Enhancing Knowledge Graphs with Data Representatives

André Pomp[1], Lucian Poth[2], Vadim Kraus[1] and Tobias Meisen[3]
[1]*Institute of Information Management in Mechanical Engineering, RWTH Aachen University, Aachen, Germany*
[2]*Computer Science, RWTH Aachen University, Aachen, Germany*
[3]*Chair of Technologies and Management of Digital Transformation, University of Wuppertal, Wuppertal, Germany*

Keywords: Semantic Model, Knowledge Graph, Ontologies, Semantic Similarity, Machine Learning.

Abstract: Due to the digitalization of many processes in companies and the increasing networking of devices, there is an ever-increasing amount of data sources and corresponding data sets. To make these data sets accessible, searchable and understandable, recent approaches focus on the creation of semantic models by domain experts, which enable the annotation of the available data attributes with meaningful semantic concepts from knowledge graphs. For simplifying the annotation process, recommendation engines based on the data attribute labels can support this process. However, as soon as the labels are incomprehensible, cryptic or ambiguous, the domain expert will not receive any support. In this paper, we propose a semantic concept recommendation for data attributes based on the data values rather than on the label. Therefore, we extend knowledge graphs to learn different dedicated data representations by including data instances. Using different approaches, such as machine learning, rules or statistical methods, enables us to recommend semantic concepts based on the content of data points rather than on the labels. Our evaluation with public available data sets shows that the accuracy improves when using our flexible and dedicated classification approach. Further, we present shortcomings and extension points that we received from the analysis of our evaluation.

1 INTRODUCTION

Growing digitalization in companies results in an increasing amount of generated data. For using the potential of machine learning approaches and realizing visions such as the Internet of Production (IoP) (RWTH Aachen University, 2017), which pursues the goal of guaranteeing the availability of (application dependant) real-time information at any time and place, companies start to collect and store these data sets. With growing amounts of sensors and applications within a production facility, such visions will lead to an improved quality management as goods can be checked much faster, easier and more reliably in the production line (Li et al., 2017).

In order to store and process the generated amount of data, recent solutions rely on data lakes, which allow to store structured and unstructured data by following a schema-on-read approach. Compared to previous approaches like data warehouses, data lakes offer the flexibility that modern data analytics and machine learning processes require. However, to enable data scientists to access, find and understand the stored data sets, metadata has to be created for every data set as there exists no global schema like in data warehouses. (Terrizzano et al., 2015) propose to use a curated data lake where the data does not need to be fully transformed and cleansed before integration, but has a data manager for annotating the data with meta-information to improve the processing steps and make the data valuable. In order to keep a consistent semantic meaning and common understanding of the metadata over multiple data sets, companies start to establish knowledge bases in the form of ontologies, business glossaries or knowledge graphs. The purpose of such a knowledge base is to offer semantic concepts that can be annotated as metadata to a data set. By adding these concepts to data sets and putting them in relation to each other, one can describe a data set in more detail. This process is called semantic modeling.

However, creating semantic models is a cumbersome task that is very time consuming and requires a lot of domain knowledge. The amount of concepts in ontologies can easily extend a size feasible to be manageable by humans as shown by the compared ontologies of (Groß et al., 2016) under consideration of the bio-medical domain. With numbers over 300,000

Pomp, A., Poth, L., Kraus, V. and Meisen, T.
Enhancing Knowledge Graphs with Data Representatives.
DOI: 10.5220/0007677400490060
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 49-60
ISBN: 978-989-758-372-8

concepts, it becomes increasingly difficult for a data steward to select the right concept for the data attributes of a data set. To support the data steward, concepts can be recommended based on a deep analysis of the data. For instance, (Paulus et al., 2018) proposed a solution for recommending semantic concepts that uses the combination of data labels from a data set together with additional information extracted from other knowledge bases.

Although Paulus et al. achieve very good results on data sets that have conventional data labels, they also identified a class of data labels, called random labels, in which their recommendation framework will never be able to identify the correct concept by just analyzing the labels of a data set. In these cases, the label may be just a random number of characters that is auto-generated, e.g., by a database management system. This implies that the concept recommendation has to be performed based on the data instances rather than on the data labels.

Thus, the goal of our work is to exploit which data-driven approaches are suitable for different types of data and how those approaches can be used to learn the data representation of semantic concepts. We therefore develop a data-driven recommendation approach which is based on the knowledge graph provided by the semantic data platform ESKAPE (Pomp et al., 2017). We show that our approach is capable of learning multiple data representations, called *data representatives*, for concepts and is able to use those learned representations to recommend concepts. Each data representative is based on a defined classification approach, which can be a machine learning model, a rule or a statistical method. The idea of learning multiple representatives and equip them with different classification approaches for a single concept is motivated by an analysis of public data sets in which we identified different types of *Data Classes*. These classes are built on a combination of different metrics, such as data type, number of possible values, etc. We expect multiple recommendation strategies to achieve different accuracies based on the data class in which a set of values is grouped.

To analyze the accuracy of our data-driven recommendation approach, we annotated real-world data sets with semantic models and evaluated the quality of the semantic concept recommendation under the condition of the different data classes. Altogether, the main contributions of this paper are

1. an approach that makes it possible to learn semantic concept representations for dynamic knowledge bases which are not yet considered by related work, since knowledge bases are always considered static here,
2. an extension of the knowledge graph model proposed by (Pomp et al., 2017) for equipping knowledge graphs with data representations,
3. an approach that maintains multiple data representatives and classification approaches for a single semantic concept,
4. a more detailed identification of different types of data classes compared to related work which just considers numbers or text,
5. an evaluation that compares the performance of the different classification approaches for the identified data classes.

The remainder of this paper is organized as follows: Section 2 motivates the necessity to develop a recommendation approach that uses data instances and can deal with knowledge bases that are extended at runtime. Afterwards, Section 3 provides an overview of related work and Section 4 defines the data classes which we identified when reviewing real-world data sets. Based on these classes, we present our identified classification approaches and the implementation of our approach in Section 5. Section 6 gives an evaluation of our approach before we conclude with a summary and a short outlook in Section 7.

