User-Driven Filtering and Ranking of Topical Datasets Based on Overall Data Quality

Wenze Xia, Zhuoming Xu (✉), Chengwang Mao
College of Computer and Information
Hohai University
Nanjing, 210098, China
E-mail: {xiawenze, zmxu, cwmao}@hhu.edu.cn

***Abstract*—Finding relevant and high-quality data is the eternal needs for data consumers (i.e., users). Many open data portals have been providing users with simple ways of finding datasets on a particular topic (i.e., topical datasets), which are not a way of filtering and ranking topical datasets based on data quality. Despite the recent advances in the development and standardization of data quality models and vocabulary, there is a lack of systematic research on approaches and tools for user-driven data quality-based filtering and ranking of topical datasets. In this paper we address the problem of user-driven filtering and ranking of topical datasets based on the overall data quality of datasets by developing a generic software architecture and the corresponding approach, called ODQFiRD, for filtering and ranking topical datasets according to user-specified data quality assessment criteria. Additionally, we use our implemented prototype of ODQFiRD to conduct a case study experiment on the U.S. Government's open data portal. The prototype implementation and experimental results show that our proposed ODQFiRD is achievable and effective.**

Keywords—data quality-based filtering and ranking; topical datasets; overall data quality; data quality assessment; Data Quality Vocabulary (DQV); open data portal

I. INTRODUCTION

Finding relevant and high-quality data is the eternal needs for *data consumers* (hereinafter also referred to as users). The boom in open data portal technologies in recent years has been providing great possibilities to meet such user needs. Existing open data portals, such as the U.S. Government's open data portal (`data.gov`) and the European Union Open Data Portal (`data.europa.eu`), have published a large number of datasets that are free to use for data consumers. Data portals usually provide the users with two kind of ways of finding topical datasets: (1) browsing in the topic classification and selecting desired datasets; (2) searching for datasets using keywords and filtering the search results through simple facets. However, such ways of finding topical datasets are not data quality-based filtering of topical datasets. As stated in the W3C's Data on the Web Best Practices Recommendation [1], "Data quality might seriously affect the suitability of data for specific applications," and "the inclusion of data quality information in data publishing and consumption pipelines is of primary importance."

Despite the significant progress in the development and standardization of the data quality model and vocabulary such as the ISO/IEC 25012 Data Quality model [2], the W3C's Data on the Web Best Practices Recommendation [1] and the W3C's Data Quality Vocabulary (DQV) [3], there is a lack of systematic research on approaches and tools for user-driven data quality-based filtering and ranking of topical datasets [4].

Data quality is commonly defined as "fitness for use" for a specific application or use case [1,5-8], and actually, "quality lies in the eye of the beholder" [3]. Therefore, quality assessment and quality-based selection of datasets should be a *user-driven process* during which users can specify their own assessment criteria. Furthermore, users typically want to filter and rank the retrieved datasets based on the *overall data quality* of the datasets, rather than just based on a single quality category, dimension, or metric.

In the field of data quality assessment, especially of linked data quality assessment [5,6], several data quality models [2,3,5,6,9] have been recently developed, and subsequently, dozens of quality assessment approaches and tools [5,6] have been built on the data quality models. However, most of existing approaches and tools do not provide support for *user-driven* data quality assessment and *quality-based* filtering and ranking of datasets. For example, the Luzzu framework [7], which represents the state of the art in this field and provides user-driven quality-based weighted ranking of datasets, only enables users to assign weights to their preferred quality categories, dimensions, or metrics, but does not enable users to define quality measurement thresholds for dataset filtering and to achieve overall data quality based ranking of topical datasets.

