Chapter 5
C-Idea:
A Fast Algorithm for Computing Emerging Closed Datacubes

Mickaël Martin-Nevot
Aix-Marseille Université, France

Sébastien Nedjar
Aix-Marseille Université, France

Lotfi Lakhal
Aix-Marseille Université, France

Rosine Cicchetti
Aix-Marseille Université, France

ABSTRACT

Discovering trend reversals between two data cubes provides users with novel and interesting knowledge when the real-world context fluctuates: What is new? Which trends appear or emerge? With the concept of emerging cube, the authors capture such trend reversals by enforcing an emergence constraint. In a big data context, trend reversal predictions promote a just-in-time reaction to these strategic phenomena. In addition to prediction, a business intelligence approach aids to understand observed phenomena origins. In order to exhibit them, the proposal must be as fast as possible, without redundancy but with ideally an incremental computation. Moreover, the authors propose an algorithm called C-Idea to compute reduced and lossless representations of the emerging cube by using the concept of cube closure. This approach aims to improve efficiency and scalability while preserving integration capability. The C-Idea algorithm works à la Buc and takes the specific features of emerging cubes into account. The proposals are validated by various experiments for which we measure the size of representations.

DOI: 10.4018/978-1-5225-4963-5.ch005

1. INTRODUCTION AND MOTIVATIONS

Decision makers are generally interested in discovering relevant trends by using a data warehouse to analyze data collected from a "population". The data warehouse contains data concerning various measures which are observed with respect to different attributes called dimensions. More precisely, all the possible combinations of dimensions can be relevant and considered at all possible granularity levels. In order to meet this need, the concept of data cube was introduced (Gray et al., 1997). It groups the tuples according to all the dimension combinations along with their associated measures. The main interest of this structure is to support an interactive analysis of data because all the possible trends are yet computed. Of course, due to its very nature (the very great volume of original data and the exponential number of dimension combinations), a data cube is especially voluminous.

Let us assume that we have a data cube costly computed from a set of data accumulated until now in a data warehouse. Let us imagine that a refreshment operation has to be performed in order to insert new collected data. A particularly interesting knowledge can be exhibited from the comparison between the cubes of these two data sets: which novelties does the refreshment bring? which trends, unknown until now, appear? or in contrast, which existing trends disappear? Similar knowledge can be exhibited every time that two semantically comparable data cubes have to be compared. For instance, if two data sets are collected in two different geographical areas or for two population samples, it is possible to highlight the behavior modifications, the contrast between their characteristics or the deviations with respect to a witness sample.

In order to capture trend reversals in data warehouses, we have proposed the concept of Emerg- ing Cube (Nedjar et al., 2013). It results from coupling two interesting structures: the data cube (Gray et al., 1997) and the emerging patterns (Dong & Li, 2005, 1999). From the cube of two database relations, the Emerging Cube gathers all the tuples satisfying a twofold emergence constraint: the value of their measure is weak in a relation (C_1 constraint) and significant in the other relation (C_2 constraint). Computing an Emerging Cube is a difficult problem because two data cubes have to be computed and then compared. As above-mentioned, the computation of the cubes is costly and their comparison has likely a significant cost because their size is really tremen- dous. Then, to really take advantage of the new knowledge captured by the Emerging Cube, it is critical to avoid the computation of the two data cubes.

Although the Emerging Cube limits the results to the ones potentially relevant, its size remains enormous in part because it encompasses a lot of redundancies. In order to discard such superfluous information, we propose the Emerging Closed