International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182 Print ISSN: 2349-5979
www.allsubjectjournal.com
Volume 3; Issue 2; February 2016; Page No. 209-210

An introduction to information extraction

Naveen Dahiya

MSIT, New Delhi, India.

Abstract
The data warehouse development process follows an incremental approach starting from requirement gathering, design, quality evaluation and information extraction. This paper discuss the basic introduction to information extraction phase of data warehouse development and the qualities of an efficient information extraction system.

Keywords: Data warehouse, Information extraction, requirement gathering, quality evaluation.

1. Introduction

Organisations nowadays are employing efficient information delivery systems for prediction of futuristic trends about sale of products, customer buying patterns, customer needs and preferences, product quality. An efficient information delivery system helps organisations to face fierce market competition and gain competitive advantage. Efficient information delivery system must give maximum useful information in minimum possible time in response to the queries thrown on it. Query response time is one of the major factors affecting the quality of efficient information extraction from an information delivery system. There exist a number of query optimization techniques to get efficient information from a data warehouse as fast as possible. The next section presents a literature review of few query optimization techniques for efficient information extraction from an information delivery system.

2. Literature Review

The section gives a literature review conducted towards efficient information extraction from data warehouses.

Intelligent decision Support Systems (Harinarayanan *et al*. 1996 [1]. Aldea *et al*. 2012) [2] are supported by data warehouses at beck end. Complex queries are thrown on data warehouses to get results in response to the queries. Query response time is one of the major factors affecting the quality of information delivery systems. Query optimization is further dependent on optimal selection of materialized views (Bellahsene *et al*. 2010) [3] for which several view selection techniques have been proposed and implemented (Harinarayan *et al*. 1996 [1]. Dhote and Ali 2009 [4]. Halevy 2010) [5].

The basic view selection technique using greedy approach was proposed by Harinarayan *et al*, 1996 [1]. He discussed lattice framework, cost model, benefit metric and greedy approach for materialized view selection.

A pick aggregates algorithm for optimal view selection was proposed by Shukla *et al*. 2000. The algorithm selects aggregate views based on pre computed benefits using greedy approach. Many researchers have used A* algorithm (Nilsson, 1971) [6] based approach to materialize view indexes.

Dhote and Ali, (2009) [4] discuss and analyze the application of various methodologies used for materialized view selection in information delivery systems.

Miami and Bellahsene, 2012 [7] categorized materialized view selection along various dimensions like: frameworks and resource constraints and categorized algorithms used for view selection as: deterministic algorithms, randomized algorithms, hybrid algorithms or constraint programming. These algorithms differ in their approach to solve materialized view selection problem and have different time and space complexities.

3. Introduction to basic terms

The section introduces few basic terms associated with information extraction defined as follows:

- **View**: A result in response to a query. It is defined in terms of base relation and/or combination of attributes (Miami and Bellahsene, 2012) [7].
- **Materialized View**: A view is materialized if its result in response to query is stored in memory (Miami and Bellahsene, 2012). Optimal selection of materialized views improves query optimization.
- **View Selection**: The process of selecting materialized views from a database to optimize query response time (Miami and Bellahsene, 2012) [7].
- **Data Cubes**: The data in a data warehouse is stored along multiple dimensions. The graphical representation of multidimensional analysis is represented as data cubes (Miami and Bellahsene, 2012) [7].
- **Lattice**: A Lattice (Shukla *et al*., 2000) [8] is a graphical representation to show dependencies among multiple views of a multi-dimensional database.

Consider the example of a business data warehouse that analyse sales along three dimensions/attributes part (p), supplier(s), customer(c). There are 3 dimensions and likewise $8(2^3=8)$ possible grouping of dimensions.

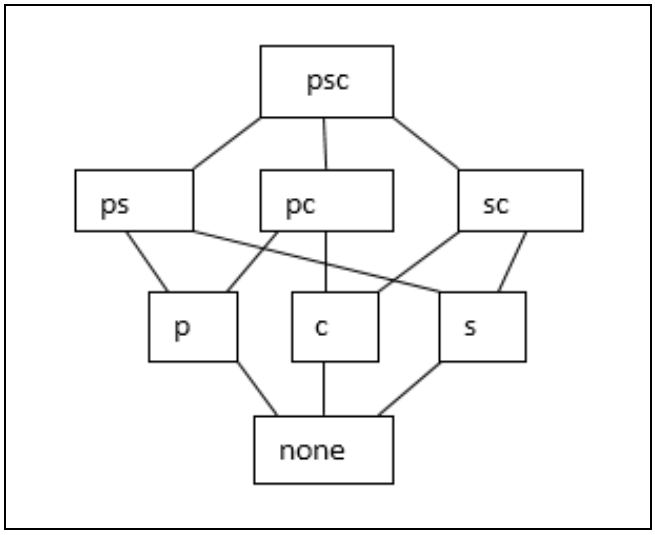

Fig 1: Lattice Framework (Source: Harinarayan *et al*., 1996) [1]

Lattice framework for business data warehouse is shown by Figure 1. The lines connecting the boxes show dependencies. The three lines from psc to ps, pc, sc show that views ps, pc, sc can be generated from psc or are dependent on psc. The same holds for other lines in the lattice framework.

The use of lattice framework for materialized view selection offers following advantages to users:

- The lattice framework gives a user friendly graphical view of the multidimensional data warehouse.
- Data dependencies and hierarchies can be visualized and understood at a glance.
- The users can analyze data quickly and efficiently.

4. Optimal Selection of Materialized Views

For making an optimal selection of views from a database to improve query response time for efficient information extraction, some aspects must be given due consideration (Harinarayan *et al*., 1996) [1]:

- The time taken to respond to a query should be as small as possible.
- The second aspect deals with optimal selection of fixed number of materialized views with no space constraint.
- The third aspect deals with optimal selection of fixed number of materialized views with limited space.

5. Conclusion

The paper presents an introduction to efficient information extraction from an information delivery system. Efficient information extraction depends on query optimization which in turn depends on optimal selection of materialized views. The paper gives a basic understanding of the phenomena of efficient information extraction to novice users.

The future work can be focused on designing newer techniques for optimal view selection and query optimization. The newer techniques can be implemented to real world applications to prove their validity and significance.

6. References

1. Harinarayan V, Rajaraman A, Ulmann J. Implementing Data Cubes efficiently. SIGMOD Conference, 1996, 205-16.
2. Aldea A, Alcantara R, Skrzypczak S. Managing information to support the decision making process. JIKM World scientific publishers, 2012, 11(3).
3. Bellahsene Z, Cart M, Kadi N. A cooperative approach to view selection and placement in P2P systems. OTM, 2010, 515-22.
4. Dhote C, Ali M. Materialized view selection in Data Warehousing: A Survey. Journal of Applied Sciences. 2009, 401-14.
5. Halevy A. Answering queries using views: A survey. VLDB Journal. 2001; 10(4):270-94.
6. Nilsson J. Problem solving methods in artificial intelligence. McGraw–Hill publishing company Ltd, 1971.
7. Miami I, Bellahsene Z. A survey of view selection methods. SIGMOD Record, 2012; 41(1):20-29.
8. Shukla A, Deshpande P, Naughton J. Materialized view selection for multi-cube data models. LNCS, 1777, Springer Verlag, 2000, 269-84.