In this paper we address the problem of user-driven filtering and ranking of topical datasets based on overall data quality by developing a generic software architecture and the corresponding approach, hereinafter referred to as ODQFiRD, for filtering and ranking topical datasets according to user-specified data quality assessment criteria (Sect. III). The core of ODQFiRD lies in: (1) user's quality assessment criteria inquiry, and (2) overall data quality calculation. We have used our implemented prototype of ODQFiRD to conduct a case study experiment on a popular open data portal, and the results show the implementability and effectiveness of our proposed ODQFiRD (Sect. IV).

978-1-5386-4806-3/17 $31.00 © 2017 IEEE
DOI 10.1109/WISA.2017.24

II. Data Quality in a Nutshell

A. Data Quality Models

The international standard ISO/IEC 25012 [2] defines *data quality* as the "degree to which the characteristics of data satisfied stated and implied needs when used under specified conditions." A *data quality model* is therefore the "defined set of characteristics which provides a framework for specifying data quality requirements and evaluating data quality" [2]; it is generally a *hierarchical* organization of quality *categories*, *dimensions*, and *metrics*. As a very general model, the ISO/IEC 25012 Data Quality model [2] does not define any quality metrics (i.e., the third level) and is composed of 15 quality characteristics (dimensions, e.g., Accuracy) classified into two quality categories: Inherent data quality and System-dependent data quality. Zaveri et al. [5] proposed a three-tier model for linked data quality assessment, which consists of 69 metrics grouped by 18 dimensions within 4 categories. The linked data quality model (LDQM) recently proposed by Radulovic et al. [9] is composed of 124 metrics (e.g., datatype syntax error) grouped by 15 dimensions within 3 categories.

B. Data Quality Vocabulary (DQV)

As an extension of the W3C's Data Catalog Vocabulary (DCAT) [10], the Data Quality Vocabulary (DQV) [3] provides a framework for describing the quality of published datasets in a data portal (data catalog). It defines various RDF classes and properties used to implement a customized quality model for the data portal based on one or more established data quality models (such as the ISO/IEC 25012 model and LDQM) and allows the dataset publishers or a broader community of users to express quality metadata.

III. Proposed Software Architecture and Approach

On the basis of our ideas of user-driven, overall data quality-based filtering and ranking of topical datasets in an open data portal, here we present ODQFiRD's generic software architecture and corresponding approach for filtering and ranking topical datasets according to user-specified data quality assessment criteria.

A. Generic Software Architecture of ODQFiRD

The generic software architecture of ODQFiRD is depicted in Fig. 1. ODQFiRD uses multiple software modules to extend a traditional data portal. Its input is a list of topical datasets, which are produced by the Dataset Search Engine through user's topical search over the Open Data Portal. After three key processing steps (user's quality assessment criteria inquiry, dataset filtering, dataset ranking based on overall data quality), ODQFiRD uses the Result Output module to display the information about ranked topical datasets on the User Interface. ODQFiRD's three core software modules are briefly described below:

1) User's Quality Assessment Criteria Inquiry: This module enables the tool to inquire about the user's customized quality assessment criteria which comprise user-chosen quality metrics together with the weights assigned to the quality metrics, as well as user-defined thresholds of quality measurements on some quality metrics.

2) Dataset Filtering based on Measurement Thresholds: This module enables the tool to filter out those topical datasets that do not satisfy the user-defined measurement thresholds.

3) Overall Data Quality Calculation and Dataset Ranking: This module enables the tool to calculate the

