Statistical analysis of the `owl:sameAs` network for aligning concepts in the linking open data cloud

Gianluca Correndo, Antonio Penta, Nicholas Gibbins, and Nigel Shadbolt

Electronics and Computer Science, University of Southampton, UK,
`[gc3,ap7,nmg,nrs]@ecs.soton.ac.uk`,
WWW home page: `http://www.ecs.soton.ac.uk/people/[gc3,ap7,nmg,nrs]`

Abstract. The massively distributed publication of linked data has brought to the attention of scientific community the limitations of classic methods for achieving data integration and the opportunities of pushing the boundaries of the field by experimenting this collective enterprise that is the linking open data cloud. While reusing existing ontologies is the choice of preference, the exploitation of ontology alignments still is a required step for easing the burden of integrating heterogeneous data sets. Alignments, even between the most used vocabularies, is still poorly supported in systems nowadays whereas links between instances are the most widely used means for bridging the gap between different data sets. We provide in this paper an account of our statistical and qualitative analysis of the network of instance level equivalences in the Linking Open Data Cloud (i.e. the `sameAs` network) in order to automatically compute alignments at the conceptual level. Moreover, we explore the effect of ontological information when adopting classical Jaccard methods to the ontology alignment task. Automating such task will allow in fact to achieve a clearer conceptual description of the data at the cloud level, while improving the level of integration between datasets.

Keywords: Linked Data, ontology alignment, owl:sameAs

1 Introduction

The increasing amount of structured information published on the Web in linked data is rapidly creating a voluminous collective information space formed of inter-connected data sets; the Linking Open Data cloud (LOD henceforth). The last version of the LOD diagram (2011/09/19) included 295 data sets, ranging from topics like encyclopaedic knowledge, to e-government, music, books, biology, and academic publications. These data sets are linked, most of the times, at the instance level where URIs representing entities are reused or aligned towards external URIs using `owl:sameAs` properties to link equivalent entities. According to OWL semantics [2], all entities within the closure set of the `owl:sameAs` relation are indistinguishable, thus every statement including one entity can be rewritten by replacing any of the equivalent element.

The problem of discovering "same" entities in different data sets, known as the record linkage problem, is quite well known in database community where a large body of literature can be found on the topic [20]. Semantic Web community has built upon the database research and proposed its set of solutions [7]. The discovery of equivalent entities in the Web of Data is therefore supported by automatic tools which exploit, similarly to ontology matching or record linkage tools, lexical and/or structural similarities between the entities of different data sets [10, 18]. Semi-automated approaches has been also implemented in tools like Google Refine[1], where linkages found are subject to user approval. The collaborative effort of data publishers in inter-connecting their data sets has created a network of equivalences between instances which is a matter of study on its own, the **sameAs** network [4]. Studying the properties of this **sameAs** network in conjunction with the network of Class-Level Similarity, or CLS network as defined in [4] (i.e. the network of classes which overlaps because sharing same, or equivalent, instances), can lead us to a better understanding of how heterogeneous data sets can be integrated together.

Despite of the great amount of linkages between instances and the high availability of tools for aligning vocabularies, little effort has been devoted to provide authoritative alignments between the ontologies present in the LOD. As a representative example, in DBpedia the only alignments between ontologies, retrieved by querying the public endpoint, have been published by using `owl:sameAs` properties between concepts in `opencyc.org` [2], and `owl:equivalentClass` properties between `schema.org` [3] concepts.

The availability of ontology alignments in the LOD would allow the use of tools that exploit schema level mappings for achieving data integration [19], fuelling in this way a wider use of published linked data. The work described in this paper starts from the above consideration and attempts to exploit the available **sameAs** network in order to deduce statistically sound dependencies between concepts which have common instances taking into consideration the semantics attributed to the `owl:sameAs` property [2].

The work we presented in this paper is an account of our first attempts to adopt a well known instance-based technique (i.e. Jaccard coefficient) in discovering alignments between concepts in the LOD cloud. The vast amount of entity alignments present in the LOD cloud, under form of `owl:sameAs` statements, provides a good asset to experiment such an approach. Although, applying statistical techniques to a potentially very noisy data set for aligning heterogeneous ontologies could prove to be unreliable to some extent. This paper reports our attempts to study the behaviour of such basic technique on a real scenario. The rationales behind this approach are to be found in a previous work in the instance based ontology matching field [9], which did not addressed specifically the LOD, and an analysis on the deployment of `owl:sameAs` networks [4].

The paper starts with Section 2 which provides some background information on instance based ontology alignments, how they are implemented in this work by exploiting `owl:sameAs` alignments, and finally describes the data used in this experiment. Section 3 provides some initial analysis, quantitative as well as

[1] `http://code.google.com/p/google-refine/`

[2] `http://sw.opencyc.org/`

[3] `http://schema.org`

qualitative, on the data used and on the alignments found in the CLS network. Section 4 provides an account of the behaviour of Jaccard based measures under different hypothesis by studying the usual indices from Information Retrieval (i.e. number of alignments, precision, and recall). Section 5 provides an account of similar works in the area of Linked Data and finally our conclusions are presented in Section 6.

