The report is written regarding to the following two papers:

1. SQAK Doing More with Keywords

SQAK is developed as a solution to enable ordinary users to query today’ s large and complex enterprise databases, which often relate hundreds of entities, and derive value from them. With SQAK, users can pose aggregate queries using simple keywords with little or no knowledge of the schema.

The architecture for SQAK is straightforward: A keyword query in SQAK is simply a set of words (terms) with at least one of them being an aggregate function (such as count, number, sum, min, or max). Terms in the query may correspond to words in the schema (names of tables or columns) or to data elements in the database.

The SQAK system consists of three major components - the Parser/Analyzer, the SQN-Builder, and the Scorer. The Parser/Analyzer in SQAK parses the query and transforms it into a set of Candidate Interpretations. A Candidate Interpretation can be thought of as an interpretation of the keyword query posed by the user in the context of the schema and the data in the database. An intuitive way of understanding a CI is to think of it as supplying just the SELECT clause of the SQL statement. The SQN Builder takes a CI as input and computes the smallest valid SQN with respect to the CI. A Simple Query Network is a connected subgraph of the schema graph. A valid SQN statement completely specifies the query by supplying the FROM, WHERE, and GROUP BY clauses to the SELECT clause supplied by the CI. By carefully choosing a subset of SQL, which we call reduced SQL, SQAK achieves a judicious tradeoff that allows keyword queries to be translated to aggregate queries in this subset while controlling the amount of ambiguity in the system. The scorer uses the match scores from the parser/analyzer to determine the weights of the nodes. In fact, the minimal SQN problem is NP-Complete. We can provide a brief sketch of the proof: The basic idea of this proof is by reduction from the Exact 3-Cover problem. Therefore we can use greedy backtracking heuristic to solve it. Having found the SQN, next task is to translate it to the corresponding rSQL query. It has been proved that the overhead of translating the keyword query should be small compared to the cost of actually executing the SQL query.

This paper shows that SQAK is a novel approach that allows ordinary users to perform sophisticated queries on complex databases that would not have been possible earlier without detailed knowledge of the schema and SQL skills. It is likely that SQAK will bring vastly enhanced querying abilities to non-experts. The remaining question is whether the techniques may be used to extend SQAK to work over multiple data sources, which is a topic of future investigation.

2. Chabot Retrieval from a Relational Database of Images

The design and construction of Chabot system uses the relational database management system POSTGRES for storing and managing the images and their associated textual data. To implement retrieval from the current collection of 11,643 images, Chabot integrates the use of stored text and other data types with content-based analysis of the images to perform “concept queries”. Concepts should embody textual metadata about the image as well as image feature information. The Chabot project was initiated to replace the existing system with a better system that includes: An advanced relational database for images and data; Large-scale storage for images; On-line browsing and retrieval of images; A flexible, easy-to-use retrieval system; Retrieval of images by content. The schema for Chabot is a collection of technical reports and a video library. To implement concept queries, Chabot uses two capabilities that POSTGRES provides: storage of pre-computed content information about each image (a color histogram) as one of the attributes in the database, and the ability to define functions that can be called at run-time as part of the regular querying mechanism to analyze this stored information. The function “MeetsCriteria” is the underlying mechanism that is used to perform concept queries. It takes two arguments: a color criterion and a color histogram. The method for finding histograms that meet the criterion employs two metrics: compliance and count. The test provided in this paper indicates that the best result is obtained when text-based search criteria are combined with content-based criteria and when a coarse granularity is used for content analysis. Although this paper brings forward a innovatory way of using DBMS that allows us to incorporate image analysis and information retrieval tools into the system, one shortage for this paper is that it fails to give a comprehensive description about the general application of Chabot in industrial area of computer technology.