Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods

Heiko Paulheim,
Data and Web Science Group, University of Mannheim, B6 26, 68159 Mannheim, Germany
E-mail: heiko@informatik.uni-mannheim.de

Abstract. In the recent years, different web knowledge graphs, both free and commercial, have been created. While Google coined the term “Knowledge Graph” in 2012, there are also a few openly available knowledge graphs, with DBpedia, YAGO, and Freebase being among the most prominent ones. Those graphs are often constructed from semi-structured knowledge, such as Wikipedia, or harvested from the web with a combination of statistical and linguistic methods. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness. In order to further increase the utility of such knowledge graphs, various refinement methods have been proposed, which try to infer and add missing knowledge to the graph, or identify erroneous pieces of information. In this article, we provide a survey of such knowledge graph refinement approaches, with a dual look at both the methods being proposed as well as the evaluation methodologies used.

Keywords: Knowledge Graphs, Refinement, Completion, Error Detection, Evaluation

1. Introduction

Knowledge graphs on the web are a backbone of many information systems that require access to structured knowledge, be it domain-specific or domain-independent. The idea of feeding intelligent systems and agents with general, formalized knowledge of the world dates back to classic Artificial Intelligence research in the 1980s [89]. More recently, with the advent of Linked Open Data [5] sources like DBpedia [55], and by Google’s announcement of the Google Knowledge Graph in 2012 [1], representations of general world knowledge as graphs have drawn a lot of attention again.

There are various ways of building such knowledge graphs. They can be curated like Cyc [56], edited by the crowd like Freebase [9] and Wikidata [101], or extracted from large-scale, semi-structured web knowledge bases such as Wikipedia, like DBpedia [55] and YAGO [98]. Furthermore, information extraction methods for unstructured or semi-structured information are proposed, which lead to knowledge graphs like NELL [14], PROSPERA [69], or KnowledgeVault [21].

Whichever approach is taken for constructing a knowledge graph, the result will never be perfect [10]. As a model of the real world or a part thereof, formalized knowledge cannot reasonably reach full coverage, i.e., contain information about each and every entity in the real world. Furthermore, it is unlikely, in particular when heuristic methods are applied, that the knowledge graph is fully correct – there is usually a trade-off between coverage and correctness, which is addressed differently in each knowledge graph. [108]

To address those shortcomings, various methods for knowledge graph refinement have been proposed. In many cases, those methods are developed by researchers outside the organizations or communities which create the knowledge graphs. They rather take an existing knowledge graph and try to increase its...
coverage and/or correctness by various means. Thus, the focus of this survey is not knowledge graph construction, but knowledge graph refinement.

For this survey, we view knowledge graph construction as a construction from scratch, i.e., using a set of operations on one or more sources to create a knowledge graph. In contrast, refinement assumes that there is already a knowledge graph given which is improved, e.g., by adding missing knowledge or identifying and removing errors. Usually, those methods directly use the information given in a knowledge graph, e.g., as training information for automatic approaches. Thus, the methods for both construction and refinement may be similar, but not the same, since the latter work on a given graph, while the former are not.

It is important to note that for many knowledge graphs, one or more refinement steps are applied when creating and/or before publishing the graph. For example, logical reasoning is applied on some knowledge graphs for validating the consistency of statements in the graph, and removing the inconsistent statements. Such post processing operations (i.e., operations applied after the initial construction of the graph) would be considered as refinement methods for this survey, and are included in the survey.

Decoupling knowledge base construction and refinement has different advantages. First, it allows – at least in principle – for developing methods for refining arbitrary knowledge graphs, which can then be applied to improve multiple knowledge graphs. Other than fine-tuning the heuristics that create a knowledge graph, the impact of such generic refinement methods can thus be larger. Second, evaluating refinement methods in isolation of the knowledge graph construction step allows for a better understanding and a cleaner separation of effects, i.e., it facilitates more qualified statements about the effectiveness of a proposed approach.

The rest of this article is structured as follows. Section 2 gives a brief introduction into knowledge graphs in the Semantic Web. In section 3 and 4, we present a categorization of approaches and evaluation methodologies. In section 5 and 6, we present the review of methods for completion (i.e., increasing coverage) and error detection (i.e., increasing correctness) of knowledge graphs. We conclude with a critical reflection of the findings in section 7 and a summary in section 8.

2. Knowledge Graphs in the Semantic Web

From the early days, the Semantic Web has promoted a graph-based representation of knowledge, e.g., by pushing the RDF standard[1]. In such a graph-based knowledge representation, entities, which are the nodes of the graph, are connected by relations, which are the edges of the graph (e.g., Shakespeare has written Hamlet), and entities can have types, denoted by is a relations (e.g., Shakespeare is a Writer, Hamlet is a play). In many cases, the sets of possible types and relations are organized in a schema or ontology, which defines their interrelations and restrictions of their usage.

With the advent of Linked Data [5], it was proposed to interlink different datasets in the semantic web. By means of interlinking, the collection of could be understood as one large, global knowledge graph (although very heterogeneous in nature). To date, roughly 1,000 datasets are interlinked in the Linked Open Data cloud, with the majority of links connecting identical entities in two datasets [92].

The term Knowledge Graph was coined by Google in 2012, referring to their use of semantic knowledge in Web Search (“Things, not strings”), and is recently also used to refer to Semantic Web knowledge bases such as DBpedia or YAGO. From a broader perspective, any graph-based representation of some knowledge could be considered a knowledge graph (this would include any kind of RDF dataset, as well as description logic ontologies). However, there is no common definition about what a knowledge graph is and what it is not. Instead of attempting a formal definition of what a knowledge graph is, we restrict ourselves to a minimum set of characteristics of knowledge graphs, which we use to tell knowledge graphs from other collections of knowledge which we would not consider as knowledge graphs. A knowledge graph

1. mainly describes real world entities and their interrelations, organized in a graph.
2. defines possible classes and relations of entities in a schema.
3. allows for potentially interrelating arbitrary entities with each other.
4. covers various topical domains.

The first two criteria clearly define the focus of a knowledge graph to be the actual instances (A-box in

2See section 7.2 for a critical discussion.
description logic terminology), with the schema (T-box) playing only a minor role. Typically, this means that the number of instance-level statements is by several orders of magnitude larger than that of schema level statements (cf. Table 1). In contrast, the schema can remain rather shallow, at a small degree of formalization. In that sense, mere ontologies without any instances (such as DOLCE [27]) would not be considered as knowledge graphs. Likewise, we do not consider WordNet [66] as a knowledge graph, since it mainly concerned with common nouns and words and their relations (although a few proper nouns, i.e., instances are also included).

The third criterion introduces the possibility to define arbitrary relations between instances, which are not restricted in their domain and/or range. This is a property which is hardly found in relational databases, which follow a strict schema.

Furthermore, knowledge graphs are supposed to cover at least a major portion of the domains that exist in the real world, and are not supposed to be restricted to only one domain (such as geographic entities). In that sense, large, but single-domain datasets, such as GeoNames, would not be considered a knowledge graph.

Knowledge graphs on the Semantic Web are typically provided using Linked Data [5] as a standard. They can be built using different methods: they can be curated by an organization or a small, closed group of people, crowd-sourced by a large, open group of individuals, or created with heuristic, automatic or semi-automatic means. In the following, we give an overview of existing knowledge graphs, both open and company-owned.

2.1. OpenCyc

OpenCyc is a freely available version of the Cyc knowledge base, which is one of the oldest knowledge graphs, dating back to the 1980s [56]. Rooted in traditional artificial intelligence research, it is a curated knowledge graph, developed and maintained by Cyc-Corp Inc. A semantic web endpoint to OpenCyc also exists, containing links to DBpedia and other LOD datasets.

OpenCyc contains roughly 120,000 instances and 2.5 million facts defined for those instances; its schema comprises a type hierarchy of roughly 45,000 types, and 19,000 possible relations.

