Future Generation Computer Systems 102 (2020) 53–65

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Semantic-aware data quality assessment for image big data

Yu Liu, Yangtao Wang, Ke Zhou *, Yujuan Yang, Yifei Liu

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

HIGHLIGHTS

- Data quality assessment is essential for realizing the promise of big data.
- Relevance can arouse the user's interest in exploiting the data source.
- IDSTH algorithm can extract semantic features with generalization ability.
- SHR algorithm calculates the importance score (rank) for each node (image).
- SDQA architecture can help assess the value of image big data.

ARTICLE INFO

Article history:
Received 28 January 2019
Received in revised form 11 June 2019
Accepted 25 July 2019
Available online 6 August 2019

Keywords:
Semantic-aware
Quality assessment
Image big data
IDSTH
SHR

ABSTRACT

Data quality (DQ) assessment is essential for realizing the promise of big data by judging the value of data in advance. Relevance, an indispensable dimension of DQ, focusing on "fitness for requirement", can arouse the user's interest in exploiting the data source. It has two-level evaluations: (1) the amount of data that meets the user's requirements; (2) the matching degree of these relevant data. However, there lack works of DQ assessment at dimension of relevance, especially for unstructured image data which focus on semantic similarity. When we try to evaluate semantic relevance between an image data source and a query (requirement), there are three challenges: (1) how to extract semantic information with generalization ability for all image data? (2) how to quantify relevance by fusing the quantity of relevant data and the degree of similarity comprehensively? (3) how to improve assessing efficiency of relevance in a big data scenario by design of an effective architecture?

To overcome these challenges, we propose a semantic-aware data quality assessment (SDQA) architecture which includes off-line analysis and on-line assessment. In off-line analysis, for an image data source, we first transform all images into hash codes using our improved Deep Self-taught Hashing (IDSTH) algorithm which can extract semantic features with generalization ability, then construct a graph using hash codes and restricted Hamming distance, next use our designed Semantic Hash Ranking (SHR) algorithm to calculate the importance score (rank) for each node (image), which takes both the quantity of relevant images and the degree of semantic similarity into consideration, and finally rank all images in descending order of score. During on-line assessment, we first convert the user's query into hash codes using IDSTH model, then retrieve matched images to collate their importance scores, and finally help the user determine whether the image data source is fit for his requirement. The results on public dataset and real-world dataset show effectiveness, superiority and on-line efficiency of our SDQA architecture.

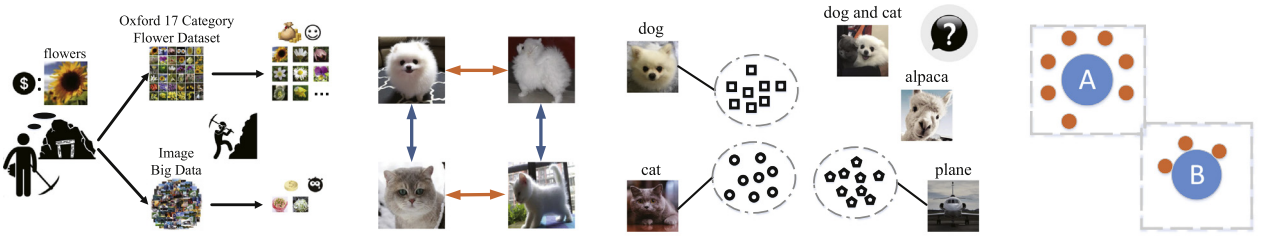
1. Introduction

With social platforms rapidly developing, the amount of image data has been growing exponentially. However, most of image data stored in the device fail to play their value, because data consumers who lack awareness of their contents, dare not audaciously use them. The reason for "lack of awareness" is not only the lack of labels and associations in storage, but also the limitations on the processing of unstructured data. Especially in a big data scenario, the implementation of data cleaning approaches is not feasible due to the size and the streaming nature of the data source, making the use of image big data more difficult [1].

Data quality assessment is a key to realize the promise of big data by judging the value of data in advance. Data Quality (DQ), generally defined as "fitness for use", is evaluated by five dimensions that include availability, reliability, relevance, usability and presentation quality [2]. Availability and reliability have been well studied in big data integration and fusion [3], where availability is defined as the degree of convenience for data accessing

* Corresponding author.
E-mail addresses: liu_yu@hust.edu.cn (Y. Liu), ytwbruce@hust.edu.cn (Y. Wang), k.zhou@hust.edu.cn (K. Zhou), gracee@hust.edu.cn (Y. Yang), yifeiliu@hust.edu.cn (Y. Liu).

https://doi.org/10.1016/j.future.2019.07.063
0167-739X/



(a) mining and wealth (b) semantic-aware and data-aware (c) out-of-sample problem in classification (d) evaluation dilemma

Fig. 1. (a) Mining does not mean wealth, because the data source may not be relevant. However, what if the user can perceive the relevance in advance? (b) The orange arrows represent the semantic-aware results, while the blue arrows represent the data-aware results. (c) The model is hard to classify "dog and cat" and "alpaca". (d) The orange nodes represent the data source. The blue nodes represent the user's query. The distances between different nodes represent the degree of similarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(i.e. accessibility, timeliness and authorization), while reliability refers to whether the user can trust the data (i.e. accuracy, integrity, consistency, completeness and auditability). However, neither of them can help the user determine whether the data source meets his demands. Paying all attention to availability and reliability is not enough to arouse the user's interest in exploiting the data source.