2 MOTIVATING EXAMPLE

In this section, we provide a motivating example illustrating the necessity for developing a recommendation framework that supports a user during the creation of semantic models with giving recommendations which are not solely relying on data labels but also on data values.

In the following scenario, we consider a large global enterprise that is active in many different industries with multiple sites in different countries. For instance, it develops mobility solutions as well as consumer goods. Thus, the company has a lot of different production processes. Due to the digitalization strategy of the company, for some years now, the company is already storing the collected data in their data lake. For each production process, the company is planning to setup a digital twin which virtually models the behaviour of the involved machines. Those digital twins are later used for optimizing the operation and maintenance of the production processes. Since the developers, such as data scientists, who are involved in the construction of the digital twins, have to find and access the data they are looking for, the company established a metadata-management solution. This approach uses an enterprise ontology and user-defined semantic model to describe the data sources in more

Table 1: Exemplary data set for which a data steward wants to create a semantic model.

prodID	tmp	tinms	e-mail	sta
313-16	856	1476912300	x@cp.com	OK
215-14	857	1476922300	y@cp.com	NOK
513-31	845	1476932300	x@cp.com	OK
...	...	...	...	...

detail. The semantic models are created by *data stewards*, who are employees that are domain experts and know the semantics of the data very well.

Since data sources in industrial environments can contain hundreds of data attributes (Paulus et al., 2018), creating sophisticated semantic models is a complex and time-consuming task. In order to improve the quality of the semantic models, it is therefore important to support the data steward during the creation process. We assume that a data steward wants to create a semantic model for the simplified data set in Table 1, which shows a product that has been dried in an oven at a certain temperature and was later manually checked for proper functioning by an employee. The data steward selects concepts from the underlying ontology and maps them to the columns of the table. For instance, the column *prodID* would be mapped to the concept *Identifier* and the column *tinms* to the concept *Timestamp*. In addition, the data steward can specify more details, like the unit *Seconds* in which the timestamp is measured. Later on, these detailed information will help the developers of the digital twin application to understand the data and implement their application correctly (e.g., considering the correct unit for timestamps).

Examining this table and the presented task of the data steward shows that, if the data sets become much larger, the creation of semantic models will be a complex and time-consuming task, which requires important domain knowledge. Supporting the user with recommendations based on labels would already help to reliably identify concepts, like E-Mail for the column *e-mail* but would fail for all other columns. Examining the data instances of the columns raises the suspicion that a recommendation based on the data instances can lead to better results. For instance, the columns *prodID* as well as *e-mail* follow a fixed pattern whereas the temperature values in column *temp* are limited to a certain range (845-857 °C) and the *sta* column only consists of two valid values (OK, NOK). However, the diversity of the different kinds of data instances also shows that it is not possible to develop a solution that solely relies on a single classification approach. For instance, training a machine learning classifier that detects a valid temperature is possible whereas the training for product identifier will not work since each entry is unique. However, defining rules for the product identifier, e.g., by using regular expressions, would result in a valid data representation for this concept. Nevertheless, only assigning a single data representative to a semantic concept will also lead to wrong results. Depending on the production process and its context, the data instances for a semantic concept may differ. For example, there might exist production processes in which the values of valid temperatures are not between 845-857 °C but between 12-15 °C. In these cases, it will be necessary to learn a different data representation for the same semantic concept.

Beside these examples, the scenario also relies on a fixed underlying vocabulary provided by the underlying ontology. In this case, one could also try to convert the use case with all its data to a multi-class classification problem where each data representation for each semantic concept will result in one class. While this would be a first solution, it leads to a very static scenario. In cases where the underlying vocabulary of the ontology will be extended or where new machines with different data representatives for the same concept are introduced, it becomes necessary to re-train the whole machine learning model which is very time-consuming and not manageable in a company with multiple sites and a very diverse product portfolio.

It is therefore necessary to develop an approach that takes into account not only the names of the data attributes, but also the data instances. In addition, this approach should be independent of the number of available concepts. If new concepts are introduced or data instances with different representation forms are added, it must be possible to learn these as well.

3 RELATED WORK

The research areas of semantic annotation, labeling or modeling have the goal to assign semantic concepts to data attributes of structured data sources. In order to simplify the finding, accessing and storing of data, several attempts have been made to support the data annotation. Strategies on how data sets can be analyzed to suggest concepts fitting to their attributes are elaborated by different researchers. First approaches focused on the analysis of the labels attached to the attributes in order to suggest concepts. Other approaches take the relations of those labels into account and recent papers also focus on the analysis of the instances the attributes are assigned to. Current approaches perform this task by either suggesting semantic concepts based on the label or by exploiting the data instances of the corresponding data attributes.

For instance, (Syed et al., 2010) or (Wang et al., 2012) suggest semantic concepts based on external knowledge bases. Therefore, they evaluate if the data label matches a semantic concept in one of the knowledge bases WordNet, Wikipedia or Probase. Another approach that also relies on external knowledge bases is presented by (Paulus et al., 2018). They improve the semantic concept suggestion by examining not only single data labels but by considering them in their complete context. Therefore, the authors send all data attributes to multiple knowledge bases. The returned results are then automatically merged and rated with a corresponding algorithm. Another approach which also solely focuses on the analysis of data labels is presented by (Goel et al., 2011) and (Goel et al., 2012). Goel et al. make use of conditional random fields (CRF) to annotate the labels of a data set with semantic labels. In order to improve the semantic labeling, Goel et al. do not only make use of the labels themselves but divide them into tokens, based on the instances in the fields of the data set (e.g, a value $70°F$ with the key *TempF* is divided into the tokens *70*, $°$ and *F*, which all separately get a token label). The newly created tokens are combined with the original field labels in order to create a sequence CRF graph. Results of experiments in which data from three different domains was used showed an accuracy between 89% and 98%. In distinction to our work, both approaches exploit the same classification strategy (CRF) for the annotation process whereas our approach considers multiple classification approaches per semantic concept. In addition, the authors always rely on meaningful labels whereas (Paulus et al., 2018) argued that there exist many data sets where a label-based strategy cannot work.

