Explaining data patterns using background knowledge from Linked Data

Ilaria Tiddi

Knowledge Media Institute, The Open University, United Kingdom
ilaria.tiddi@open.ac.uk

Abstract. When using data mining to find regularities in data, the obtained results (or patterns) need to be interpreted. The explanation of such patterns is achieved using the background knowledge which might be scattered among different sources. This intensive process is usually committed to the experts in the domain. With the rise of Linked Data and the increasing number of connected datasets, we assume that the access to this knowledge can be easier, faster and more automated. This PhD research aims to demonstrate whether Linked Data can be used to provide the background knowledge for pattern interpretation and how.

Keywords: Linked Data, Data Mining, Knowledge Discovery, Data Interpretation

1 Problem Statement

Knowledge Discovery in Databases (KDD) can be defined as the process of detecting hidden patterns and regularities in large amounts of data [?]. To be interpreted and understood, these patterns require the use of some background knowledge, which is not always straightforward to find. In most real world contexts, providing the background knowledge is committed to the experts, whose work is to analyse the results of a data mining process, give them a meaning and refine them. The interpretation turns out to be an intensive and time-consuming process, where part of knowledge can remain unrevealed or unexplained.

Our problem is illustrated with a real-world example we will use throughout this paper. The *Reading Experience Database* (RED)[1] is a record of people's reading experiences, including metadata regarding the reader, author, and book involved in the experience as well as its date and location. Several kinds of data mining processes could be applied on such a dataset. Here, for example, we look at how people can be clustered based on the similarity of what they read. Considering one such cluster, the question then becomes: "is there a reason why these people read the same kind of books?" and "where and how to find this information?". Given for instance a cluster of people having extensively read Jane Austen, an expert might consider it pertinent to point out that many of them are Anglican women, since, for a number of reasons, Jane Austen was more significantly popular with this particular audience.

Our hypothesis is that the access to the required background knowledge (in our example, that readers are Anglican and female) can be made easier with

[1] http://www.open.ac.uk/Arts/RED/index.html

Linked Data[2]. In fact, while in the documentary Web the information used to be hard to detect, hidden or even unreachable, the rise of Linked Data has made possible to directly access it. In the last decades, people have been putting efforts together in order to openly publish and link their knowledge in the form of domain-specific concepts and relationships. While Tim Berner's Lee's "Web of Data" [**?**] is still evolving and taking form, this structure and interoperability of data can already be exploited for the knowledge interpretation process.

2 Relevancy

In many real-world domains, background knowledge plays a central role for the analysis of trends or common behaviours. Generally, this knowledge is provided by experts interpreting the results and assisting the Knowledge Discovery process, which proves to be intensive and time-consuming.

In **Business Intelligence** (BI), the regularities emerging from raw data using data analytics are explained and transformed into meaningful information by an expert for business purposes, such as decision making or predictive analytics.

The young field of **Learning Analytics** aims at identifying trends and patterns from educational data using data mining, BI and Human-Computer Interaction techniques. The explanation of behaviours is crucial to assist people's learning, help teachers to support students, improve courses, as well as support the staff in planning and taking decisions.

In **Medical Informatics**, computer technologies are applied to process medical information. The explanation of trends and anomalies might come from some external knowledge, which the expert might not be aware of. A typical example is the environmental changes affecting the spread of diseases.

The analysis of data is also central in the field of **Humanities**, where researchers attempt to explain facts by finding hidden connections with some external sources. The RED example of the paper comes from this field.

These examples show on the one hand how background knowledge is required to explain the regularities in data, and on the other hand how this explanation can sometimes come from very different domains, not related to each other.

