. McCain (1989) Bibliometrics. Annual Review of Information Science and Technology, Vol. 24, pp. 119-186.

[8] D. Gibson, Jon M. Kleinberg, and P. Raghavan, Inferring Web Com-munities from Link Topology, Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, pp. 225-234.

[9] J. Hopcroft, O. Khan, and B. Selman (2003), Tracking evolving communities in large linked networks, Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 101, pp. 5249-5253.

[10] W. Baker and R Faulkner (1993) The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. Am. Social. Rev., Vol. 58, pp. 837-860.