Thus, in (Ramnandan et al., 2015), the authors present a further enhancement of the approach that has been made by Goel in (Goel et al., 2011). In comparison to the previously presented papers, the focus for the assignment of semantic labels is shifted from the analysis of attribute and key names to the instances of a data set. They do not take every value of an attribute by itself into context but the set of all instances mapped to that attribute as a whole to analyze which characteristics describe that attribute. Those characteristics are then linked to semantic concepts of an ontology. The authors differentiate between textual and numerical data as they use different analysis techniques on them. To suggest a label for the attribute of a new text document, the *cosine similarity* between all indexed documents and the new document is calculated to present the top k elements with the highest similarity score. For numerical instances, a *statistical hypothesis testing* is used, which is performed between a new data set that should be labeled and every numerical data sample used for training. As a new data set is compared with every already learned characteristics, the system is adoptable to new semantic labels without any effort. Ramnandan et al. evaluate their implementation on data from five different domains (museum, city, weather, flight status, and phone directory). With the size of the test data sets being quite small (9 different labels in the flight status domain), they achieve a maximal Mean Reciprocal Rank score ranging from 0.421 up to 0.943 for the first four concept suggestions. The basic idea of this approach is similar to ours as it is also rather focused on the exploitation of the data instances than the attribute names. However, the authors only differentiate between text and numerical instances whereas our approach focuses on more fine-granular data classes. In addition, we allow to learn multiple data representations for the same semantic concept.

In (Pham et al., 2016), the authors present an approach that does not focus on the pure data instances but the similarity of the metadata (respectively similarity metrics) of these instances, whereby their approach becomes independent of the domain the model was trained upon. As similarity metrics, the authors use *Attribute Name Similarity*, *Value Similarity*, *Distribution Similarity* for numerical instances, *Histogram Similarity* for textual instances and *Ratio* for cases where a mixture of numerical and textual instances is available. For every attribute of a set $\{a_1, a_2, ...a_n\}$, a feature vector $f_{ij}(i \neq j)$ is computed. Every similarity metric in dimension k is thereby calculated for itself, with $f[k]$ representing the similarity of a_i and a_j under metric k. Every calculated vector is annotated manually, whether the attributes are semantically similar or not. Based on the identified feature vectors, they train a machine learning classifier. They evaluated their approach for data sets of four different domains under the usage of two different classifiers, respectively Logistic Regression and Random Forest, with Logistic Regression showing the better performance. Compared to our approach, Pham et al. only train one classifier whereas our work permits multiple data representations per concept where each representation can be based on different classification approaches. Similar to our approach, Pham et al. already take the data class into account to calculate similarity metrics differently, but they restrict their differentiation merely to textual and numerical instances, with the classifier handling them similar. The data classes evaluated in our work offer a broader variety and the impact of different classification approaches is evaluated on all data classes.

While these approaches provide good results for

extending the recommendation of semantic concepts for data attributes, we believe that it is not possible to cover all different concepts with a single approach just as label recommendation, similarity measures or machine learning. Instead, we believe that the quality of the recommendation for data instances depends on the data class the instances belong to.

4 DATA CLASSES

Since our approach is targeting semantic concept suggestion based on the concrete data values, we need to consider the basic properties of data values. We identified a certain categorization which groups these properties in data classes. These classes are later used as an additional feature or criteria for the selection of the suggestion method. This section gives an overview of the data classes we identified and we discuss their specialties. The overview is based on a selection of data sets also used by (Paulus et al., 2018), mainly in the domain of publicly available, municipal data sets as well as the data set used by (Pham et al., 2016) from the domain of soccer and museums. Pham et al. already divide their data sets into textual and numerical values to perform different suggestion algorithms based on the data type, but we expect that further gradations of the data properties might improve the accuracy of suggestion strategies.

We do not claim that the following list of properties and classes is exhaustive, however, all assumptions can be validated and they cover the most seen properties. Also, video, picture, sound and any more complex data is not covered by this classification, as the assumption is made that they can be transformed into numerical values for semantic analysis.

The following data properties were observed. The first defining property of a data point is the **Data Type**. A data type defines the most basic limitation to the expressiveness to one single data value. The observed types range from numbers (discrete and floating) to a sequence of any characters. Other properties are defining the relation between two data points. A **Scale** provides the possibility to interpret the distance between two data points. Scales can be either categorical or numerical, whereby the categorical scale can be divided into nominal or ordinal scales. In a nominal scale there is no relation between the instances, they are comparable to labels. In an ordinal scale, the difference between two values is still not quantifiable or has no specific meaning, however, values can be **Sorted**. Finally, in a numerical scale, even the distance between two values has a meaning.

The following data classes are a result of the combination of the previously described properties:

Full Text: This class (*text*-class) is the least restrictive. Many of the reviewed data sets contained attributes whose values consisted of longer texts (e.g., a description or an abstract). As most of the machine learning techniques we use are based on numerical values, their application for values that consist of longer texts is quite challenging. Also, histograms are insufficient as the variety of the texts is too high. One possible solution for this issue is assumed to be a dictionary of n-grams in order to have a look up for certain phrases occurring in a longer text.

Identifier: This class (*id*-class) restricts the first class in two aspects. First, it contains less characters and second, the characters follow a more or less strict pattern. Names are a typical representative of this class. However, determining whether an arbitrary sequence of characters is conceptually a name is hard, as names do not follow any predictable pattern. Same holds true for other identification strings or numbers, which can also occur in a combination of characters and numbers. Other identifiers like an IBAN or ISBN on the other hand have a fixed pattern by which they are created. Email addresses make up a set in between as all of the contain an "@", which is rarely used in a different context, but the rest of the address is quite arbitrary.

