YAGO+F - Aligning Freebase with the YAGO Ontology

Elena Demidova, Irina Oelze, and Wolfgang Nejdl

L3S Research Center and Leibniz University of Hanover, Germany
{demidova, oelze, nejdl}@L3S.de

Technical Report, May 2013

Abstract. Linked Open Data (LOD) has emerged as the de-facto standard for publishing data on the Web. The cross-domain large scale Freebase and YAGO datasets represent central hubs and reference points for the LOD cloud. Freebase is an open-world large scale dataset which contains about 22 million entities and more than 350 million facts in more than 100 domains. The scale of Freebase makes it difficult for the users to get an overview of the data and efficiently retrieve the desired information. Integration of Freebase with the YAGO ontology that contains more than 360,000 concepts enables us to provide more semantic information for Freebase and to facilitate novel applications, such as efficient query construction, over large scale data. In this paper we analyze the structure of YAGO in more depth and show how to match YAGO and Freebase categories. The new YAGO+F structure that results from our matching tightly connects both datasets and provides an important next step to systematically interconnect LOD subcollections. We make our YAGO+F structure available online in the hope that it can provide a good starting point for future applications, which can build upon a wide variety of Freebase data clearly arranged in the semantic categories of YAGO.

1 Introduction and Motivation

Linked Open Data (LOD) is a method of publishing on the Semantic Web that connects pieces of structured data, information, and knowledge to build the Web of Data. This way of publishing enables data from distributed Web sources to be connected and queried, potentially enabling a wide variety of applications to take advantage of distributed information and knowledge. The number of datasets included in LOD has grown exponentially over the last years, currently including 295 datasets with more than 31 billion entities and facts from a variety of domains[1]. The cross-domain Freebase and YAGO datasets represent central hubs and reference points for the LOD cloud.

Freebase [4] is a large scale dataset that contains about 22 million entities that belong to more than 7,500 categories in more than 100 domains. In addition, Freebase contains more than 350 million facts about these entities. The users of Freebase can collaboratively create, structure and maintain database content over an open platform. In addition, automatic imports from the external data sources such as Wikipedia , MusicBrainz[2], and others enable further growth of the data size. Given the scale of Freebase, it becomes crucial to provide effective and efficient structures that give users a

[1] The LOD cloud diagram: **http://richard.cyganiak.de/2007/10/lod**
[2] **http://musicbrainz.org**

quick and informative overview of the data available. Ontologies are typically used for organizing large scale information and knowledge in a wide variety of domains. The YAGO ontology [15] is a lexical resource that contains entities, categories, and their relations automatically extracted from Wikipedia. YAGO unifies the extracted Wikipedia categories with the concepts of the WordNet thesaurus [9], and arranges these concepts into a taxonomic hierarchy. Among other ontologies, YAGO is a natural choice for organizing Freebase data, as both YAGO and Freebase share a large number of entities originating from Wikipedia.

Fig. 1 exemplifies the elements of the Freebase and YAGO hierarchies. For example, in YAGO an instance *"Stephen King"* is associated with the leaf Wikipedia categories *"American Novelist"*, *"Writers from Maine"*, and *"People from Country Dublin"*, which are the sub-concepts of *"Writer"*, *"Communicator"*, *"Person"*, and *"Entity"* in WordNet. In Freebase, the same instance *"Stephen King"* belongs to the category *"Author"* located in the *"Books"* domain that is further categorized in the *"Arts & Entertainment"* top level domain of Freebase.

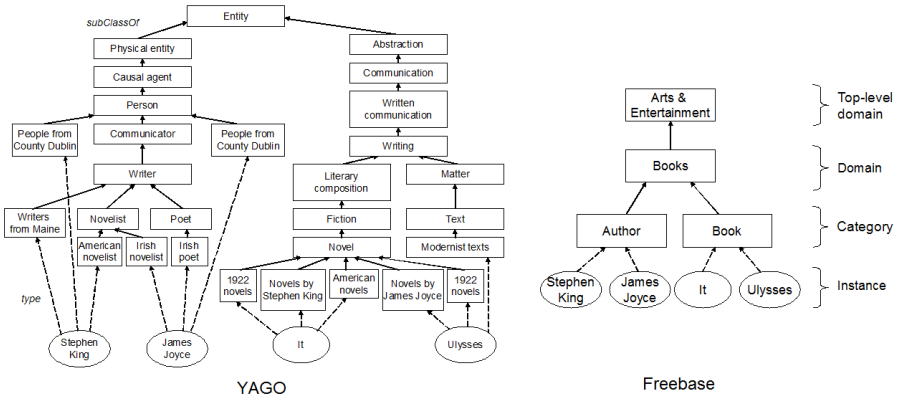

Fig. 1: YAGO and Freebase Structure

In this paper we focus on the problem of enrichment of Freebase categories and entities with the conceptual categories of YAGO. Our contributions are as follows: First, we analyze the initial structure of the large scale YAGO hierarchy which contains more than 360,000 concepts. Second, we describe a matching algorithm, which identifies the most suitable YAGO category for every category of Freebase. Third, we compare the structure of the sub-hierarchy of YAGO that is relevant for the Freebase mapping, which we call YAGO+F, with the original YAGO hierarchy and show that only a small part of the YAGO ontology is required to describe a large scale real-world multi-domain dataset like Freebase. Finally, we evaluate and discuss the matching quality and make the matching results available to the community[3].