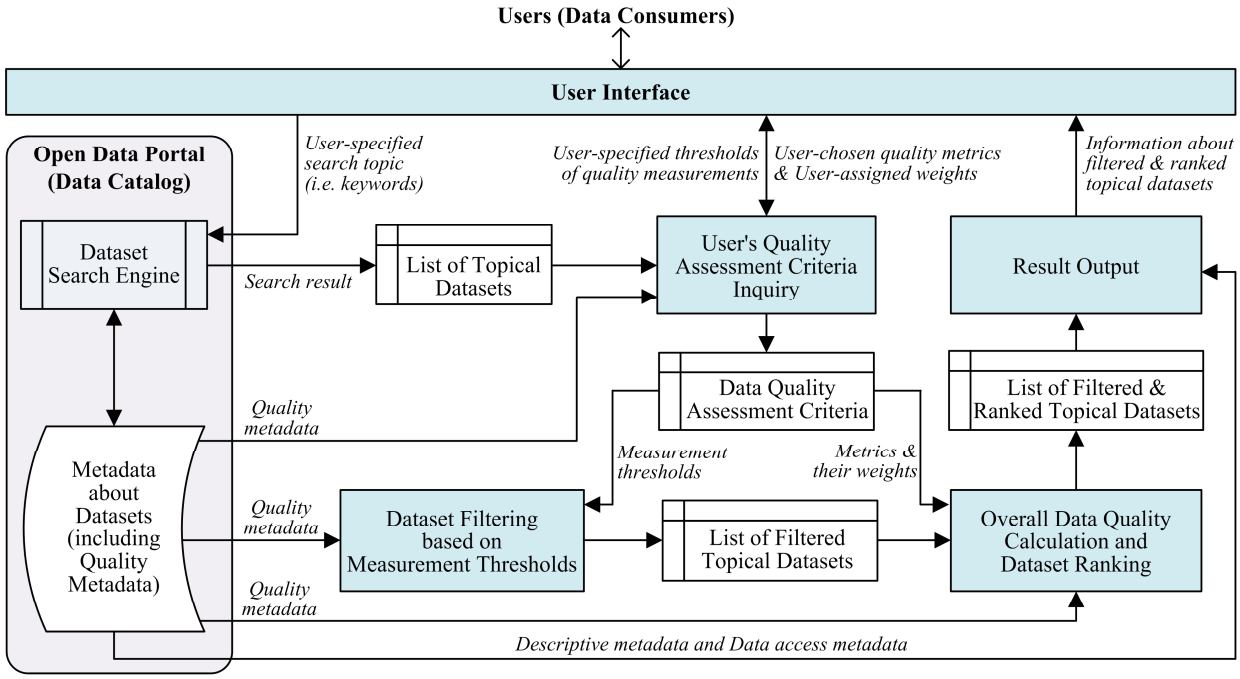

Figure 1. The generic software architecture of ODQFiRD.

overall data quality scores of the remaining (filtered) topical datasets according to the user-chosen quality metrics and user-assigned weights to the quality metrics and then rank the datasets using the overall data quality scores.

B. Assessment Criteria Inquiry and Dataset Filtering

According to the "fitness for use" view of data quality [1,5-8] and the idea that "quality lies in the eye of the beholder" [3], we design a user interface for inquiring about user's quality assessment criteria, allowing the user to make his/her own judgment about datasets' fitness for the specific use case or application at hand. As shown in Fig. 2, the user interface consists of the left part and the right part:

The left part shows information about quality metadata of the retrieved topical datasets to the user, which includes:

- The quality model hierarchy relevant to all retrieved topical datasets: The information on this quality Category-Dimension-Metric hierarchy, as part of the full hierarchical quality model implemented in the data portal, can be easily extracted from the existing quality metadata about the retrieved topical datasets.
- The inherent ranges of quality metrics: The range of a quality metric, including a *worst measurement* and a *best measurement*, is predefined by the quality management expert and stored as quality metadata in the data catalog. For a numeric metric, its value (i.e., quality measurement) is a finite nonnegative real number. For a Boolean metric, its value is 'true' or 'false'. Note that Boolean values are always converted to real numbers ('true' to 1; 'false' to 0) when calculating the overall data quality.

