A Hyperknowledge Approach to Support Dataset Engineering

Marcio Moreno, Polyana Bezerra, Rodrigo Costa, Vítor Nascimento, Elton Soares, and Marcelo Machado

IBM Research, Brazil, Av Pasteur 146 Rio de Janeiro - RJ, Brazil
mmoreno@br.ibm.com, {polyana.bezerra, rodrigo.costa, vitor.nascimento, eltons, marcelo.machado}@ibm.com

Abstract. The use of machine learning has become a common approach for solving complex problems across multiple application domains. As its usage often requires training and validation of models with large and heterogeneous datasets, the engineering of these datasets becomes a critical task, although in many cases it does not follow any well-defined process. In this demonstration paper, we present a novel approach to dataset engineering, which comprises the construction, structuring, understanding, and reuse of datasets from a semantic perspective. Our approach uses a hybrid conceptual model called Hyperknowledge, which can semantically describe both symbolic and non-symbolic nodes, including representing the datasets' structure and enabling dataset retrieval/creation queries.

Keywords: Hyperknowledge; Hybrid Knowledge Representation; Hyperlinked Knowledge Graph; HyQL; Multimodal data.

1 Introduction

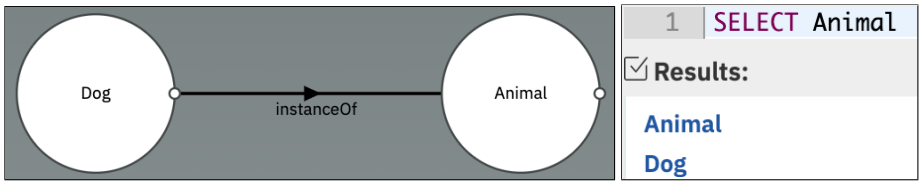
As the popularity of Machine Learning (ML) tasks increases, so does the amount of data used to train and test them. The effectiveness of such algorithms is related to the quality and variety of the data applied during the training stage [2]. The continuous growth in size and heterogeneity of datasets used in ML tasks makes what we call *Dataset Engineering* (DE) a key step for effective data exploitation, leading to more effective models. Here we define DE as the process of handling data through structuring and traceability. The process of DE can be roughly divided into three main tasks: (i) *representation*; (ii) *retrieval*; and (iii) *creation*. The motivation behind these tasks is to prepare data for further use in an ML task. Few works have approached some of the aforementioned tasks [1, 2, 4]. However, to the best of our knowledge, none of them deals with the lifecycle of data, and few of them tackle the structuring of datasets in the ML context. Hence, we propose a knowledge-oriented approach for DE in the ML domain

and describe how this approach can support DE tasks in this domain, as well as its advantages. For the sake of this discussion, we rely on Hyperknowledge [5] (HKW) for data structuring and HyQL (Hyperknowledge Query Language) for both the retrieval and creation of datasets.

HKW is a conceptual model that can represent, in the same description framework, high-level semantic concepts and unstructured data. By semantic concepts, we mean high-level description regarding linked data, such as facts about subjects, or specialized knowledge from a given domain, formally described by ontologies [5]. By unstructured data, we mean raw data, such as multimedia content (image, audio, text, video) or executable content (ML models, programs, etc). Traditionally, those two modalities of information are represented separately, and engineers somehow create links to combine them in a system when needed. HKW fills this representational gap by providing a higher-level framework while also promoting reasoning over cross-modal types of information [5].

2 Dataset Engineering with Hyperknowledge

To support the understanding of the examples illustrated in the remainder of this paper, we present a simplified formalization of HyQL grammar [2] and its processing engine. First, it is important to understand the basic behavior of HyQL when selecting HKW entities using the SELECT clause. A basic selection returns the concept passed in the clause and any other concept that is linked to this concept by a link that contains a connector of type *hierarchical* [3]. Figure 1(a) illustrates the *instanceOf* (i.e., the connector is declared as hierarchical) relationship between Dog and Animal. Figure 1(b) depicts the result of a query to select animals and returns the *Animal* node and also the *Dog* node.