3 Related Work

While ontologies have been widely explored in the data mining context since the early 2000s, the last years have seen an increasing number of researches aiming at exploiting the potential of Linked Data. The overall idea behind the two trends is to exploit the datasets' structure and semantics and combine them with the Machine Learning algorithms to produce more accurate results. Earlier works proposed the use of ontologies as a support for data preparation [**?**,**?**,**?**] or to constrain the algorithms search [**?**,**?**,**?**,**?**]. Linked Data-driven approaches can be found in [**?**,**?**,**?**]. On the other hand, few works [**?**,**?**] had been addressed so far on using ontologies to assist the interpretation of the results. Recently, the idea has been considered in [**?**,**?**], where the authors stress the importance of

[2] http://linkeddata.org/

capturing useful knowledge from ontologies to reduce the user's workload for the interpretation process. While the motivation of both these works and ours is to combine ontologies and data mining in view of a complete virtuous KDD process, we also intend to further the use of Linked Data. This idea can be found in [**?**], where Linked Data are used to understand the results of a Sequence Pattern Mining process in the context of Learning Analytics. Linked Data are here only a navigation support to the user (who can easily visualise the results), while the interpretation is still based on his previously acquired expertise.

Ontologies for hypothesis generation have been treated in the clinical domain (see survey in [**?**]), and combined with Logic Programming in the fields of Description Logic Programming [**?**,**?**], as well as in the Onto-Relational Learning domain [**?**]. Particularly, this last approach exploits the unary and binary predicates of ontologies, to provide a strong background knowledge and combine it with Inductive Logic Programming in order to produce rules or hypotheses from observations.

4 Research Questions

The main research question we address is this work is: "*how do we explain patterns in data using the background knowledge from Linked Data?*". If, on one side, this "explanation" means the generation of some *hypotheses* (or rules) interpreting the data patterns, on the other side, these hypotheses should rely on some background knowledge that needs to be somehow retrieved, and we assume that at least some of it might be available through Linked Data. To answer this, we articulated our space in a specific set of subquestions, possible solutions and expected risks, which are illustrated below.

Q1 – Finding the data. Our first question is *how to find* the right background knowledge in Linked Data. This is our major question, and is articulated in:

1. *Dataset selection.* Does the Linked Data cloud contains the right datasets describing our data? Where and how to find them?
2. *Data detection.* Once we have found the datasets, how to detect the correct data into them? Do the data have enough information? In other words, how do we find the *correct pieces of knowledge*, in terms of predicates about our data?

Initial Solution. The question here concerns the exploration of the Linked Data cloud and the knowledge herein represented. While technical solutions such as the CKAN API[3], the Semantic Web indexers[4] or the SPARQL endpoint lists[5] are already popular in the community, our objective is to automatise the process of selecting the important bits of information required for the explanation. Whether we choose a top-down approach, where the search space is first defined by deeply analysing the datasets and then narrowed using the initial data to detect the salient bits of information for hypotheses generation, or a bottom-up

[3] http://docs.ckan.org/en/latest/api.html

[4] such as Sindice: http://sindice.com/

[5] http://www.w3.org/wiki/SparqlEndpoints

approach, that exploits the initial data to iteratively add pieces of (Linked Data) background knowledge to produce more and more refined hypotheses, the key of the process are the available connections in the Linked Data cloud. Exploiting such connections to make emerge underlying knowledge in order to maximise the automatisation of this selection process will be our major contribution.
Expected risks. The search for background knowledge could be unsuccessful as the patterns might not be described enough in Linked Data (*lack of information* or *lack of datasets*).

Q2 – Generating the hypotheses. Assuming that the background knowledge about the data has been found, we will have to answer the question: *how do we use it to explain* the data patterns. What kind of mechanisms can generate explanations, that we previously called hypotheses?
Initial Solution. We identified as a possible solution the use of Inductive Logic Programming to produce hypotheses from both data patterns and Linked Data background knowledge.
Expected risks. The chosen mechanism to generate explanations might not be scalable and might lead to computational problems (*data deluge*).

Q3 – Evaluating the hypotheses. Once the hypotheses have been generated, the last questions is: how do we know that they are *good*? That is, what is the significance of a rule? This evaluation step is also two-folded:

1. *Hypotheses evaluation.* Which are the criteria to assess the interestingness of a hypothesis?
2. *Method evaluation.* How do we evaluate that our method is efficient when compared to those of the domain(s) experts?

