LODS: A Linked Open Data Based Similarity Measure

Nasredine Cheniki, Abdelkader Belkhir, Yacine Sam, Nizar Messai

LODS : A Linked Open Data Based Similarity Measure

Nasredine Cheniki
University of Sciences and
Technology Houari
Boumediène (USTHB), Algeria
Email: n.cheniki@usthb.dz

Abdelkader Belkhir
Laboratoire des Systmes
Informatiques (LSI), USTHB, Algeria
Email: Kaderbelkhir@hotmail.com

Yacine Sam
and Nizar Messai
Francois Rabelais University
Tours, France
Email: yacine.sam@univ-tours.fr
nizar.messai@univ-tours.fr

***Abstract*—With the rapid evolution of Linked Open Data (LOD), researchers are exploiting it to solve particular problems such as semantic similarity assessment. Existing LOD-based semantic similarity approaches attach compared data (terms or concepts) to LOD resources to exploit their semantic descriptions and relationships with other resources and estimate the degree of overlap between resources. Current approaches suffer from two limitations: they focus on the analysis of links between resources and ignore the important taxonomic structure of concepts and categories used to describe resources. On the other hand, they do not exploit interlinks between LOD resources in order to enrich data used to compute the similarity score. In this paper, we overcome the above limitations by proposing a new LOD-based similarity measure based on the combination of ontological, classification and property dimensions of LOD resources.**

I. Introduction

LOD initiative aims at making a gigantic interlinked database of interlinked datasets. Last few years, it has known a striking achievement where billions of linked facts (i.e. semantically connected entities) are available on the Web[1]. With that massive amounts of semantically interlinked data, it is now time to build – on top of this wealth source of data – applications that access, consume, and integrate the provided knowledge base. Semantic similarity measures are such applications that could benefit from LOD data.

Semantic similarity measures are developed to calculate the degree of matching between pairs of terms (for example, *Food* and *Fruit*). They are applied in several domains such as Web services discovery and classification, pervasive computing, information retrieval and recommendation systems.

Traditional semantic-based similarity measures usually rely on hand-crafted ontologies such as WordNet. These ontologies cover a restricted number of domains and require significant efforts to be maintained and kept up to date [1]. In contrast, LOD provides a rich semantic data for a large number of domains which are constantly updated. It constitutes a great source of semantic information that complements domain-specific knowledge bases.

LOD can be exploited by similarity measures to estimate the degree of overlap between compared concepts. As reported in [2], the informativeness of LOD is high and can hence be useful in many applications in which similarity measures are important, such as information retrieval.

In this paper, we propose a new similarity measure that assesses the matching degree between pairs of terms represented with Linked Open Data resources. We exploit for that the taxonomic structure of ontological concepts and classification schemata, in addition to the semantic properties that semantically describe LOD resources. Besides, our approach strives to take benefit of LOD by traversing interlinking relationships between resources defined in same or different datasets to glean richer information about compared resources. By doing this, we reduce the impact of missing information within a single dataset. Experiments show that our similarity measure provides better results when data is augmented from distinct LOD datasets.

The remainder of the paper is organized as follows. We introduce some useful background notions in Section II before the presentation of our proposed similarity measure in Section III. The evaluation results are reported in Section IV. In Section V, we discuss and compare our approach to related works. Section VI concludes with a summary and some ideas and directions for future work.

II. Background

This section provides main necessary constructions and characteristics of LOD, DBpedia and its interlinked datasets.

A. Linked Open Data

RDF triples. Consider a set of URIs $\mathcal{U}$ and literals $\mathcal{L}$, an rdf triple t is defined as $t = \langle s, p, o \rangle$, where the subject $s \in \mathcal{U}$, the property $p \in \mathcal{U}$ and the object $o \in \mathcal{U} \cup \mathcal{L}$.

