The Journal of Systems and Software 118 (2016) 221–233

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

Efficient quality-driven source selection from massive data sources

Yiming Lin[a], Hongzhi Wang[a,*], Shuo Zhang[b], Jianzhong Li[c], Hong Gao[c]

[a] Harbin Institute of Technology, Harbin, 150001, China
[b] IBM Research China, Beijing, 100193, China
[c] Harbin Institute of Technology, Harbin, 150001, China

ARTICLE INFO

Article history:
Received 4 September 2015
Revised 23 February 2016
Accepted 16 May 2016
Available online 17 May 2016

Keywords:
Information integration
Data source selection
Data quality

ABSTRACT

The query based on massive database is time-consuming and difficult. And the uneven quality of data source makes the multiple source selection more challenging. The low-quality data source can even make the result of the information unexpected. How to efficiently select quality-driven data sources on massive database remains a hard problem. In this paper, we study the efficient source selection problem on massive data set considering the quality of data sources. Our approach evaluates the quality of data source and balances the limitation of resources and the completeness of data source. For data source selection for a specific query, our method could select the data sources with the number of keywords larger than a given threshold. And the selected sources are ranked according to the values of information in data sources. Experimental results demonstrate that our method can scale to millions of data sources and perform pretty efficiently.

1. Introduction

In the era of big data, many heterogeneous data sources emerge. To make sufficient usage of the data sources, information integration is in demand. That is to access all the required data sources for a given query and combine the returned results from data sources.

Currently, the number of data sources becomes large. Different from traditional one, millions of data sources are to be handled for current information integration. Google announced that the number of the Web pages reached one trillion. Such a large number of Web pages also mean the existence of massive data sources. In 1998, the index number is only 26 million and increased to 1 billion in 2000. Now there are more than tens of billions of new index page in one day. With the large amount of data sources, the time of schema mapping and result merging could not be ignored. Additionally, since the data sources may be in form of deep web or even web pages, the accessing of data sources may be time-consuming. Thus, for a given query, all the data sources should not be accessed and a small share of data sources related to the query should be selected. Thus in information integration for massive data sources, data source selection is a crucial step. Even though data source selection has been widely studied for traditional information integration on a small amount of data sources, it is not suitable for that on massive data sources, which brings the following new challenges.

First, the efficiency should be taken into consideration. Since the amount of data sources could be large and each of data source may contain many attributes, source selection has to handle massive data. Thus efficient data source selection approach suitable for large-scale data set is in demand.

Second, as the cost of accessing data sources could be large if many data sources are selected, the number of selected data sources should be very limited. Since many data sources may contain content related to the given query, the more data sources are selected, the completer the query results are. Thus, it is a problem to balance the limitation of resources and the completeness of data sources.

Last but not least, the quality of the data sources could be low. Traditional information integration approaches always assume that the quality of each data source is high. Current concerns include only the freshness of data sources and the errors led by integration such as inconsistency and duplications. However, since the data sources are heterogeneous, such assumption may not hold. To obtain the high quality information for integration, during source selection, higher quality data source should be selected in higher priority. The problem is how to take the data quality issues into the information integration.

For these challenges, some techniques have been proposed. Yu et al. (2007) considers the inner link between keywords but it is only applied to limited amount of data, and their work does not

* Corresponding author.
E-mail address: wangzh@hit.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.jss.2016.05.026
0164-1212/

evaluate the effect of data source quality. For the evaluation of source quality, many other techniques also did a lot of great work. Dong et al. (2012) studies the static quality-driven data source selection, but it is pretty time-consuming on massive data. And their work does not consider the association between source selection and specific queries and only obtains an static local optimal solution. Rehatsinas et al. takes freshness of data source as the key point without a comprehensive consideration about the quality of data source and the constraint of resources. For the data sources amount reaching around hundreds of millions, the efficiency is not sufficient.

Unlike previous work, they study source selection in small amount of datasets. We focus on the scalability and efficiency issues, that is, we seek to select data sources on datasets with up to millions of data sources and at the same time the quality of results can be guaranteed. Our technique can obtain a quasi-optimal solution in massive data. We have a comprehensive consideration of the data quality evaluation and the keyword-based selection of relational databases. Further more, we propose an efficient technique to filter the data sources from massive databases combining several factors above. Our system can finally get a ranked source list according to the information value of data source, and it will be proved to have great scalability in massive data and fine flexibility for system.