[3] The YAGO+F mapping is available at: **`http://iqp.l3s.uni-hannover.de`**

The advantages of the YAGO+F mapping are twofold. First, a hierarchical structure of YAGO can enable an efficient navigation over the large-scale Freebase dataset. To this extent, in FreeQ [5], we presented a novel query construction approach that enables novice users to create structured queries over Freebase and takes advantage of YAGO+F to shorten the process of user-computer interaction. Second, YAGO+F increases the number of entities available in YAGO by more than an order of magnitude. These additional entities can possibly be used in the future work to enhance YAGO-based applications such as e.g. question answering systems (see YAGO-QA [1]).

2 Concepts and Instances in YAGO

To enable an effective matching between YAGO and Freebase, we perform several steps. In this section, we first analyze the distribution of concepts and instances within the original YAGO hierarchy. Then, we consider an overlap of the YAGO and Freebase instances and examine the distribution of the instances shared between YAGO and Freebase within the YAGO ontology.

2.1 Concept Structure of YAGO

The common elements of the YAGO ontology are the concepts, e.g. *"Entity"* and *"Person"*. The concepts represent semantic categories and are hierarchically organized using the *"subClassOf"* relation. A concept can be associated with a set of instances, e.g. the concept *"Person"* is associated with the instances *"Stephen King"* and *"James Joyce"*. The relation *"type"* links together a concept with its associated instances.

To facilitate our analysis, we assign each concept within the YAGO hierarchy a *depth* value. The concepts at the top level of the hierarchy (depth=0) do not possess any parent concepts. Then, the depth of the concept C is determined as the length of the path from C to the top level concept that is associated with C using the *"subClassOf"* relation.

A YAGO instance can be associated with multiple YAGO concepts connected using the *"subClassOf"* relation. In order to differentiate the most specific concepts that have instances directly assigned to them from the concepts that are only indirectly connected to instances, we introduce the notion of a *leaf category*. A *leaf category* L is a category that is associated with an instance I and has the highest depth value across all the categories associated with I and connected to L using the *"subClassOf"* relation. For example, an instance *"Alexander the Great"* is associated with the Wikipedia leaf categories *"4th-century BC Greek people"*, *"Macedonian monarchs"*, *"Ancient Macedonian generals"*, and *"Monarchs of Persia"*, which are connected to the more general WordNet concepts *"Person"*, *"Head of state"*, and *"General"*.

As of March 2011, YAGO possesses 361,211 semantic categories that are organized in a hierarchical structure with 20 levels [10]. The backbone of YAGO is build from the Wikipedia and WordNet categories. YAGO includes 292,070 Wikipedia categories located at the depth of 1-19 of the hierarchy. These categories can be very specific and include e.g. *"Burials at Kensico Cemetery", "1729 essays", "Multidirectional shooters", "Burials at Montmartre Cemetery", "Paris", "Founders of utopian communities", "59 crimes", and "1st-century executions"*. Further, YAGO includes 68,446 more general

Table 1: Distribution of Categories in YAGO

Depth	Categories in YAGO						Leaf Categories in YAGO				
	Total	Total, n.	WordNet	Wikipedia	YAGO	Geo	Total	WordNet	Wikipedia	YAGO	Geo
0	25	0.00007	24	0	1	0	2	1	0	1	0
1	221	0.00061	16	180	22	3	123	0	113	10	0
2	79	0.00022	35	10	6	28	1	0	1	0	0
3	3,726	0.01032	419	3,286	17	4	3,293	19	3,271	1	2
4	42,979	0.11899	2,465	40,390	5	119	40,262	60	40,134	0	68
5	27,635	0.07651	5,926	21,561	1	147	21,586	165	21,333	0	88
6	67,928	0.18806	9,819	57,950	0	159	57,732	298	57,352	0	82
7	91,455	0.25319	14,954	76,388	0	113	75,234	317	74,857	0	60
8	63,015	0.17445	11,171	51,802	0	42	51,460	154	51,283	0	23
9	28,901	0.08001	8,006	20,880	0	15	19,912	72	19,835	0	5
10	16,115	0.04461	6,257	9,849	0	9	9,498	33	9,461	0	4
11	8,520	0.02359	3,953	4,565	0	2	4,520	9	4,510	0	1
12	4,603	0.01274	2,349	2,253	0	1	2,192	4	2,188	0	0
13	3,294	0.00912	1,270	2,023	0	1	1,910	0	1,909	0	1
14	1,443	0.00399	761	682	0	0	620	0	620	0	0
15	571	0.00158	459	112	0	0	102	0	102	0	0
16	461	0.00128	375	86	0	0	72	0	72	0	0
17	191	0.00053	163	28	0	0	25	0	25	0	0
18	47	0.00013	23	24	0	0	24	0	24	0	0
19	2	0.00001	1	1	0	0	1	0	1	0	0
Total	361,211	1	68,446	292,070	52	643	288,569	1,132	287,091	12	334

WordNet categories such as *"Parent", "Life", "City university", "Clip art", "Logic diagram", "Routine", and "Call"*. The WordNet categories are spread over the depth 0-19 of the YAGO hierarchy. In addition, 642 Geo categories such as *"Water mill", "Copper mine", "Phosphate works", "Factory", and "Research institute"*, are located at the depth of 1-13. Finally, YAGO offers a set of 53 its own categories located at the depth of 0-5, e.g. *"Length", "Number", "Weight", "ISBN", and "Monetary value"*.