Linked Open Data. A dataset that follows a linked data principles [3] is a graph $G = (\mathcal{R}, L)$, where $\mathcal{R} = \{r_1, r_2, ..., r_{|\mathcal{R}|}\}$ is a set of resources and $L = \{l_1, l_2, ..., l_{|L|}\}$ a set of links. l_i is defined as $l_i = \langle r, p, r' \rangle \vee \langle r, p, v \rangle$, where p is a property that interlinks the resource r with the internal/external resource r' or with a literal attribute v, which is a basic value (string, date, number ...). So, Linked Open Data is a set of interlinked open datasets: $LOD = \bigcup_i G_i$.

Ontology. Ontology is a graph of triples that describe domain concepts and their relations. In LOD, it is preferred

[1] http://stats.lod2.eu/

to use concepts from widely used ontologies to instantiate its resources [4]. This allows an efficient data integration and reuse by LOD-based applications. We denote by O all ontologies used to describe the whole LOD resources.

Classification schema. Classification schemata are similar to ontologies; they classify resources into categories. However, unlike ontologies, they are richer and may contain cycles in their taxonomic structure, so we need to limit the level of extracted categories to avoid retrieving useless ones. We denote by C all classification schemata used to classify the entire LOD resources.

Triple patterns, Basic Graph Patterns (BGP). Triple patterns are like RDF triples except that each of the subject, predicate or object may be a variable (started with '?'). BGP are constructed from a set of triple patterns. We adopt SPARQL BGP[2] to represent queries over RDF datasets.

Properties types. An infinite set of properties could be used to describe LOD resources; we specify here those that will be used in our proposed similarity measure:

- Instantiation properties (IP). An IP $\tau \in \mathcal{U}$ is a property that attributes a concept c from an ontology $o \in O$ to a particular resource r, we write $\langle r, \tau, c \rangle$. Usually the property *rdf:type*[3] is used as IP.
- Classification properties (CP). A CP $\varsigma \in \mathcal{U}$ is a property used to classify a resource into a particular category in a rich schema. Usually, *dcterm:subject* is used as CP.
- Linking properties (LP). A LP $\xi \in \mathcal{U}$ is a property used to interconnect two equivalent resources r and r' belong generally to distinct LOD datasets. We write $\langle r, \xi, r' \rangle$ to express that r' is an equivalent resource of r. The property *owl:sameAs* is commonly used in that regards.
- Subsumption properties (SP). Consider two concepts c_i and c_j from an ontology or a classification schemata. The triple $\langle c_i, \delta, c_j \rangle$ denotes that concept c_i is a specialization or a subclass of concept c_j.

We call the remaining properties as characterization properties (denoted by P) since they distinguish LOD resources. Properties that have a resource r as subject, i.e. $\langle r, p, ?o \rangle$, are called outgoing properties and those having r as object, i.e. $\langle ?o, p, r \rangle$, ingoing properties.

Property paths and paths patterns. We adopt Sparql 1.1 property path[2] notations such as *ZeroOrMorePath*, *ZeroOrOnePath*, *OneOrMorePath*, *SequencePath*, *AlternativePath*, denoted respectively $*, +, ?, /$ and $|$. They are used to navigate between resources and reach particular data inside a single or distinct LOD datasets.

Using property paths in a triple that contains variables expresses a path pattern that retrieves all triples satisfying it. For instance, the pattern : $\langle r, \xi/\tau, ?c \rangle$ retrieves from equivalent resources of r, all possible instantiation concepts.

B. DBpedia and its interlinked LOD datasets

DBpedia [5] is the semantic counterpart of Wikipedia that realizes LOD vision by structuring its content and interlinking it with external datasets such as Wikidata[4] and YAGO[5]. DBpedia transforms every Wikipedia article into a resource, annotated with a set of properties extracted from the article Web page. Many LOD datasets are producing data pointing to DBpedia resources making it as the kernel of LOD cloud[6]. So, our approach relies on DBpedia as starting point to glean initial data to compute similarity degree between compared concepts. Afterward, it follows interlinks to enrich data from other datasets; we call this process of aggregating data from related datasets as the data augmentation process. Important LOD resources properties used by our measure are presented in the next subsections:

Instantiation. LOD resources are described with various multi-domain ontologies such as UMBEL[7] Reference Concept ontology (O_{umbel}), Schema.org[8] ontology (O_{schema}), YAGO classes (O_{yago}), DBpedia ontology (O_{dbo}).

