# Explaining Clusters with Inductive Logic Programming and Linked Data

Ilaria Tiddi, Mathieu d'Aquin, Enrico Motta

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

**Abstract.** Knowledge Discovery consists in discovering hidden regularities in large amounts of data using data mining techniques. The obtained patterns require an interpretation that is usually achieved using some background knowledge given by experts from several domains. On the other hand, the rise of Linked Data has increased the number of connected cross-disciplinary knowledge, in the form of RDF datasets, classes and relationships. Here we show how Linked Data can be used in an Inductive Logic Programming process, where they provide background knowledge for finding hypotheses regarding the unrevealed connections between items of a cluster. By using an example with clusters of books, we show how different Linked Data sources can be used to automatically generate rules giving an underlying explanation to such clusters.

## 1 Introduction

Knowledge Discovery in Databases (KDD) is the process of detecting hidden patterns in large amounts of data [2]. In many real-world contexts, the explanation of such patterns is provided by experts, whose work is to analyse, visualise and interpret the results obtained out of a data mining process in order to reuse them. For instance, in Business Intelligence, the analyst uses such interpretation for decision making; in Learning Analytics, the detected patterns are used to assists people's learning; in Medical Informatics, trends can be useful for anomalies detection. This production of explanation becomes an intensive and time-consuming process, particularly when the background knowledge needs to be gathered from different domains and sources.

In a practical example, the university of Huddersfield provides books recommendations within its library catalogues[1], where records of books transactions over a decade can be used for stock management and students recommendation systems. Here, we are interested in explaining why groups of books, obtained from a clustering process, have been borrowed by the same students. Considering one such cluster, the question is: "why these books have been borrowed by those particular students?" and "where and how to find this information?".

Our hypothesis is that this answer can be given with Linked Data[2], which provide the required background knowledge (in our example, a trivial explanation for a pattern can be that authors of the books borrowed by the students enrolled in English Literature are from England). While works into data preparation and data mining using Linked Data have already been presented (see

[1] http://library.hud.ac.uk/data/usagedata/_readme.html
[2] http://linkeddata.org/

the ones of [4, 6, 8]), few works have considered Linked Data for results interpretation (some preliminary attempts are to be found in [1, 7]). However, the former uses Linked Data only to support the user's navigation, and the latter does not take into account the whole knowledge discovery process and focuses on the interpretation of statistical data. For this reason, we aim to exploit the interconnected knowledge from Linked Data to explain patterns resulting from a clustering process, by combining the existing semantic technologies with a Machine Learning technique, i.e. *Inductive Logic Programming* [3], to automatically produce underlying explanations for the formation of such patterns.

## 2 Approach

### 2.1 On Inductive Logic Programming

Inductive Logic Programming (ILP) is a research field at the intersection of Machine Learning and Logic Programming, investigating the inductive construction of first-order clausal theories starting from a set of examples $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$ [3]. While $\mathcal{E}^+$ represents the relation to be learnt, $\mathcal{E}^-$ are the facts where the relation does not hold. The distinguished feature of ILP is the use of some additional background knowledge $\mathcal{B}$ about the examples in $\mathcal{E}$. Believing $\mathcal{B}$, and faced with the facts in $\mathcal{E}$, the induction process derives an hypotheses space $\mathcal{H}$. The success of the induction requires that $\mathcal{H}$ covers all the positive examples ($\mathcal{H}$ is *complete*) and none of the negative ones ($\mathcal{H}$ is *consistent*), with respect to $\mathcal{B}$ (i.e., there is no contradiction with the facts written in $\mathcal{B}$).

### 2.2 Proposed approach

Assuming that we have retrieved some clusters, our approach is articulated as follows (see Fig. 1):

**1. Linked Data Selection**. We retrieve information about the data contained in each cluster from the Linked Data cloud, across several datasets.

**2. Hypotheses Generation**. We generate some hypotheses using ILP. A hypothesis is an explanation ("why those items are part of that particular cluster").

**3. Hypotheses Evaluation**. We validate the hypotheses using two rules evaluation measures: the *Weighted Relative Accuracy* ($WR_{acc}$, as described in [5]), providing a trade-off between coverage and relative accuracy, that we exploit to obtain explanations for small clusters, and the very well known and Information Retrieval F-measure ($\mathcal{F}$).

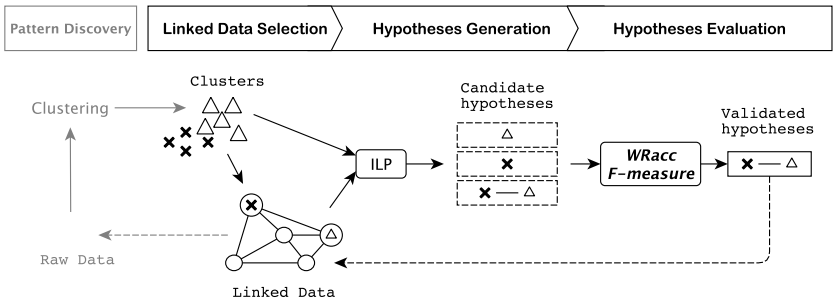


**Fig. 1.** Structure of the ILP approach for clusters explanation.

## 3 Experiments

We ran our experiments on the Huddersfield's books usage dataset introduced in the first section. Our target problem is defined as: *considering some clustered books borrowed by students from the Humanities faculty, explain what those books have in common and why they belong to a particular cluster.* The manual analysis of each cluster's centroid shows that each cluster represents the books borrowed by students from the same course, such as *Music Technologies, Politics* or *English Literature.*

For each book, we retrieve some information from the Linked Data cloud. We first use `bibo:isbn10` as an equivalence property to navigate from the Huddersfield dataset to the British National Bibliography one[3]. From there, we retrieve information about the book using the existing Linked Data vocabularies: Dublin Core[4] for topic and author, the Event Ontology[5] for the publication time, place and publisher. Finally, we exploit the `owl:sameAs` property to navigate to the Library of Congress Subject headings[6] and retrieve the broader concepts of each topic using the `skos:broader` property.

