Analyzing Linked Data Quality with LiQuate

Edna Ruckhaus(✉), Maria-Esther Vidal, Simón Castillo, Oscar Burguillos, and Oriana Baldizan

Universidad Simón Bolívar, Caracas, Venezuela
{eruckhaus,mvidal,scastillo,oburguillos,obaldizan}@ldc.usb.ve

Abstract. The number of datasets in the Linking Open Data (LOD) cloud as well as LOD-based applications have exploded in the last years. However, because of data source heterogeneity, published data may suffer of redundancy, inconsistencies, or may be incomplete; thus, results generated by LOD-based applications may be imprecise, ambiguous, or unreliable. We demonstrate the capabilities of LiQuate (Linked Data Quality Assessment), a tool that relies on Bayesian Networks to analyze the quality of data and links in the LOD cloud.

1 Introduction

Linking Open Data initiatives have made a diversity of collections available, and facilitate scientists the mining of linked datasets to discover patterns or suggest potential new associations. To ensure trustworthy results, linked data must meet high quality standards. However, data in the LOD cloud has not been necessarily curated, and tools are required to detect possible quality problems and ambiguities produced by redundancy, inconsistencies, and incompleteness of both data and links [2]. We developed LiQuate, a tool able to identify potential quality problems and ambiguities among data and links. LiQuate relies on statistical reasoning to analyze the quality of data based on completeness and potential redundancies or inconsistencies. A Bayesian Network models the dependencies among resources that belong to a set of linked datasets [1,3]; conditional probability tables annotate the nodes of the network and represent joint probability distributions of relationships among resources. Queries against the Bayesian Network represent the probability that different resources have redundant labels or that a link between two resources is missing; thus, the returned probabilities can suggest ambiguities or possible incompleteness in the data or links. We demonstrate the data quality validation capabilities of LiQuate and the benefits of the approach on the Biomedical datasets: *Drugbank Website*[1], *LinkedCT*[2], *D2R Diseasome*[3], *D2R Dailymed*[4], *D2R Drugbank*[5], *Bio2RDF Drugbank*[6], and

[1] http://www.drugbank.ca/

[2] http://linkedct.org/

[3] http://wifo5-04.informatik.uni-mannheim.de/diseasome

[4] http://wifo5-03.informatik.uni-mannheim.de/dailymed/

[5] http://wifo5-04.informatik.uni-mannheim.de/drugbank/

[6] http://download.bio2rdf.org/current/drugbank/drugbank.html

V. Presutti et al. (Eds.): ESWC Satellite Events 2014, LNCS 8798, pp. 488–493, 2014.
DOI: 10.1007/978-3-319-11955-7_72

DBPedia[7]. This demo illustrates how queries to a Bayesian Network that models RDF data and dependencies among properties, can be used to study quality problems related to both incompleteness of links, and ambiguities among labels and links. We show the following key issues: redundancy among drug labels in the *LinkedCT* dataset, and incompleteness and inconsistencies of links in Biomedical datasets. The demo is published at http://liquate.ldc.usb.ve.

2 The LiQuate System

As a proof of concept, LiQuate has been built on top of the Biomedical linked datasets that maintain data related to clinical trials, interventions, conditions, drugs, diseases, and the relationships among them. LiQuate exploits visualization services implemented by the D3.js JavaScript library[8]. Figure 1 illustrates the LiQuate architecture. LiQuate receives a *quality validation request* which is expressed as one or more evidence queries against the Bayesian Network. The answer of a *quality validation request* is a number in the range [0.0:1.0] that indicates the probability that a given quality problem occurs among the data. Currently, three types of quality validation requests can be expressed: (*i*) probability that labels or names of a given (type of) resource are redundant, (*ii*) probability of incomplete links among a given set of resources, and (*iii*) probability of inconsistent links. LiQuate is comprised of two components: the LiQuate Bayesian Network Builder and the Ambiguity Detector. The LiQuate Bayesian Network Builder is a semi-automatic off-line process; it relies on an expert's knowledge about the properties in the RDF linked datasets that are going to be represented in the Bayesian Network. Relevant data is retrieved from SPARQL endpoints, and stored in a relational database to compute the histograms that implement the conditional probability tables (CPTs) associated with the nodes of the network. The demo is focused on the Ambiguity Detector: a probabilistic model that supports the analysis of the three above mentioned linked data quality problems. The Ambiguity Detector is in turn comprised of three components: (*1*) the Quality Validation Request Analyzer, (*2*) the Bayesian Network Query Translator, and (*3*) the Bayesian Network Inference Engine. The Quality Validation Request Analyzer receives a user request and determines if it can be satisfied with the existing Bayesian Network. The Bayesian Network Query Translator considers the user request and generates the set of queries that must be posed against the Bayesian Network. It also gathers the answers of these queries and generates an answer to the user request. Finally, the Bayesian Network Inference Engine is responsible of performing the inference process required to answer each of the queries posed against the Bayesian Network. This engine is implemented by the *SamIam* Bayesian Inference Tool[9].