Some ontologies have a very rich taxonomic structure between concepts and may consequently well describe resources; it is the case of O_{yago}. Others are however poorly structured; it is the case of O_{dbo} and O_{schema}.

Classifications. LOD resources can be arranged into categories. For instance, DBpedia semantically organizes Wikipedia categories into a rich taxonomy used to classify resources. We denote these classifications by C_{dbp} for Dbpedia English, and $C_{dbp\_\{lang\}}$ for other language chapters.

Properties. LOD resources are described with different properties which distinguish and characterize resources. For instance, DBpedia describe resources by properties (denoted P_{dbo}) extracted from Wikipedia infoboxes/templates. These properties are semantically described in the ontology O_{dbo}.

Interlinking. LOD resources are usually interlinked to each other using owl:sameAs. Each resource from DBpedia (*DBp*) is interlinked to the following datasets:

- DBpedia chapters such as Dutch (DBp_{de}), Italian (DBp_{it}) and French (DBp_{fr}).
- Wikidata dataset (*WD*). It is a collaboratively edited knowledge base that provides a richer taxonomic structure (O_{wd}) between its entities. Its entities are also described with a set of properties (P_{wd}).
- *YAGO* knowledge base extracts facts from Wikipedia and combines it with WordNet[9] to produce a rich ontology with high coverage and quality. Its classes O_{yago} are used to instantiate DBpedia resources.

Based on data provided by DBpedia and its interlinked datasets, our similarity measure is defined in the next section.

[2] http://www.w3.org/TR/sparql11-query/

[3] Henceforth, prefixes defined in http://dbpedia.org/sparql?nsdecl are reused

[4] https://www.wikidata.org/

[5] https://www.mpi-inf.mpg.de/yago-naga/yago/

[6] You can refer to http://lod-cloud.net/ to see the LOD cloud

[7] http://umbel.org/

[8] http://schema.org/

[9] https://wordnet.princeton.edu

III. Linked Open Data based Similarity Measure

The proposed similarity measure is composed of three sub-measures: (i) the first one exploits the taxonomic structure of ontological concepts used to instantiate resources, (ii) the second one operates on classification schemata used to categorize resources and, (iii) the third one uses resources characterization properties.

The reason for combining all these sub-measures is to reduce the negative impact of poorly described LOD resources since it is not the case that all resource are instantiated with ontological, classified into categories or well described with distinguishing properties. Also, we try to enrich data by following links that exist between different datasets. As a consequence, the proposed similarity measures operates on a comprehensive collection of information.

We maintain feature-based similarity approach based on Tversky [6] model instead of distance-based methods for the following reasons: (i) LOD resources are usually described with multi-domain ontologies where taxonomic relations do not necessary represent uniform distance. (ii) Moreover, one resource may be described with concepts from multiple ontologies; feature-based methods are preferred since edge-counting methods cannot be directly applied [7]. (iii) In addition, feature-based measures can be computed very fast compared to other approaches.

We provide hereafter by the mean of some definitions the theoretical foundations of our proposal. All measures are normalized in the range [0...1], where score equals 0 means that compared resources are dissimilar, and 1 means that resources are identical.