2 Alignment based on sameAs network analysis

The ontology alignment task has been widely studied in the last decade by the scientific community [7]. Ontology matching tools usually exploit a number of information sources such as lexical or structural similarity applied to the ontologies alone in order to produce a measure of the semantic distance between concepts. In recent years, methods based on statistical information (e.g. machine learning, bayes, etc.) have been also studied and proved to produce promising results [5, 9].

The high level of inter-linking within the LOD cloud induces us to consider statistical techniques for ontology alignment as a promising approach to resolve semantic heterogeneity. The assumption we adopted in this work is that `owl:sameAs` equivalence bundles [8] can be treated as singleton instances whose interpretation is provided by following `owl:sameAs` semantics. Therefore all equivalent instances, hosted by different data sets, will be considered as a unique instance which is classified differently in different data sets (as seen in Rule 1 where *type* is `rdf:type` and *sameas* is `owl:sameAs`).

$$\mathit{type}(?x, ?xt) \wedge \mathit{sameas}(?x, ?y) \wedge \mathit{type}(?y, ?yt) \rightarrow \mathit{type}(?b, ?xt) \wedge \mathit{type}(?b, ?yt) \quad (1)$$

Leveraging the `owl:sameAs` inference we are then able to treat equivalence bundles as instances and compute the degree of overlapping between concepts by processing the typing statements (i.e. statements in the form $\mathit{type}(?b, \mathit{class})$). In our approach we used the Jaccard coefficient [11] ($J(A, B)$ in Equation 2) in order to measure the similarity between two concepts when interpreted as sets of instances.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (2)$$

Here the cardinality of the intersection set $|A \cap B|$ is computed in our triple store by counting the cardinality of the set $\{\langle x, y\rangle : \mathit{type}(x, A) \wedge \mathit{same}(x, y) \wedge \mathit{type}(y, B)\}$. The cardinality of the union set is then computed by summing the cardinality of the set of instances for the two concepts A and B (i.e. $\{x : \mathit{type}(x, A)\}$ and $\{x : \mathit{type}(x, B)\}$ respectively) and then subtracting the intersection as previously defined.

2.1 Definition of ontology alignment

In the work here described we reused and modified the framework proposed in Isaac et al. [9] for representing instance-based alignments. In [9] an alignment between a source ontology S and a target ontology T is a triple $t \in S \times T \times R$

where R is a relation taken from the set $\{\equiv, \sqsubseteq, \sqcap, \perp\}$ which expresses respectively equivalence, subsumption, overlap and disjointness. Such definition fits a scenario where describing some informal degree of relatedness, measurable by sets overlapping, is acceptable and even desirable. Given the target objective of our work, data integration, we set for a less richer framework where it is possible to distinguish only between $\{\equiv, \sqsubseteq, \perp\}$ since we could not make any use of information about overlapping concepts for integrating different data sets into an homogeneous vocabulary. Moreover, when we state that two concepts are equivalent (i.e. $A \equiv B$), since we are not taking into account the concepts definitions but merely the possibility of them covering the same set of instances, we will intend that the two concepts are in `owl:equivalentClass` relationship, and not `owl:sameAs`. Hence, the two concepts can still have different definitions without causing any inconsistency.

In the subsequent evaluation of the alignments (see Section 4) we will consider a successful alignment, a true positive, one which correctly correlate a couple of concepts which are equivalent or in a relation of subsumption (i.e. one subsumes non trivially the other). Any alignment provided which includes two disjoint concepts is a false positive. This shrink in the power of discrimination is due to the nature of the Jaccard measure itself which has bees devised to measure concepts equivalence only.

2.2 Experimental setup based on LOD entities

In order to experiment the usefulness of the Jaccard coefficient as a means for measuring the semantic similarity between concepts in the LOD cloud a source data set and a number of target data sets, aligned to the selected source by `owl:sameAs` links, have been considered.

Because of its centrality in the LOD cloud, `DBpedia` is the natural candidate as a source data set while a number of target data sets has been selected based on their abundance of instance alignments and diversity of size in terms of concepts to align. The target data sets considered for our experiments are described in Table 1 where for the source data set is reported the number of info box concepts used to classify the DBpedia instances[4], and for each one of the target data set is reported the number of equivalence links connecting it to DBpedia and the number of concepts contained. It noteworthy that for for the `nytimes` data set, although containing a rich hierarchy of terms, only two concepts are found. This is due to the fact that all the entities aligned are instances of `skos:Concept` and some of `geonames:Feature`. Therefore, not knowing any background information about the dataset, it is not possible to recognize a valid OWL concept hierarchy since it is encoded in a vocabulary (i.e. SKOS [15]) which encodes concept hierarchies only between instances and not between concepts. This implies that every instance mapping from an instance in `DBpedia` to one in `nytimes` will support a correspondence between an OWL concept to the concept `skos:Concept` which is not very informative as an alignment between ontologies.

[4] The dump used in this experiment is the DBpedia version 3.7.

