TODQA: Efficient Task-Oriented Data Quality Assessment

Anran Li*, Lan Zhang*, Jianwen Qian†, Xiang Xiao*, Xiang-Yang Li*, Yunting Xie*
*Department of Computer Science and Technology, University of Science and Technology of China
† Department of Computer Science and Technology, Illinois Institute of Technology

***Abstract*—Data quality assessment is vital for many information services ranging from sensor networks to smart city systems. The current data quality assessments, however, are often derived from intrinsic data characteristics, disconnected from specific application contexts, or are not applicable or efficient for large datasets. In this work, we propose a novel task-oriented data quality assessment framework, which balances between the intrinsic and contextual quality. We carefully craft the assessment metrics, quantify them, and fuse them to rank candidate datasets by quality given specific tasks. To improve the system efficiency, two fast calculation algorithms are designed to quantify the relationship between datasets and the task, and the distribution of data items. We conduct extensive evaluations on six public image datasets (with $460,247$ images in total) and four text document datasets (with $37,372$ documents in total) to evaluate the efficacy and efficiency of our design. Experimental results show that our algorithms can save about 90% computing time with little accuracy loss which validates the feasibility and effectiveness of our framework for large datasets.**

***Index Terms*—Data Quality Assessment; Sampling; Locality Sensitive Hashing; Rank Aggregation**

I. Introduction

The rapid development of networking technologies, such as mobile networks, sensor networks and crowdsensing technologies, has made it possible to aggregate massive diverse data. Recently, there is a significantly growing trend to improve quality of various information services by taking advantages of large amounts of data. The quality of data, therefore, plays a vital role in those information systems from the following aspects. 1) High-quality data provides adequate and accurate information to fulfill a specific task, such as training a high-quality machine learning model and making wise decisions in a smart city system. 2) A large number of services provide data itself as the product on demand to the user, *e.g.*, crowdsensing services. For those services, the quality of data determines user satisfaction. 3) Measuring quality of data also helps to optimize system resources utilization. Limited resources (*e.g.*, bandwidth, storage and computing resources) should be allocated to high quality data first to guarantee the system performance and service quality. Taking a crowdsensing application [1] as an example, where a large number of participants upload images collected by their mobile phones, effective data quality assessment, especially efficient quality assessment of large sets of images, can significantly facilitate the selection of uploaded images and even save bandwidth by avoiding transmission of low-quality data.

Data quality assessment has gained much attention from researchers [2], [3], [4], [5], [6]. Existing methods towards data quality assessment, however, have several limitations facing diversified tasks and large-scale datasets. **First**, most previous attempts focus on assessing inherent quality of data, while few consider contextual factors such as the target tasks or services, which requires the data content should be relevant to the task. Contextual factors have been shown to strongly influence perceptions of data quality [4]. **Second**, existing works mainly measure quality of an individual data piece [5] other than that of a collection of data pieces, while the latter is more commonly used by present services, e.g., crowdsensing services. As an example, 1000 high-quality images with the same content dont necessarily make a high-quality dataset, which requires the content of the dataset to be diverse. **Third**, though various dimensions of data quality have been proposed, given a task, how to fuse those dimensions to yield a quality-based ranking of different datasets remains a challenge.

In this work, we aim to enable data consumers to select high-quality datasets for a given task in an efficient and interpretable way. To achieve this goal, we propose a novel framework named Task-Oriented Data Quality Assessment model (TODQA), which comprehensively and efficiently measures quality of datasets for a given task (*e.g.*, image classification or sentiment analysis [6]) and ranks datasets by quality. The framework design supports datasets composed of a large number of unstructured data, e.g., image, text or video. Specifically, we design TODQA to answer the following challenging questions. 1) How to comprehensively quantify the quality of a dataset, especially adaptively quantify the task-oriented quality for various tasks? Based on the quality assessment of each data piece, the framework should not only characterize the relationship among data pieces but also the relationship between a dataset and a given task. 2) How to efficiently and accurately measure the quality of a large-scale dataset? Nowadays, it is common that there are millions of data pieces in one dataset. Considering the large computational cost of assessing a single unstructured data item [5], it will cause unacceptable overhead to measure quality of every data piece and the relationship of all data pairs. 3) How to reasonably fuse multiple dimensions of data quality and obtain an interpretable ranking of datasets? Measurements of different dimensions are not comparable, which makes methods like weighted average inapplicable here.

By conquering the aforementioned challenges, contributions

978-1-7281-5212-7/19/$31.00 ©2019 IEEE
DOI 10.1109/MSN48538.2019.00028

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

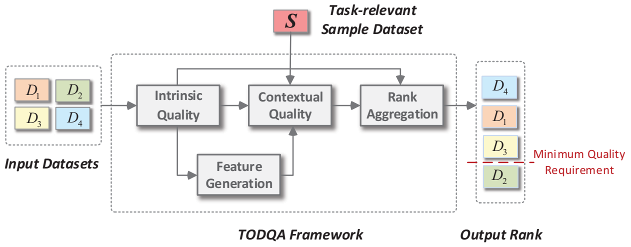


Fig. 1. System overview.

of this work can be summarized as follows.

- We propose a novel framework TODQA, which to the best of our knowledge is so far the most comprehensive large-scale datasets quality assessment framework incorporating both task-independent intrinsic quality and task-dependent contextual quality.
- We propose two novel dimensions "task relevancy" and "content diversity" to assess quality of a large-scale dataset by respectively characterizing its relevancy to a given task and the relationship among data pieces. To achieve highly efficient and accurate assessments of these two dimensions, we design two fast calculation algorithms based on sampling and locality sensitive hashing (LSH). A rank aggregation algorithm is also designed to fuse multiple disparate quality dimensions, so as to rank datasets by comprehensive quality.
- To validate our design, we conducted extensive evaluations on 6 popular image datasets (with $460,247$ images in total) and 4 popular text datasets (with $37,372$ documents in total) to thoroughly investigate their quality and rank them for different tasks. Experimental results show that our two fast calculation algorithms can save about 90% runtime with little accuracy loss.

The rest of this paper is organized as follows. We present an overview of our system in Section II. Section III gives detailed definitions and quantification methods of intrinsic and contextual quality dimensions. In Section III-C, we present the concrete rank aggregation method. Comprehensive evaluations are introduced in Section IV. We review related work in Section V and conclude this work in Section VI.