Bag-of-words: The next class also consists of limited character sequences, however, in this case the structure of the character sequence is of secondary importance. The defining criteria of this class is that the attributes are composed only from a fixed set of allowed values called a bag-of-words (*bow*). Examples for this are soccer player positions ("Goalkeeper", "Left-back", "Centre midfield", "Striker", ...) or nominal scales ("good", "average", "bad"). A direct transfer from values with a similar semantic meaning but different representations is not possible. For example a fixed set of words cannot be recognized by the abbreviation of the words (e.g., for soccer "GK", "LB", "CM", "ST") or one nominal scale by another (e.g., English to German grades). The new set has to be attached to the corresponding concept in order to suggest the concept to new data.

Numerical Values: The last class consists of only numbers (discrete or floating point). Numerical values (*num*-class) occur in any form of measurement or calculated results. The possible values in a data set with real values cannot be grouped with a fixed set as there are arbitrary many valid values. In difference to the identifiers that may also be composed of numbers, real numbers can be put in a relation to each other which allows different evaluations as they can be placed on numerical scale.

5 DATA REPRESENTATION

In this section, we present our developed approach which we integrated into the semantic data platform ESKAPE (Pomp et al., 2017), which offers Ontology-Based Data Access (OBDA). The idea of ESKAPE is that data stewards publish data sources and describe the additional required and interesting meta-information with semantic models. Based on the semantic models, other users like data scientists, can query and access these data sources later on. In order to support the data steward during the semantic model creation, ESKAPE identifies semantic concept recommendations based on the framework presented by (Paulus et al., 2018). Data stewards can then create the semantic models with the graphical user interface of ESKAPE (Pomp et al., 2018) where they can choose from the recommendations or browse through all existing semantic concepts and relations provided by ESKAPE's underlying knowledge graph, which encourages them to make choices upon that shared terminology. Compared to traditional OBDA approaches, users can extend the knowledge graph's vocabulary by introducing new concepts or relations on-demand directly to the knowledge graph to make them available for others. This circumstance is very crucial for the design of our approach as it excludes the possibility of using a multi-class classification approach. If we would use this method, the approach would require re-training after a new semantic concept has been added to the knowledge base. However, this would be very time- and resource-intensive. Hence, the goal of our approach is to enhance each semantic concept with data representatives that capture or describe the characteristics of the data instances that are annotated with this concept.

We therefore identified different approaches, such as statistical methods, regular expressions and machine learning methods that can be used for representing the instances of a semantic concept. The goal of these approaches is to evaluate later on if a number of data instances are valid or invalid representatives of this semantic concept. We call these approaches *classification approaches*. Each semantic concept can be represented by one or more classification approaches. For instance, one could have different machine learning models or statistic methods that are used for identifying if a data value is a representative of this semantic concept. We decided to link a semantic concept to different classification approaches as the instances of data attributes can belong to different data classes. We then integrated our framework into the semantic data platform ESKAPE and modified the semantic modeling approach of ESKAPE. In the following, we give an overview of the identified classification approaches (cf. Section 5.1), the modifications which we made for the knowledge graph model provided by ESKAPE (cf. Section 5.2) and the process of how the recommendation and training works (cf. Section 5.3 and 5.4).

5.1 Classification Approaches

Classification approaches represent the idea that we can describe a semantic concept based on the sum of all of its data instances. However, since the same semantic concept can be represented by different kinds of data instances (e.g., as a textual representation or as a number), we identified that it is not possible to solely find a single classification approach that is capable of capturing all characteristics of the semantic concept. Based on the data classes and their characteristics that we identified in Section 4, we selected classification approaches which match these. As classification approaches, we chose a rule-based approach based on regular expressions, a histogram approach as statistical method and one machine learning approach in the form of a one-class classifier.

Regular Expression: The analysis of the given data sets showed that certain attributes contain instances, which follow a fixed set of syntactic rules. For example, the attribute *dim* of a data set with information to paintings contain instances like 135.7x55.3cm, 44.2x55.0cm and 62.8x81.9cm. The concept of the attribute is "dimension", which can be described by a representative based on a regular expression (RegEx) as the instances follow a fixed pattern. The RegEx sufficient to characterize this instances would be `(\d)*.\d x (\d)*.\dcm`. The disadvantage of this approach is that the user has to come up with the RegEx by hand in order to describe the instances. It is also necessary to keep the RegEx as specific as possible in order to prevent it fitting to other concepts. The RegEx `.*` would also cover the top example, but every other value would also be verified by this pattern.

Histogram: Opposed to the rule-based approach, which takes every value separately into account, the histogram-based method is chosen to use the frequency of every value as metadata. This frequency is intended to provide a better insight into the data set, as it provides an opportunity to distinguish concepts that inherit similar instances, but are semantically different. As a histogram does not consider the order of the instances, a distribution of the instances is not taken into account. As mentioned in Section 4,

the order of the instances might include further information on a data set if the instances are denoted in a fixed order which is often encoded in the instances of a different attribute. For real numbers we also consider binning strategies based on a rounding factor. For textual instances, this binning is not implemented (e.g., to group instances with a similar meaning).

One-Class Classifier: In order to detect boundaries between instances that represent one concept and instances representing another one, a one class classifier (OCC) is used. The usage of an OCC provides the advantage over other classification techniques that it is restricted to a training set containing only valid instances. For the classification approaches of the concepts, this is useful as each classification approach becomes independent for the other. New classification approaches can be added without the need for re-training of previously trained models. Currently, we use a support vector machine (SVM) as one-class classifier. However, other solutions, like AutoEncoder would also be possible.

5.2 Knowledge Graph Model

For our implementation, we first extended an existing knowledge graph model provided by the semantic data platform ESKAPE (Pomp et al., 2017) to be capable of linking semantic concepts to data representations.

