A Statistical Approach to Rule Selection in Semantic Query Optimisation

Barry G. T. Lowden and Jerome Robinson

Department of Computer Science, The University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, Essex, United Kingdom

Abstract. Semantic Query Optimisation makes use of the semantic knowledge of a database (rules) to perform query transformation. Rules are normally learned from former queries fired by the user. Over time, however, this can result in the rule set becoming very large thereby degrading the efficiency of the system as a whole. Such a problem is known as the utility problem. This paper seeks to provide a solution to the utility problem through the use of statistical techniques in selecting and maintaining an optimal rule set. Statistical methods have, in fact, been used widely in the field of Knowledge Discovery to identify and measure relationships between attributes. Here we extend the approach to Semantic Query Optimisation using the Chi-square statistical method which is integrated into a prototype query optimiser developed by the authors. We also present a new technique for calculating Chi-square, which is faster and more efficient than the traditional method in this situation.

1 Introduction

Semantic query optimization is the process of transforming a query into an alternative query that is semantically equivalent but executes in a shorter time. A semantically equivalent query returns the same result set as the original query and is generated from it by the application of rules.

Rules can consist of ‘integrity constraints’ that hold for all states of the database or ‘dynamic rules’ that hold only for a given state. Generally, rules can be supplied by the database engineer or derived automatically. Automatic rule generation methods include heuristic based systems [13], logic based systems [1], graph based systems [12] and data driven systems [7, 11].

With the development of automatic rule derivation, it becomes increasingly important to filter out ineffective rules which can lead to a deterioration in optimisation performance. The overall size of the rule set (n) is also of crucial importance since some optimisation algorithms run in $O(n^2)$ time.

The problem of a large rule set degrading the efficiency of the reformulation process is referred to as the utility problem [4, 8, 10, 14]. Earlier work by the authors [6] partially addressed the utility problem by time stamping rules when created,
modified or used by the reformulation process. When the number of rules exceeds a
specified limit, those which are recorded as unused for more than a given period of
time or have a low usage count are removed from the rule set.

To further explore this problem we have examined a range of statistical tests as
applied in knowledge discovery research [5, 8, 2]. In this paper we propose the use of
Chi-square for the analysis and selection of effective rules for query optimisation.

In section 2 we briefly describe a prototype optimiser, developed by the authors,
on which we base our experiments. We then discuss a method of calculating Chi-
square which is particularly appropriate to the present application and in section 4 we
present our computational results.

2 The ARDOR prototype

In most rule derivation systems, the characteristics of worthwhile rules are first
identified and then a query issued to the database in an attempt to derive such rules.
The ARDOR system [7] takes a different approach whereby actual queries to the
system are used to learn new rules. For each query (q) an optimum alternative (q') is
constructed, using a simple inductive process. The equality \( q \Leftrightarrow q' \) is then used to
deduce the minimum set of rules needed to perform the transformation
algorithmically and these rules added to the rule base.

The system is illustrated in Figure 1 and consists of (1) query reformulation, and
(2) the rule derivation and learning modules. The query reformulation process takes
the original query submitted by the user and attempts to reformulate it into a more
cost effective query by matching the conditions of the query to the rule set using a
‘best first’ search strategy. For issues relating to bounded optimality of search
strategies see [9].

By this means, the original query is transformed into a semantically equivalent
query known as a reformulated query using the standard transformation rules of
constraint introduction and removal [3]. This query will give the same result as the
original query but with normally a shorter execution time.

![Fig. 1. Overview of Query Optimiser](image)