The right part of the user interface is used for obtaining user-specified quality assessment criteria, which includes:

- User-chosen quality metrics: The user can choose some quality metrics (e.g., Quality Metric 111 in Fig. 2) from the quality model hierarchy, as deemed suitable or important for his/her intended specific use case.
- Weights on user-chosen quality metrics: The user should assign a weight (importance) to each chosen metric (e.g., 0.30 to Quality Metric 111 in Fig. 2), which will later be used to calculate the overall data quality of datasets. The sum of all weights must be 1.
- Measurement thresholds: The user may also define measurement thresholds for several chosen quality metrics (e.g., 0.8 for Quality Metric 111 in Fig. 2), which will later be used to filter out datasets.

The tool is now able to filter out topical datasets according to the user-defined measurement thresholds. Any dataset without quality measurement on one of the user-chosen quality metrics with user-defined measurement thresholds is filtered out. Also, any dataset that one of its quality measurements on a quality metric is "worse than" the corresponding measurement threshold is filtered out. Suppose the user has chosen k quality metrics $M_1,\ldots,M_k$, and topical dataset d_j has a measurement ms_{ij} on quality metric M_i with its worst measurement being m_{iw}, best measurement m_{ib}, and user-defined measurement threshold τ_i. The "worse than" rules we apply are as follows:

- For Boolean metric M_i, if $ms_{ij} \neq \tau_i$ then measurement ms_{ij} is worse than τ_i;
- For numeric metric M_i such that $m_{iw} < m_{ib}$, if $ms_{ij} < \tau_i$ then measurement ms_{ij} is worse than τ_i;
- For numeric metric M_i such that $m_{iw} > m_{ib}$, if $ms_{ij} > \tau_i$ then measurement ms_{ij} is worse than τ_i.

After the filtering of datasets, the tool produces a list of remaining n datasets $FTDL = (d_1,\ldots,d_n)$.

C. Overall Data Quality Calculation and Dataset Ranking

The method and steps for calculating the overall data quality of datasets comprise: (1) The k quality metrics chosen by the user are treated as a k-dimensional real

Information about Quality Metadata of the Retrieved Topical Datasets				**User-Specified Quality Assessment Criteria**			
Quality Model Hierarchy Relevant to the Topical Datasets		**Inherent Ranges of Quality Metrics**		**User-Chosen Metrics**	**Automatic Number**	**Measurement Thresholds**	**Weights on Metrics**
		Worst Measurement	**Best Measurement**				
Quality Category 1							
Quality Dimension 11							
	Quality Metric 111	0.0	1.0	⊙Yes ○No	1	0.8	0.30
	Quality Metric 112	<maximum value>	0	○Yes ⊙No			
	Quality Metric 113	false	true	⊙Yes ○No	2	true	0.25
Quality Dimension 12							
	Quality Metric 121	1.0	0.0	⊙Yes ○No	3		0.05
	Quality Metric 122	0	100	⊙Yes ○No	4		0.35
Quality Category 2							
Quality Dimension 21							
	Quality Metric 211	0	<maximum value >	○Yes ⊙No			
	Quality Metric 212	true	false	⊙Yes ○No	5	false	0.05
						Sum of All Weights	**1.00**

Figure 2. ODQFiRD's user interface for inquiring about user's quality assessment criteria.

coordinate space $\mathbb{R}^k$; (2) The k-tuple $(m_{1b},\ldots,m_{kb}) \in \mathbb{R}^k$, as a representation of the **best data quality** BQ, is considered as a *point* in the coordinate space , where each measurement $m_{ib}, i=1,\ldots,k$ is the best measurement of quality metric M_i; (3) For each topical dataset $d_j \in FTDL, j=1,\ldots,n$, another k-tuple $(m_{1j},\ldots,m_{kj}) \in \mathbb{R}^k$ is constructed by using all of this dataset's quality measurements $ms_{ij}, i=1,\ldots,k$ on the quality metrics, which represents the **dataset's data quality** of d_j, denoted DQ_j, and is also considered as a *point* in the coordinate space; (4) The **overall data quality** of each dataset $d_j \in FTDL$ can be calculated by exploiting a weighted version of Canberra distance between points BQ and DQ_j in the coordinate space.