Table 1. Data sets considered for the experiments

source		number of concepts
DBpedia		9237320
target	number of `owl:sameAs`	number of concepts
opencyc	20362	314671
nytimes	9678	2
drugbank	4845	4
diseasome	2300	2
factbook	233	1
dailymed	43	1

Once identified the source and target data sets, we proceeded to download from the respective websites the triples belonging to: the **sameAs** network; the **type** network; and the **concepts hierarchy**. As we already mentioned, the **sameAs** network of a data set D is the set of triples contained in D which connect two entities by the property `owl:sameAs` (i.e. consistently with the notation already used: $\mathit{sameas}(D) = \{\mathit{same}(s, p) \in D\}$). The **type** network of a dataset D is the set of triples contained in D which connect every entity with one or more concept by the property `rdf:type` (i.e. $\mathit{type}(D) = \{\mathit{type}(a, b) \in D\}$). Finally, the **`concepts hierarchy`** is the set of triples contained in D wich connect two concepts by the property `rdfs:subClassOf` (i.e. $\mathit{hierarchy}(D) = \{\mathit{subclassof}(a, b) \in D\}$).

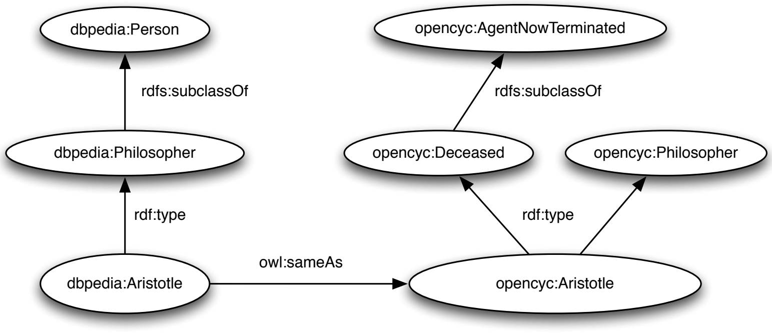

Fig. 1. Example of different networks extracted from DBpedia neighbours

An example of the networks taken into consideration in this paper are depicted in Figure 1. For the sake of the statistical analysis conducted in the experiments we did not take into consideration any other property which could describe the entities and the concepts (e.g. labels, abstracts or other properties) since only the `owl:sameAs` bundles are considered. In Figure 1 we can see how the bundle {`dbpedia:Aristotle, opencyc:Aristotle`} belongs to the intersection of the concepts `dbpedia:Philosopher` and `opencyc:Philosopher`, and `dbpedia:Philosopher` and `opencyc:Deceased`. By computing the number of co-occurrences of concepts connected by a common bundle in the way showed by Figure 1 we are able to compute the size of the intersection set and then to compute the Jaccard measure for each couple of concepts.

3 Experiment scenario analysis

In order to better understand the characteristics of such collected network, we decided to study a couple of aspects before processing the data in trying to discover concept alignments. The first thing we decided to look into is the size of the sameas bundles collected. Since the number of concepts reached from a single DBpedia entity can be reasonably related to the size of its equivalent class computed via `owl:sameAs` links, studying the distribution of such parameter can give us an insight about the variance we can expect in processing such bundles. The distribution of the frequency of the bundles' size is depicted in the graph in Figure 2, where the dimension of the y axes is reported in logarithmic scale. Considering the distribution in Figure 2 we can see that the size of bundles follows a logarithmic distribution where the more frequent size is 2, i.e. only one other entity except the source entity, and where bundles of size greater than 10 are very infrequent.

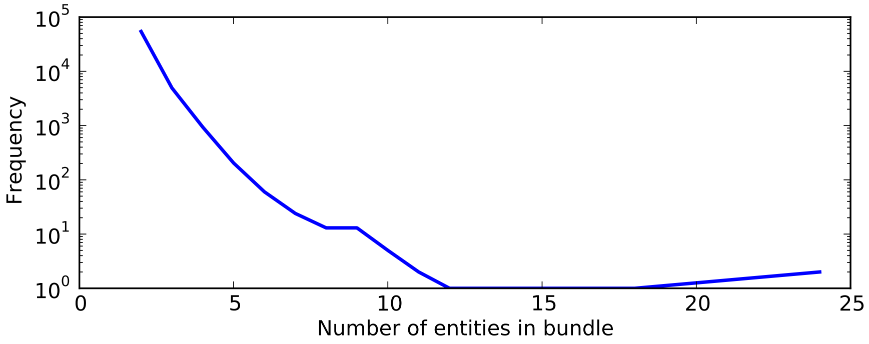

Fig. 2. Frequency of sameas bundles by size

The second aspect we studied, once we computed the Jaccard coefficient from the collected data, is the ratio of the cardinalities between aligned concepts. The hypothesis we formulated, given our past experience in handling linked data, is that the cardinality of overlapping concepts in the LOD cloud would be highly heterogeneous and therefore we would have a high level of asymmetry between the aligned data sets. In Figure 3 is reported the frequency of alignments plotted against the ratio (expressed in percentage) of the cardinality of the two concepts aligned[5]. Looking at Figure 3 we can say that concepts with similar cardinality would be nearer the right end of the graph, while concepts dissimilar in cardinality would be nearer to the left end of the graph; the graph reported makes clear that the vast majority of alignments produced are between concepts dissimilar in cardinality. Although this result is particular to the scenario under scrutiny, it is also true that DBpedia is a typical example of a general domain

[5] The ratio has been normalised between [0,1].

hub data set whose behaviour in terms of inter linkages is likely to be seen in other hub data sets.