Relevance, as another indispensable dimension of DQ, is more crucial to judge "fitness for requirements". It is defined as the degree of correlation between the content of data and the user's expectations or demands. There are two-level evaluations for the relevance: (1) the amount of data that meets the user's requirements; (2) the matching degree of these relevant data. However, it is tricky to achieve these evaluations, especially for unstructured image data which focus on semantic similarity. Previous attempts at image data quality assessment pay attention to evaluate the quality of image presentation or retrieval results, which is inconsistent with our topic. Besides, existing DQ assessment jobs prefer to adopt data-aware methods to achieve evaluation at other dimensions. As a result, there lacks jobs evaluating relevance by semantic-aware methods, especially for image big data. On the other hand, for similarity comparison, semantic-aware feature extraction methods have been well studied in image retrieval, but those techniques have not been integrated into DQ assessment.

In this paper, we propose a semantic-aware data quality assessment (SDQA) architecture for image big data, where we design our improved deep self-taught hashing (IDSTH) algorithm, semantic hash ranking (SHR) algorithm and semantic-aware assessment process. Compared with the traditional DQ assessment, we focus on the relevance and bring three improvements: (1) we extract semantic feature with generalization ability and complete hash mapping for large-scale unlabeled image data; (2) we evaluate and quantify the relevance based on both the quantity of relevant images and the degree of semantic similarity; (3) we introduce deep learning and image retrieval technologies into the architecture, and combine off-line analysis with on-line evaluation to complete relevance assessment. Based on this, we ensure the efficiency of assessment in large-scale scenarios. To evaluate our work, we implement our SDQA on public datasets with classification supervision information. Our results show SDQA is sensitive to semantic content and gives higher scores to images whose semantic contents account for higher proportion in the whole dataset. At last, we show the relevance assessment result on real-world datasets for given requirements.

2. Motivation

In practical applications, data mining is like a gamble. There is a high chance that we have spent huge mining cost but fail to get corresponding value. This is not because the data mining algorithms are not good enough, but because the data cannot necessarily bring value to the demands (applications). As shown in Fig. 1(a), faced with the demand of "searching flowers", Oxford 17 Category Flower Dataset could provide enough value but Image Big Data failed in this aspect. This inherent value relationship between the data and demands can never be changed by data mining algorithms, but the degree of this relationship can be judged and measured in advance to avoid the gamble behavior as much as possible.

For judging and measuring this kind of relationship, it requires a unified parsing mechanism to extract both the semantic information of data and demands. However, semantic extraction for unstructured image data is restricted by the generalization ability and thus researchers doubt about the practicality of this conduct in a big data scenario. As a result, existing DQ assessment lacks exploration in the relevance of data. When we try to evaluate semantic relevance between an image data source and a query (requirement), there are three challenges: (1) how to extract semantic information with generalization ability for all image data? (2) how to quantify relevance by fusing the quantity of relevant data and the degree of similarity comprehensively? (3) how to improve assessing efficiency of relevance in a big data scenario by design of an effective architecture? We now discuss these issues in further detail.

(1) Semantic feature with generalization ability: semantic features are data representations that express human cognition. Different from data-aware feature extraction (commonly used in conventional DQ assessment) which entirely relies on the data distribution itself, semantic-aware methods rely more on hand-crafted labels and their results are suppose to be more meaningful. As shown in Fig. 1(b), according to what the blue arrows connect, dogs and cats have the same pixel distribution, so data-aware methods consider them to be the same category. However, dogs and cats have completely different semantic contents, as connected by the orange arrows. This is because data-aware methods care about what the data (pixel) looks like, while semantic-aware ones focus on what the data itself is.

Semantic-aware feature extraction has been well studied in image classification, which can ensure the distance between same categories data is getting smaller and smaller while the one between different categories is getting bigger and bigger. However, this way loses the generalization ability to cognize objects, which leads to serious out-of-sample problems. As shown in Fig. 1(c), the model that has learned single dog, cat and plane can classify images of single cat, dog and plane well, but it fails to classify "dog and cat" and recognize "alpaca". It is impossible that a model can learn all entities in real world, so semantic feature extraction methods without generalization ability cannot apply to big data scenarios.