The original graph model described by (Pomp et al., 2017) describes a basic model for semantic concepts, called Entity Concepts, and the relations between them. To additionally enable the learning of data representations for those Entity Concepts, we extended the graph model. Therefore, we extended each Entity Concept node ec_x with an additional data representation node drn_x. The data representation node drn_x is linked to the different data representatives drc_i that were trained for the entity concept ec_x. We introduce for each new data representative j, a new node drc_j and attach it to the corresponding drn node. Depending on the type of the data representative (Histogram, Regex, OCC, etc.), each of those drn nodes has different properties. For histograms, we save the *data class* for which this histogram was created, a *location* where we store the histogram model and the *rounding* which describes on how many decimal places we round, or -1 if it is a text-based histogram. For the one class classifier, we store the *location* where the trained model is stored and for which data class it was trained. Finally, for a regular expression, we store the *data class*, the *pattern* and a *fit percentage* which describes the percentage of data instances of the annotated data attributes on which this regular expression matched.

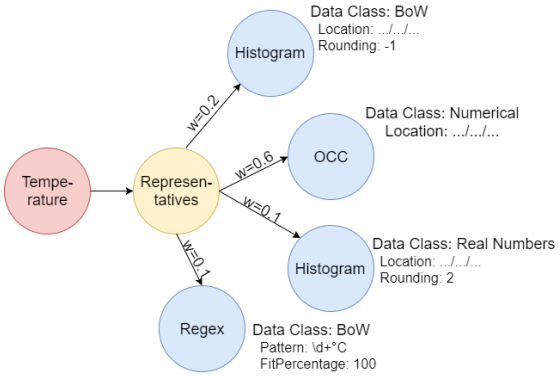

Figure 1: Example for the concept Temperature. Based on the currently annotated data, four different data representations were created. The weights w show that most of the annotated data attributes were trained with the One Class Classifier.

Since the number of possible data representations per semantic concept potentially increases with the number of new data sources, we added a weighting factor w to the edge e_{ij} between the data representation node $drn_x i$ and the corresponding data representative drc_j. This weighting factor describes how many percent of all annotated data attributes were used with this concept to train this data representative. Figure 1 shows an example. For the semantic concept *Temperature*, ten data attributes were annotated with this concept, but only four different data representatives were trained. Two of those representatives are Histograms, one OCC and one regular expression. The weighting factor shows how many data attribute columns were used for the different data representatives. For instance, the 0.6 for the OCC indicates that the instances of six out of ten data attributes were currently used to train the OCC model, whereas for the regular expression representative the instances of only one data attribute were used. In the later recommendation process, the data representatives with the higher weight will be evaluated first.

5.3 Recommending Concepts

The recommendation and the training or updating of existing data representatives is done during the schema analysis and the semantic model creation in the ESKAPE platform. During the schema analysis, ESKAPE identifies the best fitting semantic concept for a corresponding data attribute. Previously, this process was based on the recommendation framework provided by (Paulus et al., 2018). We extended this process to additionally evaluate the already existing data representatives. First, we randomly sample ten

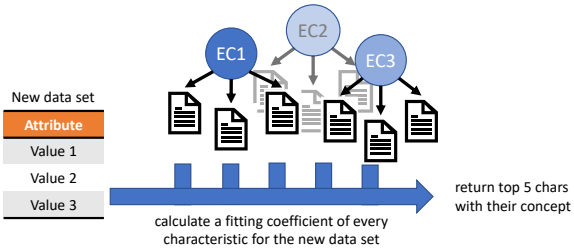

Figure 2: Model of the concept suggestion based on different classification approaches. The instances of a new data set are analyzed and compared with every existing classification approach and a similarity score is calculated. The concepts with the highest score are returned.

percent of the available data instances *dv* from the data attribute *da*. For each available concept that has already attached data representatives in the knowledge graph, we check how well this concept represents the data instances. Therefore, we calculate for each data representative *drc* a similarity score *s* between the set of randomly extracted data instances and the *drc*. This similarity score defines how well the data instances fit the current representation of the semantic concept. To speed-up the finding of a match, we first evaluate the data representatives that have a higher weight. If a data representative with a high similarity score, i.e., at least 80% , was identified, the concept is marked as relevant and the other data representatives of this concept are not evaluated anymore. As soon as all semantic concepts were evaluated, the top five concepts with the highest similarity score are returned as recommendations. Figure 2 shows this approach for a small example.

Similarity Score: The calculation of the similarity score depends on the different classification approaches. For the rule-based approach based on regular expressions and for the one class classifier, the strategies are quite similar. In both cases, for every data value of the attribute that is about to be annotated, it is checked whether they fit into the given data representation or not. For the rule-based approach, the RegEx pattern is evaluated on every value and for the OCC-based method every value is classified by the OCC algorithm and the percentage of instances fitting the class is used for evaluation. The higher the percentage of fitting instances is, the more it is assumed that the concept associated with the respective data representation fits to the attribute of the instances. To suggest a concept based on a histogram classification approach, the histogram for the instances of the data set is calculated. The histogram is normalized as the amount of instances should not effect the calculation. Afterwards, the histogram of every available representative is compared with the one from the new data set and the intersection for them is calculated. Due to the normalization, results between 0 and 1 occur with the highest rated intersections being used for the suggestion of their respective concepts.

5.4 Training Concepts

With each new data set and the corresponding created semantic model, the concept representations are updated and/or extended. For training classification approaches for semantic concepts, we have to differentiate between different cases.

No Representation Available: In this case, we allow the user, who is creating the semantic model, to select a classification approach that should be used for learning the semantic concept. Hence, if the user selects the histogram or one class classification approach, the system will calculate a histogram or train a SVM classifier for the annotated data attribute and store them on the underlying storage system. For the histogram, the user has to additionally define the rounding factor. For the SVM classifier, the user does not have to define anything. Currently, all SVM parameters are set to default. However, in the future we plan to perform an automatic grid search with a different parameter set. With the help of cross-validation, we will then select the best fitting parameters for the available data instances. If the user selects a regular expression, he has to define it and provide it to the system. Our approach will then create a new data representative for the used semantic concept.