Finally, can the evaluation method affect the data selection? Can a hypothesis help in pruning the selected data, and support the Knowledge Discovery process?
Initial Solution. Currently, we are exploiting the ILP evaluation measures to score the significance of a hypothesis. However, we are aware that this preliminary solution will need to be further investigated. We also intend to investigate genetic algorithms to verify if the evaluation method can affect the data selection.
Expected risks. A clear evidence for some of the generated rules could be missing (*lack of background knowledge*). Moreover, some of the hypotheses might iteratively require a new piece of knowledge to explain the patterns (*recursion issue*).

5 Hypothesis

Our hypothesis is: "*Linked Data can be used as background knowledge to explain data patterns*". The main idea is that using Linked Data as background knowledge will reduce the efforts put into explaining the data patterns. Assuming this, Knowledge Discovery can leverage Linked Data as they will assist the experts and reduce their commitment into the KDD process, as explained below.

Time gaining. The expert will require **less time** to explain patterns. The connections between datasets of different areas will make emerge new information for the explanations requiring external knowledge.
Efficiency. Linked Data will show the expert the information which is not from his domain. Our method can be **more efficient** than a group of experts.
Completeness. The expert can be less specialised as Linked Data can bring the missing information, in order to have a **more complete** explanation.

6 Approach

The approach is structured according to our research questions (see also Fig. **??**).
1. *Data Selection.* Assuming some patterns obtained from a data mining process (clusters, association rules, sequence patterns...), we search in the Linked Data cloud information about the data in the patterns.
2. *Hypotheses Generation.* We use Inductive Logic Programming to represent both the data patterns and the Linked Data information, and generate hypothesis from them.
3. *Hypotheses Evaluation.* We evaluate the hypotheses in order to rank them and select the best rules. These are presented to the experts for interpretation, but also used to refine the data selection of the first step and to start a new cycle.

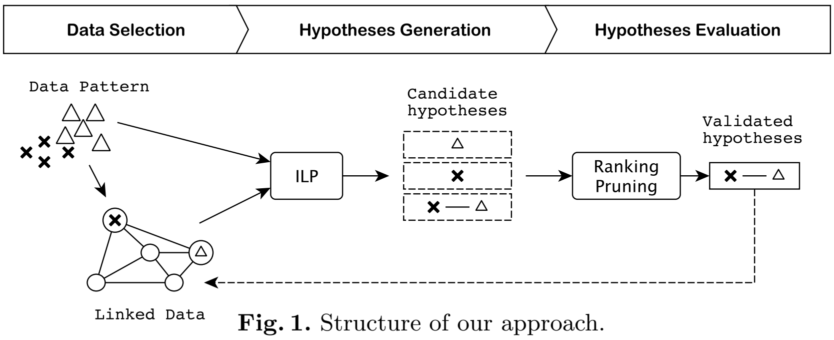

Fig. 1. Structure of our approach.

Data Selection. We introduced in the first section the RED example that we use to illustrate our approach. Once we obtain clusters of readers, we proceed with the search for information about them in Linked Data. For the purpose of a preliminary study, we started with the manual selection of some properties from DBpedia[6].

Hypotheses Generation. The step concerns the problem formulation in the ILP framework. Inductive Logic Programming is a research field at the intersection of Machine Learning and Logic Programming, investigating the inductive

Table 1. Prolog-encoded examples. Gordon Byron and Samuel Coleridge are examples of readers belonging to the same cluster `c`.

clusters	`c('Gordon Byron'). c('Samuel Coleridge').`
RDF predicates	`originCountry('Gordon Byron','England').`
RDF is-a relations	`country('England').`

[6] http://dbpedia.org/About

construction of first-order clausal theories (Logic Programming heritage) starting from a set of examples (Machine Learning heritage) [?]. Its distinguished feature is the use of some additional background knowledge to derive the hypotheses. In such framework, the data patterns represent the negative and positive examples, while information from Linked Data is the background knowledge required to generate hypotheses. Therefore, we encode them into Prolog clauses, as follows: The hypotheses are generated using the Aleph[7] system, and take the form of:

```
[Pos cover=14, Neg cover=308] c(A):-female(A)∧originCountry(A,‘England’)
```

which is interpreted as: “the reader `A` is part of the cluster `c` because of being `female` and from `England`”. `Pos cover` is the number of examples e^+ covered by the rule r included in the cluster c ($e^+ \in c$), while `Neg cover` is the number of examples e^- covered by r, where $e^- \notin c$.