(a) (b)

Fig. 1. Representation of the relationship between concepts Animal and Dog (a) and a query to select all animals (including Animal) from the base (b).

Dataset Representation: The dataset representation task stands for structuring data and its metadata in a common knowledge framework (*e.g.*, ontology). It promotes the enhancement of dataset querying and reuses capabilities, which are important properties for supporting the other DE tasks. As an illustrative example, consider the dataset ontology depicted in Figure 2. It shows the concept

[2] Accessible at https://ibm.box.com/v/iswc2021-hyql-grammar. For clarity, we have suppressed the handling of spaces, comments, and lower case keywords.

[3] HKW connectors have types. The *hierarchical* type can be used to represent taxonomical relationships, such as instantiations and specializations of concepts.

dataset modelled within its data and metadata concepts, i.e., *datatype* and *class* (red arrows) and possible instances of each represented concept (blue arrows).

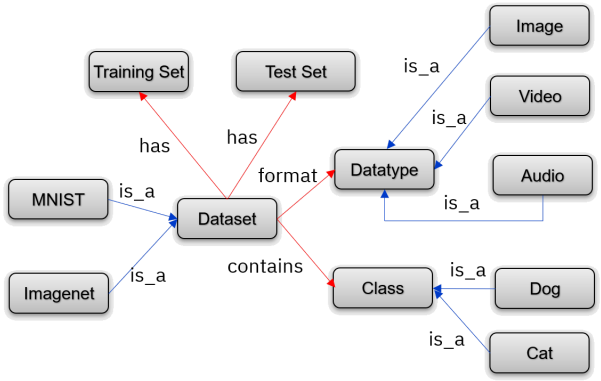


Fig. 2. A simple ontology for dataset description.

HKW can be used for describing this ontology. Nodes can represent the ontology's classes and instances (round rectangles in the figure) and links can represent the relationships among these entities (arrows in the figure). The description of each dataset could be organized in contexts, promoting a better organization of the data. Likewise, Figure 3 illustrates an example of how HKW can be used for structuring the information of datasets. In this example, the *Cleansed Dataset A* is described as a context having other three nested contexts: *Training*, *Validation*, and *Testing*. In each of these contexts, there are descriptions of their content using nodes and links. For instance, it describes that the training dataset contains an image (*Image1*) that has a cat and a dog.

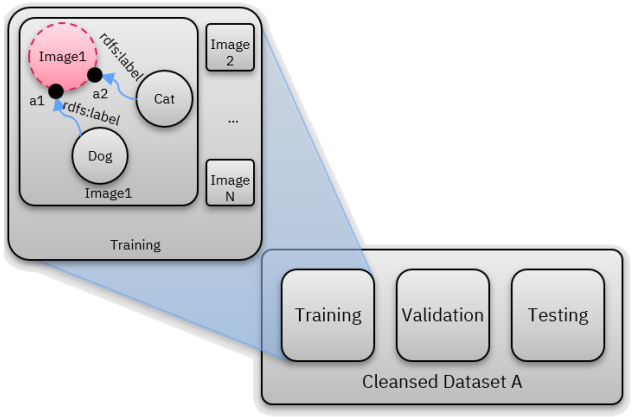


Fig. 3. An example of how to use HKW for structuring datasets.

Dataset Retrieval: Suppose one is given the task of classifying wild animals in images. For performing such a task, one could query the KB for retrieving all datasets that contain images using lines 1 to 2 of the query depicted in Figure 4. However, such a query is too broad, as it returns datasets having any type of image. An HKW entity can have anchors that represent part of its content. For instance, a node representing an image may have several anchors and one can define relationships using these anchors, that is, defining a meaning to fragments of the image. If we assume that our KB has images of animals and that information is expressed in terms of HKW relationships, one could define an additional constraint to this query, which is depicted in line 3 of Figure 4. If we have a more detailed ontology, one can be even more specific in the construction of the query.

```
1  SELECT Dataset
2   WHERE Image FROM Dataset
3     AND Image has Animal
```

Fig. 4. Query to return all contexts defined as dataset, that contain images with animals. FROM clause reduces the search scope to entities contained in HKW contexts that are instances of Dataset.