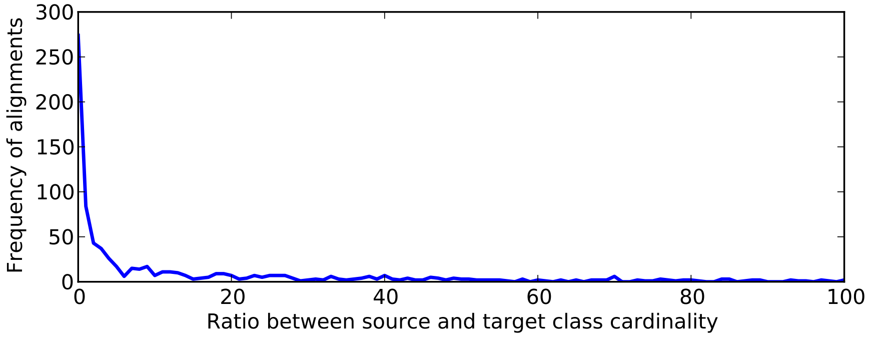

Fig. 3. Ratio of cardinalities between aligned concepts

3.1 Qualitative analysis of Jaccard alignments

Before discussing into detail the quality of the Jaccard coefficient as a means for producing ontology alignments in the LOD cloud (topic covered in Section 4), we would like to provide a qualitative analysis of the first batch of results. Among the alignments we have manually checked for judging their quality, in order to compute the precision and recall of the procedure, we noticed that some of the alignments produced, which were supported by statistical evidence, were quite interesting in nature.

Many of the alignments produced, even with a high value of Jaccard coefficient, had the `owl:Thing` as a source concept. This is due to the fact that many DBpedia instances have multiple types associated, and the top of the hierarchy is directly, and quite frequently, mentioned in the type network, whatever the level of abstraction of the entity is. This fact hinders us in identifying a canonical classification of entities and introduces some noise in discovering concept alignments. The root of the OWL hierarchy is in fact non trivially equivalent to any other concept in any other ontology, at the same time is the superclass of all OWL based hierarchies, therefore any mapping would provide very little information gain.

Similarly, as mentioned earlier for the `nytimes` data set, encoding all entities as `skos:Concept` instances, implies that the only mapping one can find within the data set is to that concept, rendering useless any alignment effort. We may expect to find the same results every time we try to exploit entities alignments between domain concept instances and knowledge organization systems as thesauri and classification schemes.

The last consideration we did on the alignments found is on some related patterns that seem to be quite common and which could be justified by a cultural and contextual interpretation of the data, and alignments. An account of some of the unusual patterns discovered by processing the concepts co-occurrences is provided in Table 2. As we can see, the first two alignments are indicative of

Table 2. Concept alignment patterns

source[`dbpedia`]	target[`opencyc`]
Model	Woman
Writer	Male
Philosopher	Dead organism
Monument	Bell

a statistical preference of representing female models and male writers[6]. The alignment between concept *Philosopher* and *Dead organism* is proposed as less likely than the correct alignment (i.e. opencyc *Philosopher* concept), and it is probably due to the fact that the vast majority of the philosophers described in DBpedia are actually deceased. The last alignment is due to the fact that in DBpedia, listed as entities of type *Monument* are just historical bells (e.g. the Liberty Bell in Philadelphia). Therefore, although odd, the wrong alignment reflects the extensional definition of the concept which clearly conflicts with the semantics we would expect from the *Monument* concept.

4 Evaluation

In order to study the behaviour of Jaccard alignments we collected the usual measures from Information Retrieval under different conditions. We proposed in fact two scenarios that affect either the way the alignments are produced or the way the alignments are used, and we measured the performances of Jaccard for each scenario. The measures under scrutiny are: the **Number of alignments** computed (either correct or incorrect), the **Precision** of the alignments computed, **Recall** of the alignments computed, the **F-measure**[7] of the results, and finally the **Precision at** n^{th} of the alignments[8].

The first scenario explored the gain we have when we take into account (or not) the concept hierarchy when we compute the cardinality of the two sets, A and B respectively, as defined in Section 2. For doing this we used 4sr, a reasoner that efficiently implements the reasoning over `rdfs:subClassOf` axioms (sc_0 and sc_1 rules in [17]).

The second scenario studied the different performances we gain when relaxing the acceptance criteria from **equivalence** only (i.e. an alignment is considered correct if the two concepts are equivalent) to **equivalence** or **subsumption** (i.e. an alignment is considered correct even when two concepts are not trivially subsuming one another). Around a thousands of the generated concept alignments have been manually checked and classified as: **e**rroneous, **s**ubclass/ uperclass, or **c**orrect. The precisions have been computed by considering as successful either **s**ubclass/uperclass and **c**orrect or **c**orrect only. The legends of Figure 4, 5, 6, 7,

[6] Note, this is not due only to the particular source data considered but also to the instance alignment performed on the target data set.