2.2. Freebase

Curating a universal knowledge graph is an endeavor which is infeasible for most individuals and organizations. To date, more than 900 person years have been invested in the creation of Cyc [90], with gaps still existing. Thus, distributing that effort on as many shoulders as possible through crowdsourcing is a way taken by Freebase, a public, editable knowledge graph with schema templates for most kinds of possible entities (i.e., persons, cities, movies, etc.). After MetaWeb, the company running Freebase, was acquired by Google, Freebase was shut down on March 31st, 2015.

The last version of Freebase contains roughly 50 million entities and 3 billion facts. Freebase’s schema comprises roughly 27,000 entity types and 38,000 relation types.

2.3. Wikidata

Like Freebase, Wikidata is a collaboratively edited knowledge graph, operated by the Wikimedia foundation that also hosts the various language editions of Wikipedia. After the shutdown of Freebase, the data contained in Freebase is subsequently moved to Wikidata. A particularity of Wikidata is that for each axiom, provenance metadata can be included – such as the source and date for the population figure of a city.

...
To date, Wikidata contains roughly 16 million instances and 66 million statements. Its schema defines roughly 23,000 types and 1,600 relations.

2.4. DBpedia

DBpedia is a knowledge graph which is extracted from structured data in Wikipedia. The main source for this extraction are the key-value pairs in the Wikipedia infoboxes. In a crowd-sourced process, types of infoboxes are mapped to the DBpedia ontology, and keys used in those infoboxes are mapped to properties in that ontology. Based on those mappings, a knowledge graph can be extracted.

The most recent version of the main DBpedia (i.e., DBpedia 2015-04, extracted from the English Wikipedia based on dumps from February/March 2015) contains 4.8 million entities and 176 million statements about those entities. The ontology comprises 735 classes and 2,800 relations.

2.5. YAGO

Like DBpedia, YAGO is also extracted from DBpedia. YAGO builds its classification implicitly from the category system in Wikipedia and the lexical resource WordNet, with infobox properties manually mapped to a fixed set of attributes. While DBpedia creates different interlinked knowledge graphs for each language edition of Wikipedia, YAGO aims at an automatic fusion of knowledge extracted from various Wikipedia language editions, using different heuristics.

The latest release of YAGO, i.e., YAGO3, contains 4.6 million entities and 26 million facts about those types. The schema comprises roughly 488,000 types and 77 relations.

2.6. NELL

While DBpedia and YAGO use semi-structured content as a base, methods for extracting knowledge graphs from unstructured data have been proposed as well. One of the earliest approaches working at web-scale was the Never Ending Language Learning (NELL) project. The project works on a large-scale corpus of web sites and exploits a coupled process which learns text patterns corresponding type and relation assertions, as well as applies them to extract new entities and relations. Reasoning is applied for consistency checking and removing inconsistent axioms. The system is still running today, continuously extending its knowledge base. While not published using Semantic Web standards, it has been shown that the data in NELL can be transformed to RDF and provided as Linked Open Data as well.

In its most recent version (i.e., the 945th iteration), NELL contains roughly 2 million entities and 433,000 relations between those. The NELL ontology defines 285 classes and 425 relations.

2.7. Google’s Knowledge Graph

Google’s Knowledge Graph was introduced to the public in 2012, which was also when the term knowledge graph as such was coined. Google itself is rather secretive about how their Knowledge Graph is constructed; there are only a few external sources that discuss some of the mechanisms of information flow into the Knowledge Graph based on experience. From those, it can be assumed that major semi-structured web sources, such as Wikipedia, contribute to the knowledge graph, as well as structured markup (like schema.org Microdata) on web pages and contents from Google’s online social network Google+.

According to, Google’s Knowledge Graph contains 18 billion statements about 570 million entities, with a schema of 1,500 entity types and 35,000 relation types.

These numbers have been derived from the promotion heatmap at http://rtw.ml.cmu.edu/resources/results/08m/NELL.08m.945.heatmap.html.

E.g., http://www.techwyse.com/blog/search-engine-optimization/seo-efforts-to-get-listed-in-google-knowledge-graph/
2.8. Google’s Knowledge Vault

The Knowledge Vault is another project by Google. It extracts knowledge from different sources, such as text documents, HTML tables, and structured annotations on the Web with Microdata or MicroFormats. Extracted facts are combined using both the extractor’s confidence values, as well as prior probabilities for the statements, which are computed using the Freebase knowledge graph (see above). From those components, a confidence value for each fact is computed, and only the confident facts are taken into Knowledge Vault [21].

According to [21], the Knowledge Vault contains roughly 45 million entities and 271 million fact statements, using 1,100 entity types and 4,500 relation types.

2.9. Yahoo!’s Knowledge Graph

Like Google, Yahoo! also has their internal knowledge graph, which is used to improve search results. The knowledge graph builds on both public data (e.g., Wikipedia and Freebase), as well as closed commercial sources for various domains. It uses wrappers for different sources and monitors evolving sources, such as Wikipedia, for constant updates.

Yahoo’s knowledge graph contains roughly 3.5 million entities and 1.4 billion relations. Its schema, which is aligned with schema.org, comprises 250 types of entities and 800 types of relations. [6]

2.10. Microsoft’s Satori

Satori is Microsoft’s equivalent to Google’s Knowledge Graph[20]. Although almost no public information on the construction, the schema, or the data volume of Satori is available, it has been said to consist of 300 million entities and 800 million relations in 2012, and its data representation format to be RDF[21].

2.11. Facebook’s Entities Graph

Although the majority of the data in the online social network Facebook[22] is perceived as connections between people, Facebook also works on extracting a knowledge graph which contains many more entities. The information people provide as personal information (e.g., their home town, the schools they went to), as well as their likes (movies, bands, books, etc.), often represent entities, which can be linked both to people as well as among each other. By parsing textual information and linking to Wikipedia, the graph also contains links among entities, e.g., the writer of a book. Although not many public numbers about Facebook’s Entities Graph exist, it is said to contain more than 100 Billion connections between entities[23].

2.12. Summary

Table 1 summarizes the characteristics of the knowledge graphs discussed above. It can be observed that the graphs differ in the basic measures, such as the number of entities and relations, as well as in the size of the schema they use, i.e., the number of classes and relations. From these differences, it can be concluded that the knowledge graphs must differ in other characteristics as well, such as average node degree, density, or connectivity.

3. Categorization of Knowledge Graph Refinement Approaches

Knowledge graph refinement methods can differ along different dimensions. For this survey, we distinguish the overall goal of the method, i.e., completion vs. correction of the knowledge graph, the refinement target (e.g., entity types, relations between entities, or literal values), as well as the data used by the approach (i.e., only the knowledge graph itself, or further external sources). All three dimensions are orthogonal.

There are a few research fields which are related to Knowledge Graph refinement: Ontology learning mainly deals with learning a concept level description of a domain, such as a hierarchy (e.g., Cities are Places) [13,63]. Likewise, description logic learning is mainly concerned with refining that concept level description [54]. As stated above, the focus of knowledge graphs, in contrast, is rather the instance (A-box) level, not so much the concept (T-box) level. Following that notion, we consider those works as knowl-

---

Table 1
Overview of Popular Knowledge Graphs. The table depicts the number of instances and facts; as well as the number of different types and relations defined in their schema. *Instances* denotes the number of instances or A-box concepts defined in the graph, *Facts* denotes the number of statements about those instances, *Entity types* denotes the number of different types or classes defined in the schema, and *Relation types* denotes the number of different relations defined in the schema. Microsoft’s Satori and Facebook’s Entities Graph is not shown, because to the best of our knowledge, no detailed recent numbers on the graph are publicly available.