Definition 1: Let $O_r \subseteq O$ be the subset of ontologies containing concepts that instantiate a resource r. We define a function $\phi_o(r)$ that returns all taxonomic features of a resource r, i.e. all concepts and their subsumers in an ontology $o \in O_r$. Formally,

$$\phi_o(r) = \{?c \in o | \langle r, \tau | \delta^* | \tau/\delta^*, ?c \rangle \} \tag{1}$$

To enrich taxonomic features or increase the size of instantiation ontologies space of a resource. We follow equivalent resources belonging to the same or different LOD datasets. We consider that LP properties are transitive, and we define a path length Π that limits the number of levels we follow to get all possible equivalent resources. After getting all equivalent resources, we define $\phi^*_o(r)$, an augmented function as follows:

$$\begin{aligned} \phi^*_o(r) = \phi(r) \cup \{?c' \in o \mid \langle r, \xi, r' \rangle \wedge \\ \langle r', \tau | \delta^* | \tau/\delta^*, ?c' \rangle, \ ?c' \notin \phi(r) \} \end{aligned} \tag{2}$$

After applying $\phi^*_o(r)$, the set of instantiation ontologies of r will be so augmented. We denote this new set by O^*_r.

Definition 2: Let $O^*_{a,b} \subseteq O^*_a \cap O^*_b$ be the set of shared augmented ontologies between two resources a and b. The instantiation similarity $SimI^*_o(a,b)$ of two resources described with concepts from an ontology $o_i \in O^*_{a,b}$ is computed based on the cardinalities of differential and common taxonomic features of compared resources. $SimI^*_o(a,b)$ is calculated as follows:

$$SimI^*_{o_i}(a,b) = \frac{|\phi^*_{o_i}(a) \cap \phi^*_{o_i}(b)|}{|\phi^*_{o_i}(a) \cap \phi^*_{o_i}(b)| + |\phi^*_{o_i}(a) \setminus \phi^*_{o_i}(b)| + |\phi^*_{o_i}(b) \setminus \phi^*_{o_i}(a)|} \tag{3}$$

So, we formally define the overall instantiation similarity $SimI^*$ as follows:

$$SimI^*_{\forall o_i \in O^*_{a,b}}(a,b) = \frac{\sum_{o_i \in O^*_{a,b}} SimI^*_{o_i}(a,b)}{|O^*_{a,b}|} \tag{4}$$

The average function is applied to all score values calculated from shared augmented ontologies. This keeps the balance between ontologies having a rich taxonomic structure and those with a poor taxonomic structure.

Definition 3: Let $C_r \subseteq C$ be the subset of classification schemata that arrange r into categories. We define $\Delta^\ell_t(r)$ as the function that returns, for a resource r, all the classification categories and their super-categories in a schema $t \in C_r$. The parameter ℓ limits the deepness of the hierarchical level in which categories are retrieved. Formally,

$$\begin{aligned} \Delta^\ell_t(r) = \{?c_1, \ldots, ?c_\ell \in t | (\langle r, \varsigma, ?c_1 \rangle \wedge \\ \langle ?c_1, \delta, ?c_2 \rangle) \vee ... \langle ?c_{\ell-1}, \delta, ?c_\ell \rangle \} \end{aligned} \tag{5}$$

In contrast, the function $\nabla^{\ell'}_t(r)$ returns all the classification categories and their sub-categories, for a resource r, in a schema $t \in C_r$, with respect to ℓ' levels:

$$\begin{aligned} \nabla^{\ell'}_t(r) = \{?c_1, .., ?c_{\ell'} \in t | (\langle r, \varsigma, ?c_1 \rangle \wedge \langle ?c_2, \delta, ?c_1 \rangle) \vee \\ ... \langle ?c_{\ell'}, \delta, ?c_{\ell'-1} \rangle \} \end{aligned} \tag{6}$$

We combine, the two functions to obtain all classification features of a resource r :

$$\varphi^{\ell,\ell'}_t(r) = \Delta^\ell_t(r) \cup \nabla^{\ell'}_t(r) \tag{7}$$

Following the same principle of expanding the space of instantiation concepts of a resource r, we can define augmented classification features function denoted by ${\varphi^*_t}^{\ell,\ell'}(r)$. We can as well obtain its set of augmented classification schemata, denoted by C^*_r.