[7] The harmonic mean of precision and recall

[8] The precision computed for the first n^{th} alignments

and 8 reports **Jaccard** when no hierarchy information is used and only equivalent entities are considered as correct, **h Jaccard** when hierarchy information is used and equivalent concepts are considered, **s Jaccard** when no hierarchy information is used and with subclasses considered as correct, and finally **hs Jaccard** when hierarchy information is used and with sub concepts considered as correct. Finally, the recall of the respective measures have been computed by taking the maximum number of correct alignments found as the reference limit. That is why in Figure 6 the legend reports **Relative** recall in the label.

4.1 Number of alignments

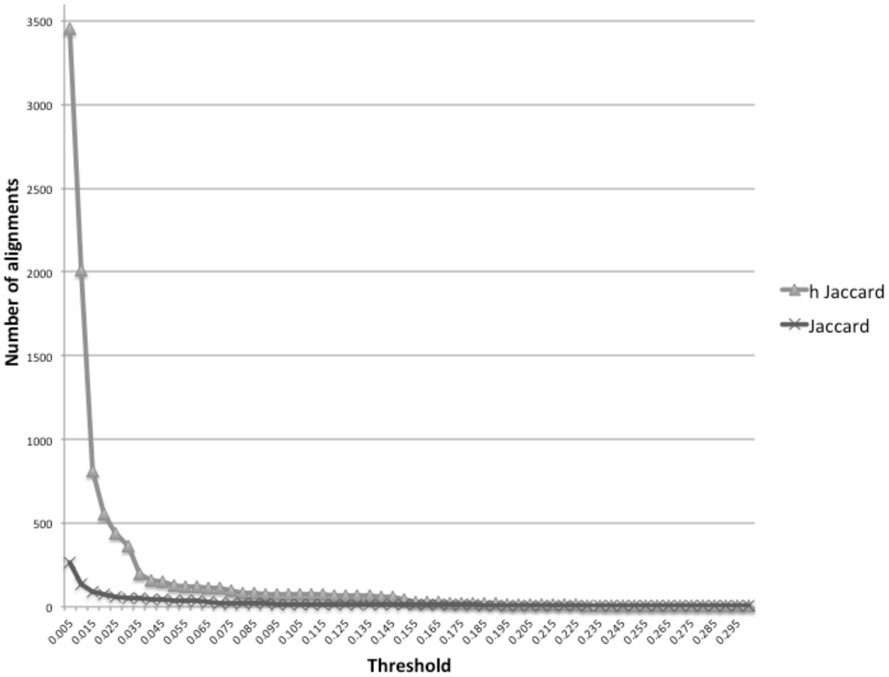

Fig. 4. Number of alignments found per threshold value

The comparison here is made by using or not the concepts hierarchy in the computation of the Jaccard coefficients, since the acceptance of the produced alignments does not influence their generation. A first superficial analysis of the distribution of the number of alignments found per different values of thresholds (see Figure 4), the number of alignments produced increases exponentially when lowering the threshold value and it is noteworthy the fact that the most of the alignments are produced with very little values of threshold. This implies that it is important to maintain a good quality of the alignments even at low values of thresholds since the amount of false positives could hinder the usability of the produced alignments to a point where human intervention could not be feasible any more.

Moreover, comparing the distribution of generated alignments we can notice that, even if both distributions are inversely exponential, including hierarchical information increases drastically the number of produced alignments. A superficial analysis showed that the rate between the two distributions increases from 1, for higher values of thresholds, to 15, for lower thresholds.