II. System overview.

A. Problem Formulation

The aggregation of high-quality datasets is important for many applications. In this work, we aim to design a task-oriented data quality assessment framework, which measures multi-dimensional quality of different datasets for a given task and ranks these datasets by quality. We consider three roles involved in the quality assessment process: *data owners*, *data consumers*, and the *data evaluator* (e.g., the cloud). Data owners provide the data evaluator with a collection of m datasets $\mathbb{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_m\}$. A data consumer desires to find the high-quality dataset to fulfill a target task $\mathcal{T}$. The task-oriented data requirement of the data consumer can be

TABLE I
Some important notations.

Notation	Description
$\mathbb{D}$	a set of datasets. Each $\mathcal{D} \in \mathbb{D}$ is a dataset;
d_i	the i-th data piece in dataset $\mathcal{D}$;
$\mathcal{T}$	a target task;
$\mathcal{S}$	a set of sample data pieces;
$\mathcal{Q}$	a collection of quality dimensions. Each $q \in \mathcal{Q}$ is one specific quality dimension;
τ_i	the ranking list of all datasets in $\mathbb{D}$ on quality dimension q_i;
$q_c^{\mathcal{D}}$, $q_p^{\mathcal{D}}$	the quantified value of dimension correctness and precision;
$q_{tr}^{\mathcal{D}}$, $q_{cd}^{\mathcal{D}}$, $q_{co}^{\mathcal{D}}$, $q_a^{\mathcal{D}}$, $q_t^{\mathcal{D}}$	the quantified value of dimension task relevancy, content diversity, completeness, appropriate amount and timeliness.

expressed by a set of sample data pieces $\mathcal{S} = \{s_1, s_2, \dots, s_l\}$, which are known to be very suitable for the task $\mathcal{T}$. The presentation of the sample set is necessary for measuring the relevancy between the task $\mathcal{T}$ and datasets in $\mathbb{D}$. The sample set can be provided by the data consumer or the data evaluator. For each dataset in $\mathbb{D}$, the data evaluator quantifies a collection of quality dimensions $\mathcal{Q} = \{q_1, q_2, \dots, q_n\}$, including both *task-independent intrinsic quality* and *task-dependent contextual quality* with respect to $\mathcal{T}$. In the end, the evaluator returns a ranking σ of m datasets. The higher the ranking of a dataset is, the higher quality it has for the task $\mathcal{T}$.

B. Design goals

We design our task-oriented data quality assessment framework TODQA (see in Figure 1) to achieve the following three nontrivial goals:

- *Rationality:* According to the assessment of TODQA, a dataset with higher task-oriented quality should empower the task $\mathcal{T}$ to achieve better performance (*e.g.*, higher accuracy and efficiency) with high probability than a lower quality dataset.
- *Interpretability:* When different datasets perform quite diversely on the same task, the assessment results should explain what differences of their quality dimensions result in the performance/rank differences.
- *Efficiency and accuracy:* TODQA should be efficient enough to assess the quality of large-scale datasets, while guarantee the accuracy of assessment.

III. Task-oriented Dataset Quality Assessment.

For a specific task, a high-quality dataset should be not only intrinsically good, but also contextually appropriate for the task [2]. In this section, we present the concrete definitions and quantification methods for each dimension of intrinsic quality and contextual quality. We also propose two fast computation algorithms for large-scale datasets measurements. For better illustration, some notations are summarized in Table I.

A. Intrinsic Quality Assessment

Intrinsic quality evaluates task-independent internal characteristics of data. Based on an extensive review of existing work, generally, intrinsic quality can be defined in multiple

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

dimensions including correctness, precision [2], [3]. For a dataset $\mathcal{D}$, let the quantified value of these two dimensions be $q_c^{\mathcal{D}}$, $q_p^{\mathcal{D}}$. The minimum quality requirements for two dimensions are θ_c, θ_p. $\mathcal{D}$ must fulfill the minimum intrinsic quality requirement R, where

$$R = (q_c^{\mathcal{D}} \geq \theta_c) \wedge (q_p^{\mathcal{D}} \geq \theta_p). \tag{1}$$

Poor-quality datasets failing to meed R will be put in the bottom of the output rank, needing no further assessment. **Correctness** evaluates the extent to which data is correct and reliable [2]. For example, the grammatical correctness of text documents [7], and the correctness of labels for images. The correctness of a dataset $\mathcal{D}$ is denoted by $q_c^{\mathcal{D}}$, which is quantified as $q_c^{\mathcal{D}} = \frac{1}{|\mathcal{D}|}\sum_{d_i \in \mathcal{D}} q_c^{d_i}$. **Precision** refers to the accuracy of data acquisition and storage, such as the accuracy of sensor readings, the sharpness or the compression ratio of images and videos, *etc*. We quantify the dataset precision by $q_p^{\mathcal{D}} = \frac{1}{|\mathcal{D}|}\sum_{d_i \in \mathcal{D}} q_p^{d_i}$, where $q_p^{d_i}$ can be obtained by existing quality assessments [5], [8].

B. Contextual Quality Assessment.

Contextual data quality highlights the demand that data quality must be considered within the context of the data usage [2]. Few works study the quantification method of contextual quality, not to mention the contextual quality of large datasets. In this work, we propose a series of novel contextual quality dimensions including task relevancy, content diversity, appropriate amount and design efficient methods to quantify these proposed dimensions and two existing dimensions, *e.g.*, completeness and timeliness.

1) Task relevancy: Task relevancy measures the extent to which the data content is relevant to the requirement of the task. Since the requirement can be expressed by a collection of sample data pieces $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$, we can quantify the similarity between a dataset $\mathcal{D}$ and the sample dataset to indicate the task relevancy [9]. For example, if the data consumer wants to train a scene classifier, the sample set $\mathcal{S}$ should be several images of different scenes. The choice of the sample set (*e.g.*, its size and content distribution) will affect the evaluation of task relevancy. We will explore the impact of the sample set by detailed experiments in Section IV. Note that, the size of $\mathcal{D}$ is usually much larger than the size of $\mathcal{S}$.

