A Framework for Summarizing a Log of OLAP Queries

Julien Aligon  
Université François Rabelais Tours  
Laboratoire d'Informatique  
Email: julien.aligon@etu.univ-tours.fr

Patrick Marcel  
Université François Rabelais Tours  
Laboratoire d'Informatique  
Email: patrick.marcel@univ-tours.fr

Elsa Negre  
Université François Rabelais Tours  
Laboratoire d'Informatique  
Email: elsa.negre@univ-tours.fr

Abstract—Leveraging query logs benefits the users analyzing large data warehouses. But so far nothing exists to allow the user to have concise and usable representation of what is in the log. In this paper, we propose a framework for summarizing OLAP query logs. This framework is based on the idea that a query can summarize another query and that a log can summarize another log. It includes a simple language to declaratively specify a summary, a measure to assess the quality of a summary and an algorithm for automatically computing a good quality summary of a query log.

I. INTRODUCTION

It is becoming accepted that leveraging query log would help the user analyzing large databases or data warehouses [1]. This is particularly relevant in a collaborative context for instance to issue recommendations [2], [3], [4], [5]. But to the best of our knowledge, even the simple problem of providing the user with a concise representation of what is inside a large log has rarely been addressed [1]. Using such a summary, that avoids overwhelming the user, would have many advantages, including:

• helping an administrator to manage and tune the OLAP server if the summary indicates the frequently accessed members,
• allowing a user to have a rough idea of the queries launched by other users,
• assisting the user to perform new analysis sessions, for instance by not redoing what has been done, or to write new queries, by looking for what has been done.

In this paper, we propose to summarize automatically an OLAP query log. In this context, we work with the following assumptions:

1) A log is a (often very large) sequence of queries that must be summarized with a concise representation,
2) The way summaries are obtained should be very flexible (i.e., various relevant summaries may be computed from one query log),
3) In addition, the quality of the summary must be evaluated, in terms of how faithful it is to the log.

To answer point 1, we consider that a log will be summarized by a sequence of queries, i.e., by another (much shorter) log. Our answer to point 1 entails that a query will summarize other queries. For point 2, we propose that summaries are specified in a declarative fashion, i.e., summaries are the results expressions constructed with a dedicated manipulation language. This language allows to transform one or more queries into another query for which its quality as a summary is measured (point 3).

Our contributions include:

• A query manipulation language (called QML) tailored for OLAP queries. This language is composed by binary and unary operators that allow to summarize queries,
• A quality measure based on the classical notions of precision and recall that allows to measure to which extent a query is a good summary of another query,
• A greedy algorithm for automatically constructing, using QML, a summary of a query log w.r.t. a quality threshold given by the user.

This paper is organized as follows. The next section motivates the approach with a toy example. In Section III, we propose a method to summarize queries by introducing the language for manipulating queries and in Section IV we introduce a quality measure for the summarized queries. In Section V we present the algorithm that automatically computes a summary of a query log. Section VI briefly discusses related work. We conclude and draw perspectives in Section VII.

II. MOTIVATION

In this section, we illustrate with a toy example our approach for summarizing a log of OLAP queries. The context of this example is that of a user navigating a data warehouse. In our example, the data warehouse records sales of beverages in different locations at different time. The dimensions of this data warehouse are given Figure 1. Consider a sequence of queries $s = (q_1, q_2, q_3)$ where $q_1$ is the first query launched, $q_2$ the second one and $q_3$ the last one. Suppose these queries are logged in a log $L$ and ask respectively for:

1) The sales of Pepsi and Coke for July 2008, in cities Paris or Marseille,
2) The sales of Coke for July 2008, in regions North or South,
3) The sales of Orangina for the second semester 2008, in regions North or South.

Assume we want to summarize these queries by another query. Various solutions are possible. First, we can summarize
the queries by retaining for each dimension the most frequent members. In that case, the resulting query would ask for sales of Coke in regions North or South during July 2008.

A second alternative would be to summarize the queries with another query having for each dimension the members that cover all members present in the initial queries. For example, note that 'Pepsi', 'Coke' and 'Orangina' are sodas, cities 'Paris' and 'Marseille' and regions 'North' and 'South' are in France and all three queries concern year '2008'. The query summarizing the log L would then ask for the sales of soda in France in 2008.