4.2 Precision

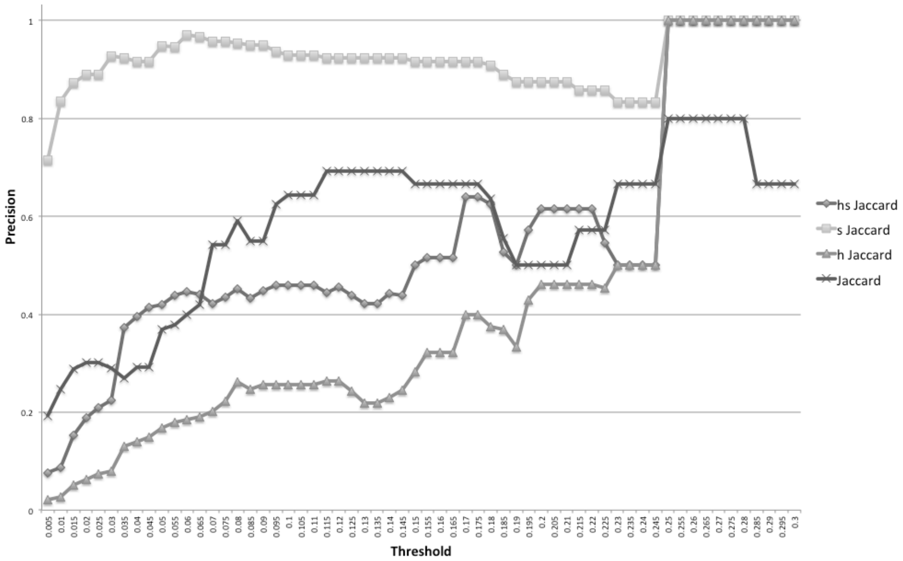

Fig. 5. Precision per threshold value

In Figure 5 it is shown the graph of the distribution of the precision of the alignments, under different conditions as described earlier, by varying the value of threshold. The comparison of Jaccard performances (acceptance for equivalence) with and without hierarchical information shows that when decreasing the threshold level, the more informative measure (the one with the hierarchical information) drops its precision level drastically and from that point on its precision is always worse than the less informative measure (i.e. the one without hierarchy information).

Comparing instead the two acceptance criteria, the one for equivalence and the one with equivalence or subsumption (see Figure 5 **Jaccard** and **s Jaccard**), we can notice that Jaccard provides increasingly imprecise equivalence alignments when lowering the threshold, while the precision of the method is steadily high if we are satisfied with alignments that we can refine later on. Even then though, by using hierarchical information (see Figure 5 **hs Jaccard**) the precision of the method drops quickly to unacceptable levels.

One striking fact from all the precision distributions depicted in Figure 5 is that such distributions are not monotone non-decreasing as one would expect. In fact for all distributions there are frequent local maxima and only the general trend is, for all plots, increasing. This strange behaviour could be caused by the high level of noise within the **sameAs** network, although further experiments are needed to confirm that.

4.3 Relative recall

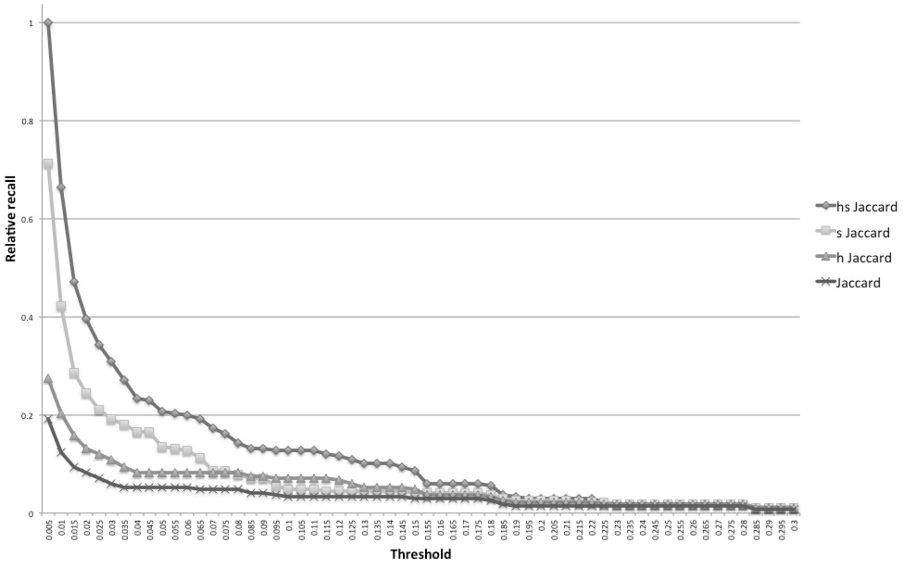

Fig. 6. Relative recall per threshold value

In Figure 6 we can see the graph for the relative recall of each measure. Not surprisingly we can see that all distributions are monotonic decreasing and that, roughly for all levels of thresholds, **hs Jaccard** provides the highest recall, followed by **s Jaccard**, **h Jaccard**, and finally **Jaccard** which is the less prolific method. For lowest levels of thresholds we can see that measures that share the same acceptance criteria provides more similar recall values while the use of hierarchical information, although it increases the recall of a method, it affects less heavily the overall behaviour of a method.

4.4 F-measure

The value of F-measure computed as the harmonic mean of precision and (relative) recall depicted respectively in Figure 5 and Figure 6 are reported in Figure 7. The most remarkable thing when considering the plots in Figure 7 is that the two less informed measures (i.e. **s Jaccard** and **Jaccard**) shows the same monotonic non-increasing trend while the two most informed measures (i.e. **hs Jaccard** and **h Jaccard**) have a local max before decreasing.