Clusters and the Linked Data extracted knowledge are encoded as Prolog clauses as follows:

| | | |
|---|---|---|
| $\mathcal{E}^+$ | clusters | $\texttt{cl}_{MT}$(‘book_1’). |
| $\mathcal{E}^-$ | | $\texttt{cl}_{MT}$(‘book_4’). $\texttt{cl}_{MT}$(‘book_5’). |
| $\mathcal{B}$ | RDF predicates | subject(‘book_1’,‘electronic music’). |
| | RDF is-a relations | book(‘book_1’). topic(‘electronic music’). |

Here we search the hypothesis space $\mathcal{H}$ specific to the *Music Technologies* cluster ($\texttt{cl}_{MT}$). $\mathcal{E}^+$ is composed by books in $\texttt{cl}_{MT}$ (as `book_1`), while books in other clusters (such as `book_4` and `book_5`) form $\mathcal{E}^-$. The process is repeated for each cluster. Both the RDF binary relations (`hud:book_1 dc:subject ‘electronic music’`) and the unary ones (`hud:book_1 a bibo:book`) are also transformed into Prolog clauses and then added to $\mathcal{B}$.

We ran several experiments combining different properties (in different $\mathcal{B}$s), in order to see the properties impact on the hypotheses generation. These are shown in Table 1. Other hypotheses demonstrated the relations between different predicates, such as the relation between a publisher and a specific topic (see Table 2).

## 4 Conclusion and future work

We showed how ILP can be good in generating hypotheses to explain patterns, e.g. "*books borrowed by students of Music Technologies are clustered together because they talk about music*". Although it is a trivial example, the automation of such a process is not an easy task. We demonstrated how the use of Linked Data is important to generate such hypotheses, and how combining different sources

[3] http://bnb.data.bl.uk/
[4] http://dublincore.org/documents/dcmi-terms/
[5] http://motools.sourceforge.net/event/event.html
[6] http://id.loc.gov/authorities/subjects.html

**Table 1.** Expanding $\mathcal{B}_1$ with the LCSH knowledge ($\mathcal{B}_2$) improves the hypotheses. Those are read as follows: the item A belongs to the cluster `cl` because it has some properties, which appear in the body ("A's topic is mass media" or "A's broader topic is publicity").

| Centroid | $\mathcal{B}$ | Hypothesis | $\mathcal{F}(\%)$ | $WR_{acc}$ |
|---|---|---|---|---|
| Media& Journalism | $\mathcal{B}_1$ | `cl(A):-subject(A,'mass media,social aspects')` | 10.8 | 0.004 |
| | $\mathcal{B}_2$ | `cl(A):-broader(A,'publicity')` | 16.4 | 0.007 |
| Humanities | $\mathcal{B}_1$ | `cl(A):-subject(A,'criminology')` | 11.3 | 0.003 |
| | $\mathcal{B}_2$ | `cl(A):-broader(A,'social sciences')∧broader(A,'auxiliary sciences')` | 15.5 | 0.003 |
| Music Technologies | $\mathcal{B}_1$ | `cl(A):-subject(A,'sound, recording and reproducing')` | 10.6 | 0.003 |
| | $\mathcal{B}_2$ | `cl(A):-broader(A,'digital electronics')` | 18.8 | 0.006 |
| | $\mathcal{B}_1$ | `cl(A):-subject(A,'popular music, history and criticism')` | 14.5 | 0.005 |
| | $\mathcal{B}_2$ | `cl(A):-broader(A,'music')` | 21.2 | 0.008 |
| English& Media | $\mathcal{B}_1$ | `cl(A):-subject(A,'language acquisition')` | 11.6 | 0.005 |
| | $\mathcal{B}_2$ | `cl(A):-broader(A,'child development')∧broader(A,'philology')` | 13.7 | 0.006 |

**Table 2.** Hypotheses revealing hidden connections between properties.

| Centroid | Hypothesis | $\mathcal{F}(\%)$ | $WR_{acc}$ |
|---|---|---|---|
| Media | `cl(A):-broader(A,'psychology')∧pubPlace(A,'oxford')` | 10.3 | 0.004 |
| English Literature | `cl(A):-publisher(A,'routledge')∧broader(A,'literature')∧broader(A,'philology')` | 11.1 | 0.003 |
| Politics | `cl(A):-publisher(A,'macmillan')∧broader(A,'political science')∧broader(A,'social sciences')` | 4.3 | 0.001 |

of background knowledge (i.e., different datasets) produces better explanations of patterns of data. The future work concerns the automatic selection of the datasets from Linked Data, the use of a more appropriate evaluation measure and the generalisation of the approach to other data mining techniques.

## References

1. 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.
2. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
3. Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19, 629-679.
4. Narasimha, V., Kappara, P., Ichise, R., & Vyas, O. P. (2011). LiDDM: A Data Mining System for Linked Data.
5. Lavrač N., Flach P., and Zupan B. (1999). Rule Evaluation Measures: A Unifying View. In Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP '99). Springer-Verlag, London, UK, 174-185. .
6. 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.
7. Paulheim, H. (2012). Generating possible interpretations for statistics from Linked Open Data. In The Semantic Web: Research and Applications. Springer, pp. 560574.
8. Verborgh, R., Van Deursen, D., Mannens, E., & Van de Walle, R. (2010). Enabling advanced context-based multimedia interpretation using linked data.