Data Patterns Explained with Linked Data

Ilaria Tiddi (✉), Mathieu d'Aquin, and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
{ilaria.tiddi,mathieu.daquin,enrico.motta}@open.ac.uk

Abstract. In this paper we present the system Dedalo, whose aim is to generate explanations for data patterns using background knowledge retrieved from Linked Data. In many real-world scenarios, patterns are generally manually interpreted by the experts that have to use their own background knowledge to explain and refine them, while their workload could be relieved by exploiting the open and machine-readable knowledge existing on the Web nowadays. In the light of this, we devised an automatic system that, given some patterns and some background knowledge extracted from Linked Data, reasons upon those and creates well-structured candidate explanations for their grouping. In our demo, we show how the system provides a step towards automatising the interpretation process in KDD, by presenting scenarios in different domains, data and patterns.

1 Introduction

In Knowledge Discovery in Databases (KDD), patterns are defined as "statements or expressions describing an interesting relationship among a subset of the analysed data, which is typically resulted from a data mining process (classification, cluster, sequence-pattern mining, association rules mining and so on)" [1]. Our work focuses on the KDD step following the data mining one, i.e. the process of pattern interpretation.

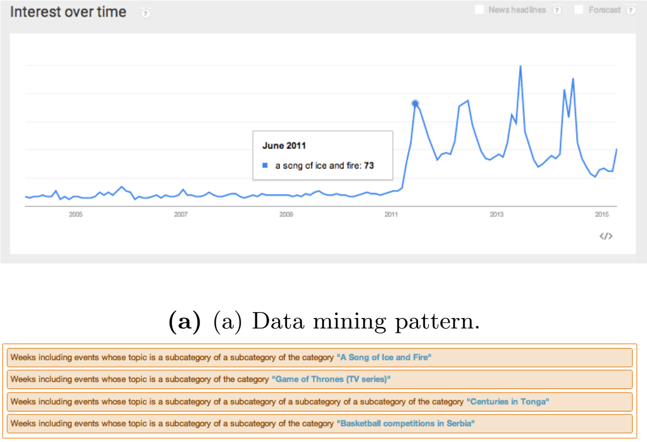
Let us imagine that we aim at explaining why a term such as *A Song of Ice and Fire* is searched over the Web only at specific times of the year: this is shown in Figure 1a, where one can observe regular popularity peaks. Such a pattern can only be explained by someone who, having background knowledge about the fantasy novels, can explain that the popularity increases in those periods in which a new Game of Thrones TV season or a new novel is released. In many other real-world contexts, the revealed patterns are generally provided to experts that analyse, refine and interpret them in order to reuse them for further purposes. For instance, patterns can be used in Business Intelligence for decision making, in E-commerce for item recommendation, in Learning Analytics for assisting people's learning. Producing pattern explanations becomes then an intensive and time-consuming activity, particularly when the background knowledge needs to be gathered from different domains and sources.

With that said, we state that the Web knowledge in the form of Linked Data can facilitate the problem of interpreting Knowledge Discovery patterns. Linked

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 271–275, 2015.
DOI: 10.1007/978-3-319-23461-8_28

Data refer to a set of best practices for publishing and connecting structured data on the Web [2], with the purpose of fostering reuse, linkage and consumption of data. Thanks to their well-established principles (use of HTTP URIs for naming, provision of useful information about data, and inclusion of links to connect to external resources), Linked Data consist nowadays in a globally available knowledge graph, where several datasets are represented in RDF standards, can be accessed and understood by machines and, most importantly, are connected across disciplines. In our example, it is possible to use information about events (e.g., times and topics) to detect that the peaks of popularity correspond to moments where there have been events somehow related to the book series *A Song of Ice and Fire*.

In this demo, we present Dedalo, a tool to generate Linked Data candidate explanations, as in Figure 1b, from data mining patterns such as the one of Figure 1a. We will show how Dedalo can be applied to patterns and scenarios of different nature, thanks to the variety of domains existing within Linked Data. Our work aims at proving that Linked Data can help turning the interpretation process into an automatic process relieving the manual effort of the experts.



(a) (a) Data mining pattern.

Weeks including events whose topic is a subcategory of a subcategory of the category "A Song of Ice and Fire"

Weeks including events whose topic is a subcategory of the category "Game of Thrones (TV series)"

Weeks including events whose topic is a subcategory of a subcategory of a subcategory of a subcategory of the category "Centuries in Tonga"

Weeks including events whose topic is a subcategory of a subcategory of the category "Basketball competitions in Serbia"

(b) (b) Candidate explanations.

Fig. 1. Dedalo Explanation Visualizer. In (a) the pattern *A Song of Ice and Fire* chosen by the user. In (b) Dedalo gives the best candidate explanations based on the Linked Data knowledge.