Hypotheses Evaluation. In this preliminary study, the hypotheses evaluation is performed using the weighted relative accuracy function (WR_{acc}) provided by Aleph and described in [?]. WR_{acc} measures the unusualness of a rule and expresses it in terms of number of positive and negative examples covered. By providing a trade off between of a rule's coverage and relative accuracy, WR_{acc} allows us to obtain explanations which are valid for patterns of small sizes. Given a rule r and a cluster c, WR_{acc} is defined as:

$$WR_{acc} = \frac{e_r^+ + e_r^-}{\mathcal{E}_c^+ + \mathcal{E}_c^-}\left(\frac{e_r^+}{e_r^+ + e_r^-} - \frac{\mathcal{E}_c^+}{\mathcal{E}_c^+ + \mathcal{E}_c^-}\right) \tag{1}$$

where e_r^+ and e_r^- the number of positive and negative examples covered by r, $\mathcal{E}^+$ the size of c and $\mathcal{E}^-$ the number of examples provided outside c. Using this formula, we obtained a preliminary ranking of the generated hypotheses. Examples of rules with the best scores are presented in Table **??**.

Table 2. Examples of generated hypothesis with their WR_{acc} score.

cluster	size	hypothesis	WR_{acc}
Austen J.	110	`c(A):- religion(A,‘Anglican’)`	0.025
		`c(A):- female(A)`	0.02

Pepys S.	13	`c(A):- religion(A,‘Anglican’)∧male(A)`
		`∧country(A,‘England’)`
			0.025

7 Reflections

The previous table presents some promising results for the hypotheses evaluation, ranking and selection. The results for the first cluster are fairly strong when compared to the sample set ($\mathcal{E}^+ \cup \mathcal{E}^-$=1230), and show how ILP is a good approach to explain data patterns, e.g. “*people reading Jane Austen were Anglican women*”. This initial test also confirms our intuition that the proposed approach could naturally combine different sources of background knowledge (i.e., different datasets) to produce explanations of found patterns in the data. Here for example, information about the gender of readers come from the RED data, while the information about their religion is present in DBpedia. However,

[7] http://www.cs.ox.ac.uk/activities/machlearn/Aleph/

as expected, triggering a new background knowledge search process is required to make the explication more understandable. In practice, we might require a more specific answer to the question "what connects readers of Jane Austen?" than that they are Anglican women. We are also aware that finding a more adequate scoring measure to check the validity of a hypothesis is necessary. The WR_{acc} might be a good starting point but we will have to find an evaluation measure which takes into account aspects such as the lack of information or a smaller cluster size. This will, in fact, have a direct impact on the data selection. Finally, in order to detect what strongly connects the data in a pattern, we need to find a good way to detect valid background knowledge. Most of this PhD work will be focused on this issue.

8 Evaluation plan

(1) Hypothesis validity evaluation. We aim at finding the good rules using background knowledge from Linked Data. For instance, is "people reading Jane Austen were Anglican women" good, or good enough? Depending on the use-case we will be working on, a manual evaluation of the rules will be asked to the relevant domains experts.

(2) Experts support. How much our approach reduces the efforts needed from an expert? Does the explanation about the readers of Jane Austen bring any new knowledge to the expert, that he can exploit for the interpretation process? We will compare the results of a full KDD process achieved with and without our method to see whether the later can effectively reduce the expert's involvement.