Definition 1 (Task relevancy.). *Given a dataset $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$ to be assessed and a sample dataset $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$ of a specific task $\mathcal{T}$, the task relevancy of the dataset $\mathcal{D}$ to the task $\mathcal{T}$ is*

$$q_{tr}^{\mathcal{D}|\mathcal{S}} = \frac{X(\mathcal{D}, \mathcal{S})}{|\mathcal{D}|} \tag{2}$$

$X(\cdot, \cdot)$ evaluates the size of intersection of two sets, which is defined as follows:

$$X(\mathcal{D}, \mathcal{S}) = \sum_{d_i \in \mathcal{D}} I(\min_{s_j \in \mathcal{S}} Dis(d_i, s_j), \delta) \tag{3}$$

$Dis(\cdot, \cdot)$ measures the distance of two data content, and

$$I(\min_{s_j \in \mathcal{S}} Dis(d_i, s_j), \delta) = \begin{cases} 1, & \min\limits_{s_j \in \mathcal{S}} Dis(d_i, s_j) \leq \delta \\ 0, & \min\limits_{s_j \in \mathcal{S}} Dis(d_i, s_j) > \delta \end{cases} \tag{4}$$

For the sake of generality, $Dis(\cdot, \cdot)$ is an abstract function measuring content dissimilarity between two pieces, e.g., calculating Euclidean distance or cosine distance of two extracted feature vectors (see in Section II). δ is a empirical value. The larger value of Equation (2), the greater relevancy of the dataset to the task.

The time complexity of assessing the task relevancy is $O(L \cdot l \cdot |\mathcal{D}|)$, where L is the dimension of the feature vector. Therefore, when the size of $\mathcal{D}$ is large and the dimension L is high, there will be a large computational overhead for assessment. It is desired to significantly reduce the computational cost while retain the assessment accuracy. We notice that we only care the most similar pairs whose distances are below a threshold δ. Based on this observation, our basic idea is to efficiently filter pairs that are very likely to be similar and ignore other pairs. Specifically, we design a fast calculation method based on sampling and locality-sensitive hashing (LSH) to reduce a large portion of computation with little accuracy loss. Before we present detailed design of our algorithm, we simply review the (r_1, r_2, p_1, p_2)-sensitive property of LSH [10].

Algorithm 1 Fast calculation method for Task Relevancy.

```
Input: A dataset $\mathcal{D}$, a sample dataset $\mathcal{S}$, and a hash function $h(\cdot)$ of a LSH family.
Output: $q_{tr}^{\mathcal{D}|\mathcal{S}}$: relevancy quality of dataset $\mathcal{D}$ for the task.
Partition the feature space using $k$ random hyperplanes parametrized by $w_1, w_2, \ldots, w_k$, thus obtaining the set of $k$ hash functions $H = \{h_1, h_2, \cdots, h_k\}$.
for $\mathcal{A} \in (\mathcal{D} \cup \{\mathcal{S}\})$ do
    for $a \in \mathcal{A}$ do
        for $i = 1, 2, \ldots, k$ do
            $h_i(a) = \begin{cases} 1, & w_i^T a \geq 0 \\ 0, & w_i^T a < 0 \end{cases}$
        end for
        Place $a$ in a bucket using $H(a)$ as its key/index.
    end for
end for
The set of buckets is denoted as $\mathcal{B}[H(\mathcal{D} \cup \{\mathcal{S}\})]$.
$Count = 0$.
for each bucket $\mathcal{B}[v]$ in $\mathcal{B}[H(\mathcal{D} \cup \{\mathcal{S}\})]$ do
    $\mathcal{D}' = \{d | d \in \mathcal{D}, H(d) = v\}$.
    $\mathcal{S}' = \{s | s \in \mathcal{S}, H(s) = v\}$.
    Calculate $X(\mathcal{D}', \mathcal{S}')$.
    $Count = Count + X(\mathcal{D}', \mathcal{S}')$.
end for
Calculate $q_{tr}^{\mathcal{D}|\mathcal{S}} = \frac{Count}{|\mathcal{D}|}$.
return $q_{tr}^{\mathcal{D}|\mathcal{S}}$.
```

Definition 2 ((r_1, r_2, p_1, p_2)-sensitive)**.** *A LSH family $\mathcal{H}$ of functions is defined for a metric space $\mathcal{X}$, four thresholds r_1, r_2, $p_1 \in [0, 1]$ and $p_2 \in [0, 1]$, which satisfy $r_1 < r_2$ and $p_1 > p_2$. $\mathcal{H}$ is a family of functions $h : \mathcal{X} \to \mathcal{B}$ which maps elements from the metric space to a bucket $b \in \mathcal{B}$. $\mathcal{H}$ is called (r_1, r_2, p_1, p_2)-sensitive for metric $Dis(\cdot, \cdot)$ if for any*

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

pair $p, q \in \mathcal{X}$, using a function $h \in \mathcal{H}$ which is chosen at random, we have
if $Dis(p, q) \leq r_1$, then $Pr_{\mathcal{H}}[h(p) = h(q)] \geq p_1$,
if $Dis(p, q) \geq r_2$, then $Pr_{\mathcal{H}}[h(p) = h(q)] \leq p_2$.

We use k functions $H = \{h_1, h_2, \cdots, h_k\}$ of a LSH family $\mathcal{H}$ to hash input feature vectors of data pieces, so that similar data pieces will be mapped to the same bucket while dissimilar data pieces will be mapped to different buckets with high probability. Here, the threshold r_1 is set to the distance threshold δ in Equation (4). In this way, to obtain $X(\mathcal{D}, \mathcal{S})$, we only need to calculate the distance $Dis(d_i, s_j)$ ($d_i \in \mathcal{D}$ and $s_j \in \mathcal{S}$) when feature vectors of d_i and s_i are mapped to the same bucket ($H(d_i) = H(s_j)$). The algorithm is illustrated in Algorithm 1 with the error bound $(1-p_1)$ and time complexity $O(L \cdot |\mathcal{S}| \cdot \log |\mathcal{D}|)$. When the dimension L of feature vectors is large, *e.g.*, $L = 1000$ for image feature vectors extracted by a VGG-16 network [11], we further cut down the time complexity to $O(\frac{16}{\epsilon^2} \cdot |\mathcal{S}| \cdot \log |\mathcal{D}|)$ by reducing the dimension of feature vectors to l^* ($l^* > 8\frac{\log |\mathcal{D}|}{\epsilon^2}$) with an additional error $0 < \epsilon < 1$. At this time, the error bound is $(1 - p_1 + \epsilon)$.