Dataset Creation: As aforementioned, dataset creation is one of the DE tasks that has been drawing the most attention from the research community, mostly because of its importance and difficulty. The creation of new datasets with data from a variety of sources is difficult and can lead to errors. We argue that a full-fledged knowledge-oriented approach for DE should support the creation of datasets through queries, which can avoid representation errors by applying a common ontology to describe different instances of datasets. For example, in the query depicted in Figure 5(a), every image from *CopiedDataset* is inserted into *NewDataset*. Practical use of this query is when one wants to reuse images from a preexisting dataset into a new one. Another example is depicted in Figure 5(b), where a new dataset is created from a selection of data with specific features.

```
1  INSERT INTO NewDataset
2  VALUES {
3      SELECT Image
4      WHERE Image from CopiedDataset
5  }
```

(a)

```
1  INSERT INTO ParentContext
2  VALUES (
3      SELECT Image WHERE Image has Dog
4          AND Image FROM DatasetOfImages
5  )
```

(b)

Fig. 5. Query that copies the content of *CopiedDataset* to the a new context (a) and query that copies the images of dogs from *DatasetOfImages* (b).

3 Demonstration

The main goal of this demo is to show HKW features that support dataset engineering tasks through KES and HyQL. In this sense, we use the Pascal VOC2012 [3] and the Playing for Benchmarks [6] image datasets to show how HKW can support dataset representation, retrieval, and creation tasks. The images will be exploited to demonstrate HKW's capability of integrating symbolic knowledge with non-symbolic content. The use of the ontologies is to illustrate the support to semantic queries for dataset retrieval. Finally, queries performed using KES will demonstrate HKW's ability to answer dataset creation queries.
Demo video 1: Dataset representation and integration with HKW ontology.
`https://ibm.box.com/v/iswc2021-dataseteng-video1`
Demo video 2: Semantically enriched dataset retrieval and media visualization.
`https://ibm.box.com/v/iswc2021-dataseteng-video2`
Demo video 3: Dataset creation using HyQL dataset engineering functions.
`https://ibm.box.com/v/iswc2021-dataseteng-video3`

4 Conclusions

In this paper, we presented a knowledge-based approach for DE in the ML domain. The main contributions of this work are the proposal and demonstration of this approach, which comprises core DE tasks such as dataset representation, retrieval, creation. Data scientists could use the approach here described to find new datasets, balance, clean, and resample them, and select specific features, which saves time and promotes better data exploitation.

As the main drawback of our approach is the need to annotate datasets for achieving richer descriptions, the (semi-)automatic annotation of datasets is a future work that will allow saving more of data scientists' time.

References

1. Chun, D.X., Jun, C.J., Zhong, C.Y., Chao, T.M., Cong, P.: Data engineering in information system construction. In: 2012 IEEE Symposium on Robotics and Applications (ISRA). pp. 135–137. IEEE (2012)
2. Dong, X.L., Rekatsinas, T.: Data integration and machine learning: A natural synergy. In: Proceedings of the 2018 international conference on management of data. pp. 1645–1650 (2018)
3. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision 111(1), 98–136 (2015)
4. Kunze, S.R., Auer, S.: Dataset retrieval. In: 2013 IEEE Seventh International Conference on Semantic Computing. pp. 1–8. IEEE (2013)
5. Moreno, M.F., Brandao, R., Cerqueira, R.: Extending hypermedia conceptual models to support hyperknowledge specifications. International Journal of Semantic Computing 11(01), 43–64 (2017)
6. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 2232–2241 (2017)