From the total number of 361,211 categories in the YAGO hierarchy, about 80% (288,569 categories) are the leaf categories that have instances directly assigned to them. For example, an instance *"San Francisco"* is directly assigned to the following Wikipedia categories: *"Populated places established in 1776", ''County seats in California", "Populated coastal places in California"*, and "California counties". 7,394 categories do not have instances as direct children. Among them are: *"Accident", "Action", "Young person", "Legal actor", "Organism", "Motor", "College"*, and *"Comedian"*. Finally, 65,248 categories do not possess any instances, neither direct, nor indirect. These YAGO categories include: *"Length", "Number", "Weight", "ISBN", and "Monetary value"*.

Table 1 presents the distribution of the categories at each level of the YAGO hierarchy. As Table 1 illustrates, 60% of all YAGO categories are assigned to the depth 6-8 of the YAGO hierarchy. 90% of the categories are located within the depth of 4-10. The categories that do not possess any parent categories are located at the top level of the YAGO hierarchy (depth=0). The most important YAGO category at the top level is the WordNet *"Entity"* category. This category is a parent of the majority of the categories in YAGO. However, the YAGO hierarchy does not form a clean tree structure. This is mostly because some of the WordNet categories in YAGO are not related to the *"Entity"* category and do not possess any parent categories. These categories include: *"Administrative district", "Brahman", "Control character", "Epacris", "Evan-*

gelicalism", "Exorcist", "Faun", "Fen", "Fundamentalist", and others. Finally, the YAGO category *"Relation"* is an additional YAGO category that does not possess any parents. This explains an unusually big number of categories and instances at the depth one, as many of the categories at this depth are the child categories of the WordNet categories located at the top level of the hierarchy. For example, the top level WordNet category *"Administrative district"* has 141 subordinated categories, among them are the Wikipedia categories: *"Settlements in New Brunswick"*, *"Neolithic settlements in Crete"*, and *"Settlements established in 1312"*. This also leads to a big number of associated instances already at the second level of the YAGO hierarchy.

2.2 Instance Distribution in YAGO

In total, 2,632,948 unique instances of YAGO are assigned to the 295,963 categories using the *"type"* relation. Majority of the YAGO instances (2,632,756) originate from Wikipedia. As YAGO allows an instance to be assigned to the multiple categories in the hierarchy, the total number of instances located at the leaves of the YAGO hierarchy is 2.76 times higher than the number of unique instances (7,255,584 instances in total). For example, an instance *"Stephen King"* is assigned to the multiple categories such as *"American school teachers"*, *"People from Portland, Maine"*, *"Film director"*, and *"American horror writers"*.

As presented in Table 2, the most populated level of the YAGO hierarchy with respect to the instance distribution is located at the depth 7 and contains 29% of the instances. The levels 6-8 together contain 64% of the instances. The majority of the instances (90%) are located within the depth 4-9 of the YAGO hierarchy.

In order to better understand how good the YAGO ontology structure fits to the Freebase dataset, we analyze how the instances shared by the both datasets are distributed across the different levels of the YAGO hierarchy. Table 2 presents the distribution of the shared instances of YAGO and Freebase in the YAGO hierarchy. In total, 99% of YAGO instances are shared with Freebase. In Table 2, "YAGO Leaf Cat." is the number of the leaf categories of YAGO that contain shared instances. The majority of the shared instances (94%) are located within the depth 3-10 of the YAGO hierarchy. In total these instances are assigned to the 260,488 leaf categories of YAGO.

In Table 2 "Freebase Cat." is the number of Freebase categories that are associated with the shared instances found at a certain depth of the YAGO hierarchy. At this point we do not yet perform the matching, such that each Freebase category can be associated with multiple levels of the YAGO hierarchy. For this reason, the total number of the Freebase categories presented in Table 2 (8,745) is much higher than the number of Freebase tables containing the shared instances (1,391). This is because the instances of one Freebase category are, on average, distributed across 6.3 YAGO categories.

2.3 Instance-Based Overlap between YAGO and Freebase

The aim of the matching function is to map each Freebase category to the most similar category of YAGO. To assess similarity of the categories, matching techniques often make use of the instances these concepts share [7], [2]. In order to create an effective

Table 2: Distribution of Instances in YAGO

Depth	All Instances in YAGO						Shared Instances	
	Total	Total, n.	WordNet	Wikipedia	YAGO	Geo	YAGO Leaf Cat.	Freebase Cat.
0	32	0.00001	3	0	29	0	0	0
1	31,140	0.00429	0	895	30,245	0	106	225
2	5	0.00001	0	5	0	0	1	4
3	124,679	0.01718	26,039	98,594	37	9	2,752	549
4	1,251,280	0.17246	4,450	1,234,315	0	12,515	37,735	833
5	402,104	0.05542	11,249	380,542	0	10,313	20,045	947
6	1,561,758	0.21525	26,181	1,529,220	0	6,357	51,969	979
7	2,152,886	0.29672	12,819	2,128,606	0	11,461	68,706	1,129
8	1,051,682	0.14495	4,985	940,515	0	106,182	45,404	1,033
9	305,063	0.04205	2,712	302,314	0	37	17,888	861
10	200,300	0.02761	85	200,173	0	42	8,555	769
11	85,298	0.01176	58	85,232	0	8	3,511	477
12	50,678	0.00698	92	50,586	0	0	1,852	336
13	27,193	0.00375	0	27,192	0	1	1,309	191
14	8,470	0.00117	0	8,470	0	0	447	156
15	1,451	0.00020	0	1,451	0	0	95	82
16	862	0.00012	0	862	0	0	67	82
17	536	0.00007	0	536	0	0	22	53
18	155	0.00002	0	155	0	0	23	28
19	12	0.00001	0	12	0	0	1	11
Total	7,255,584	1	88,673	6,989,675	30,311	146,925	260,488	8,745