Lemma 1. *The fast calculation method with dimensionality reduction of feature vectors for task relevancy has the time complexity $O(\frac{16}{\epsilon^2} \cdot |\mathcal{S}| \cdot \log |\mathcal{D}|)$.*

Proof. In Algorithm 1, line 4-6 costs $L \cdot k$, line 7 costs $|\mathcal{D}|/2^k$ in expectation (because there are $|\mathcal{D}|$ points in dataset $\mathcal{D}$ and 2^k regions in our partitioned space). Thus, the total cost is $Lk + \log |\mathcal{D}|/2^k$. When k is taken to be about $\log |\mathcal{D}|$, we get the desired $O(L \cdot \log |\mathcal{D}|)$. Thus for points in $\mathcal{S}$, the total expectation cost is $O(L \cdot |\mathcal{S}| \cdot \log |\mathcal{D}|)$. Given $0 < \epsilon < 1$, a dataset $\mathcal{D}$ whose data pieces have feature vectors in R^L and a number $l^* > 8\frac{\log |\mathcal{D}|}{\epsilon^2}$, according to Johnson-Lindenstrauss lemma, there is a linear projection $f : R^L \rightarrow R^{l^*}$ such that,$(1-\epsilon)||d_i - d_j||^2 \leq ||f(d_i) - f(d_j)||^2 \leq (1+\epsilon)||d_i - d_j||^2$, for all $d_i, d_j \in \mathcal{D}$. By reducing the dimension of feature vectors from L to l^*, the total cost for calculating task relevancy is $O(\frac{16}{\epsilon^2} \cdot |\mathcal{S}| \cdot \log |\mathcal{D}|)$.

□

2) Content diversity: Here, we want to answer the question that does a dataset with high intrinsic quality and task relevancy necessarily make a high-quality data for the task? The answer depends on the nature of different tasks. As an example, in a task of training a face recognition model, images of many different persons' faces are more preferred than images of one person's face, although they may have the same relevancy quality. Images of one person's face, however, could be more preferred by a task of training face unlock model for his/her phone. Various tasks have divergent requirements for data content diversity. For tasks like training machine learning models, proper data diversity can mitigate overfitting and improve the generalization ability of models. Therefore, we consider content diversity as an influential contextual quality dimension. A data consumer can express his/her requirement by setting the range of diversity or simply provide a sample dataset to imply his/her preferred diversity.

There are various ways to define diversity of a set [12]. Here, we employ the average pairwise distance among feature vectors of data pieces in a dataset [12]. Formally, the content diversity $q_{cd}^{\mathcal{D}}$ of a dataset $\mathcal{D}$ is

$$q_{cd}^{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{d_i \in \mathcal{D}} \sum_{d_j \in \mathcal{D}} Dis(d_i, d_j), \tag{5}$$

where $Dis(\cdot, \cdot)$ is the same function as that in Equation (4).

The time complexity of Equation (5) is $O(|\mathcal{D}|^2)$. When the dataset $\mathcal{D}$ is large, *e.g.*, millions of images, the quadratic time cost is very expensive. To improve the efficiency, we need a fast calculation algorithm. The existing algorithm achieves an approximation of the average distance with a multiplicative error $(1 + \eta)$ in time $O(|\mathcal{D}|/\eta^{7/2})$ with high probability [13]. In this work, we propose a sampling based algorithm to achieve comparable average distance approximations while with smaller time cost (see evaluation in Section IV). Assuming that the pairwise distances satisfy a certain distribution, our basic idea is to sample a small number of pairwise distances following this distribution. We leverage LSH to characterize the implicit distribution. Using k hash functions $H = \{h_1, h_2, \cdots, h_k\}$ of a LSH family $\mathcal{H}$, data pieces can be mapped to buckets. So that, the numbers of data pairs in different buckets imply the distribution of pairwise distances of this dataset. By randomly sampling a certain proportion (*e.g.*, the sampling rate r) of data pairs from each bucket, we can obtain the sampling set $\mathcal{G}$, whose average pairwise distance $q_{cd}^{\mathcal{G}}$ is the approximation of $q_{cd}^{\mathcal{D}}$. The details are presented in Algorithm 2. The time complexity of Algorithm 2 is $O(|\mathcal{G}|^2)$. As the sampling rate increases, the approximation error gets smaller. According to Lemma 2, in practice, we can set a proper sampling rate to keep $|\mathcal{G}|^2 < |\mathcal{D}|$ while retain a good approximation.

Algorithm 2 Fast calculation method for Content Diversity.

```
Input: A dataset $\mathcal{D}$, $k$ hash function $H = \{h_1, h_2, \cdots, h_k\}$ of a LSH
    family, and the sampling rate $r \in (0, 1)$.
Output: $p_{cd}^{\mathcal{G}}$: diversity approximation.
$\mathcal{G} \leftarrow \emptyset$.
for each $d_i \in \mathcal{D}$ do
    Calculate $H(d_i)$ using the equation defined in Algorithm line 4-6.
    Add $d_i$ to the bucket $\mathcal{B}[h(d_i)]$.
end for
for each $\mathcal{B}[j]$ in $\mathcal{B}[H(\mathcal{D})]$ do
    $\mathcal{G} \leftarrow$ randomly selected $\lceil r \cdot |\mathcal{B}[j]| \rceil$ data pairs in $\mathcal{B}[j]$.
end for
$p_{cd}^{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{d_i, d_j \in \mathcal{G}, i \neq j} Dis(d_i, d_j)$.
return $p_{cd}^{\mathcal{G}}$.
```

Lemma 2. *$Pr(|q_{cd}^{\mathcal{G}} - q_{cd}^{\mathcal{D}}| \geq \theta) \leq 2exp(-2M \cdot \theta^2)$ for $\theta > 0$, here $M = |\mathcal{G}|^2$, which is the number of sampled distances.*

Proof. The sampled M distances are independent random variables bounded by the interval $[a_i, b_i], : a_i \leq l_i \leq b_i$. Without loss of generality, we assume the interval is $[0, 1]$. $q_{cd}^{\mathcal{G}}$ is the mean of these variables. According to the Hoeffding's inequality, we can obtain $Pr(|q_{cd}^{\mathcal{G}} - q_{cd}^{\mathcal{D}}| \geq \theta) \leq 2exp(-2M \cdot \theta^2)$ for $\theta > 0$. □

