

Retrieval of Large-Scale Semantic Data: Investigating a Hybrid DB+IR Approach

Hany Azzam

No Institute Given

1 Introduction

Keyword retrieval has emerged and become an established method for efficient retrieval of large-scale data. However, this method of retrieval either depends on specific encoding of available information or on simple full-text analysis. Those two approaches suffer from certain shortcomings. First of all, both completely rely on the input vocabulary of the user and secondly a specific encoding significantly reduces the recall of a query.

Consequently, information retrieval (IR) over semantic data extracted from the Internet or Intranet started to receive an increasing amount of interest in both industry and academia. In particular, explicating the vocabulary through the usage of ontologies and utilising formal query languages such as SQL that describe the semantics and refers to the schema of the underlying data sources became a potential solution for the shortcomings of keyword retrieval.

There are at least three problems with formal queries: Query formulation, efficient retrieval (scalability) and relevance-based ranking. Keyword retrieval easily outperforms in these three criteria, but from point of view of expressiveness, the need to answer "semantic" queries justifies the usage of formal query languages. Thus, I work on the problems facing formal-queries and how I can make use of IR and database-centric techniques to solve these problems.

2 Query Formulation

The first problem facing semantic data retrieval is query formulation. It can be tackled by an efficient query translation layer that is based on probabilistic retrieval models and retrieval strategies that ensure results in optimal time. This layer is based on modelling layers for efficient semantic queries execution that shields how the underlying services, like query processing and optimisation techniques are implemented and executed. The layer performs language-independent translation of semantic queries to lower level languages like probabilistic SQL (PSQL) and conduct implicit selection of probabilistic retrieval strategies for efficient query execution.

3 Efficient Retrieval

Efficient retrieval becomes the second problem facing semantic data retrieval. I classically distinguish between semantic optimisation (SO), algebraic optimisa-

tion (AO), and processing optimisation (PO). SO is concerned with the early detection of empty results and the removal of redundant query parts. AO focuses on finding the optimal algebraic expression, selection pushing and magic set rewriting are some of the standard techniques. PO exploits the underlying indexes, and eventually executes in parallel with the other optimisation techniques.

4 Evaluation

I evaluate the query formulation layer and the optimisation techniques to see their effect on retrieval time of semantic data. System A a database-centric retrieval system capable of performing full analysis of semantic relations during the loading of data. HySpirit, a retrieval framework that provides a database-centric access to large-scale data with the data being stored in files and indexing structures (file-system storage), similar to the inverted list. The analysis of semantic relations is done per-query basis.

The queries used in evaluation are based on the Lehigh University Benchmark, which includes three data sets containing OWL files for 1, 5, and 10 universities respectively, the largest one having over 1,100,000 triples in total. Figure 1 compares the query response time of System A and HySpirit with differ-

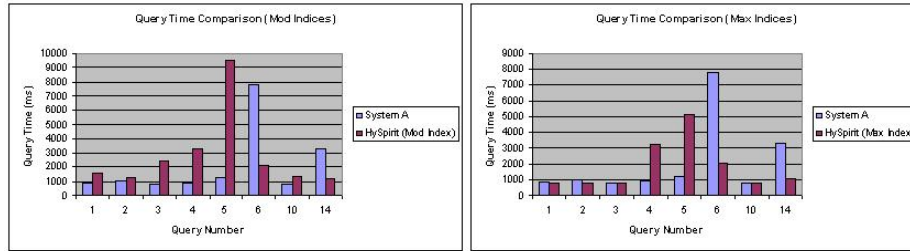


Fig. 1. Query time comparison between System A and the hybrid system

ent indexing schemes being utilised. With the moderate indexes scheme i.e. only selected columns are indexed the overall response time of System A is better. Conversely, in the same figure the response time improves in maximum indexes scheme i.e. all columns are indexed and it becomes substantially better than System A's even with large datasets.

5 Summary

I am investigating the efficiency and scalability challenges facing semantic data retrieval. I started with query optimisation techniques as a potential solution for the retrieval of large-scale semantic data. The next step would be to utilise ranking of semantic data to improve the overall retrieval times.