Finally, note that so far, we have illustrated the problem of summarizing queries by another query having for each dimension the members that cover all members present in the initial queries. For instance, consider again the queries 1 and 2 as in Fig. 1.

An n-dimensional cube $C = \langle D_1, ..., D_n, F \rangle$ is defined as the classical $n + 1$ relation instances of a star schema, with one relation instance for each of the n dimensions $D_i$ and one relation instance for the fact table $F$. For a dimension $D_i$ having schema $S = \{ L_1^i, ..., L_{d_i}^i \}$, a member $m$ is any constant in $\bigcup_{L_j^i \in S} \Pi_{L_j^i}(D_i)$. For a dimension $D_i$, we consider that members are arranged into a hierarchy $\langle i \rangle$ and we note $m <_{\langle i \rangle} m'$ the fact that the member $m'$ is the ancestor of $m$ in this hierarchy.

Fig. 1. Dimensions of our toy example

Given such a cube, a cell reference (or reference for short) is an n-tuple $(m_1, ..., m_n)$ where $m_i$ is a member of dimension $D_i$, $\forall i \in [1, n]$. We define multidimensional queries as sets of references that can be expressed as Cartesian products of multisets. The reason for having multisets is to be able to define operators that count members’ occurrences.

**Definition 3.1:** Given an n-dimensional cube $C = \langle D_1, ..., D_n, F \rangle$, let $R_i$ be a multiset of members of dimension $D_i, \forall i \in [1, n]$. A query $q$ over $C$ is the multiset of references $q = R_1 \times \cdots \times R_n$.

In what follows, we assume an n-dimensional cube $C = \langle D_1, ..., D_n, F \rangle$. In the following definitions, $i$ ranges from 1 to n. For a query $q$, $m_i(q)$ denotes its multiset of members in dimension $D_i$. A query $q$ will be noted $\times_i m_i(q)$, and multiset $m_i(q)$ will be noted $\langle S_i, f_i \rangle$, where $S_i$ is a set and $f_i$ is a function giving the occurrences of each element of $S_i$.

**Example 3.1:** Consider the three queries $q_1$, $q_2$ and $q_3$ of the toy example described in the previous section. $q_2$ can be expressed in the MDX query language:

```sql
SELECT
{[Drink],[DrinkAll],[Soda],[Pepsi],
[Drink],[DrinkAll],[Soda],[Coke]}
ON COLUMNS
CROSS ON
{[Country],[CountryAll],[France],[N orth],[Paris],
[Country],[CountryAll],[France],[South],[Marseille],
{[Date],[DateAll],[2008],[S2-08],[July08]})
ON ROWS
FROM Cube
```

We have $m_1(q_1) = \{ \text{Pepsi, Coke} \}$, $m_2(q_1) = \{ \text{July08} \}$, $m_3(q_1) = \{ \text{Paris, Marseille} \}$. This query is the set of references: $q_1 = \{ \text{Pepsi, Coke} \} \times \{ \text{July08} \} \times \{ \text{Paris, Marseille} \}$. The set of references for $q_2$ and $q_3$ are:

- $q_2 = \{ \text{Coke} \} \times \{ \text{July08} \} \times \{ \text{North, South} \}$
• \(q_i = \{\text{Orangina}\} \times \{\text{S2-08}\} \times \{\text{North, South}\}\)

The language we propose is composed by unary operators and binary operators that manipulate queries and output a query, that is called a summary query (or simply summary for short). The main idea behind the definition of these operators is that they operate dimension-wise: They construct a new query from the one(s) in parameter by treating each dimension independently. We now present formally these operators, starting with the binary operators.

### B. The binary operators of QML

The first operators are the classical bag operators [6] extended to multiple dimensions.

**Definition 3.2:** (Bag operators) Given two queries \(q_A\) and \(q_B\) and \(\oplus \in \{\cup, \cap, \setminus, \setminus\}\), \(q_A \oplus q_B = x_i(m_i(q_A) \oplus m_i(q_B))\).

**Example 3.2:** Consider the first two queries of Example 3.1, we have:

\[q_1 = q_B \cup Bq = \{\text{Pepsi, Coke, Coke}\} \times \{\text{July 08, July 08}\} \times \{\text{Paris, Marseille, North, South}\}\]

\[q_2 = q_B \cap Bq = \{\text{Coke}\} \times \{\text{July 08}\} \times \emptyset = \emptyset\]

The next operators give priority to one query over the other.