Definition 4: Let $C^*_{a,b} \subseteq C^*_a \cap C^*_b$ be the set of shared augmented classification schemata between two resources a and b. The classification similarity ${SimC^*_t}^{\ell,\ell'}(a,b)$ of two resources, described with categories from a classification schema $t \in C^*_{a,b}$ with respect to limited super-categories and sub-categories hierarchy levels ℓ and ℓ', is computed based on the cardinalities of differential and common categories features of compared resources. Formaly,

$$SimC_t^{*\ell,\ell'}(a,b) = \frac{|{\varphi_t^*}^{\ell,\ell'}(a) \cap {\varphi_t^*}^{\ell,\ell'}(b)|}{|{\varphi_t^*}^{\ell,\ell'}(a) \cap {\varphi_t^*}^{\ell,\ell'}(b)| + |{\varphi_t^*}^{\ell,\ell'}(a) \setminus {\varphi_t^*}^{\ell,\ell'}(b)| + |{\varphi_t^*}^{\ell,\ell'}(b) \setminus {\varphi_t^*}^{\ell,\ell'}(a)|} \tag{8}$$

The measure that computes the overall score for the whole augmented classifications space is then defined as:

$$SimC_{\forall t_i \in C^*_{a,b}}^{*\ell,\ell'}(a,b) = \frac{\sum_{t_i \in C^*_{a,b}} SimC_{t_i}^{*\ell,\ell'}(a,b)}{|C^*_{a,b}|} \tag{9}$$

Definition 5: Let the subset $P_r \subseteq P$ be the sub-space of properties that contains characterizing attributes of a resource r. We define $\Omega_{P_i}(r)$ (respectively, $\Omega'_{P_i}(r)$) the function that returns all ingoing (respectively outgoing) characterizing properties (called properties features) of a resource r in a particular space $P_i \in P_r$. The function Ψ combines the results of the two functions. Formally,

$$\Omega_{P_i}(r) = \{(?p, IN), ?p \in P_i | \langle ?x, ?p, r \rangle\} \tag{10}$$

$$\Omega'_{P_i}(r) = \{(?p, OUT), ?p \in P_i | \langle r, ?p, ?x \rangle\} \tag{11}$$

$$\Psi_{P_i}(r) = \Omega_{P_i}(r) \cup \Omega'_{P_i}(r) \tag{12}$$

Following interlinked equivalent resources of r, we can also define as in equation 2 the function $\Psi^*_{P_i}$ that returns the set of augmented properties describing r. The augmented space of properties P^*_r can be hence obtained.

Definition 6: Let $P^*_{a,b} \subseteq P^*_a \cap P^*_b$ be the space of shared augmented characterizing properties of two resources a and b, we define characterizing properties similarity $SimP^*_{P_i}(a,b)$ of a and b as:

$$SimP^*_{P_i}(a,b) = \frac{\mu(\Psi^*_{P_i}(a) \cap \Psi^*_{P_i}(b))}{\mu(\Psi^*_{P_i}(a) \cap \Psi^*_{P_i}(b)) + \mu(\Psi^*_{P_i}(a) \setminus \Psi^*_{P_i}(b)) + \mu(\Psi^*_{P_i}(b) \setminus \Psi^*_{P_i}(a))} \tag{13}$$

Where μ is the partial information content of characterizing properties [8]. This function gives more importance to specific properties based on their occurrences in LOD datasets. Formally,

$$\mu(\Psi^*_{P_i}(r)) = \sum_{\forall \rho \in \Psi^*_{P_i}(r)} -\log\left(\frac{Freq(\rho)}{N}\right) \tag{14}$$

where $Freq$ is a function that counts the occurrence frequency of the property feature ρ in the description of LOD resources, and N the total number of resources in the underlying dataset.