[7] http://wiki.dbpedia.org/Downloads32

[8] http://d3js.org/

[9] http://reasoning.cs.ucla.edu/samiam/help/recursiveconditioning.html

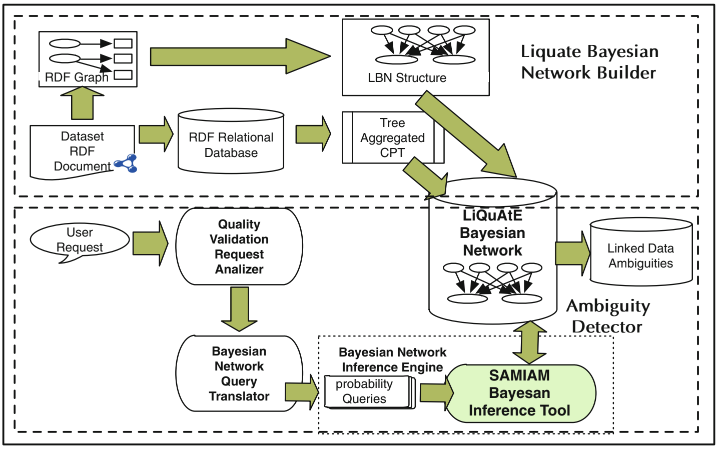


Fig. 1. The LiQuate system architecture.

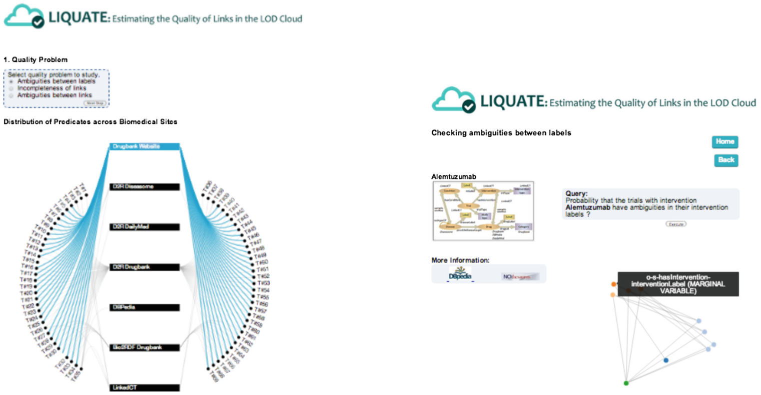
3 Demonstration of Use Cases

As of September 2011, LinkedCT contains 106,308 trials, 2.7 million entities and over 25 million RDF triples. Additionally, we consider the following datasets that are linked to LinkedCT: (*i*) Drugbank (over 765,936 triples), (*ii*) Diseasome (around 91,182 triples), and (*iii*) DBPedia (links from LinkedCT 25,476). We built local RDF storage with LinkedCT triples and the triples from these three datasets that are related to LinkedCT. The Bayesian network and its corresponding CPT's were computed and stored in the *SamIam* Bayesian Inference Tool. The generated network is comprised of 17 nodes and the aggregated CPTs are of up to 167, 616 entries; for the cases to be shown, the average response time of LiQuate is 4, 715 ms. Figure 2(a) illustrates the description of Biomedical linked datasets, and Fig. 2(b) presents the Bayesian Network that represents the dependencies between these properties and links. `Concept Network Browser` plots[10] and `Force-Directed Graphs`[11] are used for visualization.

We demonstrate the following use cases:

Ambiguities between labels of Interventions or Drugs: Starting with Alemtuzumab as an exemplar, we retrieve the intersection of Monoclonal antibodies and Antineoplastic agents. This creates a dataset of 12 drugs: Alemtuzumab, Bevacizumab, Brentuximab vedotin, Cetuximab, Catumaxomab, Edrecolomab, Gemtuzumab, Ipilimumab, Ofatumumab, Panitumumab, Rituximab, and Trastuzumab. These drugs are frequently tested in clinical trials, and there are up to 723 clinical trials with a given intervention, e.g., the intervention that

[10] http://www.findtheconversation.com/concept-map
[11] http://bl.ocks.org/mbostock/4062045


LIQUATE: Estimating the Quality of Links in the LOD Cloud
Next Step
Drugbank Website
D2R Diseasome
D2R DailyMed
D2R Drugbank
DBPedia
Bio2RDF Drugbank
LinkedCT
LIQUATE: Estimating the Quality of Links in the LOD Cloud
Home
Back
Execute
DBpedia
NCI thesaurus
o-s-hasIntervention-interventionLabel (MARGINAL VARIABLE)

1. Quality Problem

Select quality problem to study.
● Ambiguities between labels
○ Incompleteness of links
○ Ambiguities between links

Distribution of Predicates across Biomedical Sites

Checking ambiguities between labels

Alemtuzumab

Query:
Probability that the trials with intervention **Alemtuzumab** have ambiguities in their intervention labels ?