Although F-measure is only an indication of the overall performances of an information retrieval method, the results of the experiments conducted by the authors seem to suggest that, when the more alignments are retrieved by lowering the threshold value it is best not to use hierarchical information. In this way in

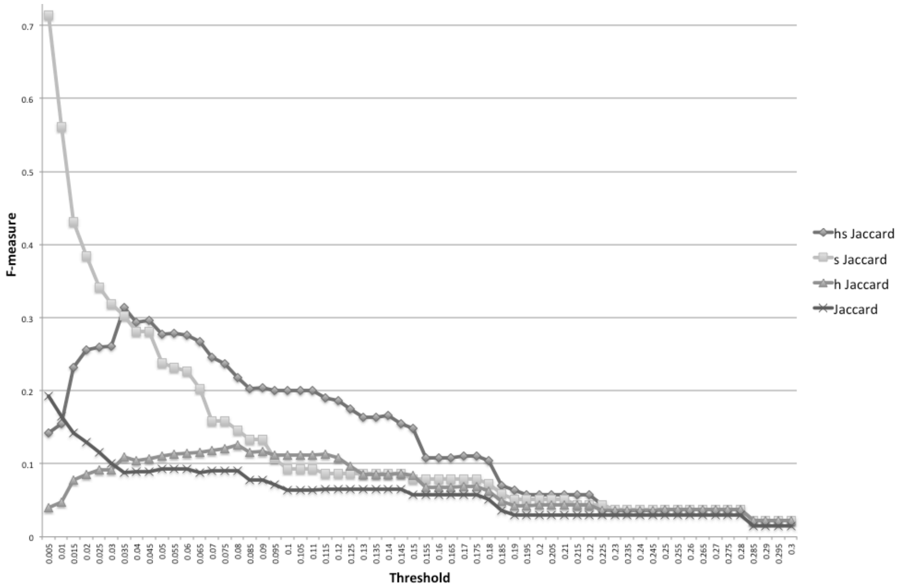

Fig. 7. F-measure per threshold value

fact the overall performances, precision and recall wise, seem improving steadily making still rewarding the consideration of alignments even for low levels of thresholds (i.e. when more noise is expected).

4.5 Precision at n

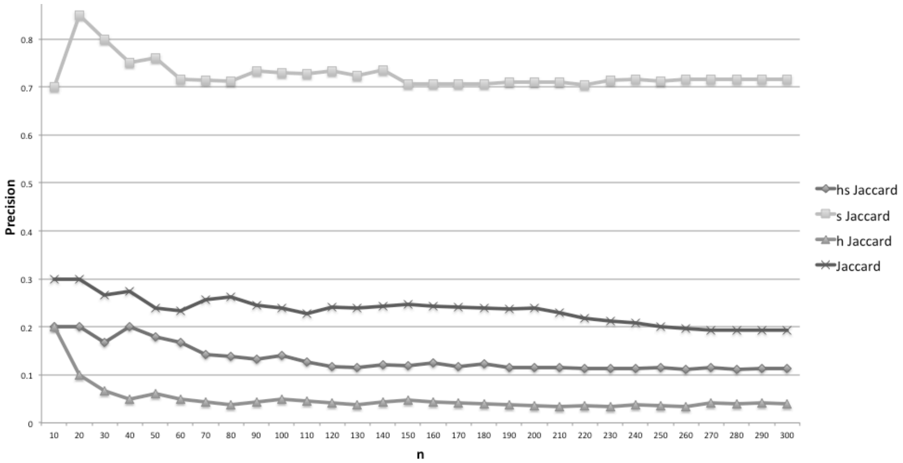

Fig. 8. Average precision at n^{th} alignment

The actual usability of Jaccard alignment in a user engaging scenario can also be judged by looking at the average precision of the measures for the first n alignments. The precision at n for all the considered scenarios is plotted in Figure

8 where we can clearly see that for all scenarios the average precision quickly stabilize after the first 50 alignments. **s Jaccard** provides the best precision with 7 good alignments out of 10. All the other scenarios perform quite poorly: **Jaccard** with 2 good alignments out of 10, **hs Jaccard** with 1 out of 10, and finally **h Jaccard** with 1 out of 20 correct alignments.

5 Related Work

The alignment problem, which is finding correspondences among concepts in ontologies, is well studied research problem in the Semantic Web community [7], as well as in the Database community to support the integration of heterogeneous sources [13] or to solve record linkage problems [6]. In particular, a lot of work is done by researchers in the last years in the context of Ontology Matching [7], where the alignments are the output of these systems. Different techniques have been proposed to address this problem in order to improve the performance related to the Ontology Merging, Integration and Translation systems. An evaluation competition [9] is also proposed to compare those matching systems using common testbeds. These techniques can be described in four main categories : i) *lexical*, which means that are based on detecting similarities between the concept descriptions such as labels; ii) *structural*, which means that are based on the knowledge descriptions; iii) *instance mapping*, which means that are based on the knowledge expressed in the ABox. Most of the proposed techniques comes from the Machine Learning research area [7, 16, 5, 14]. In literature [3, 1] also different measures are proposed to evaluate the semantic similarities among concepts that takes into account :i) the expressive power of the description logic used by the knowledge bases, ii) the information content assigned to the observed classes. Despite the different studies in the theoretical background, we observe a marginal effort in evaluate these approaches in the Linked Data Cloud, which is the most concrete realization of the Semantic Web vision nowadays and they are a valuable resources for different application domain. In literature, the most recent works that use Linked Data in the simalar context of our paper but with different purposes are [4, 12]. In particular the first evaluates the implications of `owl:sameAs` assertions in Linked Data data sets and the second uses the LOD to evaluate an ontology matching system. In our paper we boost some of these previous studies in order to give a real evaluation in the context of Linked Data Cloud. From the instance-based matching techniques, the closest paper is [9], our contribution differs from the previous one in the richness of measures adopted and in evaluation proposed in the Linked Data environment.