matching function for YAGO and Freebase, we first analyze the instance overlap in the both sources.

The categories of Freebase and the corresponding data instances are available for download directly as SQL tables. Freebase data dump used in our experiments contains 1,578 categories [11]. On the one hand, we observed that 88% (1,391) of the Freebase categories contain Wikipedia instances that are also found in the YAGO ontology. For example, *"Stephen I of Hungary"* from Freebase and *"King Stephen"* from YAGO denote the same person, as they share the same Wikipedia identifier. On the other hand, as majority of the YAGO instances come from Wikipedia, these instances are also contained in the Freebase dataset. The instances coming from Wikipedia are uniquely distinguished by their Wikipedia identifiers in the both datasets and can be directly used for the category matching.

Each Freebase category is associated with a set of instances that can be partly shared with YAGO. Freebase dataset contains 22,542,665 unique instances, from which about 16% (3,661,329 instances) originate from Wikipedia. As an instance can be associated with multiple Freebase categories, for example *"Stephen King"* with *"Author", "Award winner", "Fictional character creator"*, and *"Pet owner"*, the overall number of Freebase instances is 2.47 times higher than the number of unique instances (55,684,113 instances in total). In total, 2,632,756 unique instances that originate from Wikipedia

are shared between the YAGO and Freebase datasets, which is about 12% of the Freebase instances.

The pie chart presented in Fig. 2 illustrates the number of the Freebase categories associated with a given percentage of shared instances. For example, we can see that for 72% (1,150 out of 1,578) Freebase categories, more than 20% of the associated instances are within the shared set. The Freebase categories with the highest instance overlap (80-100%) include: *"Astronomy asteroid", "Baseball coach", "Chess player", "Film critic", "Geography mountain", "US president", "Olympic games", "Religion Monastery", "Royalty kingdom", "Sports boxer"*, and *"Theater actor"*. For these categories, given a compatible YAGO concept structure, an instance-based matching should already provide good results.

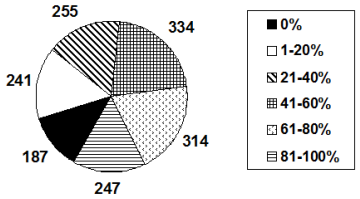

Fig. 2: Number of Freebase Categories with a Given Percentage of Shared Instances

Another part of the Freebase categories (about 16%) is associated with less than 20% shared instances. These are, for instance, *"Astronomy comet", "Business location", "Business industry", "Business job title", "Film editor", "Food drinking establishment"*, and *"Public library"*. In these cases, an additional evidence can be necessary to perform an effective matching. Finally, 12% of the Freebase categories are not associated with any shared instances. These are, for example, *"Luminous flux unit", "Astronomy galaxy classification code", "Book technical report"*, and *"Law US patent type"*. In order to include these categories in the mapping, an extension of the YAGO concept structure might be required.

3 Matching YAGO and Freebase

As Freebase and YAGO share a significant number of instances coming from Wikipedia as discussed in Section 2.3, instance-based matching techniques appear to be the most suitable to align these ontologies. *Instance-based matching techniques* assess similarity of the concepts based on the instances these concepts share [7], [2]. The real-world entities like a film, a car, or a person often coincide across ontologies. Given a set of corresponding instances, the similarity of the concepts can be measured as the instance overlap using e.g. Jaccard coefficient.

Instance-based matching will enable us to match about 90% of the Freebase categories. In the future, we plan to investigate adding further matching techniques such

as element-based matching and structure similarity [8] to further increase the number of Freebase categories that can be mapped to the YAGO ontology and to improve the quality of the matching incrementally.

Matching Design

Freebase dataset is a collection of categories that describe the real-world entities, like *"Person", "Book", "Location", "Airline"*, and *"Award"*, as well as facts associated with these entities.

In the matching process, we automatically assign each Freebase category to the most similar YAGO category. For example, we assign the Freebase category *"Author"* to the YAGO category *"Writer"*. An example of the matching between YAGO and Freebase categories is illustrated in Fig. 3. The output of the matching process is a

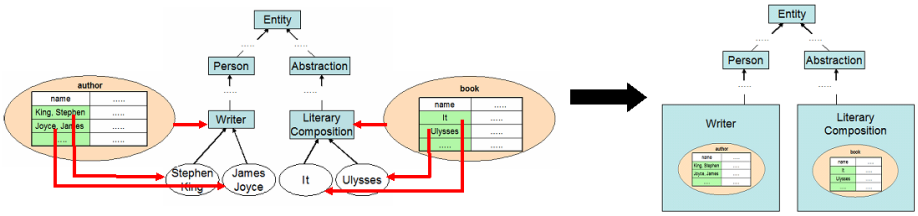

Fig. 3: Matching of YAGO and Freebase Categories

mapping of each Freebase category to the most likely semantic category of YAGO. Each Freebase concept is mapped only to the most likely YAGO category, whereas a YAGO category can be associated with several Freebase categories. For example, the YAGO category *"Writer"* may unify the Freebase categories *"Author"* and *"Comic book author"*.

