Sequence Datalog: Declarative String Manipulation in Databases *

Anthony Bonner[1]
bonner@db.toronto.edu

Giansalvatore Mecca[2]
mecca@dis.uniroma1.it

[1] Department of Computer Science – University of Toronto – Toronto, Ont. – Canada
[2] D.I.F.A. – Università della Basilicata – via della Tecnica, 3 – 85100 Potenza, Italy

Abstract. We investigate logic-based query languages for sequence databases, that is, databases in which strings of symbols over a fixed alphabet can occur. We discuss different approaches to querying strings, including Prolog and Datalog with function symbols, and argue that all of them have important limitations. We then present the semantics of *Sequence Datalog*, a logic for querying sequence databases, and show how this language can be used to perform structural recursion over sequences.

1 Introduction

In the last decade, new applications (*e.g.* CAD and CASE) have been a powerful driving force in the development of new database technology. New paradigms have arisen (*e.g.* object oriented databases [4]) that provide greater flexibility than the traditional relational model. However, in some application areas, such as *genome databases* [14] and *text databases* [16], there is still a need for greater flexibility, both in data representation and manipulation. These applications are characterized by a need to store, manipulate and transform strings.

Strings are a particularly interesting domain for query languages. In contrast to sets, computations over strings can easily become infinite, even when the underlying alphabet is finite. This is because symbols are allowed to repeat; so the number of possible strings over any finite alphabet is infinite. The researcher thus faces an interesting challenge: on the one hand, the language should provide powerful primitives for restructuring strings; on the other hand, the expressive power of the language should be carefully limited, to avoid infinite computations.

In [21], we introduced a logic called *Sequence Datalog* for querying sequence databases. *Sequence Datalog* is a Horn-like logic with special, interpreted function symbols that allow for *structural recursion* over strings. Sequence Datalog has both a clear declarative semantics and an operational semantics. The semantics are based on fixpoint theory, as in classical logic programming [19], and on a new notion of active domain, called the *extended active domain*, introduced to

* The first author was partially supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The second author was partially supported by MURST and Consiglio Nazionale delle Ricerche (CNR).