We then calculate the similarity of all properties space as follows:

$$SimP^*_{\forall P_i \in P^*_{a,b}}(a,b) = \frac{\sum_{P_i \in P^*_{a,b}} SimP^*_{P_i}(a,b)}{|P^*_{a,b}|} \tag{15}$$

Definition 7: Given a pair of two compared resources a and b that are: (i) instantiated with concepts from augmented shared ontologies $o_i \in O^*_{a,b}$, (ii) classified into categories from augmented shared classification schemata $t_i \in C^*_{a,b}$, and (iii) characterized with a set of augmented shared space of properties $P_i \in P^*_{a,b}$. We define the Linked Open Data Similarity (*LODS*) between two resources a and b as follows:

$$LODS^{\ell,\ell'}(a,b) = AVG(SimI^*_{\forall o_i \in O^*_{a,b}}(a,b), SimC^{*\ell,\ell'}_{\forall t_i \in C^*_{a,b}}(a,b), SimP^*_{\forall P_i \in P}(a,b)) \tag{16}$$

The final measure combines all previous measures by computing the average of their resulted scores. If a sub-measure could not take similarity judgment due to data missing, it will not be included in LODS. So, the average function is applied only to measures that return a score based on available semantic data.

IV. Evaluation

Our proposed similarity measure is implemented using Java and Jena framework[10]. We query data directly from provided SPARQL endpoints as much as possible. Unreachable data via SPARQL endpoints or HTTP are downloaded and hosted on a local endpoint.

A. Experimental LOD datasets

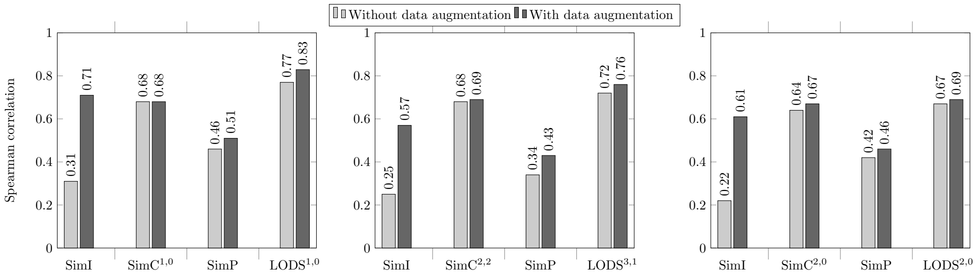
We rely on DBpedia knowledge base as a primary source of semantic linked data. We then use interlink relationships to navigate through and get more data from Wikidata, YAGO, and three different DBpedia chapters; we deem most active ones: DBpedia Dutch, Italian and French. So we can consider that our LOD dataset is: $LOD = \{DBp, WD, YAGO, DBp_{de}, DBp_{it}, DBp_{fr}\}$. Our subset ontology space is: $O = \{O_{dbo}, O_{yago}, O_{schema}, O_{umbel}, O_{wd}\}$. The used classification schemata are: $C = \{C_{dbp}, C_{dbp\_de}, C_{dbp\_it}, C_{dbp\_fr}\}$. Finally, we apply properties of DBpedia ontology (P_{dbo}) which are enriched from other chapters ($P_{dbo\_de}, P_{dbo\_it}, P_{dbo\_fr}$) since they used the same ontology to describe resources properties. We also use Wikidata properties (P_{wd}), so: $P = \{P_{dbo}, P_{dbo\_de}, P_{dbo\_it}, P_{dbo\_fr}, P_{wd}\}$

We note that the ontology Cyc/OpenCyc[11] is not considered since it is derived by the ontology O_{umbel}. Also, Freebase is now read-only, and it will be shut-down; its data will be transferred to Wikidata[12]. Consequently, it is not considered in our evaluation datasets. Moreover, we did not take *YAGO* resources properties, because used benchmarks do not contain corresponding resources.

[10] https://jena.apache.org/

[11] http://sw.opencyc.org/

[12] https://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc



(a) MC-30

(b) RG-65

(c) WSIM-353

Figure 1: Evaluation of different proposed similarities based on MC-30 (a) , RG-65 (b) and WSIM-353 (c) benchmarks.