<table>
<thead>
<tr>
<th>Name</th>
<th>Instances</th>
<th>Facts</th>
<th>Types</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia (English)</td>
<td>4,806,150</td>
<td>176,043,129</td>
<td>735</td>
<td>2,813</td>
</tr>
<tr>
<td>YAGO</td>
<td>4,595,906</td>
<td>25,946,870</td>
<td>488,469</td>
<td>77</td>
</tr>
<tr>
<td>Freebase</td>
<td>49,947,845</td>
<td>3,041,722,635</td>
<td>26,507</td>
<td>37,781</td>
</tr>
<tr>
<td>Wikidata</td>
<td>15,602,060</td>
<td>65,993,797</td>
<td>23,157</td>
<td>1,673</td>
</tr>
<tr>
<td>NELL</td>
<td>2,006,896</td>
<td>432,845</td>
<td>285</td>
<td>425</td>
</tr>
<tr>
<td>OpenCyc</td>
<td>118,499</td>
<td>2,413,894</td>
<td>45,153</td>
<td>18,526</td>
</tr>
<tr>
<td>Google’s Knowledge Graph</td>
<td>570,000,000</td>
<td>18,000,000,000</td>
<td>1,500</td>
<td>35,000</td>
</tr>
<tr>
<td>Google’s Knowledge Vault</td>
<td>45,000,000</td>
<td>271,000,000</td>
<td>1,100</td>
<td>4,469</td>
</tr>
<tr>
<td>Yahoo! Knowledge Graph</td>
<td>3,443,743</td>
<td>1,391,054,990</td>
<td>250</td>
<td>800</td>
</tr>
</tbody>
</table>

edge graph refinement approaches which focus on refining the A-box. Approaches that only focus on the T-box are not considered for this survey, however, if the schema or ontology is refined as a means to ultimately improve the A-box, those works are included in the survey.

3.1. Completion vs. Error Detection

There are two main goals of knowledge graph refinement: (a) adding missing knowledge to the graph, i.e., completion, and (b) identifying wrong information in the graph, i.e., error detection. From a data quality perspective, those goals relate to the data quality dimensions free-of-error and completeness [8].

3.2. Target of Refinement

Both completion and error detection approaches can be further distinguished by the targeted kind of information in the knowledge graph. For example, some approaches are targeted towards completing/correcting entity type information, while others are targeted to (either specific or any) relations between entities, or interlinks between different knowledge graphs, or literal values, such as numbers. While the latter can be of any datatype (strings, numbers, dates, etc.), most research focuses on numerical or date-valued literal values.

Another strand of research targets the extension of the schema used by the knowledge graph (i.e., the T-box), not the data (the A-box). However, as discussed above, approaches focusing purely on the schema without an impact on the instance level are not considered for this survey.

3.3. Internal vs. External Methods

A third distinguishing property is the data used by an approach. While internal approaches only use the knowledge graph itself as input, external methods use additional data, such as text corpora. In the widest sense, approaches making use of human knowledge, such as crowdsourcing [11] or games with a purpose [102], can also be viewed as external methods.

4. Categorization of Evaluation Methods

There are different possible ways to evaluate knowledge graph refinement. On a high level, we can distinguish methodologies that use only the knowledge graph at hand, and methodologies that use external knowledge, such as human annotation.

4.1. Partial Gold Standard

One common evaluation strategy is to use a partial gold standard. In this methodology, a subset of graph entities or relations are selected and labeled manually. Other evaluations use external knowledge graphs and/or databases as partial gold standards.
For completion tasks, this means that all axioms that should be there are collected, whereas for correction tasks, a set of axioms in the graph is manually labeled as correct or incorrect. The quality of completion approaches is usually measured in recall, precision, and F-measure, whereas for correction methods, accuracy and/or area under the ROC curve (AUC) are often used alternatively or in addition.

Sourcing partial gold standards from humans can lead to high quality data (given that the knowledge graph and the ontology it uses are not overly complex), but is costly, so that those gold standards are usually small. Exploiting other knowledge graphs based on knowledge graph interlinks is sometimes proposed to yield larger-scale gold standards, but has two sources of errors: errors in the target knowledge graph, and errors in the linkage between the two. For example, it has been reported that 20% of the interlinks between DBpedia and Freebase are incorrect [107], and that roughly half of the owl:sameAs links between knowledge graphs connect two things which are related, but not exactly the same (such as the company Starbucks and a particular Starbucks coffee shop) [33].

4.2. Knowledge Graph as Silver Standard

Another evaluation strategy is to use the given knowledge graph itself as a test dataset. Since the knowledge graph is not perfect (otherwise, refinement would not be necessary), it cannot be considered as a gold standard. However, assuming that the given knowledge graph is already of reasonable quality, we call this method silver standard evaluation, as already proposed in other works [32,44,73].

The silver standard method is usually applied to measure the performance of knowledge graph completion approaches, where it is analyzed how well relations in a knowledge graph can be replicated by a knowledge graph completion method. As for gold standard evaluations, the result quality is usually measured in recall, precision, and F-measure. In contrast to using human annotations, large-scale evaluations are easily possible. The silver standard method is not suitable for error detection, since it assumes the knowledge graph to be correct.

There are two variants of silver standard evaluations: in the more common ones, the entire knowledge graph is taken as input to the approach at hand, and the evaluation is then also carried out on the entire knowledge graph. As this may lead to an overfitting effect (in particular for internal methods), some works also foresee the splitting of the graph into a training and a test partition, which, however, is not as straightforward as, e.g., for propositional classification tasks [71], which is why most papers use the former method. Furthermore, split and cross validation do not fully solve the overfitting effect. For example, if a knowledge graph, by construction, has a bias towards certain kinds of information (e.g., relations are more complete for some classes than for others), approaches overadapting to that bias will be rated better than those which do not (and which may actually perform better).

A problem with this approach is that the knowledge graph itself is not perfect (otherwise, it would not need refinement), thus, this evaluation method may sometimes underrate the evaluated approach. More precisely, most knowledge graphs follow the open world assumption, i.e., an axiom not present in the knowledge graph may or may not hold. Thus, if a completion approach correctly predicts the existence of an axiom missing in the knowledge graph, this would count as a false positive and thus lower precision.

4.3. Retrospective Evaluation

For retrospective evaluations, the output of a given approach is given to human judges for annotation, who then label completions or flagged errors as correct and incorrect. The quality metric is usually accuracy or precision, along with a statement about the total number of completions or errors found with the approach, and ideally also with a statement about the agreement of the human judges.

In many cases, automatic refinement methods lead to a very large number of findings, e.g., lists of tens of thousands of axioms which are potentially erroneous. Thus, retrospective evaluations are often carried out only on samples of the results. For some approaches which produce higher level artifacts – such as error patterns or completion rules – as intermediate results, a feasible alternative is to evaluate those artifacts instead of the actually affected axioms.

While partial gold standards can be reused for comparing different methods, this is not the case for retrospective evaluations. On the other hand, retrospective evaluations may make sense in cases where the interesting class is rare. For example, when evaluating error detection methods, a sample for a partial gold standard from a high-quality graph is likely not to contain a meaningful number of errors. In those cases, retrospective evaluation methodologies are often preferred over partial gold standards.
Another advantage of retrospective evaluations is that they allow a very detailed analysis of an approach’s results. In particular, inspecting the errors made by an approach often reveals valuable findings about the advantages and limitations of a particular approach.

Table 2 sums up the different evaluation methodologies and contrasts their advantages and disadvantages.

4.4. Computational Performance

In addition to the performance w.r.t. correctness and/or completeness of results, computational performance considerations become more important as knowledge graphs become larger. Typical performance measures for this aspect are runtime measurements, as well as memory consumption.

Besides explicit measurement of computational performance, a “soft” indicator for computational performance is whether an approach has been evaluated (or at least the results have been materialized) on an entire large-scale knowledge graph, or only on a subgraph. The latter is often done when applying evaluations on a partial gold standard, where the respective approach is only executed on entities contained in that partial gold standard.

5. Approaches for Completion of Knowledge Graphs

Completion of knowledge graphs aims at increasing the coverage of a knowledge graph. Depending on the target information, methods for knowledge graph completion either predict missing entities, missing types for entities, and/or missing relations that hold between entities.