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

3) Completeness denotes the extent to which data is of sufficient breadth, depth, and scope for the task at hand [2], e.g., if the image dataset $\mathcal{D}$ has one label for an classification task, it is complete for the task with $q_{co}^{\mathcal{D}}$ being 1.0. **Appropriate amount**. In most cases (e.g., deep learning), the more data pieces, the higher the performance (e.g., performance on vision tasks increases logarithmically based on volume of training data) [14]. However, it has been found in practice that when the size of dataset exceeds a certain optimal point, it will not only cause unnecessary computational and storage resources waste but also decreases the decision making performance [15]. The appropriate amount of a dataset (denoted as $q_a^{\mathcal{D}}$) for a task should be adequate but do not cause extra overhead. **Timeless** (denoted as $q_t^{\mathcal{D}}$) evaluates the extent to which the age of the data is suitable for the task, which can be quantified as the maximum of two terms: 0 and $1 - currency/volatility$ [16].

C. Rank aggregation.

Given a task, for dataset $\mathcal{D}_i \in \mathbb{D}$, now we can obtain the value of each quality dimension $q_j^{\mathcal{D}_i} \in \mathcal{Q}$ introduced in Section III. Since quality values of different dimensions are incomparable, we propose a rank aggregation method to obtain the overall ranking based on their ranks in quality dimensions.

Specifically, for each quality dimension q_j, a ranking τ_j can be obtained by ordering all datasets by their quality in this dimension in descending order. Then we have a set of rankings $\tau = \{\tau_1, \tau_2, \ldots, \tau_n\}, n = |Q|$, where each $|\tau_j| = |\mathbb{D}| = m$. Let $\tau_j(\mathcal{D}_i)$ denote the position of $\mathcal{D}_i$ in the ranking τ_j. We address the optimal rank aggregation problem by finding the optimal overall ranking σ that minimizes the extent of disagreement among rankings $\tau_1, \tau_2, \ldots, \tau_n$. The optimization objective function is

$$\min F(\sigma, \tau_1, \tau_2, \ldots, \tau_n) = (1/n)\sum_{i=1}^{n} \mu_i K(\sigma, \tau_i), \tag{6}$$

$$K(\sigma, \tau_i) = |\{(j,k)|j \neq k, \sigma(\mathcal{D}_j) < \sigma(\mathcal{D}_k), \tau(\mathcal{D}_j) > \tau(\mathcal{D}_k)\}|.$$

Here, $K(\sigma, \tau_i)$ is the Kendall tau distance, which counts the number of pairwise disagreements between two lists. The weight μ_i represents the influence of the quality dimension q_i on the overall quality, which can be determined by the data consumer. This problem is NP-hard, and there exists a 2-approximation algorithm [17]. It demonstrates that, Kendall tau distance can be approximated via the Spearman footrule distance, which can be computed in polynomial time via a minimum cost matching [17]. We adopted and modified this 2-approximation algorithm to efficiently obtain the overall ranking of n quality dimensions.

IV. Evaluation

TODQA evaluates task-oriented data quality from two aspects, intrinsic quality and contextual quality. Given a specific task, a sample set, and some candidate datasets, TODQA ranks candidate data-sets by quality in descending order. In this section, we validate the feasibility and efficiency of TODQA on two types of data, images and text documents.

A. Experiment Configuration

We acquired image and text datasets from publicly available sources. Table II gives the details of our datasets.

Image datasets. We adopted six popular image datasets *ImageNet* [18], *COCO test* [19], *VOC2012* [20], *COIL-100* [21], Labeled Faces in the Wild (LFW) [22] and Large Logo Dataset (LLD) [23], which are denoted by $\mathbb{D}_I = \{\mathcal{D}_{I_1}, \mathcal{D}_{I_2}, \ldots, \mathcal{D}_{I_6}\}$.

Text datasets. We adopted four popular text datasets *MR* [24], *SST-1* [25], *Subj* [26] and *CR* [27], which are denoted by $\mathbb{D}_T = \{\mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \mathcal{D}_{T_3}, \mathcal{D}_{T_4}\}$.

Tasks. For image data, the task $\mathcal{T}_I$ in our experiments is to train a 10-category image classification model. The sample set S_I for this task is composed of images randomly chosen from 10 categories with 100 images for each from ImageNet. For text data, the task $\mathcal{T}_T$ in our experiments is sentiment analysis for text documents. The sample set S_T for this task is composed of 500 positive and 500 negative reviews randomly chosen from these four datasets (See Section IV-C for detailed analyses of the impact of sample sets on evaluations). The image classification and text sentiment analysis models are trained with TensorFlow 1.10.1. All evaluation experiments are conducted on a server equipped with a 12-core i7 Intel CPU, 64G of RAM and 4 Titan X GPUs.

TABLE II
Overview of our datasets.

Dataset type	Notations	Dataset name	Size	Total Size
Image	$\mathcal{D}_{I_1}$	ImageNet	200,000	460,247
	$\mathcal{D}_{I_2}$	COCO test	100,000	
	$\mathcal{D}_{I_3}$	VOC2012	17,125	
	$\mathcal{D}_{I_4}$	COIL-100	7202	
	$\mathcal{D}_{I_5}$	LFW	13,000	
	$\mathcal{D}_{I_6}$	LLD	122,920	
Text	$\mathcal{D}_{T_1}$	MR	10,662	37,372
	$\mathcal{D}_{T_2}$	SST-1	10,605	
	$\mathcal{D}_{T_3}$	Subj	5500	
	$\mathcal{D}_{T_4}$	CR	10,605	

B. Data Feature Extraction

To measure the contextual quality of datasets (*e.g.*, task relevancy and content diversity), we need to extract feature of each image and text document to represent the data content. For images, we adopt the well-known VGG-16 [11] model to extract feature vectors and classify images. Studies have shown that the feature activations of the eighth fully-connected layer of VGG-16 can serve as a good abstract of image content [18]. For text datasets, we adopt the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) [28] to extract feature representations.