2 Dedalo's Implementation

We implemented Dedalo as a system that integrates the following modules: an *Explanation Generator*, which produces explanations based on an Inductive Logic Programming (ILP) strategy; a *Background Knowledge Builder*, which builds the background knowledge using an A* strategy to traverse the Linked

Data graph and collect the salient information, and an *Explanation Visualiser* which finally presents the users both the pattern to explain and the generated candidate explanations.

Explanation Generator. This module is designed with the idea that it is possible to learn why some items, considered as the positive examples, belong to a pattern, while some others, the negative examples, do not. Following common Machine Learning approaches, the initial data are therefore organised in positive and negative observations to learn from. More specifically, given a pattern to be explained which is selected by a user, the items belonging to it will be considered as positive examples, while the ones not belonging to it will be the negative examples. In the example of the search term *A Song of Ice and Fire*, each search rate evaluated as a peak is considered as positive example, while the remaining are considered as negative examples. The aim of this module is to derive candidate explanations which cover a maximum number of positive examples and a minimum number of negative examples, e.g., high search rates correspond to events somehow related to the fantasy series. As in Inductive Logic Programming, explanations are derived by reasoning upon the background knowledge about both the positive and negative examples, which is built using statements extracted directly from Linked Data.

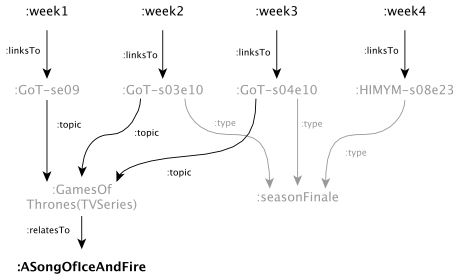


Fig. 2. Dedalo's built graph.

Background Knowledge Builder. This module automatically and iteratively builds the background knowledge from Linked Data. The assumption here is that it is not feasible to import the whole knowledge represented in Linked Data (also, most of the knowledge might indeed not be relevant). On the other hand, it is possible to iteratively extend the background knowledge about the data, with the aim of deriving explanations which represent a bigger portion of positive examples (i.e., the pattern to explain). Starting from the URI representation of the items in the dataset, which in our case consists in weeks of a year in which the term is searched, a graph is iteratively built by following the URIs links and exploring ("dereferencing") the new discovered entities. In this way, no a priori knowledge is introduced in the process: The graph exploration is simply carried out by following existing links between Linked Data URIs. The Linked Data traversal relies on the assumption that data are connected and therefore data sources can be easily and naturally spanned to gather new, unknown knowledge to reason upon. For instance in Figure 2 many of the weeks are linked to aired TV episodes (through the relation *:linksTo*), and some of those are further linked to the Game of Thrones TV series (following the relation *:topic*), which in turn is linked to *A Song of Ice and Fire* through *:relatesTo.*

Explanation Visualiser. The final module consists in presenting to the user the candidate explanations that have been found for the pattern he had initially chosen. Candidate explanations consists in a path of RDF properties and one final entity that a subset of items have in common: In the example of Figure 2, one of the explanations we can derive is shown ase_1=$\langle$:linksTo.:topic.:relatesTo$\rightarrow$:ASo- ngOfIceAndFire$\rangle$. The evaluation of the candidate explanations is shaped as in a classification task, where the classifier prediction is represented by the number of positive examples that the explanation covers, and the external judgment consists in the whole set of positive examples. The closer those two sets are, the better the explanation represents the pattern to explain, and the better it is evaluated. The best explanations are then visualised and presented to the user in natural language, as in Figure 1b, where we show the candidate explanations generated for the search term *A Song of Ice and Fire.*

3 Demo Scenarios

During the demonstration we will present to the audience scenarios of different nature, following use-cases of our previous works [3]. Users will be allowed to choose a pattern, that will be visualised in the way it is provided to Dedalo, and will be also shown the candidate explanations that Dedalo has derived with the information from Linked Data. We present them below.

(a) KMiData - Clusters of researchers grouped according to their co-authorship, for which Dedalo explains the reasons for which the authors are working together;
(b) WorldMaps - Worldbank[1] maps of countries grouped according to different economic indicators, for which Dedalo finds socio-economical reasons explaining the countries similarity;
(c) Education - Clusters of words grouped according to their semantic similarity, for which Dedalo is able to find which topic relates them;
(d) Trends - Trends of topics searched over the last 10 years, for which Dedalo explains the peaks of popularity.

We will encourage the audience in understanding the efforts required to give an explanation for a pattern and will show the benefits of using Linked Data to assist the explanation process. By using different scenarios we intend to make a step forward in the automatisation of the pattern interpretation process.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine **17**(3), 37 (1996)

[1] http://www.worldbank.org/

2. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology 1(1), 1–136, (2011). Semantic Web: Research and Applications (pp. 560–574). Springer, Berlin Heidelberg
3. Tiddi, I., d'Aquin, M., Motta, E.: Dedalo: Looking for clusters explanations in a labyrinth of linked data. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 333–348. Springer, Heidelberg (2014)