In this section, we survey methods for knowledge graph completion. We distinguish internal and external methods, and further group the approaches by the completion target.

5.1. Internal Methods

Internal methods use only the knowledge contained in the knowledge graph itself to predict missing information.

5.1.1. Methods for Completing Type Assertions

Predicting a type or class for an entity given some characteristics of the entity is a very common problem in machine learning, known as classification. The classification problem is supervised, i.e., it learns a classification model based on labeled training data, typically the set of entities in a knowledge graph (or a subset thereof) which have types attached. In machine learning, binary and multi-class prediction problems are distinguished. In the context of knowledge graphs, in particular the latter are interesting, since most knowledge graphs contain entities of more than two different types. Depending on the graph at hand, it might be worthwhile distinguishing multi-label classification, which allows for assigning more than one class to an instance (e.g., Arnold Schwarzenegger being both an Actor and a Politician), and single-label classification, which only assigns one class to an instance [100].

For internal methods, the features used for classification are usually the relations which connect an entity to other entities [80,86], i.e., they are a variant of link-based classification problems [31]. For example, an entity which has a director relation is likely to be a Movie.

In [78,79], we propose a probabilistic method, which is based on conditional probabilities, e.g., the probability of a node being of type Actor is high if there are ingoing edges of type cast. Such probabilities are exploited by the SDType algorithm, which is currently deployed for DBpedia and adds around 3.4 million additional type statements to the knowledge graph.

In [95], the use of Support Vector Machines (SVMs) has been proposed to type entities in DBpedia and Freebase. The authors also exploit interlinks between the knowledge graphs and classify instances in one knowledge graph based on properties present in the other, in order to increase coverage and precision. Nickel et al. [72] propose the use of matrix factorization to predict entity types in YAGO.

Since many knowledge graphs come with a class hierarchy, e.g., defined in a formal ontology, the type prediction problem could also be understood as a hierarchical classification problem. Despite a larger body of work existing on methods for hierarchical classification [93], there are, to the best of our knowledge, no applications of those methods to knowledge graph completion.

In data mining, association rule mining [37] is a method that analyzes the co-occurrence of items in itemsets and derives association rules from those co-
### Table 2: Overview on evaluation methods with their advantages and disadvantages

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Gold Standard</td>
<td>highly reliable results</td>
<td>costly to produce</td>
</tr>
<tr>
<td></td>
<td>reusable</td>
<td>balancing problems</td>
</tr>
<tr>
<td>Knowledge Graph as Silver Standard</td>
<td>large-scale evaluation feasible</td>
<td>less reliable results</td>
</tr>
<tr>
<td></td>
<td>subjectiveness is minimized</td>
<td>prone to overfitting</td>
</tr>
<tr>
<td>Retrospective Evaluation</td>
<td>applicable to disbalanced problems</td>
<td>not reusable</td>
</tr>
<tr>
<td></td>
<td>allows for more detailed analysis of approaches</td>
<td>approaches cannot be compared directly</td>
</tr>
</tbody>
</table>

Occurrences. For predicting missing information in knowledge graphs, those methods can be exploited, e.g., in the presence of redundant information. For example, in DBpedia, different type systems (i.e., the DBpedia ontology and YAGO, among others) are used in parallel, which are populated with different methods (Wikipedia infoboxes and categories, respectively). This ensures both enough overlap to learn suitable association rules, as well as a number of entities that only have a type in one of the systems, to which the rules can be applied. In [76], we exploit such association rules to predict missing types in DBpedia based on such redundancies.

In [96], the use of topic modeling for type prediction is proposed. Entities in a knowledge graph are represented as documents, on which Latent Dirichlet Allocation (LDA) [7] is applied for finding topics. By analyzing the co-occurrence of topics and entity types, new types can be assigned to entities based on the topics detected for those entities.

#### 5.1.2. Methods for Predicting Relations

While primary used for adding missing type assertions, classification methods can also be used to predict the existence of relations. To that end, Socher et al. [97] propose to train a tensor neural network to predict relations based on chains of other relations, e.g., if a person is born in a city in Germany, then the approach can predict (with a high probability) that the nationality of that person is German. The approach is applied to Freebase and WordNet. A similar approach is presented in [49], where the authors show that refining such a problem with schema knowledge – either defined or induced – can significantly improve the performance of link prediction. In [48], an approach similar to association rule mining is used to find meaningful chains of relations for relation prediction. Similarly, in [109], an embedding of pairwise entity relations into a lower dimensional space is learned, which is then used to predict the existence of relations in Freebase.

Likewise, association rule mining can be used for predicting relations as well. In [45], the mining of association rules which predict relations between entities in DBpedia from Wikipedia categories is proposed.

#### 5.2. External Methods

External methods use sources of knowledge – such as text corpora or other knowledge graphs – which are not part of the knowledge graph itself. Those external sources can be linked from the knowledge graph, such as knowledge graph interlinks or links to web pages, e.g., Wikipedia pages describing an entity, or exist without any relation to the knowledge graph at hand, such as large text corpora.

##### 5.2.1. Methods for Completing Type Assertions

For type prediction, there are also classification methods that use external data. In contrast to the internal classification methods described above, external data is used to create a feature representation of an entity.

Nuzzolese et al. [74] propose the usage of the Wikipedia link graph to predict types in a knowledge graph using a k-nearest neighbors classifier. Given that a knowledge graph contains links to Wikipedia, interlinks between Wikipedia pages are exploited to create feature vectors, e.g., based on the categories of the related pages. Since links between Wikipedia pages are not constrained, there are typically more interlinks between Wikipedia pages than between the corresponding entities in the knowledge graph.

Apriosio et al. [3] use types of entities in different DBpedia language editions (each of which can be understood as a knowledge graph connected to the others) as features for predicting missing types. The authors use a k-NN classifier with different distance mea-

---

Note that since Wikipedia categories are part of the DBpedia knowledge graph, we consider this approach an internal one.
sures (i.e., kernel functions), such as overlap in article categories. In their setting, a combination of different distance measures is reported to provide the best results.

Another set of approaches uses abstracts in DBpedia to extract definitionary clauses, e.g., using Hearst patterns [35]. Such approaches have been proposed by Gangemi et al. [28] and Klieg [46], where the latter uses abstracts in the different languages in order to increase coverage and precision.

### 5.2.2. Methods for Predicting Relations

Lange et al. [51] learn patterns on Wikipedia abstracts using Conditional Random Fields [50]. A similar approach, but on entire Wikipedia articles, is proposed by [106][25].

Another common method for the prediction of a relation between two entities is distant supervision. Typically, such approaches use large text corpora. As a first step, entities in the knowledge graph are linked to the text corpus by means of Named Entity Recognition [39, 88]. Then, based on the relations in the knowledge graph, those approaches seek for text pattern which correspond to relation types (such as: \( Y \)'s book X being a pattern for the relation author holding between X and Y), and apply those patterns to find additional relations in the text corpus. Such methods have been proposed by Mintz et al. [67] for Freebase, and by Aprosio et al. [4] for DBpedia. In both cases, Wikipedia is used as a text corpus. In [80], a similar setting with DBpedia and two text corpora – the English Wikipedia and an English-language news corpus – is used, the latter showing less reliable results. A similar approach is followed in the RdfLiveNews prototype, where RSS feeds of news companies are used to address the aspect of timeliness in DBpedia, i.e., extracting new information that is either outdated or missing in DBpedia [29].

West et al. [104] propose the use of web search engines to fill gaps in knowledge graphs. Like in the works discussed above, they first discover lexicalizations for relations. Then, they use those lexicalizations to formulate search engine queries for filling missing relation values. Thus, they use the whole Web as a corpus, and combine information retrieval and extraction for knowledge graph completion.

While text is unstructured, some approaches have been proposed that use semi-structured data for completing knowledge graphs. In particular, approaches leveraging on structured data in Wikipedia are found in the literature. Those are most often used together with DBpedia, so that there are already links between the entities and the corpus of background knowledge, i.e., no Named Entity Recognition has to be performed, in contrast to the distant supervision approaches discussed above.