C. Data Quality Assessment

1) Correctness. For each image dataset $\mathcal{D}_{I_i}$, we randomly sample items from it to calculate the correctness quality $q_c^{\mathcal{D}_{I_i}}$. All sampled items are labeled correctly, thus we set $q_c^{\mathcal{D}_{I_i}} = 1, i \in [1, 6]$ and put all datasets in the same position of the ranking τ_c^I. For text datasets, we evaluate the correctness of spelling and grammar using the exiting method [7]. The result is shown in Figure 2. **Precision.** For images, we

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

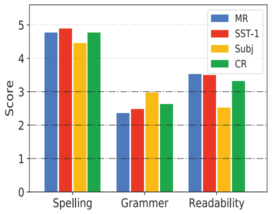


Fig. 2. Correctness & Precision scores for text datasets.

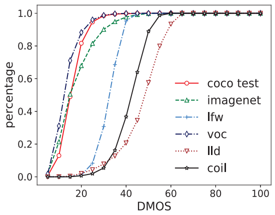


Fig. 3. The cumulative distribution of DMOS for image datasets.

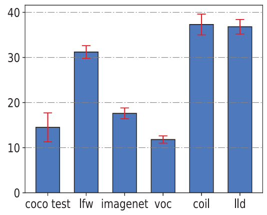


Fig. 4. Average DMOS for image datasets.

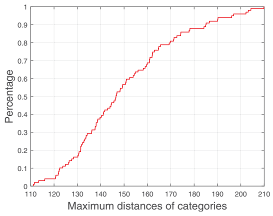


Fig. 5. Maximum distances within each category.

adopted a CNN model [5] to indicate the precision of image datasets by the Differential Mean Opinion Score (DMOS) ranging in $[0, 100]$. The larger the DMOS, the lower the precision. For each image dataset, the cumulative distribution of DMOS is shown in Figure 3 and the average DMOS is shown in Figure 4. The ranking of datasets by precision is $\tau_p^I = \{\mathcal{D}_{I_3}, \mathcal{D}_{I_2}, \mathcal{D}_{I_1}, \mathcal{D}_{I_5}, \mathcal{D}_{I_6}, \mathcal{D}_{I_4}\}$. For text datasets, we quantify precision by their readability [8]. The result is shown in Figure 2. The ranking of datasets in terms of precision is $\tau_p^T = \{\mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \mathcal{D}_{T_4}, \mathcal{D}_{T_3}\}$.

2) Task relevancy. After obtaining feature vectors for items in all datasets, we calculate the task relevancy using the proposed sampling based algorithm (Algorithm 1). For the sample dataset $\mathcal{S}_I/\mathcal{S}_T$ and each dataset $\mathcal{D}_{I_i}/\mathcal{D}_{T_i}$, we leverage k hash functions to hash all high-dimensional feature vectors (*e.g.*, 1000 dimensions for images and 768 dimensions for text) to k-dimensional (*e.g.*, $k = 7$) Hamming vectors. Then, we only calculate the distance between items from $\mathcal{D}_{I_i}/\mathcal{D}_{T_i}$ and items from $\mathcal{S}_I/\mathcal{S}_T$ in the same bucket, and finally obtain the task relevancy. We validate the accuracy and efficiency of our sampling based algorithm first, then use this algorithm to evaluate task relevancy of datasets.

•**Algorithm accuracy.** We measure the accuracy loss of our sampling based algorithm from two aspects: the accuracy of the LSH function and the accuracy of the estimate dataset relevancy quality. Two parameters determine the accuracy of the LSH function: the distance threshold δ in Eq. (4) and the dimension k of the hash space. Specifically, we use three indicators to characterize the accuracy of the LSH function: (1) precision: the ratio of the number of distances smaller than δ in all buckets to the total number of distances smaller than δ; (2) recall: the ratio of the number of distances smaller than δ in all buckets to the total number of distances in all buckets; (3) F-score: $\frac{2 \cdot precision \cdot recall}{precision + recall}$. Taking images as an example, we calculate the maximum distance of images within each category (with 100 categories, 500 images for each one). The distribution of 100 maximum distances is presented in Figure 5. The precision, recall and F-score against to different δ and k are illustrated in Figure 6-8. The results suggest that the LSH function achieves the highest accuracy when $\delta = 130$ and $k = 7$, and the precision is 84.09%, recall is 78.86% and F-score is 81.39%. Further, we investigate the estimation error and time cost of task relevancy using sample sets with different sizes (shown in Figure 9). We set the sample size of each category to be 100 to achieve both low evaluation error and low time cost. Now we compare the true relevancy quality directly calculated according to the Definition 1 with the result estimated by our sampling based algorithm. Here, we consider two types of datasets to be evaluated, which are composed of images from the same categories and different categories with the sample dataset (results are shown in Figure 10(a), 11(a)). Obliviously, the relevancy of datasets composed of images of the same categories as S_I is much higher than that of datasets composed of images of different categories with S_I. More importantly, the estimated relevancy using our algorithm is very close to the true relevancy, especially when images are from same categories as S_I. *The average estimation error for two types of datasets is only* 0.034 *when the relevancy score is with the range* $[0, 1]$. For text documents, we hash 768-dimensional feature vectors to 7-dimension Hamming vectors. *The results are similar and the average estimation error is only* 0.0019.

•**Algorithm Efficiency.** For the aforementioned experiments, we compare the runtime of our sampling based algorithm with that of the original method. As shown in Figure 10(b) and Figure 11(b), our algorithm significantly reduce the runtime. As an example, when the size of the dataset to be evaluated is $100,000$, the average runtime the original method is 320.767s, while the average runtime of our method is only 24.169s, *saving 92.5% runtime*.

•**Task relevancy of datasets.** With high accuracy and efficiency guarantees, we use our proposed algorithm to measure the task relevancy of each dataset to the sample dataset. The relevancy scores of 6 image datasets are 0.863, 0.121, 0.194, 0.205, 0.185 and 0.049 respectively, Thus the ranking by task relevancy is $\tau_{tr}^I = \{\mathcal{D}_{I_1}, \mathcal{D}_{I_4}, \mathcal{D}_{I_3}, \mathcal{D}_{I_5}, \mathcal{D}_{I_2}, \mathcal{D}_{I_6}\}$. The relevancy scores of four text datasets are 0.9995, 0.9984, 0.9957 and 0.9994 respectively. Thus the ranking by task relevancy is $\tau_{tr}^T = \{\mathcal{D}_1, \mathcal{D}_4, \mathcal{D}_2, \mathcal{D}_3\}$.