For the problem above, our main contributions are as follows:

1. We model the data source selection problem as a multi-object optimization problem. Unlike previous work, we consider not only the correlation between the query and sources but also the quality of data sources. In our model, multiple aspects of data quality are included (Section 2)
2. We prove that the source selection problem according to the model is an NP-hard problem. To solve this problem, we propose an approximate algorithm with rigorous theoretical guarantees with ratio bound 2 (Section 3).
3. To achieve high efficiency, we develop an efficient index for data sources and keywords, and a ranking algorithm to obtain ranked data sources according to their comprehensive value (Section 4).
4. To verify the efficiency and effectiveness of proposed model and methods, we conduct extensive experiments. Experimental results show that our model could select data sources of high precision with a given limitation of sources. Our method could select sources among millions of data sources within a few seconds (Section 5).

2. Selection of multiple sources

In this section, we first formally define our problem in Section 2.1 and describe the framework of source selection system in Section 2.2. Then we propose the methods of data source quality estimation and correlation evaluation in Section 2.3 and Section 2.4, respectively.

2.1. Definition

Before we define the problem, we make some assumptions. First, we suppose that the cost of accessing each data source is known. The cost could be that of purchasing the data or data integration operation including data cleaning, resolving conflicts, and mapping heterogeneous data items. For the former case, the cost of data is obtained by data pricing mechanisms. For the latter case, the cost is estimated from historical data.

In ideal condition, we wish to maximize the gain while minimizing the cost. Thus we have following problem definition. In this paper, to support multiple kinds of queries, we define *partial keyword query* as Definition 1.

Definition 1. A *partial keyword query* q is $\{K, \epsilon\}$, where K is a set of keywords and ϵ is a threshold. The results of q on a table is the tuples in table with each one containing the keyword number larger than ϵ.

For data source selection problem, the results of a partial keyword query q are a set of ranked data sources that contain the results of q. And the sources are ranked according to their comprehensive gain value, which will be described in detail in the remainder of this part.

The source selection problem of various types of queries could be converted to that of queries in such form.

- For SQL queries, the query is converted to multiple partial keyword queries, each of which contains one table with constraints. The range constraints is handled by treating each small range as a keyword. In such partial keyword, the threshold is set to be the number of constraint on each table.
- For keyword search on relations Li et al. (2008), the partial keyword query is used to find the data sources that contains the tuples with a large number of keywords and could be used to generate the final results in a high possibility.
- For fuzzy keyword search queries, the keywords are split in q-grams, which are the keywords in the partial keyword query. Thus with the well-defined threshold, the data sources that contain the tuples with keywords similar as those in the queries could be found.

Thus in this paper, we focus on data source selection for partial keyword queries. Before describing the definition of source selection problem, we first define some features of data source formally.

Let $gain_i$, C_i and A_i be the profit, cost and accuracy of i-th data source, and $Rel_{k,i}$ be the correlation with keyword k and i-th source, respectively. To take the above factors into consideration, we denote by G_i the comprehensive gain value of i-th source. Then given a set of data source $\mathcal{S}$, $G(\mathcal{S}) = \sum_{i=1}^{|\mathcal{S}|} G_i$, $C(\mathcal{S}) = \sum_{i=1}^{|\mathcal{S}|} C_i$. Next we define the problem as follows.

Definition 2. Given a data source set $\Omega = S_1, S_2, ..., S_n$, let τ_c be the cost budget. We try to find a subset $\mathcal{S} \subseteq \Omega$ that maximizes $G(\mathcal{S})$ under constraint $C(\mathcal{S}) \leqslant \tau_c$.

Next we make some discussions for the comprehensive gain value G_i of i-th source. On one hand, if the source owns a high accuracy, we still refuse to choose it if its correlation with keyword k is low. In other words, when $Rel_{k,i}$ is 0, G_i turns 0 surely. On the other hand, if the accuracy of a source is low, we also reject it. Here we assume that the correlation with keywords and the accuracy of data source is independent of each other. When we mention correlation, we only consider the relationship between specific query and data sources. And the accuracy of data source is determined by the completeness, consistency and coincidence of the data source in our paper. Moreover, the more $gain_i$ is, the higher G_i will be, which is consistent with our intuition. Based on the discussions above, we write this comprehensive gain value of i-th source G_i as follows:

$$G_i = gain_i \times Rel_{k,i} \times A_i \tag{1}$$

For i-th data source, we require the knowledge of $gain_i$ as input. And in Section 2.3 and Section 2.4 we will describe the computational methods for $Rel_{k,i}$ and A_i, respectively.

2.2. Framework

Our system consists of two components.

The first component determines the value of G_i for each source S_i and omits sources whose cost exceeds the cost budget τ_c. Then it describes an improved greedy-based algorithm to select data