Muñoz et al. [68] propose extraction from tables in Wikipedia. They argue that for two entities co-occurring in a Wikipedia table, it is likely that the corresponding entities should share an edge in the knowledge graph. To fill in those edges, they first extract a set of candidates from the tables, using all possible relations that hold between at least one pair of entities in two columns. Then, based on a labeled subset of that extraction, they apply classification using various features to identify those relations that should actually hold in the knowledge graph.

Ritze et al. [87] extend this approach to arbitrary HTML tables. This requires that not only that pairs of table columns have to be matched to properties in the DBpedia ontology, but also that rows in the table need to be matched to entities in DBpedia. The authors propose an interactive approach to solve those two problems. The approach is evaluated on a gold standard mapping for a sample of HTML tables from the WebDataCommons Web Table corpus [26]. Since such tables can also contain literal values (such as population figures), the approach is capable of completing both relations between entities, and literal values for entities.

In [82], we have proposed the use of list pages in Wikipedia for generating both type and relation assertions in knowledge graphs, based on statistical methods. The idea is that entities appear together in list pages for a reason, and it should be possible to identify that common pattern appearing for the majority of the instance in the list page. For example, instances linked from the page List of Jewish-American Writers should all be typed as Writer and include an edge religion to Jewish, as well as an edge nationality to United States of America. Once such patterns are found for the majority of the list items, they can be applied to the remaining ones to fill gaps in the knowledge graph.

Many knowledge graphs contain links to other knowledge graphs. Those are often created automatically [70]. Interlinks between knowledge graphs can be used to fill gaps in one knowledge graph from infor-
mation defined in another knowledge graph. If a mapping both on the instance and on the schema level is known, it can be exploited for filling gaps in knowledge graphs on both sides.

One work in this direction is presented by Bryl and Bizer [12], where different language versions of DBpedia (each of which can be seen as a knowledge graph of its own) are used to fill missing values in the English language DBpedia (the one which is usually meant when referring to DBpedia).

Dutta et al. [23] propose a probabilistic mapping between knowledge graphs. Based on distributions of types and properties, they create a mapping between knowledge graphs, which can then be used to derive additional, missing facts in the knowledge graphs. To that end, the type systems used by two knowledge graphs are mapped to one another. Then, types holding in one knowledge graph can be used to predict those that should hold in another.

6. Approaches for Error Detection in Knowledge Graphs

Like completion methods discussed in the previous section, methods for identifying errors in knowledge graphs can target various types of information, i.e., type assertions, relations between individuals, literal values, and knowledge graph interlinks.

In this section, we survey methods for detecting errors in knowledge graphs. Like for the previous section, we distinguish internal and external methods, and further group the approaches by the error detection target.

6.1. Internal Methods

Internal methods use only the information given in a knowledge graph to find out whether an axiom in the knowledge graph is plausible or not.

6.1.1. Methods for Finding Erroneous Type Assertions

In contrast to relation assertions, type assertions are most often more correct in knowledge graphs than relation assertions [79]. Hence, methods for finding erroneous type assertions are rather rare. One such method is proposed by Ma et al. [62], who use inductive logic programming for learning disjointness axioms, and then apply those disjointness axioms for identifying potentially wrong type assertions.

6.1.2. Methods for Finding Erroneous Relations

For building Knowledge Vault, Dong et al. use classification to tell relations which should hold in a knowledge graph from those which should not [21]. Like the work by Muñoz et al. discussed above, each relation is used as an instance in the classification problem, with the existence of the relation in the knowledge graph being used as a binary class. This classification is used as a cleansing step after the knowledge extraction process. While the creation of positive training examples from the knowledge graph is quite straightforward, the authors propose the creation of negative training examples by applying a Local Closed World Assumption, assuming that a relation \( r \) between two entities \( e_1 \) and \( e_2 \) does not hold if it is not present in the knowledge graph, and there is a relation \( r \) between \( e_1 \) and another \( e_3 \).

In [79], we have proposed a statistical method for finding wrong statements within a knowledge graph. For each type of relation, we compute the characteristic distribution of subject and object types for edge, i.e., each instantiation of the relation. Edges in the graph whose subject and object type strongly deviate from the characteristic distributions are then identified as potential errors.

Reasoning is a field of study in the artificial intelligence community which deals with automatically deriving proofs for theorems, and for uncovering contradictions in a set of axioms [89]. The techniques developed in this field have been widely adopted in the semantic web community, leading to the development of a larger number of ontology reasoners [19,20,61].

For exploiting reasoning for error checking in knowledge graphs, a rich ontology is required, which defines the possible types of nodes and edges in a knowledge graph, as well as the restrictions that hold on them. For example, if a person is defined to be the capital of a state, this is a contradiction, since capitals are cities, and cities and persons are disjoint, i.e., no entity can be a city and a person at the same time. Reasoning is often used at the building stage of a knowledge graph, i.e., when new axioms are about to be added. For example, NELL and PROSPERA perform reasoning at that point to determine whether the new axiom is plausible or not, and discard implausible ones [14,69]. For real-world knowledge graphs, reasoning can be difficult due to the presence of errors and noise in the data [85,42].

Works using reasoning as a refinement operation for knowledge graphs have also been proposed. However, many knowledge graphs, such as DBpedia, come with
ontologies that are not rich enough to perform reasoning for inconsistency detection – for example, they lack class disjointness assertions needed for an inference as in the example above. Therefore, approaches exploiting reasoning are typically used in conjunction with methods for enriching ontologies, such as statistical methods, as proposed in [41] and [99], or association rule mining, as in [52]. In all of those works, the ontology at hand is enriched with further axioms, which can then be used for detecting inconsistencies. For example, if a reasoner concludes that an entity should both be a person and an organization, and from the enrichment steps a disjointness axiom between the two types added, a reasoner can state that one out of a few axioms in the knowledge graph has to be wrong.

In [83], a light-weight reasoning approach is proposed to compare actual and defined domains and ranges of relations in a knowledge graph schema. The authors propose a set of heuristics for fixing the schema if the actual and the defined domain or range strongly deviate.

6.1.3. Methods for Finding Erroneous Literal Values

Outlier detection or anomaly detection methods deal aim at identifying those instances in a dataset that deviate from the majority from the data, i.e., that follow different characteristics than the rest of the data [15, 38].

As outlier detection in most cases deals with numeric data, numeric literals are a natural target for those methods. In [105], we have proposed the application of different univariate outlier detection methods (such as interquartile range or kernel density estimation) to DBpedia. Although outlier detection does not necessarily identify errors, but also natural outliers (such as the population of very large cities), it has been shown that the vast majority of outliers identified are actual errors in DBpedia, mostly resulting from mistakes made when parsing strings using various number formats and units of measurement.

To lower the influence of natural outliers, an extension of that approach has been presented in [24], where the instance set under inspection is first split into smaller subsets. For example, population values are inspected for countries, cities, and towns in isolation, thus, the distributions are more homogenous, which leads to a higher precision in error identification. Furthermore, the approach foresees cross-checking found outliers with other knowledge graphs in order to further reduce the influence of natural outliers, which makes it a mixed approach with both an internal and an external component.

6.1.4. Methods for Finding Erroneous Knowledge Graph Interlinks

In [77], we have shown that outlier detection is not only applicable to numerical values, but also to other targets, such as knowledge graph interlinks. To that end, the interlinks are represented as a multi-dimensional feature vector, e.g., with each type of the respective entity in both knowledge graphs being a binary feature. In that feature space, standard outlier detection techniques such as Local Outlier Factor [11] or cluster-based outlier detection [34] can be used to assign outlier scores. Based on those scores, implausible links, such as an owl:sameAs assertion between a person and a book, can be identified based only on the overall distribution of all links, where such a combination is infrequent.