3) Content diversity. Here we first measure the accuracy and efficiency of Algorithm 2 and then evaluate content diversity of all datasets using our algorithm.

•**Accuracy.** The accuracy of estimated content diversity is also affected by the accuracy of the adopted LSH functions, which has already been analyzed in the above subsection. Therefore, we directly compare the true content diversity obtained by the original method and the estimated value using our sampling based algorithm (Algorithm 2) and a random sampling method [13]. For image datasets, as the results depicted in Figure 12(a), the estimated diversity using our

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

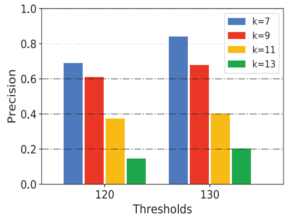


Fig. 6. Precision of LSH with different k and δ. k is the dimension of the hash vector.

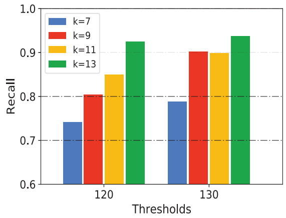


Fig. 7. Recall of LSH with different k and δ. k is the dimension of the hash vector.

algorithm is very close to the true value with a only 0.35% average relative error. For text datasets, the average relative error is only 0.113%. Although the random sampling method achieves slightly better accuracy, our method is much more efficient as presented below.

•**Efficiency.** We also compare the total runtime of our algorithm with that of the original method and the random sampling method [13], and our algorithm costs much less time (see Figure 12(b)). When the size of the image dataset is $40,000$, the runtime of the original method and random sampling method are $4,254$s and 429s. The runtime of our algorithm is only 69s, which saves 98.4% runtime compared to the original method and saves 84% runtime compared to the existing method.

•**Content diversity of datasets.** Using our algorithm, the content diversity of 6 image datasets are 150.23, 128.02, 102.84, 61.27, 51.65 and 54.63 respectively. For the task training an image classification model, a dataset with higher diversity is preferred to increase the generalization ability of model. So the ranking by content diversity is $\tau_{cd}^I = \{\mathcal{D}_{I_1}, \mathcal{D}_{I_2}, \mathcal{D}_{I_3}, \mathcal{D}_{I_4}, \mathcal{D}_{I_6}, \mathcal{D}_{I_5}\}$. For 4 text datasets, content diversity are 8.571, 8.379, 7.531 and 8.826, with the average deviation 0.113. Similarly, a dataset with higher diversity is better to train a robust sentiment classification model, thus ranking of these datasets is $\tau_{cd}^T = \{\mathcal{D}_{T_4}, \mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \mathcal{D}_{T_3}\}$.

3) Completeness. Since the ImageNet dataset has category labels, other datasets do not, so it is at the first position of the ranking τ_{co}^I, and the other five datasets are after it. For text datasets, all text reviews have corresponding sentiment labels, so they are all at the same position in the ranking τ_{co}^T. **Appropriate amount.** The size of each dataset $\mathcal{D}_{I_i} \in \mathbb{D}_I$, $\mathcal{D}_{T_i} \in \mathbb{D}_T$ does not exceed the optimal size of data that is appropriate for image classification/sentiment analysis [14], [27], thus we rank them by their sizes. That are $\tau_a^I = \{\mathcal{D}_{I_1}, \mathcal{D}_{I_6}, \mathcal{D}_{I_2}, \mathcal{D}_{I_3}, \mathcal{D}_{I_5}, \mathcal{D}_{I_4}\}$ and $\tau_a^T = \{\mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \mathcal{D}_{T_4}, \mathcal{D}_{T_3}\}$. **Timeliness.** The tasks in our experiments (*e.g.*, classification tasks for images and text documents) are insensitive to the two factors (currency and volatility), so every dataset's timeliness is 1 and it is a tie in terms of timeliness.

D. Rank Aggregation

After obtaining rankings $\tau_1^I, \tau_2^I, \ldots, \tau_n^I$ of image datasets and rankings $\tau_1^T, \tau_2^T, \ldots, \tau_n^T$ of text datasets for every quality dimension, we use the rank aggregation algorithm to find the optimal rankings σ^I and σ^T, given the tasks of training classification models for images and sentiment analysis for texts. When the weights of all quality dimensions are set to be the same (*e.g.*, $\mu_i = 1.0, i \in [1, n]$), the optimal ranking for image datasets is $\sigma^I = \{\mathcal{D}_{I_1}, \mathcal{D}_{I_2}, \mathcal{D}_{I_3}, \mathcal{D}_{I_5}, \mathcal{D}_{I_4}, \mathcal{D}_{I_6}\}$, and for text datasets is $\sigma^T = \{\mathcal{D}_{T_4}, \mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \mathcal{D}_{T_3}\}$.

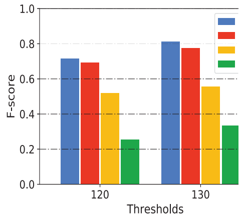


Fig. 8. F-score of LSH with different k and δ. k is the dimension of the hash vector.

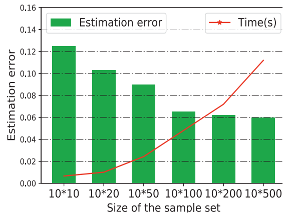
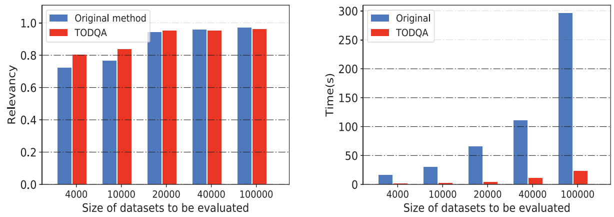


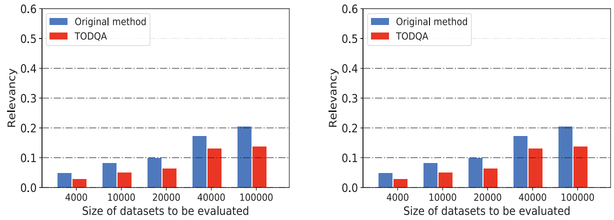
Fig. 9. The effect of the sample set S_I's size on the task relevancy evaluation.