For the above step (3), when computing coordinate $m_{ij} \in \mathbb{R}$, any quality measurement ms_{ij} of dataset d_j on metric M_i must be properly handled in the following way.

- Any Boolean ms_{ij} should be converted to a real number ('true' to 1; 'false' to 0), still denoted ms_{ij}.
- If d_j has multiple measurements on M_i, then m_{ij} takes the worst of these measurements, denoted by $\mathrm{wt}(ms_{ij})$.
- If d_j has no measurement on numeric metric M_i, then m_{ij} takes the median of all the measurements on M_i of all other datasets in $FTDL$, denoted by $\mathrm{md}(ms_i)$.
- If d_j has no measurement on Boolean metric M_i, then m_{ij} takes the worst measurement m_{iw} on the metric.

Therefore, m_{ij} can be computed by using equation (1).

$$m_{ij} = \begin{cases} ms_{ij}, & d_j \text{ has exactly one measurement on } M_i \\ \mathrm{wt}(ms_{ij}), & d_j \text{ has multiple measurements on } M_i \\ \mathrm{md}(ms_i), & d_j \text{ has no measurement on numeric } M_i \\ m_{iw}, & d_j \text{ has no measurement on boolean } M_i \end{cases} \quad (1)$$

When calculating the overall data quality in the above step (4), we use Canberra distance [11,12], the weighted L_1 distance that yields a real number distance between a pair of points in a real coordinate space. Given the coordinates of two points X and Y in $\mathbb{R}^k$: $X=(x_1,\ldots,x_k)$ and $Y=(y_1,\ldots,y_k)$, the Canberra distance $CD(X,Y) \in \mathbb{R}$ between X and Y is computed using equation (2) [11,12].

$$CD(X,Y) = \sum_{i=1}^{k} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (2)$$

For each dataset $d_j \in FTDL$, its overall data quality ODQ_j can now be calculated using equation (3), which uses a weighted version of the Canberra distance between the **dataset's data quality** point $DQ_j = (m_{1j},\ldots,m_{kj}) \in \mathbb{R}^k$ and the **best data quality** point $BQ = (m_{1b},\ldots,m_{kb}) \in \mathbb{R}^k$, where all the coordinates are non-negative real numbers (so the absolute value in the denominator can be omitted here).

$$ODQ_j = 1 - \sum_{i=1}^{k} w_i \cdot \frac{|m_{ib} - m_{ij}|}{m_{ib} + m_{ij}} \quad (3)$$

where all w_i are user-assigned weights to the quality metrics, and they satisfy $\sum_{i=1}^{k} w_i = 1$.

Finally, the overall data quality scores (0~1) computed by equation (3) can be used to produce a list of ranked topical datasets $RFTDL$ by sorting the scores in reverse order.

IV. Experiments

We have implemented an ODQFiRD prototype as a proof of concept, which is a Web application developed using Java EE 8.0 and Apache Jena API and deployed on the Apache Tomcat 7.0.55 server. To validate the effectiveness of ODQFiRD, we have used the prototype to conduct a case study experiment on the U.S. Government's open data portal (`data.gov`). Because so far there are no data portals providing quality metadata about their published datasets, we had to construct mock quality metadata about some of the topical datasets in the search results retrieved from the data portal.

First, we searched for topical datasets in `data.gov` using keyword "`unemployment statistics`" and took all of the 68 datasets from the search results. Next, we used a Java program to download metadata from the Metadata Sources of these datasets. Finally, based on the downloaded metadata, we constructed DQV-based, mock quality metadata (in the form of RDF triples) for randomly selected 20 datasets from the 68 datasets. The quality categories and dimensions were chosen from the ISO/IEC 25012 Data Quality model [2]. The quality metrics were chosen from the LDQM proposed by Radulovic et al. [9]. We also defined the ranges of these quality metrics by specifying the worst and best measurements for each quality metric. When constructing the quality metadata, we randomly generated quality measurements within the ranges for the chosen quality metrics. Table I (the left part) shows the quality model hierarchy and the ranges of the quality metrics.