**Definition 3.3:** (Priority operators) Given two queries \(q_1\) and \(q_2\), \(q_1 \prec q_2 = q_1\).

### C. The unary operators of QML

Recall that a query can be seen as a Cartesian products of multisets. Our first operator outputs, for a query \(q\) in parameter, a query for which only the most frequent members of \(q\) in each dimension are retained.

**Definition 3.4:** (Most frequent operator) Let \(q\) be a query with \(m_i(q) = (S_i, f_i)\) for all \(i\). The most frequent query \(q' = x_i(S'_i = \{m \in S_i | \exists m' \in S_i, f_i(m') > f_i(m)\}, f_{i,x}\) (\(f_{i,x}\) denotes the restriction of a function \(f_i\) to the set \(X\))

**Example 3.3:** \(\text{mostfreq}(q_1) = \{\text{Coke, Coke}\} \times \{\text{July 08, July 08}\} \times \{\text{Paris, Marseille, North, South}\}\)

Our second operator outputs, for a query \(q\) in parameter, a query for which only the most general members of \(q\) in each dimension are retained, w.r.t. the hierarchy of the dimension.

**Definition 3.5:** (Max operator) Let \(q\) be a query. \(\max(q) = x_i(S'_i = \{m \in m_i(q) | \exists m' \in m_i(q), m < m', f_{i,x}\})\)

**Example 3.4:** \(\max(q_1) = \{\text{Pepsi, Coke, Coke}\} \times \{\text{July 08, July 08}\} \times \{\text{North, South}\}\)

Our next operator outputs, for a query \(q\) in parameter, a query for which only the lowest common ancestors of the members of \(q\) in each dimension are retained, w.r.t. the hierarchy of the dimension.

**Definition 3.6:** (lca operator) Let \(q\) be a query. \(lca\. (m)\) be the function that outputs, for a given set of members \(M\) in dimension \(D_i\), their common ancestor w.r.t. \(<_i\), i.e., \(\{m \in D_i | \forall m' \in M, m' <_i m \land \exists m' <_i m', m' <_i m' m'' <_i m\}\). \(lca(q) = x_i lca(m_i(q))\).

**Example 3.5:** \(lca(q_1) = \{\text{Soda}\} \times \{\text{S2-08}\}\times \{\text{France}\}\)

### D. Expression of various summaries

We now briefly illustrate how QML can be used. For instance, consider a log \(L\) composed by 3 queries: \(L = (q_1, q_2, q_3)\). This log can be summarized by the query \(q_1^3\) that retain only the references that appear in all queries, i.e., \(q_1^3 = q_1 \cap q_2 \cap q_3\). Alternatively, \(L\) can be summarized by taking into account the frequency of the members used in the log: \(q_1^3 = \text{mostfreq}(q_1 \cup q_2 \cup q_3)\). Finally, \(L\) can be summarized by a query roughly indicating the parts of the cube that were explored: \(q_1^3 = lca(q_1 \cup q_2 \cup q_3)\). We illustrate these possibilities on our running example.

**Example 3.6:** Summarizing by retaining the common references of all queries, we obtain: \(q_1^3 = q_1 \cap q_2 \cap q_3 = \emptyset\). Summarizing by frequencies on these queries, we obtain: \(q_1^3 = \text{mostfreq}(q_1 \cup q_2 \cup q_3) = \{\text{Coke, Coke}\} \times \{\text{July 08, July 08}\} \times \{\text{North, North, South, South}\}\). Summarizing by lca, we obtain: \(q_1^3 = lca(q_1 \cup q_2 \cup q_3) = \{\text{Soda}\} \times \{2008\} \times \{\text{France}\}\).

In the following section, we introduce our measure for assessing the quality of summaries expressed with QML.

### IV. Quality Measure

In this section, we propose a measure that evaluates to which extent a query is a good summary of some other queries. The measure is based on the classical notion of precision and recall. Note that the operators of QML define summaries by adding or deleting references to their operands. It is the case, for instance, of the lca operator which summarizes by adding ancestors.

For instance, in Example 3.5, the added references in the summary are: \(A = \{\text{Soda}\} \times \{\text{S2-08}\} \times \{\text{France}\}\). And in that case, the deleted references in the summary are: \(D = \{\text{Coke, Pepsi}\} \times \{\text{July 08}\} \times \{\text{Paris, Marseille, North, South}\}\). The intuition behind this example is that we can consider this summary as a good summary with a good quality. However, if we apply the classical precision and recall measures to evaluate its quality, both are null (since \(A \cap D = \emptyset\)).