6 Conclusion

In this paper we conducted some experiment with Jaccard-based concept similarity measures based on the analysis of the instance alignments provided by the `sameAs` network that connects **DBpedia** with some of the neighbourhood data sets. Being the chosen domain very broad (i.e. **DBpedia** concepts) and

[9] http://oaei.ontologymatching.org/

the alignments not focused on any specific application, we assumed to have very noisy results which suggested the use of statistical methods a natural choice.

The first analysis on the experimental data showed the typical signs of a power-based network, where a small number of `sameAs` bundles contained many entities and the vast majority contained no more than five instances (see Figure 2). We devised four different scenarios under which analyse the behaviour of classical Jaccard similarity measure, studying the influence of hierarchical information in producing the alignments and the difference when choosing a broader acceptance criteria.

The experimental results showed that Jaccard, for this particular DBpedia experiment, provided very low values which makes it difficult to choose a good threshold value which produced a fair amount of good alignments. The results outlined that the use of hierarchical information in computing concepts similarity measures increased drastically the number of alignments found but unfortunately dropped the precision of the results as well making increasingly inconvenient to consider further alignments below a given threshold. Conversely, by considering the concepts detached by a subclass hierarchy, Jaccard measures improve steadily.

The relaxation of the acceptance criteria on the other hand, did not influence the overall performance of the measures while giving better performances of the respective more restrictive measures. This is not surprising since the alignments found and the coefficients computed are the same when hierarchy is counted in or not, and it changes only the criteria for the acceptance, and one criteria includes the other. Ultimately, a less restrictive acceptance criteria, without hierarchy information, gives us a better overall performance and stably produces a fair amount of sensible alignments.

This scenario suits best an approach where alignments can be proposed to users for classification and where more elaborate alignments (i.e. not only concept equivalence) can be exploited for integrating data.

Future work will include a better study of the sources of noise in Jaccard-based methods when applied to the in order to provide a robust methodology for aligning ontologies at Web scale.

Acknowledgments. This work was supported by the EnAKTing project funded by the Engineering and Physical Sciences Research Council under contract EP/G008493/1.

References

1. H. Al-Mubaid and H.A. Nguyen. Measuring semantic similarity between biomedical concepts within multiple ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(4):389 –398, 2009.
2. Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language Reference. W3C Recommendation, February 2004.
3. Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. A semantic similarity measure for expressive description logics. *Computing Research Repository-arxiv.org*, 2009.

4. Li Ding, Joshua Shinavier, Zhenning Shangguan, and Deborah L. McGuinness. SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl: sameAs in Linked Data. In *International Semantic Web Conference (1)*, pages 145–160, 2010.
5. Anhai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Ontology matching: A machine learning approach. In *Handbook on Ontologies in Information Systems*, pages 397–416. Springer, 2003.
6. A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1 –16, 2007.
7. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer, Berlin, Heidelberg, 2007.
8. Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009)*, April 2009.
9. Antoine Isaac, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang. An empirical study of instance-based ontology matching. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 253–266. 2007.
10. Robert Isele, Anja Jentzsch, and Christian Bizer. Silk Server - Adding missing Links while consuming Linked Data. In 1^{st} *International Workshop on Consuming Linked Data (COLD 2010), Shanghai, China*, November 2010.
11. Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.
12. Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma, and Peter Z. Yeh. Ontology alignment for linked open data. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, pages 402–417, Berlin, Heidelberg, 2010. Springer-Verlag.
13. Maurizio Lenzerini. Data integration: a theoretical perspective. *In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.
14. Ming Mao, Yefei Peng, and Michael Spring. Ontology mapping: as a binary classification problem. *Concurrency and Computation: Practice and Experience*, 23(9):1010–1025, 2011.
15. Alistair Miles and José R. Pérez-Agüera. SKOS: Simple Knowledge Organisation for the Web. *Cataloging & Classification Quarterly*, 43(3):69–83, 2007.
16. Mathias Niepert, Christian Meilicke, and Heiner Stuckenschmidt. A probabilistic-logical framework for ontology matching. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
17. Manuel Salvadores, Gianluca Correndo, Steve Harris, Nick Gibbins, and Nigel Shadbolt. The design and implementation of minimal rdfs backward reasoning in 4store. In *ESWC (2)*, pages 139–153, 2011.
18. Manuel Salvadores, Gianluca Correndo, Benedicto Rodriguez-Castro, Nicholas Gibbins, John Darlington, and Nigel Shadbolt. LinksB2N: Automatic Data Integration for the Semantic Web. In *International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, June 2009.
19. Andreas Schultz, Andrea Matteini, Robert Isele, Christian Bizer, and Christian Becker. LDIF - Linked Data Integration Framework. In 2^{nd} *International Workshop on Consuming Linked Data (COLD 2011), Bonn, Germany*, October 2011.
20. William E Winkler. Overview of record linkage and current research directions. Technical report, Bureau of the Census, 2006.