During the experiment, we operated the ODQFiRD prototype on the aforementioned dataset metadata including the mock quality metadata. The user-specified quality assessment criteria, including the user-chosen quality metrics, the weights on the quality metrics, and the measurement thresholds for some of the quality metrics, are shown in Table I (the right part). After filtering out datasets

Table I. The quality model hierarchy, ranges of quality metrics, and user-specified quality assessment criteria implemented during the experiment.

Quality Model Hierarchy Relevant to the Topical Datasets	Worst Measurement	Best Measurement	User-Chosen Metrics	Measurement Thresholds	Weights on Metrics
iso:inherentDataQuality					
iso:accuracy					
ldqm:averagePropertyDiscordance	1.0	0.0	Yes	0.3	0.02
ldqm:datatypeSyntaxError	true	false	Yes		0.20
ldqm:numberOfInvalidRules	maximum	0	Yes	15	0.03
iso:completeness					
ldqm:interlinkingDegree	0.0	1.0	Yes		0.05
ldqm:numberOfBlankNodes	100	0			
ldqm:propertyCompleteness	0.0	1.0	Yes	0.8	0.14
iso:consistency					
ldqm:averageUndefinedProperties	1.0	0.0			
ldqm:numberOfStableIRIs	0	1000	Yes		0.04
iso:systemDependentDataQuality					
iso:availability					
ldqm:multipleSerializationFormats	false	true	Yes	true	0.28
ldqm:sparqlSupport	false	true			
iso:portability					
ldqm:vocabularyReuse	false	true	Yes		0.24

according to the user-defined measurement thresholds and calculating the overall data quality scores for the remaining datasets and ranking the datasets using the quality scores, the tool produced a ranked list of the resulting datasets, as listed in Table II, and displayed the information about the ranked datasets on the user interface, as shown in Fig. 3.

The user can also use ODQFiRD's output interface, as shown in Fig. 3, to select topical datasets with high overall data quality scores. Take our experiment as an example, assuming that the user wants the overall data quality score to exceed 0.910, only the dataset titled "Local Area Unemployment Statistics: Beginning 1976" (its overall data quality score is 0.916) and the dataset titled "Local Area Unemployment Statistics" (its overall data quality score is 0.911) can meet the user's requirements.

The experimental results indicate that our proposed generic software architecture and corresponding approach (i.e., ODQFiRD) for user-driven filtering and ranking of topical datasets based on overall data quality is implementable and effective.

Table II. The list of ranked topical datasets and their overall data quality scores produced through the experiment.

Rank	Dataset title	Dataset publisher	Issue date	Access URL	Overall data quality
1	Local Area Unemployment Statistics: Beginning 1976	State of New York	2017-04-20	https://catalog.data.gov/dataset/local-area-unemployment-statistics-beginning-1976	0.916
2	Local Area Unemployment Statistics	data.wa.gov	2014-02-06	https://catalog.data.gov/dataset/local-area-unemployment-statistics-3885c	0.911
3	Quarterly Census of Employment and Wages Quarterly Data: Beginning 2000	State of New York	2017-03-13	https://catalog.data.gov/dataset/quarterly-census-of-employment-and-wages-quarterly-data-beginning-2000	0.904
4	Maryland Veterans Unemployment Rate	data.maryland.gov	2015-09-08	https://catalog.data.gov/dataset/maryland-veterans-unemployment-rate-3ea61	0.900
5	Local Area Unemployment Statistics (LAUS), Seasonally Adjusted Data: Beginning 1976	State of New York	2017-04-20	https://catalog.data.gov/dataset/local-area-unemployment-statistics-laus-seasonally-adjusted-data-beginning-1976	0.893
6	Unemployment Rate (2005-Present)	data.nola.gov	2015-09-15	https://catalog.data.gov/dataset/unemployment-rate-2005-present	0.892
7	IDES - LAUS MSAs Annual Average 1976-2011	data.illinois.gov	2012-05-14	https://catalog.data.gov/dataset/ides-laus-msas-annual-average-1976-2011-c49a6	0.889
8	Quarterly Census of Employment and Wages Annual Data: Beginning 2000	State of New York	2016-06-20	https://catalog.data.gov/dataset/quarterly-census-of-employment-and-wages-annual-data-beginning-2000	0.885
9	Tobacco Use Supplement to the Current Population Survey (TUS-CPS) Data	Centers for Disease Control and Prevention	2016-05-08	https://catalog.data.gov/dataset/tobacco-use-supplement-to-the-current-population-survey-tus-cps-data	0.879