Representation Available but Inappropriate: If a semantic concept has already assigned classification approaches but none of those were valid candidates for the recommendation, then we request the user to define the same information as if no representation would be available. This means a user has to choose a classification approach and has to define the required parameters. Our approach will then create a new data representative for the used semantic concept.

Appropriate Representation Available: If an appropriate representation is available, i.e., the similarity score was larger than 80%, this representation will be extended with the new instances. In case of a regular expression classification approach, the fit percentage value will be updated. In case of histogram, we will create a new histogram and store it. The one class classifier will be re-trained with the old and new data instances.

6 EVALUATION

This section evaluates the three implemented classifications approaches (histogram-, rule-based and OCC-based approach) and how accurate their suggestions of concepts are compared to manually annotated concepts. All evaluations are conducted with respect to the data classes (cf. Section 4). This means that we considered the four identified data classes (bag-of-words (*bow*), Numerical values (*num*), identifiers (*id*) and textual (*text*)). The results are compared with the ones achieved by (Pham et al., 2016). To achieve compatibility with Pham et al., we limit our evaluation to the data sets of the domains *museum* and *soccer*. The source data sets museum and soccer can be summarized with following distribution: The museum data set contains 4132 entries over 28 different semantic concepts in total. Most frequently, the *bow* class can be observed with 18 entries. The other concepts are evenly distributed among the other classes. For the soccer data set, the **bow** class and the numeric class are the most prominent with 14 and 13 out of 33 respectively. The other classes are also evenly distributed. There are 9443 entries in total in this set. The other domains provided either only ten data values or provided only five attributes.

6.1 Evaluation Method

All measurements are conducted on a knowledge graph, which contains all concepts that are manually assigned to the attributes of the data sets.

Two varieties of hit ranks are measured. The *Mean Reciprocal Rank* (MRR) of the top four elements (MMR4) is evaluated beside the top-rating (T1). The top four elements are evaluated as they are also used by (Pham et al., 2016) for their evaluation. The MRR is defined by (Craswell, 2009) as followed:

Mean Reciprocal Rank. *The Reciprocal Rank (RR) information retrieval measure calculates the reciprocal of the rank at which the first relevant document was retrieved. RR is 1 if a relevant document was retrieved at rank 1, if not it is 0.5 if a relevant document was retrieved at rank 2 and so on. When averaged across queries, the measure is called the Mean Reciprocal Rank (MRR). Mathematically described as followed for all queries Q:*

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}.$$

After the hit rank is evaluated, the current accuracy of the suggestion process is evaluated and a representation that is created based on the values of the current attribute is attached to manually labeled concept.

Since we propose an evolutionary learning approach, the order in which the data representatives are added to the concepts might affect the quality of the suggestion process. Therefore, we perform a ten-fold cross validation using a random shuffle of the labeled attributes.

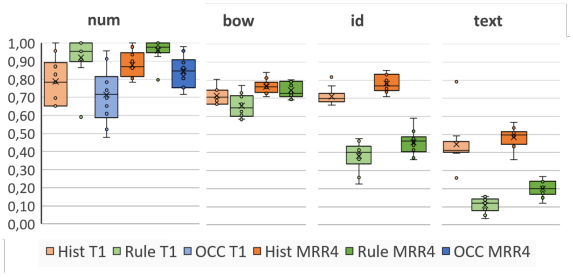

Figure 3: Box plot diagrams of the measurement series' over the different data classes with different classification approaches on the museum data set.

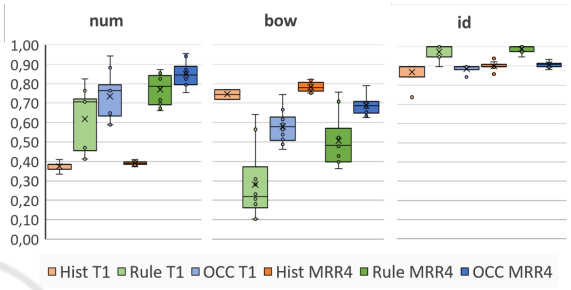

Figure 4: Box plot diagram for the evaluated accuracy on the soccer data sets. The data class *text* is left out as the amount of attributes associated with that class is too low.

6.2 Classification Methods

To evaluate which of the classification approaches suits best for which kind of data class, we first compare our suggestions in their basic form on single domains. In the next step we increase the complexity by considering multiple domains.

Single Domain. Figure 3 (museum data set) and Figure 4 (soccer data set) show the domain dependent results of the evaluation of which classification approach suits best for which kind of data class, a cross evaluation of all approaches on the data sets from the domain museum and soccer is conducted. As the number of attributes associated with the data class *text* in the soccer domain data sets are too few, they were left out of the evaluation. The data set contained only seven entries, which are all labeled with "date".

For the museum data set, the OCC approach is only evaluated on the data class *num*, since the mapping process from text values to numerical values is too expensive, i.e., the suggestion of a single concept takes on average 10 minutes, even if only a few representatives are available. The OCC accuracy is evaluated for the data set from the soccer domain, since the number of data values per attribute is much lower than those for the museum data set. However, the calculation of the OCC approach is still slower than that of the other approaches.

In the museum data, the best results for *numerical* data could be achieved through a rule-based approach with a measured accuracy of T1: 0.9182 (MRR4: 0.9575). For the soccer data set, the OCC-based approach performed better with accuracies of only 0.7353 (0.8480), however. While the histogram approach achieves reasonable results for the museum set, it has a very low accuracy in the soccer case 0.3744 (0.3902).

