

ABSTRACT

Recent years, there exist meaningful structured collections that can be exploited in search task. When searching for these structured collections, the expressiveness of structured queries allows structures to be specified at the query layer in order to obtain a more focused and precise search results. However, constructing such queries in an adhoc search environment is difficult as users need to be familiar with the syntax of the query languages. Heterogeneities of structure usages across different collections also hinder users from selecting appropriate structure or concept when writing queries.

In this thesis, we are motivated to automate the construction of these queries from keywords query which are more familiar to any user. The work of query transformation results in two main challenges. First, to propose a generic framework such as it can be easily adapted to changes in structured retrieval environment such as retrieval systems, collections, scoring models. Second, to propose a query interpretation within the framework that will handle structure complexities in collection. Since the usage of markups and structures in current structured collections can be loosely defined, these collections are now richer and more complex in their information structures, especially for text centric collection. Current works have yet to explore into these newly emerging complex structures when capturing knowledge for query interpretation.

In order to address these challenges, a flexible query transformation framework (FQT) is proposed. The flexibility feature is desired such that the framework can cater for various settings of structured retrieval environment e.g. different types of structured collections and structured query interfaces. This framework consists of a novel intermediate query representation that will be the central of the transformation process, i.e. a structure that captures the information needs of query and the syntax of query separately. Its main strength is to allow the transformation to be generic to cater for more than single type

of structure query. Supporting this intermediate query representation are the query interpretation and query construction algorithms. The former uses context-based probabilistic approach for interpreting source query, whereas the latter constructs the interpreted query into an intermediate query. Once a source query is interpreted and represented as intermediate query, it can be easily mapped to a structured query language using a set of predefined query templates in knowledge base.

Lastly, experiments are carried out at the algorithm, application and representation levels on both synthetic and real world data sets to demonstrate the feasibility and scalability of the query transformation framework. The experimental results confirm that our framework is more effective in terms of query interpretation especially dealing with collection with complex structures. The framework is also able to represent various kinds of information needs and structured query languages with its proposed intermediate query representation. Better performance in terms of precision has also been achieved when structured query generated by the framework is applied in structured retrieval task.