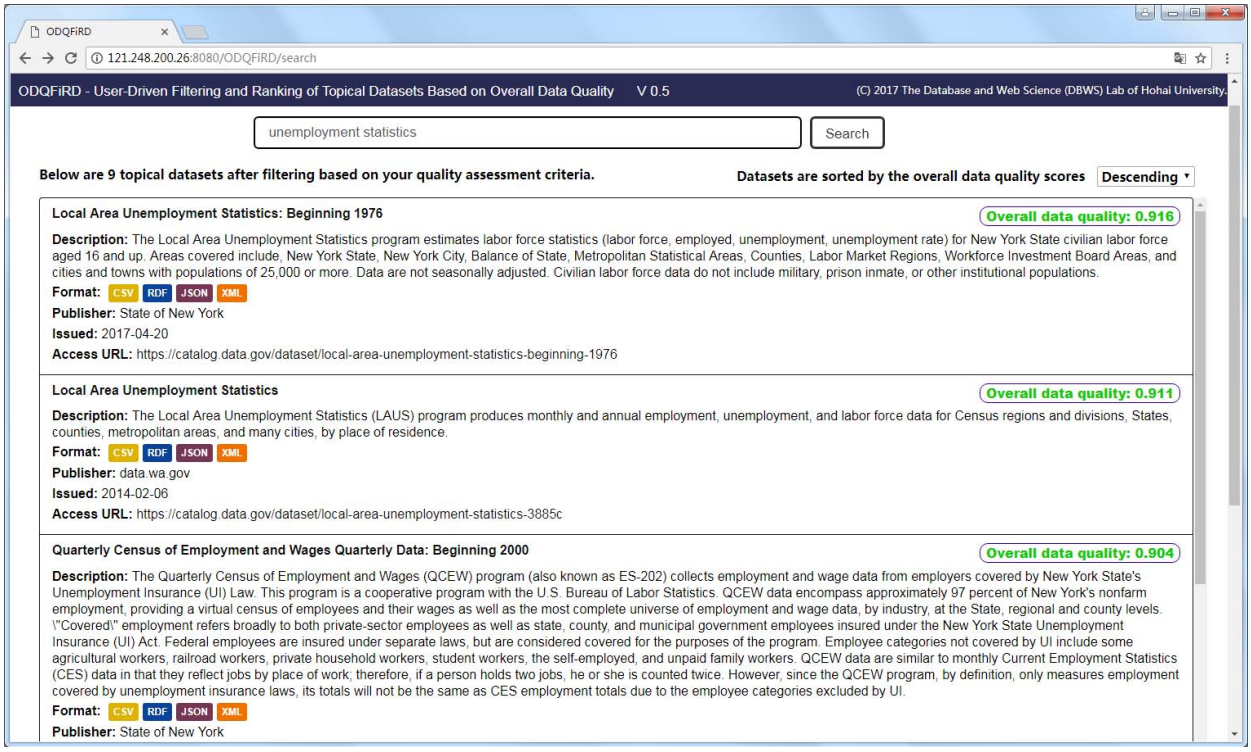

Figure 3. Screenshot of the ODQFiRD output interface for filtering and ranking datasets using the mock data quality metadata and assessment criteria.