The *bow*-class is similarly well predicted by the histogram- and rule-based approach in the museum case, with accuracies of 0.7137 (0.7640) for the histogram-based and 0.6583 (0.7436) for the rule-based approach. The histogram result can be reproduced in the soccer case 0.7462 (0.7831), however, the rule-based approach performs badly 0.2795 (0.5057).

The histogram-based approach performed well for the *id*-class and *text*-class, while the *text*-class has generally a bad accuracy for the museum data sets. In the soccer set, the rule-based approach performs best with accuracies of 0.9737 (0.9868), while the histogram-based approach also has reasonably fine and even significantly better results than for the museum case, i.e., 0.8684 (0.8989) vs. 0.7137 (0.7640).

These results indicate that no single approach is a universal winner, the proper approach seems to be highly dependant on the individual data points.

Mutliple Domains. As the previous tests where restricted to only one domain, the evaluation of this section measured the accuracy for the mixture of soccer and museum based data sets. This enlarges the set of possible concepts that can be suggested to an attribute and creates a broader variety of concepts within one knowledge graph. The results of the evaluation are depicted in the box plot diagram of Figure 5.

The results of the analysis show that the developed framework holds up the measured accuracy for mostly all combinations of data classes and classification approaches. The only significant drops of quality occurred with the rule-based approach being conducted to the data classes *bow* and *id*. For the *bow*-class, the quality dropped from 0.6583 (0.7436) in the museums domain to 0.4020 (0.5570) in the overall approach. The range of the box plot diagram is also enlarged indicating the importance of the order in which the representatives are added to the concepts. This might be explained with the low quality of the rule-based approach achieved on the soccer domain for data sets that are assigned to the *bow*-class.

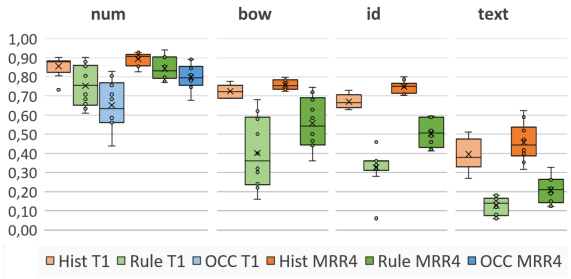

Figure 5: Box plot diagram for the accuracy measured over all data classes and classification methods with data sets from the museum and soccer domain being included into the knowledge graph and annotated randomly mixed.

One significant improvement of the accuracy could be measured for the combination of the data class *num* and the histogram-based classification approach. The quality was enhanced from 0.2795 (0.5057) in the soccer domain based data sets to 0.7244 (0.7572) in the domain mixed approach. Also the range of the box blot scale shrinks drastically for the histogram-based approach.

The results of this test show that the enlargement of the concept variety does not impact the accuracy of the classification methods as long as the boundaries of the added concepts stay disjointed.

Comparison to Pham. This section provides a discussion of the measured results. As the accuracy measured on the data sets from the domains soccer and museum is based on the same data sets as the evaluation of Pham et al., the results will be set into the context of their evaluation. Unfortunately, the results presented in their paper could not be reproduced by the provided code, so the results denoted in their outline will be used as a comparison.

Table 2: MRR scores of different classifiers when training on soccer by (Pham et al., 2016).

	soccer	museum
Logistic Regression	0.814	0.863
Random Forest	0.794	0.799

As our evaluation was only conducted of the data sets from the domains soccer and museum, only those

Table 3: MRR scores of different classifiers when training on museum by (Pham et al., 2016).

	soccer	museum
Logistic Regression	0.815	0.845
Random Forest	0.820	0.778

Table 4: MRR scores of the classification approaches on soccer. Bold marked values scored higher than the best result of Pham et al. on the soccer data.

	bow	num	id	text	average
Histogram	0.783	0.390	**0.899**	**0.954**	0.757
Rule	0.506	0.771	**0.987**	**0.983**	0.812
OCC	0.687	**0.848**	**0.907**	**0.950**	**0.848**

Table 5: MRR scores of the classification approaches on museum. Bold marked values scored higher than the best result of Pham et al. on the museum data.

	bow	num	id	text	average
Histogram	0.764	**0.877**	0.793	0.783	0.804
Rule	0.752	**0.974**	0.457	0.200	0.596
OCC		0.843			0.843

results are relevant for comparison with the results of Pham et al. and presented in Table 2 and 3. Since Pham et al. present the results of a MRR of the top four suggestions, the same measured results are presented in Table 4 and Table 5.

The results of this paper are achieved by a different approach than the one used by Pham et al., but show a similar accuracy like the one measured by Pham et al. Certain combinations of classification approach and data class showed to be more efficient, while other performed worse.

As the provided data sets and number of concepts is not that high, with 12 data sets of soccer and 28 data sets of museum data, the results can be described as equally accurate. Since the actual concepts used by Pham et al. are not all listed in their paper, the same ones could not be annotated to the attributes, evaluated by this paper. As the used concepts are unknown, the variety of annotated concepts is also slightly different. Pham et al. distinguish 20 concepts in the museum domain and 14 in the soccer domain. In this paper, for the museum domain 28 concepts and for the soccer domain 33 concepts have been identified. However, as described in Section 4 they are split up into the different data classes in this paper.

Both methods are domain independent but follow different strategies to achieve this goal. The method presented by Pham et al. relies on the training of one ML model, trained to tell whether two attributes can be associated with the same concept based on the metadata of the data instance. While Pham et al. distinguish between textual and numerical values, the here presented method focuses further on the data class a set of values can be described by. Textual data is split into the data class *bow*, *id* and *text* and numerical data is split into the data classes *num*, *bow* and *id*. The numerical values are mapped to the class *bow* or *id* if their semantic meaning does not focus on the numerical comparison of the values, but could be replaced by text (e.g., replace a rating from one to ten in a survey or replace a numerical id by a textual unique identifier). The implementation of this paper offers individual classification approaches for every concept. Similar to the method of Pham et al. the here presented implementation follows the suggestion strategy, which compares the representatives of an attribute that should be annotated with every previously integrated representative.