The work in [57] tries to learn arithmetic relations between attributes, e.g., lessThan or greaterThan, using probabilistic modeling. For example, the birth date of a person must be before her death date, the total area of a country must be larger than the area covered by water, etc. Violations of those relations are then used to identify errors.

6.2. External Methods

Purely automatic external methods for error detection in knowledge graphs are still rare. Semi-automatic approaches, which exploit human knowledge, have also been proposed.

6.2.1. Methods for Finding Erroneous Relations

Most external methods are targeted on finding erroneous relations in knowledge graphs. One of the few works is DeFacto [53]. The system uses a database of lexicalizations for predicates in DBpedia, based on those lexicalizations, it transforms statements in DBpedia to natural language sentences, and uses a web search engine to find web pages containing those sentences. Statements with no or only very few web pages supporting the corresponding sentences are then assigned a low confidence score.

Note that an error here is not a single statement, but a pair of statements that cannot be true at the same time. Thus, the approach does not trivially lead to a fully automatic repairing mechanism (unless both statements are removed, which means that most likely, one correct statement is removed as well).
Apart from fully automatic methods, semi-automatic methods involving users have been proposed for validating knowledge graphs, such as crowdsourcing with microtasks [1]. In order to increase the user involvement and motivation, game-based approaches (i.e., games with a purpose) have been proposed [36,47,94,102]. In a wider sense, those can also be viewed as external methods, with the human in the loop being the external source of information.

Generally, a crucial issue with human computation is the size of web scale knowledge graphs. In [79], it has been argued that the time needed to validate the entire DBpedia knowledge graph with the crowdsourcing approach proposed in [1] – extrapolating the task completion times reported – would take more than 3,000 years. To overcome such scaling problems, we have recently proposed a clustering of inconsistencies identified by automatic means, which allows to present only representative examples to the human for inspection [81]. We have shown that most of the clusters have a common root cause in the knowledge graph construction (e.g., a wrong mapping rule or a programming error), so that by inspecting only a few dozen examples (and addressing the respective root causes), millions of statements can be corrected.

6.2.2. Methods for Finding Erroneous Literal Values

While most of the crowdsourcing approaches above are focusing on relations in the knowledge graph, the work in [11] uses similar mechanisms for validating knowledge graph interlinks and literal values.

In [59], an automatic approach using knowledge graph interlinks for detecting wrong numerical values is proposed. The authors exploit links between identical resources and apply different matching functions between properties in the individual sources. Facts in one knowledge graph are assumed to be wrong if multiple other sources have a consensus for a conflicting fact (e.g., a radically different population figure).

7. Findings from the Survey

From the survey in the last two sections, we can observe that there are quite a few works proposed for knowledge graph refinement, both for automatic completion and for error detection. Tables 2 to 5 sum up the results from the previous section.

By taking a closer look at those results, we can derive some interesting findings, both with respect to the approaches, as well as with respect to evaluation methodologies.

7.1. Approaches

A first interesting observation is that our distinguishing into completion and error detection is a strict one. That is, there exist no approaches which do both completion and correction at the same time. The only exception we found is the pairing of the two approaches SDType and SDValidate [79], which are two closely related algorithms which share the majority of the computations and can output both completion axioms and errors.

For many of the approaches, it is not obvious why they were only used for one purpose. For example, many of the probabilistic and NLP-based completion approaches seek for evidence for missing axioms, e.g., by means of scanning text corpora. Similarly, many completion approaches ultimately compute a confidence score, which is then combined with a suitable threshold for completing a knowledge graph. In principle, they could also be used for error detection by flagging axioms for which no or only little evidence was found, or those with a low confidence score, as wrong.

Furthermore, in particular in the machine learning area, approaches exist which can be used for simultaneously creating a predictive model and creating weights for pieces of information. For example, random forests can assign weights to attributes [58], whereas boosting assign weights to instances [25], which can also be interpreted as outlier scores [16]. Such approaches could be a starting point for developing methods for simultaneous completion and error detection in knowledge graphs.

Along the same lines, there are hardly any among the error detection approaches which are also suitable for correcting errors, i.e., suggest fixes for the errors found. Here, a combination between completion and error detection methods could be of great value: once an error is detected, the erroneous axiom(s) could be removed, and a correction algorithm could try to find a new (and, in the best case, more accurate) replacement for the removed axiom(s).

In addition to the strict separation of completion and correction, we also observe that most of the approaches focus on only one target, i.e., types, relations, literals, etc. Approaches that simultaneously try to complete or correct, e.g., type and relation assertions in a knowledge graph, are also quite rare.