For a given set of queries, our idea of good summary is that the more references of the queries it has, the better quality it achieves. But additions or deletions of references should not decrease the quality if these additions cover, at best, the coverage of deleted references (for the references at the most detailed level w.r.t. the hierarchies of the dimensions). In that sense, it would be preferable that the coverage of additions introduces few references not in the coverage of the deleted references. We now formalize these notions.

**Definition 4.1:** (Coverage) A reference \(r\) covers another \(r'\) if \(r = (m_1, ..., m_n), r' = (m'_1, ..., m'_n)\) and \(\forall i \in [1, n], m_i > m'_i \lor m_i = m'_i\).

For a set of references, \(\text{cover}(R) = \{f \in \Pi_{\Pi(D)}(D_1) \times \Pi_{\Pi(D)}(D_2) \times ... \times \Pi_{\Pi(D)}(D_n) | \exists r \in R, r \text{ covers } f\}\)

In Figure 2, we note \(L\) the set of the references of some queries to be summarized, \(S\) the set of references of the summary and \(K = L \cap S\). The shaded parts are the coverages of \(D\) and \(A\) respectively, for the most detailed references.
For instance, consider Example 3.5. \( L = q_1 \cup q_2 \), \( S = \text{lca}(q_1) \) and \( L \cap S = \emptyset \) thus both \( A \) and \( D \neq \emptyset \). So, we compare \( \text{cover}(A) \) and \( \text{cover}(D) \).

\[
\text{cover}(A) = \{ \text{Peps}, \text{Coke}, \text{Orangina} \} \times \{ \text{July} 08, \text{August} 08, \text{September} 08, \text{October} 08, \text{November} 08, \text{December} 08 \}
\]

and \( \text{cover}(D) = \{ \text{Blais, Marseil} \} \times \{ \text{July} 08 \} \times \{ \text{Paris, Marseil} \} \cup \{ \text{Coke} \} \times \{ \text{July} 08 \} \times \{ \text{Blais} \} \) and \( |\text{cover}(A)| = 54.\) \( |\text{cover}(D)| = 2\times \frac{54}{54 + 1} = \frac{54}{55} \). It returns a real in \([0, 1]\).

So far, we have introduced a way to express summaries of query sets. In the following section, we present our approach for summarizing a query log, i.e., a sequence of queries.

V. HOW TO SUMMARIZE A LOG?

In this section, we present an algorithm for summarizing a log, based on QML and our quality measure. A log is a finite sequence of queries. The main idea is that a summary of a log is also a log, which is computed w.r.t. a quality threshold given by the user.

Note that, if we summarize a set of queries by one query, summary operators can be combined under the form: \( u(q_1, b q_2) \) where \( u \) is a unary operator, \( b \) is a binary operator and \( q_1, q_2 \) are two queries of the set to be summarized. The idea is to apply this form in a greedy algorithm (Algorithm Summarize detailed Figures 4 to 6). The algorithm iterates until the quality of the summary falls under a threshold or the summary stops changing. In a first part, the algorithm tests each couple of consecutive queries with a binary operator (line 6 to 17). It takes the best summary w.r.t. the quality measure (line 10 to 15). Then, with this summary query, the algorithm tests each unary operators and the best output is kept (line 19 to 26). Note that the quality is measured w.r.t. the initial query couple and not w.r.t. the output of the binary operator.

Figures 5 and 6 detail the algorithm of the quality measure. Note that the size of coverages are computed without actually computing the coverages. Indeed, we assume that the number of covered members is kept for all members of the dimension tables.

Figure 3 illustrates the algorithm with the following parameters: \( L = \{ q_1, q_2, q_3 \} \) with \( q_1, q_2, q_3 \) the three queries of the toy example, \( L \) and \( B \) are respectively the sets of unary and binary operators of QML, the dimensions of the toy example and the quality threshold of 0.7. All binary operators are tested on each couple of consecutive queries. The couple \( \{ q_1, q_2 \} \) is selected because the quality measure is the highest for the union operator. Then, with the result of binary operator of couple \( \{ q_1, q_2 \} \), all unary operators are tested. \( \max \) operator is selected. Finally, \( \{ q_1, q_2 \} \) can be summarized by \( q_3 \). The algorithm tries to summarize with the new sequence of queries \( \{ q_3, q_4 \} \) following the same principle. As the computed qualities are all below the threshold, the algorithm stops with result \( \{ q_3, q_4 \} \).
Summarization of structured data has attracted a lot of attention in various domains, covering web server logs [7], pattern mining (see e.g., [8] that includes a brief survey), sequences of event [9], database [10], multidimensional data stream [11], and datacubes [12].

