Query Exploratiweness for Integrated Search in Heterogeneous Data Sources

Ruxandra Domenig and Klaus R. Dittrich

Department of Information Technology, University of Zurich
{domenig,dittrich}@ifi.unizh.ch

Abstract. We consider query systems which allow imprecise queries and define a new property called *query exploratiweness*. This property characterizes the transformations performed by a system in order to answer imprecise queries, i.e. the system's "work" for mapping input queries into more precise target queries.

1 Introduction

Query systems today, for example Web search engines, had to learn to deal with the rapid increase of volume and diversity of on-line information and continuous changes in the number, content, and location of data sources. Several more or less suitable solutions for these problems have been proposed in recent years. One important property characterizes search engines, namely the simplicity of formulating queries. These are mostly specified by simply typing in keywords and the search engine finds the documents that may fulfill the given information need. We call such queries *fuzzy* queries. However, the price for this simplicity of querying is the quality of answers. WWW search tools often return too many results and, at the same time, results which are relevant to users are not returned.

Modern database management systems (DBMS) do a good job in managing and providing access to a large volume of record-oriented data. They store data together with its structure, which is used when formulating queries. We say that users formulate *exact* queries. The main advantage of this approach is the high quality of answers. In other words, the price for relevant results has to be payed by users through the involved formulation of queries.

These examples are the two extremes for the spectrum of sources which are available today for querying data. Systems which allow fuzzy queries are suitable for new requirements, i.e. huge amount of on-line data and of different users which want to query this data. On the other hand, there are a lot of applications which require specific, exact portions of the available data (for example financial applications), and for this reason, systems which return just relevant results of a query are still needed. In this paper, we characterize the difference between these two types of query systems through a property which we call *query exploratiweness*. A system which allows fuzzy queries needs to find the meaning of such a query, i.e. to find out what users are looking for in terms of an "equivalent" set

A. Banks Pidduck et al. (Eds.): CAiSE 2002, LNCS 2348, pp. 715–718, 2002.

of precise queries. For this reason, query exploravitiveness characterizes the special query transformation steps performed by the system in order to "explore" the query and the available data. The main aim of this extended abstract is to show the importance of systems with high exploravitiveness and the way it can be implemented.

2 Query Exploravitiveness

We consider a query system S which returns a collection of *data units* as an answer to a query and which allows fuzzy queries. This means that users may get results which do not necessary fulfill their information need. Such a system is said to expose *answer fuzziness*. For S and a query q, we use the notation Q_q for the set of data units which are an answer to q and U_q for the set of data units which the system returns for q.

Definition 1. *A query system S exposes answer fuzziness, if*

$$\exists q : (Q_q \cup U_q) \setminus (Q_q \cap U_q) \neq 0.$$

In other words, query systems expose answer fuzziness if exists a query for which the system does not return all data units that are an answer or it returns some data units that are not an answer.

Obviously, a query systems exposing answer fuzziness has to internally perform some processing steps which translate the input query into one or more *target queries*. These transformations characterize what we call *query exploravitiveness*. We call the intermediate queries generated during the transformation process *working queries*.

Definition 2. *We consider a query system S with answer fuzziness. For a query q, query exploravitiveness is the number of working queries generated for answering q.*

Query exploravitiveness and retrieval effectiveness. Query exploravitiveness and retrieval effectiveness[1] are ortogonal to each other, since there are systems which offer a high exploravitiveness, but the retrieval effectiveness is low. Since the purpose of a query system is to offer a high retrieval effectiveness, the main question is when to choose systems with a high query exploravitiveness. In our opinion, a system with a high query exploravitiveness should be used if the users' information need is not well defined, i.e. users need to "explore" for themselves the data available for querying, or to receive hints as to where and how to search in order to fulfill their information need. As we discussed in Section 1, for several applications, query systems today should allow users to formulate queries on the fly,

[1] Retrieval effectiveness "is the ability to retrieve what the user wants to see" [Sch97]. The most well-known measures for effectiveness are *recall* and *precision*. They estimate the percentage of relevant documents which have been retrieved from the given document collection and the percentage of retrieved documents which are relevant.

i.e. they should be able to issue queries without knowing the structure, location or existence of requested data. Consequently, in this case, systems with a high query exploravitiveness are suitable.