Similarity Score Computation

The intuition behind the instance-based matching is that the categories F and Y are same if F contains the same instances as Y. The more instances are shared by F and Y, the more similar the categories are. For example, *"Writer"* and *"Author"* share many instances like *"Stephen King"* and *"James Joyce"* and are considered to describe the same concept. In contrast, the categories *"Writer"* and *"Literary composition"* do not have any instances in common and thus are not similar.

For the calculation of similarity between the two sets of instances we use the similarity measure known as the Jaccard coefficient, which is based on the joint probability. We define the instance-based similarity of two categories $\alpha_i(Y, F)$ as:

$$\alpha_i(F, Y) = \frac{P(F, Y)}{P(F, Y) + P(\bar{F}, Y) + P(F, \bar{Y})}, \tag{1}$$

where $P(F, Y)$ is the part of the shared instances that belongs to the categories F and Y, $P(\bar{F}, Y)$ is the fraction that belongs to Y but not to F, and $P(F, \bar{Y})$ is the fraction that belongs to F but not to Y. $\alpha_i(F, Y)$ has the lowest value 0 when the instance sets of F and Y are disjoint, e.g. *"Literary Composition"* and *"Author"*, and the highest value 1 when F and Y contain the same set of instances and thus represent the same concept, e.g. *"Writer"* and *"Author"*.

The overall matching function is then computed as:

$$Y_{map}(F) = argmax_{y \in Y_C}\ \alpha_i(F, y). \tag{2}$$

This function assigns each Freebase category F to the most likely YAGO category Y_{map} that has the highest similarity score from all YAGO categories Y_C according to the scoring function in Equation(1).

4 Describing and Characterizing the YAGO+F Hierarchy

Using the techniques described in the previous sections, we matched the Freebase dataset from June 2011 [11] with the YAGO2 ontology [10]. The Freebase dataset includes approximately 1,578 categories containing more than 20 million entities in more than 100 domains. As described above, 88% (1,391) of the Freebase categories contain Wikipedia instances that are also found in the YAGO ontology. The hierarchy of YAGO2 possesses 361,211 categories, from which more than 80% have associated instances. In total, YAGO contains 2,632,948 unique instances, most of which are shared with Freebase. The results of our matching are available for download in the .tsv and .n3 formats from **`http://iqp.l3s.uni-hannover.de`**. In this section, we analyze the structure of the YAGO+F hierarchy that results from the matching and evaluate the matching quality.

4.1 The Concepts and the Instances in the YAGO+F Hierarchy

In this section we discuss the results of the matching between YAGO and Freebase. Specifically, we are interested in the part of the initial YAGO ontology, which is relevant to the real-world large scale dataset such as Freebase. To this end, we analyze the distribution of categories and instances in the YAGO+F hierarchy which is obtained using the matching described in Section 3 and connects Freebase and YAGO categories.

We used the matching technique described in Section 3 to assign each Freebase category to the corresponding YAGO category. We call the YAGO categories directly matched to the Freebase categories *YAGO+F leaf categories*. Then, we extracted the sub-structure of YAGO required to describe the leaf categories matched to the Freebase dataset. To this end, we extracted all paths from the top level of the YAGO hierarchy to all the YAGO+F leaf categories. We call the resulting sub-structure of the YAGO ontology *YAGO+F*. The structure of YAGO+F is presented in Table 3.

In Table 3, "Categories YAGO+F, Total" is the total number of categories that are relevant for the matching and "Categories YAGO+F, Leaf" is the number of the leaf categories directly matched to the Freebase categories. The number of the leaf categories presented in Table 3 is 1.12 times smaller than the total number of the Freebase

categories. This is because a YAGO category can group together several Freebase categories.

We observed that grouping of several Freebase categories to one YAGO category mostly happens if the Freebase category structure has a higher granularity than the YAGO category structure. In this case, the Freebase categories mapped to one YAGO category can be either sibling categories or sub-categories of each other. In the former case, the sibling categories *"Film actor"* and *"TV actor"* of Freebase are both mapped to the WordNet *"Actor"* concept of YAGO. Also, *"Location statistical region"* and *"Location citytown"* are mapped to the WordNet *"Geographical area"*. In the latter case, 99,9% instances of the Freebase category *"Music song writer"* are also contained in the Freebase category *"Music lyricist"*, such that *"Music song writer"* is a sub-category of *"Music lyricist"*. The both categories are mapped to the WordNet *"Song writer"* category of YAGO. Also, *"Tennis player"* and *"Tennis tournament champion"* are both mapped to the WordNet *"Tennis player"*, as YAGO does not possess the more specific *"Tennis tournament champion"* category. In these cases, the more specific Freebase categories can represent a possible extension to the YAGO category structure.