The results of the comparison between the different data classes and classification approaches show that for certain combinations one or the other is more efficient. While the histogram underperformed on the data class *num* on the data sets of the soccer domain, the rule-based approach showed to be inefficient on the classes *id* and *text* of the museum data.

7 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach on the process of concept suggestion for data sets in order improve the quality and duration it takes to integrate them into a semantic model. We discussed different existing strategies and methods, which revealed that only a few approaches make use of the data instances to recommend semantic concepts. Most research focuses on the evaluation of the attribute labels.

Therefore, we focus on the suggestion of concepts based on the data instances of a data set, whose semantic meaning is encapsulated in a characteristic that is added to the concept via a representative. In order to exploit the metadata of values, the attributes of the values are associated with different data classes, before concepts are suggested. The classes were determined based on the evaluation of different real-world data sets and consist of *bag-of-words*, *numerical*, *identifier* and *full text*.

The concept classification approaches developed are rule-, histogram- and one-class-classifier-based. As a foundation for the implementation, certain parts of the ESKAPE platform by (Pomp et al., 2017) are used, with the goal to extend it by the named functionality. To do so, the knowledge graph used by ESKAPE is extended by the possibility of adding data representatives to the concepts. Next, the named

classification approaches are implemented and cross tested on the discussed data classes. The results show that there are significant differences in the accuracy of the classification approaches in context of specific data classes. The use of a flexible and dedicated classification of semantic concepts, like we presented in this paper allows for better suggestions and improves the semantic labeling process massively. To improve our approach, we already started to evaluate the use of a summarized representative, using a combination of multiple classification approaches. In this context, we envision the use of a hierarchical approach for the classification, i.e., creating a global classifier which contains a selection of specialized ones. Creating specialized classifications would result in a self optimized adaptation of the look up strategies of a global method. However, at the moment, the results of our summarized representatives are only similar to the ones achieved without the creation of summarized representatives. Hence, further research is needed. Beside summarizing representatives, another direction of research includes the enhancement of the approaches applied to the different classes, such as using further machine learning methods for dealing with full texts. Another important point is the evaluation of different classification approaches. Preliminary results of using machine learning-based approaches like Autoencoders promise better classification results. In addition, it must be examined to what extent the results of our approach can be generalized in larger real applications. It will therefore be necessary to carry out evaluations with a larger number of data sets with similar yet different concepts and data attributes. In these cases it might be helpful to additionally support the classification by considering label-based suggestions with methods like (Paulus et al., 2018). Finally, the determination of the data classes is currently limited to the identified ones and is a manual process. An extension would include the implementation of a feature that allows for any granularity in data classes and an automated identification of those.

REFERENCES

Craswell, N. (2009). Mean Reciprocal Rank. In LIU, L. and ÖZSU, M. T., editors, *Encyclopedia of Database Systems*, page 1703. Springer US, Boston, MA.

Goel, A., Knoblock, C. A., and Lerman, K. (2011). Using Conditional Random Fields to Exploit Token Structure and Labels for Accurate Semantic Annotation. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI-11)*, San Francisco, CA.

Goel, A., Knoblock, C. A., and Lerman, K. (2012). Exploiting Structure within Data for Accurate Labeling Using Conditional Random Fields. In *Proceedings of the 14th International Conference on Artificial Intelligence (ICAI)*.

Groß, A., Pruski, C., and Rahm, E. (2016). Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and structural biotechnology journal*, 14:333–340.

Li, X., Tu, Z., Jia, Q., Man, X., Wang, H., and Zhang, X. (2017). Deep-level Quality Management Based on Big Data Analytics with Case Study. In *Proceedings 2017 Chinese Automation Congress (CAC)*, pages 4921–4926, Piscataway, NJ. IEEE.

Paulus, A., Pomp, A., Poth, L., Lipp, J., and Meisen, T. (2018). Gathering and Combining Semantic Concepts from Multiple Knowledge Bases. In *Proceedings of the 20th International Conference on Enterprise Information Systems*, pages 69–80. SCITEPRESS - Science and Technology Publications.

Pham, M., Alse, S., Knoblock, C. A., and Szekely, P. (2016). Semantic Labeling: A Domain-Independent Approach. In Groth, P., editor, *The semantic web - ISWC 2016*, Lecture note in computer science, pages 446–462. Springer, Cham.

Pomp, A., Paulus, A., Jeschke, S., and Meisen, T. (2017). ESKAPE: Information Platform for Enabling Semantic Data Processing. In *Proceedings of the 19th International Conference on Enterprise Information*. SCITEPRESS - Science and Technology Publications.

Pomp, A., Paulus, A., Klischies, D., Schwier, C., and Meisen, T. (2018). A Web-based UI to Enable Semantic Modeling for Everyone. In *SEMANTiCS 2018 14th International Conference on Semantic Systems*.

Ramnandan, S. K., Mittal, A., Knoblock, C. A., and Szekely, P. (2015). Assigning Semantic Labels to Data Sources. In Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., and Zimmermann, A., editors, *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 – June 4, 2015. Proceedings*, pages 403–417. Springer International Publishing, Cham.

RWTH Aachen University (2017). Digital Connected Production. https://www.rwth-campus.com/wp-content/uploads/2015/01/Broschuere-Cluster-Productionstechnik-20170508-web.pdf.

Syed, Z., Finin, T., Mulwad, V., and Joshi, A. (2010). Exploiting a web of semantic data for interpreting tables. In *Proceedings of the Second Web Science Conference*, volume 5.

Terrizzano, I. G., Schwarz, P. M., Roth, M., and Colino, J. E. (2015). Data Wrangling: The Challenging Yourney from the Wild to the Lake. In *CIDR*.

Wang, J., Wang, H., Wang, Z., and Zhu, K. (2012). Understanding tables on the web. *Conceptual Modeling*, pages 141–155.