For the approaches that perform completion, all works examined in this survey try to add missing types or for or relations between existing entities in the knowledge graph. In contrast, we have not observed
<table>
<thead>
<tr>
<th>Paper</th>
<th>Target</th>
<th>Type</th>
<th>Methods and Sources</th>
<th>Knowledge Graph(s)</th>
<th>Eval.</th>
<th>Metrics</th>
<th>Whole Comp.</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paulheim [76]</td>
<td>T</td>
<td>I</td>
<td>Association Rule Mining</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Nickel et al. [72]</td>
<td>T</td>
<td>I</td>
<td>Matrix Factorization</td>
<td>YAGO</td>
<td>KG (CV)</td>
<td>P/R, AUC-PR</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Paulheim/Bizer [78,79]</td>
<td>T</td>
<td>I</td>
<td>Likelihood based</td>
<td>DBpedia, OpenCyc, Nell</td>
<td>KG, RE</td>
<td>P/R, T</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Sleeman et al. [96]</td>
<td>T</td>
<td>I</td>
<td>Topic Modeling</td>
<td>DBpedia</td>
<td>KG (SV)</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Nuzzolese et al. [74]</td>
<td>T</td>
<td>E</td>
<td>different machine learning methods, Wikipedia link graph</td>
<td>DBpedia</td>
<td>PG (a)</td>
<td>P/R</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Gangemi et al. [28]</td>
<td>T</td>
<td>E</td>
<td>NLP on Wikipedia abstracts</td>
<td>DBpedia</td>
<td>PG (n/a)</td>
<td>P/R</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Kliegr [46]</td>
<td>T</td>
<td>E</td>
<td>NLP on Wikipedia abstracts</td>
<td>DBpedia</td>
<td>PG (n/a)</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Aprosio et al. [3]</td>
<td>T</td>
<td>E</td>
<td>K-NN classification, different DBpedia language editions</td>
<td>DBpedia</td>
<td>PG (n/a)</td>
<td>P/R</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Dutta et al. [23]</td>
<td>T</td>
<td>E</td>
<td>Knowledge graph fusion (statistical)</td>
<td>NELL, DBpedia</td>
<td>PG (a)</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Sleeman/Finin [95]</td>
<td>T</td>
<td>I, E</td>
<td>SVM, using other KGs</td>
<td>DBpedia, Freebase, Arnetminer</td>
<td>KG</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Socher et al. [97]</td>
<td>R</td>
<td>I</td>
<td>Neural Tensor Network</td>
<td>WordNet, Freebase</td>
<td>KG (SV)</td>
<td>A</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Krompaß et al. [49]</td>
<td>R</td>
<td>I</td>
<td>Latent variable models</td>
<td>DBpedia, Freebase, YAGO</td>
<td>KG (SV)</td>
<td>T</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Kolthoff and Dutta [48]</td>
<td>R</td>
<td>I</td>
<td>Rule mining</td>
<td>DBpedia, YAGO</td>
<td>KG</td>
<td>P</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Zhao et al. [109]</td>
<td>R</td>
<td>I</td>
<td>Learning Embeddings</td>
<td>WordNet, Freebase</td>
<td>KG</td>
<td>P</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
Table 4: Overview of knowledge graph completion approaches (part 2). Abbreviations used: Target (T=Types, R=Relations), Type (I=internal, E=external), Evaluation (RE=retrospective evaluation, PG=partial gold standard, either available (a) or unavailable (n/a), KG=evaluation against knowledge graph, SV=split validation, CV=cross validation), Metrics (P/R=precision and recall, A=accuracy, AUC-PR=area under precision-recall-curve, T=total new statements). Comp.: evaluation or materialization carried out on whole knowledge graph or not. Performance: computational performance reported or not.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Target</th>
<th>Type</th>
<th>Methods and Sources</th>
<th>Knowledge Graph(s)</th>
<th>Eval.</th>
<th>Metrics</th>
<th>Whole</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galárraga et al. [26]</td>
<td>R</td>
<td>I</td>
<td>Association Rule Mining</td>
<td>YAGO, DBpedia</td>
<td>KG, RE</td>
<td>A, T</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Kim et al. [45]</td>
<td>R</td>
<td>I</td>
<td>Association Rule Mining</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Bryl/Bizer [12]</td>
<td>R</td>
<td>E</td>
<td>Fusion of DBpedia language editions</td>
<td>DBpedia</td>
<td>PG (a)</td>
<td>A</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Apriosio et al. [4]</td>
<td>R</td>
<td>E</td>
<td>Distant supervision and NLP techniques on Wikipedia text corpus</td>
<td>DBpedia</td>
<td>KG</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Lange et al. [51]</td>
<td>R</td>
<td>E</td>
<td>Pattern learning on Wikipedia abstracts</td>
<td>(DBpedia)</td>
<td>KG (CV)</td>
<td>P/R</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Wu et al. [106]</td>
<td>R</td>
<td>E</td>
<td>Pattern learning on Wikipedia articles, Web search</td>
<td>(DBpedia)</td>
<td>KG (CV)</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Mintz et al. [67]</td>
<td>R</td>
<td>E</td>
<td>Distant supervision and NLP techniques on Wikipedia text corpus</td>
<td>Freebase</td>
<td>KG</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Gerber/Ngonga Ngomo [30]</td>
<td>R</td>
<td>E</td>
<td>Distant supervision and NLP techniques on two corpora</td>
<td>DBpedia</td>
<td>KG, RE</td>
<td>P/R</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Gerber et al. [29]</td>
<td>R</td>
<td>E</td>
<td>NLP and pattern mining on RSS news feeds</td>
<td>DBpedia</td>
<td>PG</td>
<td>P/R, A</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>West et al. [104]</td>
<td>R</td>
<td>E</td>
<td>search engines</td>
<td>Freebase</td>
<td>KG</td>
<td>P/R, rank</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Mu noz et al. [68]</td>
<td>R</td>
<td>E</td>
<td>Statistical measures, machine learning, using Wikipedia tables</td>
<td>DBpedia</td>
<td>PG (a)</td>
<td>P/R</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Ritze et al. [87]</td>
<td>R, L</td>
<td>E</td>
<td>Schema and instance matching using HTML tables</td>
<td>DBpedia</td>
<td>PG (a)</td>
<td>P/R</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Paulheim/Ponzetto [82]</td>
<td>T, R</td>
<td>E</td>
<td>Statistical measures, Wikipedia list pages</td>
<td>DBpedia</td>
<td>–</td>
<td>–</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Paper</td>
<td>Target</td>
<td>Type</td>
<td>Methods and Sources</td>
<td>Knowledge Graph(s)</td>
<td>Eval.</td>
<td>Metrics</td>
<td>Whole</td>
<td>Comp.</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>------</td>
<td>---------------------</td>
<td>--------------------</td>
<td>-------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Ma et al. [62]</td>
<td>T</td>
<td>I</td>
<td>Reasoning, Association Rule Mining</td>
<td>DBpedia, Zhishi.me</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Dong et al. [21]</td>
<td>R</td>
<td>I</td>
<td>Classification</td>
<td>Knowledge Vault</td>
<td>KG (SV)</td>
<td>AUC-PR</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Li et al. [57]</td>
<td>L</td>
<td>I</td>
<td>Probabilistic Modeling</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Nakashole et al. [69]</td>
<td>R</td>
<td>I</td>
<td>Reasoning</td>
<td>Prospera</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Paulheim/Bizer [79]</td>
<td>R</td>
<td>I</td>
<td>Probabilistic</td>
<td>DBpedia, NELL</td>
<td>RE</td>
<td>P</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Lehmann et al. [53]</td>
<td>R</td>
<td>E</td>
<td>Text pattern induction, Web search engines</td>
<td>DBpedia</td>
<td>KG (CV)</td>
<td>P/R, ROC, RMSE</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Töpper et al. [99]</td>
<td>R, S</td>
<td>I</td>
<td>Statistical methods, Reasoning</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Jang et al. [41]</td>
<td>R</td>
<td>I</td>
<td>Statistical methods</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, R</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Lehmann/Bühmann [52]</td>
<td>R, T</td>
<td>I</td>
<td>Reasoning, ILP</td>
<td>DBpedia, Open Cyc, seven smaller ontologies</td>
<td>RE</td>
<td>A</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Wienand/Paulheim [105]</td>
<td>L</td>
<td>I</td>
<td>Outlier Detection</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Fleischhacker et al. [24]</td>
<td>L</td>
<td>I, E</td>
<td>Outlier Detection and Data Fusion with other KG</td>
<td>DBpedia, NELL</td>
<td>RE</td>
<td>AUC-ROC</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Liu et al. [59]</td>
<td>L</td>
<td>E</td>
<td>Matching to other KGs</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Paulheim [77]</td>
<td>I</td>
<td>I</td>
<td>Outlier Detection</td>
<td>DBpedia + two linked graphs</td>
<td>PG (a)</td>
<td>P/R, ROC</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Acosta et al. [1]</td>
<td>L, I</td>
<td>E</td>
<td>Crowdsourcing</td>
<td>DBpedia</td>
<td>PG (a)</td>
<td>P</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Waitelonis et al. [102]</td>
<td>R</td>
<td>E</td>
<td>Quiz game</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Hees et al. [36]</td>
<td>R</td>
<td>E</td>
<td>Two-player game, Two-player game</td>
<td>DBpedia</td>
<td>RE</td>
<td>T</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Paulheim and Gangemi [81]</td>
<td>R</td>
<td>E</td>
<td>Reasoning, clustering, human inspection</td>
<td>DBpedia</td>
<td>RE</td>
<td>P, T</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
any approaches which populate the knowledge graph with new entities. Here, entity set expansion methods, which have been deeply investigated in the NLP field [75,91,103], would be an interesting fit to further increase the coverage of knowledge graphs, especially for less well-known long tail entities.

Another interesting observation is that, although the discussed works address knowledge graphs, only very few of them are, in the end, genuinely graph-based approaches. In many cases, simplistic transformations to a propositional problem formulation are taken. Here, methods from the graph mining literature still seek their application to knowledge graphs. In particular, for many of the methods applied in the works discussed above – such as outlier detection or association rule mining – graph-based variants have been proposed in the literature [2,43]. Likewise, graph kernel functions – which can be used in Support Vector Machines as well as other machine learning algorithms – have been proposed for RDF graphs [40,60,18] and hence could be applied to many web knowledge graphs.

7.2. Evaluation Methodologies

For evaluation methodologies, our first observation is that there are various different evaluation metrics being used in the papers examined. There is a clear tendency towards precision and recall (or precision and total number of statements for retrospective evaluations) are the most used metrics, with others – such as ROC curves, accuracy, or Root Mean Squared Error – occasionally being used as well.

With respect to the overall methodology, the results are more mixed. Evaluations using the knowledge graph as a silver standard, retrospective evaluations, and evaluations based on partial gold standards appear at equal frequency, with retrospective validations mostly used for error detection. The latter is not too surprising, since due to the high quality of most knowledge graphs used for the evaluations, partial gold standards based on random samples are likely to contain only few errors. For partial gold standards, it is crucial to point out that the majority of authors make those partial gold standards public [28], which allows for replication and comparison.