Then, Table 3 presents the number of Freebase categories assigned to each level of the YAGO+F hierarchy. The majority of the Freebase categories is assigned to the levels 4-8, which corresponds to the distribution of the shared instances in the YAGO hierarchy. Finally, Table 3 presents the number of the shared instances and the total number of instances associated with the Freebase categories located at the specific level of YAGO+F. The majority of the Freebase instances (50%) is assigned to the levels 3-8. As an instance can be associated with multiple categories of Freebase, the total number of the instances presented in Table 3 includes duplicates.

Comparing the YAGO+F structure in Table 3 with the original YAGO hierarchy presented in Table 1, we can see that the resulting sub-structure of YAGO only contains 0.4% of the leaf categories compared to the original YAGO ontology that contained 288,569 leaf categories. This is expected, as each of the 1,391 Freebase categories was assigned to only one specific YAGO leaf category. The total number of categories in the YAGO+F hierarchy is 2,141, which is only 0.6% of the total number of the all YAGO categories (361,211). As we can see, only a small proportion of YAGO categories (less than 1%) is enough to describe a large scale multi-domain database such as Freebase.

The total number of instances in the YAGO+F mapping is smaller than the total number of Freebase instances. This is because 187 Freebase categories containing 3,172,694 instances (5,7% of the overall number of the Freebase instances) are not assigned to any YAGO category as these Freebase categories do not share any instances with YAGO.

We can see that 80% of the Freebase categories and 50% of the instances are assigned to the levels 3-8 of the YAGO+F hierarchy. This distribution roughly corresponds to the distribution of the categories in the original YAGO hierarchy, where the most populated levels were located at the depth 4-9 (see Table 2). We also observe a high instance concentration at the depth 0 of the YAGO+F hierarchy. This is because a general and the most populated Freebase category *"Common topic"* that contains information about

Table 3: Distribution of the Categories and Instances in YAGO+F after the Matching

Depth	Categories in YAGO+F				Freebase		Instances in YAGO+F			
	Total	Total, n.	Leaf	Leaf, n.	Cat.	Cat., n.	Shared	Shared, n.	Freebase	Freebase, n.
0	1	0.0005	1	0.0008	1	0.0007	1,967,590	0.3075	22,577,777	0.4300
1	4	0.0019	1	0.0008	1	0.0007	463,039	0.0724	1,378,280	0.0262
2	19	0.0089	1	0.0008	2	0.0014	70,738	0.0111	100,409	0.0019
3	85	0.0397	23	0.0186	30	0.0216	1,113,000	0.1739	2,644,774	0.0504
4	235	0.1098	92	0.0746	106	0.0762	1,100,473	0.1720	2,748,704	0.0523
5	370	0.1728	170	0.1378	191	0.1373	472986	0.0739	4,155,620	0.0791
6	414	0.1934	232	0.1880	255	0.1833	698,365	0.1091	2,974,928	0.0567
7	469	0.2191	332	0.2690	379	0.2725	370,462	0.0579	13,606,038	0.2591
8	271	0.1266	182	0.1475	201	0.1445	74,510	0.0116	2,029,859	0.0387
9	137	0.0640	97	0.0786	110	0.0791	41,842	0.0065	203,555	0.0039
10	92	0.0430	72	0.0583	82	0.0590	20,597	0.0032	76,981	0.0015
11	30	0.0140	21	0.0170	22	0.0158	5,006	0.0008	12,085	0.0002
12	10	0.0047	7	0.0057	7	0.0050	549	0.0001	1,309	2.E-05
13	3	0.0014	2	0.0016	2	0.0014	93	1.E-05	1007	2.E-05
14	1	0.0005	1	0.0008	2	0.0014	10	2.E-06	93	2.E-06
Total	2,141	1	1,234	1	1,391	1	6,399,260	1	52,511,419	1

entities and their relations from the wide variety of Freebase domains is assigned to the top level *"Entity"* category of YAGO.

The levels at the depth higher than 14 did not get any Freebase categories assigned. These is because the granularity of the YAGO categories at these levels is too high compared with the Freebase structure.

4.2 Matching Results

Our evaluation compared YAGO+F against the ground truth of Freebase. The aim of the evaluation is to assess the similarity between the original classification of the instances in Freebase with the YAGO+F classification that we obtain in the matching process. To this extent, we compute the Rand Index (RI) [12] and the Jaccard Index (JI), that are the standard measures to evaluate the quality of clustering algorithms. Both RI and JI are the measures of similarity between two data clusterings. In the context of our matching, we compute RI and JI for each YAGO+F leaf category.

To this extent, we view the mapping of Freebase and YAGO categories as a series of decisions, one for each of the $N(N-1)/2$ pairs of shared instances in the both datasets. We want to assign two instances to one and the same YAGO+F category if and only if these instances belong to the same Freebase category. A *true positive decision* $TP(y, f)$ assigns two instances from one Freebase category f to one YAGO+F category y. As in the praxis multiple Freebase categories can be mapped to one YAGO+F category, to compute the Rand Index $RI(y)$ of the YAGO+F leaf category y, we take into account all Freebase categories $f \in F_y$ mapped to y. The number of true positive decisions is $TP(y, F_y)$, where y is the YAGO+F leaf category and F_y is a set of the Freebase categories mapped to y in the matching process. Further, a *true negative deci-*

sion $TN(y, F_y)$ does not assign two instances from the different Freebase categories to one YAGO+F leaf category y.