Query exploravitiveness and users feedback. A system with a high query exploravitiveness may generate queries that lower the retrieval effectiveness. Since retrieval effectiveness is dependent on a subject, i.e. a user, a good solution for this problem is to involve users in the evaluation of queries, through user feedback. The traditional way of user feedback in an information retrieval system is to evaluate the initial query, present the results, and give users the possibility to classify the results as right or wrong. Then, using the initial query and the user's feedback specification, the system can reevaluate the initial query. This method is known in information retrieval as "relevance feedback". However, for a system with a high query exploravitiveness, it makes sense to offer users the possibility to give their feedback already during the process of query transformation, so they can choose the right working and target queries which should be evaluated.

Query exploravitiveness for integrated heterogeneous query systems. Query exploravitiveness has a different meaning for integrated heterogeneous query systems, i.e. for systems which offer users the possibility to query data from a network of heterogeneous data sources in an unified way. In this case, queries entered into the system are first preprocessed, then transformed as in traditional DBMS and finally they are split into subqueries for the underlying data sources. An integrated query system which offers a high query exploravitiveness implements during query preprocessing algorithms and heuristics which translate the (fuzzy) input query into equivalent target queries. This translation is necessary, since users did not formulate the respective target queries, because they did not take into consideration the structure and query capabilities of the underlying sources. Consequently, one can also say that an integrated query system with high exploravitiveness in a first step searches for queries, instead for searching directly for data.

SINGAPORE, a system for integrated search in heterogeneous data sources. We have developed a system called SINGAPORE (**SING**le **A**ccess **PO**int for heterogeneous data **RE**positories) which can be used for querying data sources which are *structurally heterogeneous*, i.e. they contain structured, semistructured and/or unstructured data. Its main focus is on offering a unified query language, so that retrieving information can be supported by the traditional task of database querying, but also by a more vague or "fuzzy" way to query, as in information retrieval systems. Additionally the system allows on-the-fly registration of new sources, making new information available for querying at run-time.

The query language of SINGAPORE, SOQL, is an extention of OQL [Cat00], with features for unstructured and semistructured data, and for integrating data. For example, we introduced the operator contains, which takes as arguments keywords and searches for them in the data sources specified in the FROM part

of the query. Another example is the operator LIKE, which allows to formulate queries using the structure of the data, even if this is not fully known.

Since the system allows the formulation of fuzzy queries, it also offers a high query exploravtiveness. We have implemented query exploravtiveness during the preprocessing step. When a query is submitted to SINGAPORE, a parse tree is generated and the syntax is checked. The semantical check contains three special steps:

1. Handling of the special operators defined in the query language (for example contains, LIKE).
2. Generation of paths for structural specifications in the query.
3. Generation of consistent queries, eliminating inconsistencies in queries which might have been introduced during the previous steps.

For a detailed presentation of the system and the implementation of the exploravtiveness see [DD00, DD0la, DD0lb].

3 Conclusions

We have identified a new property, query exploravtiveness, which characterizes query systems and which is of high importance for querying large amounts of data in a flexible way, i.e. without knowing the structure, location or existence of requested data. Query exploravtiveness characterizes the transformations performed by the query system in order to "explore" the query and the available data.

Even if there is no direct relationship between query exploravtiveness and retrieval effectiveness, we believe that systems which implement high query exploravtiveness are needed today, since they offer users the ability to formulate queries on the fly, in the sense that they "explore" for themselves the data available for querying, or they receive hints as to where and how to search.

References

[Cat00] R. G. G. Cattell, editor. *The Object Database Standard: ODMG 3.0.* Morgan Kaufmann, 2000. 717

[DD00] Ruxandra Domenig and Klaus R. Dittrich. A query based approach for integrating heterogeneous data sources. *Proe. 9th International Conference on Information and Knowledge Management, Washington, DC,* November 2000. 718

[DD0la] Ruxandra Domenig and Klaus R. Dittrich. Query preprocessing for integrated search in heterogeneous data sources. *Proe. Datenbanksysteme in Buero, Technik und Wissenschaft, Oldenburg, Germany,* 2001. 718

[DD0lb] Ruxandra Domenig and Klaus R. Dittrich. SINGAPORE: A system for querying heterogeneous data sources. *Proc. International Conference on Data Engineering, ICDE (demo session), Heidelberg, Germany,* 2001. 718

[Sch97] Peter Schäuble. *Multimedia Information Retrieval.* Kluwer Academic Publishers, 1997. 716