The major knowledge graph used in the evaluations is DBpedia. This, in principle, makes the results comparable to a certain extent, although roughly each year, a new version of DBpedia is published, so that papers from different years are likely to be evaluated on slightly different knowledge graphs.

That being said, we have observed that roughly two out of three approaches evaluated on DBpedia are only evaluated on DBpedia. Along the same lines, about half of the approaches reviewed in this survey are only evaluated on one knowledge graph. This, in many cases, limits the significance of the results. For some works, it is clear that they can only work on a specific knowledge graph, e.g., DBpedia, by design, e.g., since they exploit the implicit linkage between a DBpedia entity and the corresponding Wikipedia page.

As discussed in section 2, knowledge graphs differ heavily in their characteristic. Thus, for an approach evaluated on only one graph, it is unclear whether it would perform similarly on another knowledge graph with different characteristics, or whether it exploits some (maybe not even obvious) characteristics of that knowledge graph, and/or overfits to particular characteristics of that graph.

Last, but not least, we have observed that only a minority of approaches have been evaluated on a whole, large-scale knowledge graph. Moreover, statements about computational performance are only rarely included in the corresponding papers [29]. In the age of large-scale knowledge graphs, we think that this is a dimension that should not be neglected.

In order to make future works on knowledge graph evolution comparable, it would be useful to have a common selection of benchmarks. This has been done in other fields of the semantic web as well, such as for schema and instance matching [22], reasoning [8], or question answering [17]. Such benchmarks could serve both for comparison in the qualitative as well as the computational performance.

8. Conclusion

In this paper, we have presented a survey on knowledge base refinement methods. We distinguish completion from error detection, and internal from external methods. We have shown that a larger body of

---

28 For this survey, we counted a partial gold standard as public if there was a working download link in the paper, but we did not make any additional efforts to search for the gold standard, such as contacting the authors.

29 Even though we were relaxed on this policy and counted also informal statements about the computational performance as a performance evaluation.
works exist which apply different methods, ranging from techniques from the machine learning field to NLP related techniques.

The survey has revealed that there are, at the moment, rarely any approaches which simultaneously try to improve completeness and correctness of knowledge graphs, and usually only address one target, such as type or relation assertions, or literal values. Holistic solutions which simultaneously improve the quality of knowledge graphs in many different aspects are currently not observed.

Looking at the evaluation methods, the picture is quite diverse. Different methods are applied, using either the knowledge graph itself as silver standard, using a partial gold standard, or performing a retrospective evaluation, are about equally distributed. Furthermore, approaches are often only evaluated on one specific knowledge graph. This makes it hard to compare approaches and make general statements on their relative performance.

In addition, scalability issues are only rarely addressed by current research works. In the light of the advent of web-scale knowledge graphs, however, this is an aspect which will be of growing importance.

To sum up, this survey shows that automatic knowledge graph refinement is a relevant and flowering research area. At the same time, this survey has pointed out some uncharted territories on the research map, which we hope will inspire researchers in the area.

References

[20] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avan-


[40] Yi Huang, Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A scalable kernel approach to learning in semantic graphs with applications to linked data. In 1st Workshop on Mining the Future Internet, 2010.


Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods


Shuangyan Liu, Mathieu DâÁZaquin, and Enrico Motta. Towards Linked Data Fact Validation through Measuring Consensus. 2015.


[108] Antoine Zimmermann, Christophe Gravier, Julien Subercaze, and Quentin Cruzzile. Nell2RDF: Read the Web, and Turn it into RDF. In Knowledge Discovery and Data Mining meets Linked Open Data, pages 2–8, 2013.
Dear editors, dear reviewers,

we would like to thank all the reviewers for the numerous constructive comments. Based on those comments, we have prepared a carefully revised version of the paper.

The following changes have been made based on the reviews:

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Comment</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 R1-R3</td>
<td>Define &quot;Knowledge Graph&quot;</td>
<td>Although we do not come up with complete definition, we define some criteria a knowledge graph should fulfill, and discuss how some kinds of knowledge bases (e.g., relational databases, general ontologies) do not meet those criteria and are therefore not considered to be KGs.</td>
</tr>
<tr>
<td>2 R1</td>
<td>Include Facebook, Microsoft and Yahoo! KGs</td>
<td>The survey of knowledge graphs now also covers those commercial developments.</td>
</tr>
<tr>
<td>3 R1</td>
<td>Define boundary between KG graph construction and refinement</td>
<td>Two additional paragraphs have been added to the introduction, defining the boundary and more crisply setting the scope of this survey.</td>
</tr>
<tr>
<td>4 R1</td>
<td>Include non-automatic methods</td>
<td>Non-automatic methods like crowdsourcing and games with a purpose have been included as &quot;full members&quot; of the survey.</td>
</tr>
<tr>
<td>5 R1</td>
<td>provide evidence that Google+ contents are used in KG</td>
<td>A footnote with a source (and a short caveat) has been inserted.</td>
</tr>
<tr>
<td>6 R1</td>
<td>Add explanations for the columns in Table 1</td>
<td>Explanations have been added to the table caption.</td>
</tr>
<tr>
<td>7 R1</td>
<td>Add sources for the data in Table 1</td>
<td>Precise sources for all numbers have been added in the text in section 2.</td>
</tr>
<tr>
<td>8 R1</td>
<td>Discuss Knowledge Vault in the text</td>
<td>For each of the knowledge graphs in Table 1, a textual description has been included</td>
</tr>
<tr>
<td>9 R1</td>
<td>Define &quot;genuine semantic web knowledge graph&quot;</td>
<td>The expression has been rephrased for clarity.</td>
</tr>
<tr>
<td>10 R1</td>
<td>Rephrase &quot;targeted kind of information&quot;</td>
<td>After a long while of thinking about an easy-to-grasp phrasing, this has been changed to &quot;Target of refinement&quot;.</td>
</tr>
<tr>
<td>11 R1</td>
<td>Explain silver standard</td>
<td>An explanation has been added.</td>
</tr>
<tr>
<td>12 R1</td>
<td>use &quot;retrospective evaluation&quot; instead of &quot;ex post&quot;</td>
<td>The term has been replaced.</td>
</tr>
<tr>
<td>13 R1</td>
<td>Rethink organization of section 5 and 6 – why introduce another taxonomy of approaches?</td>
<td>The presentation of the works is now aligned with the taxonomy in section 3.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
| 14 | R1 | Check whether the work in 6.4.1 is actually on knowledge graphs
We have checked whether the approaches have been evaluated at least on one dataset following our criteria for what a knowledge graph is (see 1). As a consequence, four works have been discarded from the survey. |
| 15 | R2 | Clarify how DLs and Knowledge Graphs differ, distinguish against DL learning literature
The characterization of knowledge graphs (see 1) now explicitly differentiates between ontologies/DLs and KGs. The difference to DL learning is now explained in the beginning of section 3. |
| 16 | R2 | Rephrase statement and replace citation for the origin of reasoning, which is not a genuine SW technique
The corresponding paragraph has been rephrased and augmented with further citations. |
| 17 | R3 | Explain difference to ontology learning
The difference to ontology learning is now explained in the beginning of section 3. Due to the explication of those differences (see also 15), a few works that have been concerned with ontology learning have been removed from the survey. |
| 18 | R3 | Include four papers by AKSW group
Three of the papers have been included in the survey. The ISWC 2013 paper on ORE has not been included since the scope of the survey has been slightly narrowed following a comment by R1, see 14 |

In addition to the reviewers' comments, the following changes were made
- Section 2 has been restructured, also briefly explaining how the knowledge graphs are constructed
- A tabular summary of evaluation approaches has been added (table 2)
- Numbers for knowledge graphs have been updated wherever newer numbers were available (e.g., for the most recent DBpedia release)
- Besides the sources named by reviewer 3, a few more additional very recent publications (that had not been published at the time the last version was submitted) have been included in the survey
- The whole paper was proofread for typos and grammar problems

We are looking forward to your feedback for the revised version.

Best regards,

Heiko Paulheim