B. Benchmarks and concepts mapping

To evaluate the proposed similarity measure, we rely on three well-known benchmarks: the MC-30 that contains 30 pairs of concepts [9], RG-65 [10] that contains 65 pairs of concepts, and similarity gold-standard provided by WSIM-353[13] collection. We try to attach a DBpedia resource for each concept in the benchmarks. If no corresponding resource is found, we take the resource to which the benchmark concept is redirected. If no corresponding resource is obtained, we search an equivalent class from YAGO classes. After applying this process, 25 pairs of concepts from MC-30 was linked to its equivalent DBpedia resources and five pairs were mapped to YAGO classes. For RG-65, we've mapped 54 pair concepts to DBpedia resources, while 11 pairs were mapped to YAGO classes. For WSIM-353, we've successfully linked 154 concepts to DBpedia resources while 43 are mapped to YAGO classes and six concepts have not mapped.

During experiments, we have found that the maximum path length $\Pi = 2$.

C. Result of different measures

We evaluate each sub-similarity measure separately, before combining them in LODS. We analyze the effect of data augmentation on each measure. The Spearman's correlation coefficient is used to compare results provided by the measures and the benchmarks.

We have conducted experiments of *SimC* measure and LODS measure with the following super and sub-categories levels $\ell, \ell' = \{(1,0),(2,0),(3,0),(1,1),(2,1),(2,2),(3,1),(3,2)\}$. Figure 1 shows results of the two measures with levels that give the best correlation.

Ontology-based measure SimI. As shown in figure 1 the augmentation has the greatest impact on $SimI$ measure than the others in the three evaluation benchmarks. This is due to the poor instantiation of DBpedia resources. So, augmenting data from other datasets such as Wikidata which contains rich ontological taxonomy of its entities (resources), contributes to the enhancement of $SimI$ results.

Classification-based measure SimC. This measure gives better results compared to other sub-measures, sometimes even without data augmentation due to the richness of taxonomic structure of categories which enables efficient comparison of resources. Nevertheless, data augmentation shows no noticeable impact; it has no impact at all in MC-30 (respectively, RG-65) benchmark when $\ell, \ell' = \{(1,0),(1,1)\}$ (respectively $\ell, \ell' = \{(2,0),(2,1),(2,2)\}$). Augmentation has adverse effect in cases where $\ell, \ell' = \{(2,0),(2,1),(2,2),(3,1)\}$ for MC-30 benchmark (correlation is decreased 1%) and $\ell, \ell' = \{(3,0),(3,1),(3,2)\}$ for RG-65 (correlation is decreased by 3% to 4%). However, it has always a positive effect on third benchmark WSIM-353 (correlation is between 0.48 to 0.64 without data augmentation, and it is between 0.56 and 0.67 with data augmentation) and the rest of the levels in the other two benchmarks. Two main reasons behind the observed results: (i) Most resources are classified, so there is no need to get missing categories from equivalent resources. (ii) DBpedia chapters categories usually have the same hierarchy because they are usually translated from those of original English Wikipedia. We reckon that including distinct classification schemata when available in LOD, will improve results of this measure.

Properties-based measure SimP. This measure returns weak results even with data augmentation (which has been improved by 5%, 9% and 4% for MC-30, RG-65 and WSIM-353 respectively). We think that the lack of distinguishing properties and the presence of commonly used properties are behind the weakness of the results. We believe that this measure will give better results when resources are well described by distinguished properties. For instance, people are the kind of resources, which are not available in used benchmarks, that are well described with distinguished properties in LOD.

Combined measure LODS. This measure balances results of sub-measures to give highest results in all benchmarks. Correlation without data augmentation ranges from 0.77 to 0.81 for MC-30, from 0.66 to 0.72 for RG-65 and from 0.63 to 0.68 to for WSIM-353. With considering data augmentation, it ranges from 0.82 to 0.83 for MC-30, from 0.73 to 0.76 for RG-65 and from 0.67 to 0.69 for WSIM-

[13] http://www.cs.cmu.edu/~mfaruqui/word-sim/EN-WS-353-SIM.txt

353.