In the matching, two types of errors can occur. A *false positive decision* $FP(y)$ assigns two instances from different Freebase categories to one YAGO+F category y. This happens if more than one Freebase category is assigned to a YAGO+F category in the matching process. A *false negative decision* $FN(F_y)$ assigns two instances from a Freebase category $f \in F_y$ to different YAGO+F categories. This can happen if the instances from one Freebase category were found in different YAGO categories before the matching. The Rand Index for a YAGO+F leaf category y measures the percentage of decisions that are correct for this category (i.e. accuracy), that is:

$$RI(y) = \frac{TP(y, F_y) + TN(y, F_y)}{TP(y, F_y) + FP(y) + FN(F_y) + TN(y, F_y)}. \quad (3)$$

If a Freebase category contains only one instance, the value of RI may be not well-defined, as the number of instance pairs in such category is zero. To this end, we apply Laplace smoothing and add one to the number of shared instances in each Freebase category.

The RI values range from $[0, 1]$, where $RI = 1$ is the best possible value. This value can be achieved in a perfect situation, where one Freebase category is exclusively matched to exactly one YAGO+F category. In the Equation 3, the TN-factor takes the size of the Freebase category into account, such that the influence of the categories that do not contain many instances is reduced. In order to get a better overview of the matching results for the categories independent of their size, we also compute the RI value without the TN-factor. This corresponds to the Jaccard Index JI that measures the level of agreement over the pairs from the Freebase categories assigned to a YAGO+F category. In the matching process we optimized the Jaccard Index directly.

$$JI(y) = \frac{TP(y, F_y)}{TP(y, F_y) + FP(y) + FN(F_y)}. \quad (4)$$

In our experiments, we also compare the RI and JI values of the leaf categories of the YAGO+F mapping with the values obtained by an alternative mapping function PARIS [14] - a recent approach to align ontologies. PARIS computes the probability of the subclass relationships among the classes of different ontologies. To facilitate our comparison, while using PARIS, we align each Freebase category with its most probable superclass in the YAGO ontology, computed using Equation 15 in [14]. Fig. 4 presents the RI values for each leaf category obtained using YAGO+F and PARIS alignments. The X-axis of Fig. 4 presents the percentage of leaf categories in the resulting mapping sorted by the RI and JI value correspondingly. The Y-axis presents the corresponding RI and JI values.

As we can observe, the YAGO+F mapping performs better than the PARIS-based mapping with respect to both, RI and JI values. This is because with PARIS, the probability of an alignment is higher for more general classes, such that more Freebase categories are jointly assigned to one YAGO category in a higher level of the hierarchy. In total, using the PARIS-based alignment, 1,391 Freebase categories are assigned to only 67 YAGO categories, whereas YAGO+F assigns these categories to 1,234

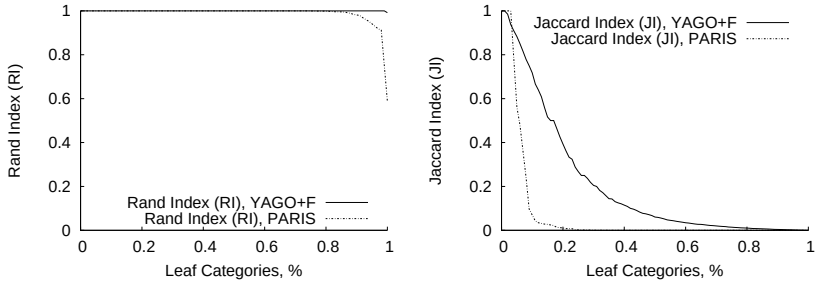

Fig. 4: Rand Index and Jaccard Index of the Leaf Categories using YAGO+F and PARIS

YAGO categories. This way, the mapping obtained by YAGO+F is more specific than the PARIS-based mapping.

With respect to the YAGO+F mappping, on the one hand, the RI values are very high for all the YAGO+F categories, which reflects the fact that the matching of the Freebase categories highly populated with shared instances is very accurate. On the other hand, the values of JI vary. 22% of the YAGO+F categories possess a value $JI \in [0.9 - 1]$, for the further 16% of the categories the JI is within [0.5-0.9). Finally, the rest of the YAGO+F categories possesses the JI values below 0.5.

The group of the YAGO+F categories with the highest JI values includes the cases where all the shared instances of a Freebase category are contained in one YAGO category. For example, *"Location tw provincial city"* is matched to the Wikipedia category *"Provincial cities of Taiwan"*, *"Medicine artery"* to the WordNet category *"Artery"*, *"Food bottled water"* to the Wikipedia category *"Bottled water brands"*, *"Luminance unit"* to the Wikipedia category *"Units of luminance"*, and *"Food culinary technique"* to the Wikipedia category *"Cooking techniques"*.

The YAGO+F categories with the JI values of (0.5-0.99] possess many shared instances, almost all of which are found in one YAGO category. In this case the YAGO category will most likely provide an adequate mapping for the Freebase category. For example, *"Location jp prefecture"* is matched to the Wikipedia category *"Prefectures of Japan"*, *"Medicine hospital"* to the WordNet category *"Medical building"*, *"Film"* to WordNet *"Movie"*, *"Book poem"* to WordNet *"Poem"*, *"Business company type"* to the Wikipedia category *"Types of companies"*, and *"Education academic"* to WordNet *"Scientist"*. This group includes 280 YAGO+F categories.