9 Conclusions

This paper presents our research aiming at using background knowledge found in the Linked Data to explain patterns and regularities in data. The main idea is to explore if and how Linked Data can assist the experts in the knowledge discovery process. The first results of our ILP-based approach are promising and revealed that the Hypotheses Generation and Evaluation steps can be improved. We identified as one of the major issues the need of a **full access** to both data and the background knowledge. This information has to be (a) **expressive** (enough properties related to the data), **consistent** (no ambiguity or contradictory facts) and **complete** (properties need to cover most of the data). The future work will investigate the Data Selection step, the core part of our project. This PhD contribution will be to set up a good method to detect relevant information in Linked Data, where "detection" concerns both the right datasets and the represented data.

References

1. Antunes, C. (2008, October). An ontology-based framework for mining patterns in the presence of background knowledge. In *Int'l Conf. on Advanced Intelligence*, Beijing, China (pp. 163-168).
2. Brisson, L., Collard, M., & Pasquier, N. (2005, November). Improving the knowledge discovery process using ontologies. In Proceedings of the IEEE MCD international workshop on Mining Complex Data (pp. 25-32).

3. D'Aquin, M., Kronberger, G., & Suárez-Figueroa, M. (2012). Combining data mining and ontology engineering to enrich ontologies and linked data. In *Workshop: Knowledge Discovery and Data Mining Meets Linked Open Data-Know@ LOD* at Extended Semantic Web Conference, ESWC.
4. D'Aquin, M., & Jay, N. (2013). Interpreting Data Mining Results with Linked Data for Learning Analytics: Motivation, Case Study and Directions. In *Third Conference in Learning Analytics and Knowledge (LAK)*, Leuven, Belgium.
5. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
6. Grosof, B. N., Horrocks, I., Volz, R., & Decker, S. (2003, May). Description logic programs: Combining logic programs with description logic. In *Proceedings of the 12th international conference on World Wide Web* (pp. 48-57). ACM.
7. Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), 1-136.
8. Hartmann, J., & Sure, Y. (2004, July). A knowledge discovery workbench for the Semantic Web. In International Workshop on Mining for and from the Semantic Web (p. 56).
9. Lisi, F. A. (2010). Inductive Logic Programming in Databases: From Datalog to DL+log. *Theory and Practice of Logic Programming*, 10(03), 331-359.
10. Liu, J., Wang, W., & Yang, J. (2004, August). A framework for ontology-driven subspace clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 623-628). ACM.
11. Marinica, C., & Guillet, F. (2010). Knowledge-based interactive postmining of association rules using ontologies. *Knowledge and Data Engineering*, IEEE Transactions on, 22(6), 784-797.
12. Moss, L., Sleeman, D., Sim, M., Booth, M., Daniel, M., Donaldson, L., & Kinsella, J. (2010). Ontology-driven hypothesis generation to explain anomalous patient responses to treatment. *Knowledge-Based Systems*, 23(4), 309-315.
13. Motik, B., & Rosati, R. (2006). Closing semantic web ontologies. Technical report, University of Manchester, UK.
14. Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19, 629-679.
15. Narasimha, V., Kappara, P., Ichise, R., & Vyas, O. P. (2011). LiDDM: A Data Mining System for Linked Data.
16. Novak, P. K., Vavpetic, A., Trajkovski, I., & Lavrac, N. (2009). Towards semantic data mining with g-segs. In Proceedings of the 11th International Multiconference Information Society, IS.
17. Paulheim, H., & Fümkranz, J. (2012, June). Unsupervised generation of data mining features from linked open data. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (p. 31). ACM.
18. Pan, D., Shen, J. Y., & Zhou, M. X. (2006). Incorporating domain knowledge into data mining process: An ontology based framework. *Wuhan University Journal of Natural Sciences*, 11(1), 165-169.
19. Phillips, J., & Buchanan, B. G. (2001, October). Ontology-guided knowledge discovery in databases. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 123-130). ACM.
20. Verborgh, R., Van Deursen, D., Mannens, E., & Van de Walle, R. (2010). Enabling advanced context-based multimedia interpretation using linked data.