More Information:

(a) LinkedCT, DrugBank (website, and two endpoints), Diseasome, and DBPedia visualized as a `Concept Network Browser` plot. Predicates published by the Drugbank Website are highlighted.

(b) Bayesian Network for LinkedCT, DrugBank, Diseasome, and DBPedia visualized by using a `Force-Directed Graph`; nodes colored in orange and in blue correspond to marginal and evidence variables, respectively

Fig. 2. Biomedical linked datasets and a LiQuate bayesian network.

corresponds to the drug Alemtuzumab is present in 112 different clinical trials, and all of these should be linked to the drug DB00087 (Alemtuzumab) in Drugbank in order for the datasets to be unambiguous. This use case illustrates the execution of a query that could indicate possible uncontrolled redundancy in the datasets. The Bayesian Network used to infer the percentage of ambiguity is visualized by using a `Force-Directed Graph`; nodes colored in orange and in blue correspond to marginal and evidence variables, respectively.

Incompleteness of links between LinkedCT, Drugbank, Diseasome, and DBPedia: We consider the family of the 12 drugs described above, and for each of the partitions induced by redundant labels we consider the `owl:sameAs` and `rdfs:seeAlso` links. A partition represents all of the clinical trials that are of interventional type and that have the same intervention (drug) label. For each intervention *id* that belongs to a partition, a query to the Bayesian Network is executed in order to determine if `owl:sameAs` links have been established for this intervention. General *results* are also presented for each of the 12 drugs. Examples of these *results* are: (*i*) a percentage of redundant labels are not linked through `owl:sameAs` to neither Drugbank or DBPedia, but 100 % of the labels are linked through `rdfs:seeAlso`, e.g., Bevacizumab; (*ii)* none of the redundant labels is linked to Drugbank or DBPedia, e.g., Brentuximab vedotin, in this case, the drug is not appear in Drugbank; and (*iii*) a percentage of redundant labels are linked to DBPedia through `owl:sameAs`, all of them are linked to DBPedia through `rdfs:seeAlso`, and none to Drugbank, e.g., Ipilimumab.

Inconsistencies of links between LinkedCT, Drugbank, Diseasome, and DBPedia: We analyze if relationships that represent diseases that are possible targets of a drug, are backed up by clinical trials. For each of the 12 drugs, the query to the Bayesian network determines if for each possible disease target of a drug, there is at least one trial with this Condition (disease) and drug intervention. Conditions and interventions should be linked by `owl:sameAs` links to their corresponding drugs and diseases, in the *Drugbank* and *Diseasome* datasets. Approximately, 10,000 probability queries were generated for each drug and disease and all the combinations of linked (through `owl:sameAs`) conditions and interventions. The marginal node is `s-s-hascondition-hasintervention`, and the evidence is a disease, drug, condition, intervention, and the existence of `owl:sameAs` links among them. The result is that 13,5 % of the drugs and targeted diseases are supported by clinical trials that can be found through `owl:sameAs` links. Similarly, another hypothesis is that drugs that can possibly treat diseases (*possibleDrug* links) are supported by the same number of clinical trials. The result is 13, 5 % and this number suggests that both links *possibleDiseaseTarget* and *possibleDrug* are the inverse of each other. Particularly, for the dataset of 12 drugs we can observe the following: the drugs Brentuximab vedotin, Ipilimumab and Ofatumumab do not appear in Drugbank while these drugs have been studied in a large number of clinical trials. The rest of these 12 drugs do appear in Drugbank, but are associated with much less diseases through the property *possibleDiseaseTarget* in Drugbank, than to conditions through a clinical trial in LinkedCT. For example, the drug Cetuximab can possibly target eighteen diseases while this drug has been tested in completed clinical trials for 82 conditions; only four of the eighteen diseases in the property *possibleDiseaseTarget* in Drugbank, are included in the list of 82 conditions in LinkedCT. This ambiguity can be also observed in the rest of the drugs.

4 Conclusions

We present LiQuate, a data and link validation tool that relies on a Bayesian Network to identify redundancies, incompleteness and inconsistencies. We demonstrate the main quality validation capabilities of LiQuate, and illustrate different quality problems that may currently occur in the LOD cloud. Particularly, we can observe some ambiguities that suggest the experts to check for uncontrolled redundancy, incompleteness or inconsistency: (*i*) the same label or name of intervention is assigned to different resources, (*ii*) incomplete `owl:sameAs` and `rdfs:seeAlso` links between datasets, and (*iii*) associations between drugs and diseases in Drugbank may not be supported by trials in LinkedCT.

References

1. Getoor, L., Taskar, B., Koller, D.: Selectivity estimation using probabilistic models. SIGMOD Rec. **30**(2), 461–472 (2001)
2. Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 87–102. Springer, Heidelberg (2012)
3. Ruckhaus, E., Vidal, M.-E.: LiQuate-estimating the quality of links in the linking open data cloud. In: Lacroix, Z., Ruckhaus, E., Vidal, M.-E. (eds.) RED 2012. LNCS, vol. 8194, pp. 56–82. Springer, Heidelberg (2013)