The lower JI values of the remaining YAGO+F categories are due to the incompatibilities in the category structures of Freebase in YAGO. First, the shared instances of a Freebase category can fall into a wide spectrum of the YAGO categories, i.e. there does not exist any clearly defined equivalent YAGO category. These Freebase categories include *"Visual art subject"*, *"Media common quotation subject"*, *"Book periodical subject"*, *"Film subject"*, and *"Amusement parks ride theme"*. Second, in comparison to Freebase, YAGO may lack a specific intermediate category, such that a Freebase category can either be assigned to the most specific parent, which is in fact too general, or to one of the children categories, which are too specific. For example, the Wikipedia cate-

gories *"Aviation accidents and incidents officially attributed to pilot error"*, and *"Airliner accidents and incidents caused by fuel exhaustion"* are direct subclasses of the WordNet category *"Accident"*. Then, the Freebase category *"Aviation accident type"* can either be mapped to one of the more specific categories (with a high FN value), or to a too general category (with a high FP value).

In summary, given a compatible category structure, our mapping provides good results for a significant number of the YAGO+F categories. Nevertheless, some Freebase categories cannot be clearly mapped to YAGO due to incompatibilities in the YAGO and Freebase structure. In future work we will investigate how to improve matching by introducing new categories that incorporate both YAGO and Freebase schema information into a richer ontology.

5 Related Work

In related work, YAGO and Freebase were brought together in the context of Linked Open Data [3]. LOD already includes Freebase and YAGO, loosely connecting their shared instances through the DBpedia references. The systematic integration of YAGO and Freebase at the schema level described in this paper takes an important step further towards tighter integration of LOD, empowering discovery of new relations across the datasets contained in different knowledge bases and enlarging the scope of the possible applications that use the Web of Data.

Schema matching plays an important role in the context of relational databases (see e.g. survey of Rahm et.al. [13]), as well as in the context of XML (e.g. [6]), and ontologies on the Semantic Web [7]. The aim of the schema matching in relational databases is to identify the most similar matching element(s) in a flat relational schema. In addition, malleable schemas enable more flexibility in schema matching by relaxing definitions of attributes or relationships [16]. In contrast, the systems for XML and ontology matching are required to identify the most-specific-parent or the most-general-child within the relevant branch of the hierarchy [7]. PARIS [14] quantifies the probability of whether the classes of two ontologies are in a subclass relation. In this work we apply the existing schema matching techniques to address a novel problem, namely to enrich a widely used YAGO ontology with large scale community-created Freebase data. We analyze and compare the initial and enriched structure of the YAGO ontology and discuss which parts of the ontology are relevant to describe a large scale real-world heterogeneous database such as Freebase.

6 Conclusion and Future Work

In this paper we considered the problem of enrichment of the Freebase dataset with the semantic categories of the YAGO ontology, to connect them not only at the instance level but also at the schema level. The resulting merged dataset YAGO+F provides a further important step towards tighter interconnection of the Linked Open Data, and will hopefully enable many future applications that can profit from a wide variety of Freebase data clearly arranged into the semantic categories of the YAGO ontology.

The results of our experiments confirm the good quality of the matching in cases where the YAGO and Freebase category structures are compatible, but also show incompatibilities between the category schemas of YAGO and Freebase. In future work we will investigate how to improve matching by introducing new categories that incorporate both YAGO and Freebase schema information into a richer ontology which will provide even richer semantic information suitable for Freebase data, avoid the difficulties caused by incompatible YAGO and Freebase hierarchy structures and incompatible class instance assignment, and potentially improve both the YAGO hierarchy as well as Freebase schema information.

Acknowledgments This work was partly funded by the European Commission under grant agreement n. 270239 (ARCOMEM).

References

1. Peter Adolphs, Martin Theobald, Ulrich Schäfer, Hans Uszkoreit, and Gerhard Weikum. Yago-qa: Answering questions by structured knowledge queries. In *Proc. of the ICSC 2011*, 2011.
2. Alexander Bilke and Felix Naumann. Schema matching using duplicates. In *Proc. of the ICDE '05*, 2005.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
4. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of the SIGMOD '08*, pages 1247–1250, 2008.
5. Elena Demidova, Xuan Zhou, and Wolfgang Nejdl. Freeq: an interactive query interface for freebase. In *Proc. of the WWW 2012*, 2012.
6. Hong-Hai Do and Erhard Rahm. COMA: a system for flexible combination of schema matching approaches. In *Proc. of the VLDB 2002*, pages 610–621, 2002.
7. AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to map between ontologies on the semantic web. In *Proc. of the WWW '02*, 2002.
8. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.
9. Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.
10. Max-Planck Institute for Informatics. Yago2 ontology. http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html, 2011.
11. Google. Freebase data dumps. http://download.freebase.com/datadumps/, 2011.
12. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
13. Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10:334–350, December 2001.
14. Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. Paris: Probabilistic alignment of relations, instances, and schema. *PVLDB*, 5(3):157–168, 2011.
15. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proc. of the WWW 2007*, New York, NY, USA, 2007. ACM Press.
16. Xuan Zhou, Julien Gaugaz, Wolf-Tilo Balke, and Wolfgang Nejdl. Query relaxation using malleable schemas. In *Proc. of the SIGMOD 2007*, pages 545–556, 2007.