(a) Task relevancy.

(b) Runtime.

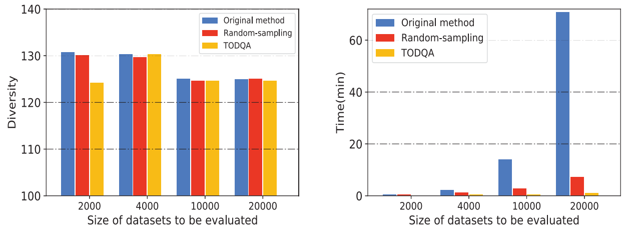
Fig. 10. The task relevancy and its runtime for datasets composed of images from same categories with the sample dataset.



(a) Task relevancy.

(b) Runtime.

Fig. 11. The task relevancy and its runtime for datasets composed of images from same categories with the sample dataset.



(a) Content diversity.

(b) Runtime.

Fig. 12. The content diversity and its runtime of calculating content diversity for images.

Rationality and interpretability analysis. For image datasets, the result ranking σ^I is consistent with the experience recognition, in that ImageNet is most suitable for the task of image classification. For text datasets, we trained and tested a convolutional neural network [6] for sentiment classification using all 4 datasets. The accuracies of four datasets are

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.

$\{80.09\%, 47.35\%, 93.16\%, 84.61\%\}$ respectively. Comparing the model accuracy ranking with dataset quality ranking, we can conclude that datasets with higher task-oriented quality can achieve better performances on the task with high probability. We further analyze that the inconsistence between the model accuracy ranking and dataset quality ranking is caused by equal weights w_i of all quality dimensions in Eq. (6). Since text sentiment classification requires high spelling and grammar correctness, we increase the weight of correctness quality to 2.0 and get the ranking $\sigma^T = \{\mathcal{D}_{T_3}, \mathcal{D}_{T_4}, \mathcal{D}_{T_1}, \mathcal{D}_{T_2}\}$, which is consistent with the model accuracy ranking.

V. Related work.

Data quality has long been a critical part of information system management. The literature on data quality management (DQM) has defined and characterized different perspectives of data quality and its management. The concept of "fitness for use" is commonly adopted as the informal definition of data quality. Empirical studies have shown that data quality is perceived as a multi-dimensional concept [1], [2], [3], [4], [5], [6]. Wang *et al.* [2] pointed out that high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to data consumers. Existing efforts, however, often focus on inherent quality of data, neglect important contextual factors such as the target tasks or services which have been shown to strongly influence perceptions of data quality [4]. Most quality assessment methods propose various measurements, *e.g.*, correctness and timeliness, for structured data. Few works consider the general quality measurements for unstructured data. Other works however mainly measure quality of an individual data piece [5] other than the quality of a collection of data pieces, while the latter is more commonly used by present services. Besides, simply averaging measurements of data pieces and neglecting relationship among them fail to capture the characteristics of a dataset. Finally, though various dimensions of data quality have been proposed [1], few works attempt to fuse those dimensions to get a comprehensive data quality assessment.

VI. Conclusion

In this work, we propose TODQA, a task-oriented data quality assessment system that assesses and ranks datasets by their overall quality. Both intrinsic and contextual quality metrics are incorporated in the system. We improve efficiency by proposing two fast calculation algorithms for two quality dimensions, task relevancy and content diversity. Our experiments on real datasets validate the feasibility and efficiency of TODQA.

Acknowledgment

Lan Zhang is the corresponding author. This work is supported by the National Key R&D Program of China 2017YFB1003003, NSF China under Grants No. 61822209, 61932016,61751211, 61572281, 61520106007, and the Fundamental Research Funds for the Central Universities.

References

[1] L. Zhang, Y. Li, and X. Xiao, "Crowdbuy: Privacy-friendly image dataset purchasing via crowdsourcing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018.

[2] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, 1996.

[3] V. Siegert, "Content-and context-related trust in open multi-agent systems using linked data," in *International Conference on Web Engineering*. Springer, 2019, pp. 541–547.

[4] M. S. Marev, E. Compatangelo, and W. Vasconcelos, "Towards a context-dependent numerical data quality evaluation framework," *arXiv preprint arXiv:1810.09399*, 2018.

[5] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR '12*, 2014.

[6] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[7] J. Tetreault, J. Foster, and M. Chodorow, "Using parse features for preposition selection and error detection," in *Proceedings of the acl 2010 conference short papers*, 2010, pp. 353–358.

[8] W. H. DuBay, *Smart Language: Readers, Readability, and the Grading of Text.* ERIC, 2007.

[9] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 380–388.

[10] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] T. Wu, L. Chen, and Hui, "Hear the whole story: Towards the diversity of opinion in crowdsourcing markets," *Proceedings of the VLDB Endowment*, 2015.

[13] P. Indyk, "Sublinear time algorithms for metric space problems," in *STOC99*. ACM, 1999, pp. 428–434.

[14] C. Sun and A. Shrivastava, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE ICCV*, 2017.

[15] D. L. Moody and P. Walsh, "Measuring the value of information-an asset valuation approach." in *ECIS*, 1999, pp. 496–512.

[16] D. Ballou and Wang, "Modeling information manufacturing systems to determine information product quality," *Management Science, 1998*.

[17] I. Caragiannis, X. Chatzigeorgiou, G. A. Krimpas, and A. A. Voudouris, "Optimizing positional scoring rules for rank aggregation," *Artificial Intelligence*, vol. 267, pp. 58–77, 2019.

[18] A. Krizhevsky and Sutskever, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[19] O. Vinyals and A. Toshev, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[20] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," *PASCAL'11*, 2011.

[21] S. A. Nene and S. K. Nayar, "object image library (coil-100," 1996.

[22] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in face detection and facial image analysis*. Springer, 2016, pp. 189–248.

[23] A. Sage and Agustsson, "Logo synthesis and manipulation with clustered generative adversarial networks," *arXiv:1712.04407*, 2017.

[24] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 115–124.

[25] R. Socher and Perelygin, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[26] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *ACL*. Association for Computational Linguistics, 2004, p. 271.

[27] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *ACM SIGKDD*, 2004.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 04,2024 at 04:09:09 UTC from IEEE Xplore. Restrictions apply.