Many of these works rely on fuzzy set theory ([7], [10]) and/or are compression techniques for which it is important that original data can be regenerated ([12], [8]). Moreover, it can be the case that the summary has not the same type as the data it summarizes. In the domain of database ([10], [11], [12]), summarizing is applied to the database instance where, for OLAP data, measure values are taken into account.

In this paper, we address the problem of summarizing an OLAP server query log. Our approach has the following characteristics:

- We do not rely on fuzzy set theory for summarizing. Instead, we leverage the hierarchies described in the dimension tables.
- The type of the summary is the same as the type of the summarized data.
- We do not address the problem of regenerating the summarized data from the summary.
- We do not summarize a database instance, but database queries.

To the best of our knowledge, no work has yet addressed the problem of summarizing a database query log in a suitable and concise representation. As pointed out in [1], many systems provide query logging, primarily for physical tuning, and allow users to view the log. But the way log is displayed, often in flat tables or files, is not suitable for browsing. In our earlier work [13], we propose to organize an OLAP query log under the form of a website. But if the log is large, browsing this website may be tedious. An effective log visualization and browsing tool is yet to be designed, and the present work is a step in that direction.

VII. CONCLUSION AND PERSPECTIVES

In this paper, we propose a framework for summarizing OLAP query logs. This framework is based on the idea that a query can summarize another query and that a log can summarize another log. Our contributions include a query manipulation language that allows to declaratively specify a summary, a measure to assess the quality of a summary and an algorithm for automatically computing a good quality summary of a query log.

Future work includes the implementation of our approach as well as various tests to assess its efficiency and effectiveness to support on-line analysis. We will also study its extension to a collaborative context where a log, composed by many sequences of queries performed by different users, can be efficiently browsed and searched.

REFERENCES

Fig. 4. Algorithm Summarize($L, U, B, D_1, \ldots, D_n, \alpha$)
INPUT:
refi, refs: set of references,
D1, ..., Dn: the dimensions

OUTPUT: The quality of ref with refi.

VARIABLES:
D, A: a set of references,
K, CD, CA, ∩, r, p: integer,
1, 2: float

1: K ← |refi ∩ refs|
2: D ← refi - refs
3: A ← refs - refi
4: D ← max(D) \{function max() returning a set of reference from D covering every references of D\}
5: A ← max(A)
6: CD ← CardinalityOfCoverage(D)
7: CA ← CardinalityOfCoverage(A)
8: Cov∩ ← 0
9: M∩ ← 0
10: C∩ ← 0
11: for each reference rD of D do
12: for each reference rA of A do
13: for each dimension d of D1, ..., Dn do
14: for each member mD of rD do
15: if mD < d mA then
16: Cov∩ = Cov∩ + |cover(mA)|
17: else \{mD ≤ d mA\}
18: Cov∩ = Cov∩ + |cover(mD)|
19: end if
20: end for
21: end for
22: M∩ = M∩ × Cov∩
23: end for
24: end for
25: Cov = Cov + M∩
26: C∩ = C∩ + Cov∩
27: end for
28: r ← K + CD
29: p ← K + CA
30: return 2 × p × r

Fig. 5. Algorithm Quality(refi, refs, D1,..., Dn)

INPUT:
Sr: set of references,
D1, ..., Dn: the dimensions

OUTPUT: |cover(Sr)|.

VARIABLES:
r: a reference, d: a dimension, m: a member,
1: Cov ← 0
2: T ← 0
3: C ← 0;
4: for each reference r of Sr do
5: for each dimension d of D1, ..., Dn do
6: for each member m of r do
7: Cov = Cov + |cover(m)|
8: end for
9: T = T × Cov
10: end for
11: C = C + T
12: end for
13: return C

Fig. 6. Algorithm CardinalityOfCoverage(Sr, D1,..., Dn)