We have noticed that correlation results are approximate. So, when using proposed measure in a particular application, it is possible to choose lower ℓ and ℓ' levels to gain some performance.

V. Related works

The remarkable growth of LOD has encouraged researchers and also developers to exploit this wealth source of knowledge to tackle a multitude of problems such as improving search engines, resolving interoperability between systems or developing efficient recommendation systems.

In [11] the author proposed a Linked Data Semantic Distance (LDSD) which relies on direct and indirect relationships between two DBpedia resources. The distance measure was employed in a music recommendation system [12]. The proposed LDSD measure was only applied on a cleaned dataset of DBpedia to compute similarity, and it does not benefit from external interlinked datasets. Moreover, it considers only ontology concepts that are used to annotate resources without taking into account concepts that could exist in the taxonomic structure. For instance, a DBpedia resource described with UMBEL concept does not include all its subsumers. Furthermore, LDSD takes into account only relations between resources; it ignores other properties that could characterize and distinguish resources.

$REWOrD$ [1] is an approach to compute semantic relatedness between entities. It is based on a Predicate Frequency, Inverse Triple Frequency (*PF/ITF*) model inspired from *TF/IDF* which is used to compute the informativeness of the paths that connect compared resources. The author has evaluated the measure separately on DBpedia and LinkedMDB based on a proposed benchmark. So, he did not benefit from links that can exist between interlinked datasets to get increase data. Moreover, *REWOrD* did not consider characterizing properties; it relies only on resources and existing paths between them in the similarity calculation process. In our approach, we take into account these two considerations.

An information content (IC) based approach has been proposed by [8] to compute the similarity between LOD resources. The proposed measure called Partitioned Information Content Semantic Similarity (PICSS), uses ingoing and outgoing edges as features to compare resources based on Tversky model [6]. An IC based measure was applied to resources features to give more importance to significant relations. During experiments, resources features have been enriched from different LOD datasets, and results showed noticeable improvements. Contrary to our approach, the approach lacks theoretical foundations and has the same shortcomings of [11] and [1].

VI. Conclusion and future work

In this paper, we have played the role of LOD consumer by querying, filtering and integrating its data to propose a new similarity measure. It relies on the taxonomic structure of ontological concepts and classification categories of LOD resources, in addition to their characterizing properties. Our approach exploits interlink relationships that exist between LOD datasets to enrich data involved in the computation of the similarity score. The goal is to reduce the problem of shortage of information within a single dataset. Experiments show that similarity measure gives good results, and data enrichment contributes to the enhancement of the similarity scores.

As future work, we intend to apply our similarity measure to improve mobile Web services discovery and recommendation. We believe that LOD is is suitable for such domain as it is open and needs multi-domain knowledge sources.

References

[1] Pirró, G, "REWOrD: Semantic Relatedness in the Web of Data," in *the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2012, pp. 129–135.

[2] R. Meymandpour and J. G. Davis, "Linked data informativeness," in *Web Technologies and Applications*. Springer, 2013, pp. 629–637.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[4] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[5] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, "DBpedia - A Large-scale , Multilingual Knowledge Base Extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[6] A. Tversky, "Features of similarity." *Psychological Review*, vol. Psychological Review, no. 4, pp. 327–352, 1977.

[7] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7718–7728, 2012.

[8] R. Meymandpour and J. G. Davis, "Enhancing Recommender Systems Using Linked Open Data-Based Semantic Analysis of Items," in *The 3rd Australasian Web Conference (AWC 2015)*, 2015, pp. 27–30.

[9] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.

[10] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[11] A. Passant, "Measuring semantic distance on linking data and using it for resources recommendations." in *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, vol. 77, 2010.

[12] ——, "dbrec Music Recommendations Using DBpedia," in *The Semantic Web–ISWC 2010*. Springer, 2010, pp. 209–224.