Query Recommendations for OLAP Discovery-Driven Analysis

Arnaud Giacometti, Université François Rabelais Tours, Laboratoire d’Informatique, France
Patrick Marcel, Université François Rabelais Tours, Laboratoire d’Informatique, France
Elsa Negre, Université François Rabelais Tours, Laboratoire d’Informatique, France
Arnaud Soulet, Université François Rabelais Tours, Laboratoire d’Informatique, France

ABSTRACT

Recommending database queries is an emerging and promising field of research and is of particular interest in the domain of OLAP systems, where the user is left with the tedious process of navigating large datacubes. In this paper, the authors present a framework for a recommender system for OLAP users that leverages former users’ investigations to enhance discovery-driven analysis. This framework recommends the discoveries detected in former sessions that investigated the same unexpected data as the current session. This task is accomplished by (1) analysing the query log to discover pairs of cells at various levels of detail for which the measure values differ significantly, and (2) analysing a current query to detect if a particular pair of cells for which the measure values differ significantly can be related to what is discovered in the log. This framework is implemented in a system that uses the open source Mondrian server and recommends MDX queries. Preliminary experiments were conducted to assess the quality of the recommendations in terms of precision and recall, as well as the efficiency of their on-line computation.

Keywords: Collaborative Filtering, Discovery-Driven Analysis, MDX Queries, OLAP, Query Recommendation

INTRODUCTION

One of the goals of recommender systems is to help users navigating large amounts of data. Existing recommender systems are usually categorized into content-based methods and collaborative filtering methods (Adomavicius et al., 2005). Content-based methods recommend to the user items similar to the ones that interested him in the past, whereas collaborative filtering methods recommend to the user items that interested similar users.

Applying recommendation technology to database, especially for recommending queries, is an emerging and promising topic (Khousainova et al., 2009; Chatzopoulou et al., 2009; Stefanidis et al., 2009). It is of particular relevance to the domain of multidimensional databases, where OLAP analysis is inherently tedious since the user has to navigate large datacubes to find valuable information, often having no idea on what her forthcoming queries...
should be. This is often the case in discovery-driven analysis (Sarawagi et al., 1998) where the user investigates a particular surprising drop or increase in the data.

In our earlier works (Giacometti et al., 2008, 2009a) we proposed to adapt techniques stemming from collaborative filtering to recommend OLAP queries to the user. The basic idea is to compute a similarity between the current user’s sequence of queries (a session) and the former sequences of queries logged by the server. In these works, similarity between sessions is only based on the query text, irrespective of the query results. In this present article, to take into consideration what the users were looking for, we leverage query results to compute recommendations. Our approach is inspired by what is done in web search and e-commerce applications (Parikh et al., 2008) where inferred properties of former sessions are used to support the current session.

The present work improves on Giacometti et al. (2009b), where we proposed a framework tailored for recommending queries in the context of discovery driven analysis of OLAP cubes. The basic idea is to infer, for every former session on the OLAP system, what the user was investigating. As it is the case in discovery-driven analysis, this has the form of a pair of cells showing a significant unexpected difference in the data. We proposed a framework for detecting in the log of an OLAP server such pairs, arranging them into a specialisation relation, and recording per session the queries at various levels of detail that contain the pairs detected. During subsequent analyses, if a difference is found that was investigated in a former session, then the discoveries of this former session are suggested to the current user.

The goal of the present paper is to demonstrate the validity of this approach for recommending query in the particular context of discovery-driven analysis of OLAP cubes. To this end, we extend the work of Giacometti et al. (2009b) in the following ways: First the framework has been slightly changed to better take into account sessions investigating the same difference pair. This means that discoveries are no more recorded only for a particular session but can span across sessions. Second, the framework has been implemented and we undertook a few experiments to assess the effectiveness and the efficiency of our approach. Finally, we propose a dedicated architecture for implementing the approach beyond a prototypical setting.

This paper is organized as follows. The section discusses our approach with a simple yet realistic example. The third section reviews related work. Preliminary definitions on OLAP data model and query model are recalled in the fourth section. The framework of our recommender system is formally presented in the fifth section, and the algorithms are presented in the sixth section. In these sections, the example given in the second section is used as a running example to illustrate the framework. The seventh section introduces our prototypical implementation of the framework, and the eighth section presents some preliminary experiments. Finally, before concluding, we briefly discuss the feasibility of our approach in a real context and propose an architecture thereof.

MOTIVATING EXAMPLE

In this section we illustrate our approach with an artificial yet realistic motivating example that will be used as a running example throughout this paper. This example uses typical discovery-driven analysis sessions of a simple datacube containing sale results of various products in various locations at different times. These sessions are sequences of queries, the result of which are depicted in Figure 1. On the left are three sessions from the log of the OLAP server and on the right is a current session. This current session consists of only one query (q) that asks for the aggregated sales of cheese in 2007 and 2008 in Europe and USA. The current user may wonder where to navigate the cube further. We will now show how the information in the log can be exploited to provide her with some suggestions.

Let us first describe the sessions contained in the log of the OLAP server. In what follows,
Related Content

Advanced Dimensionality Reduction Method for Big Data
[www.igi-global.com/chapter/advanced-dimensionality-reduction-method-for-big-data/150270?camid=4v1a](www.igi-global.com/chapter/advanced-dimensionality-reduction-method-for-big-data/150270?camid=4v1a)

Multidimensional Business Benchmarking Analysis on Data Warehouses
[www.igi-global.com/article/multidimensional-business-benchmarking-analysis-on-data-warehouses/173706?camid=4v1a](www.igi-global.com/article/multidimensional-business-benchmarking-analysis-on-data-warehouses/173706?camid=4v1a)
Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Li and Deren Li (2013). *Developments in Data Extraction, Management, and Analysis* (pp. 303-324).
www.igi-global.com/chapter/data-field-hierarchical-clustering/70803?camid=4v1a

An UML Profile and SOLAP Datacubes Multidimensional Schemas Transformation Process for Datacubes Risk-Aware Design
www.igi-global.com/article/an-uml-profile-and-solap-datacubes-multidimensional-schemas-transformation-process-for-datacubes-risk-aware-design/130667?camid=4v1a