V. CONCLUSIONS

To address the problem of user-driven filtering and ranking of retrieved topical datasets based on overall data quality, we have proposed a generic software architecture and corresponding approach (i.e., ODQFiRD) for filtering and ranking datasets according to user-specified data quality assessment criteria. ODQFiRD filters topical datasets by using user-defined measurement thresholds for user-chosen quality metrics; it ranks datasets based on the overall data quality computed by using a weighted version of Canberra distance and user-assigned weights. The implementation and experiment show that our proposed generic software architecture and approach are achievable and effective.

ACKNOWLEDGMENT

This work was supported by: (1) Research Grant No. BK20141420 from the Natural Science Foundation of Jiangsu Province, China, and (2) The Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions, China.

REFERENCES

[1] B. F. Lóscio, C. Burle, and N. Calegari (Eds.), "Data on the Web Best Practices," W3C Recommendation 31 January 2017. https://www.w3.org/TR/dwbp/. [Accessed 6 June 2017]

[2] The ISO and the IEC, "ISO/IEC 25012 -- Data quality model," developed by Joint Technical Committee ISO/IEC JTC 1/SC 7, December 2008. http://iso25000.com/index.php/en/iso-25000-standards/iso-25012. [Accessed 6 June 2017]

[3] R. Albertoni and A. Isaac (Eds.), "Data on the Web Best Practices: Data Quality Vocabulary," Working Group Note 15 December 2016. https://www.w3.org/TR/vocab-dqv/. [Accessed 6 June 2017]

[4] W. Xia, Z. Xu, J. Wei, and H. Tian, "DQFIRD: Towards data quality-based filtering and ranking of datasets for data portals," Proc. of 13th Web Information Systems and Applications Conference, WISA 2016, pp. 18-23, IEEE 2016. DOI: 10.1109/WISA.2016.14.

[5] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," Semantic Web, Vol. 7, No. 1, 2016, pp. 63-93. DOI: 10.3233/SW-150175.

[6] C. Batini and M. Scannapieco, "Data and Information Quality: Dimensions, Principles and Techniques," Springer International Publishing Switzerland 2016. DOI: 10.1007/978-3-319-24106-7.

[7] J. Debattista, S. Auer, and C. Lange, "Luzzu—A methodology and framework for linked data quality assessment," ACM Journal of Data and Information Quality (JDIQ), Vol. 8, No. 1, October 2016, Article 4, pp. 4:1-4:32. DOI: 10.1145/2992786.

[8] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and framework for data and information quality research," ACM Journal of Data and Information Quality (JDIQ), Vol. 1, No. 1, June 2009, Article 2, pp. 2:1-2:22. DOI: 10.1145/1515693.1516680.

[9] F. Radulovic, N. Mihindukulasooriya, R. García-Castro, and A. Gómez-Pérez, "A comprehensive quality model for linked data," Semantic Web, IOS Press Journal, to appear. http://www.semantic-web-journal.net/content/comprehensive-quality-model-linked-data-1.

[10] F. Maali and J. Erickson (Eds.), "Data Catalog Vocabulary (DCAT)," W3C Recommendation 16 January 2014. https://www.w3.org/TR/vocab-dcat/. [Accessed 6 June 2017]

[11] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello, "Canberra distance on ranked lists," Proc. of the 2009 Conference on Neural Information Processing Systems (NIPS), Advances in Ranking Workshop at NIPS, Decmber 2009, pp. 22-27.

[12] M. M. Deza and E. Deza, Chapter 17. Distances and similarities in data analysis, in: Encyclopedia of Distances, 4th Edition, Springer 2016, pp. 327-345. DOI: 10.1007/978-3-662-52844-0_17.