IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8038

Chapter X

Fuzzy Query Languages for Multimedia Data

Paolo Ciaccia and Wilma Penzo
DEIS-CSITE-CNR, Bologna, Italy

Danilo Montesi
DCS, Bologna, Italy

Alberto Trombetta
DSI, Milano, Italy

This chapter presents a fuzzy-based algebra, called SAME^W, to query multimedia objects. SAME^W allows for dealing within a common framework with several aspects relevant to similarity query processing as well as with the inherent imprecision that characterizes data, user requests and query results. Non-Boolean, namely fuzzy and similarity, predicates are used to rank tuples according to specific criteria. Complex multi-predicate queries can be formed by means of logical connectives, whose semantics is parameterized in order to adapt to specific scenarios. The same holds for the semantics of algebraic operators. These include properly extended traditional relational operators and new operators which allow threshold and best-matches queries to be easily expressed. A further important feature of SAME^W is the possibility of weighting both predicates and operands of algebraic operators so as to better fit user preferences/requirements. A working example dealing with Web data is used throughout the chapter to show the potentialities of SAME^W. Optimization issues are also briefly discussed.

INTRODUCTION

The advent of the World Wide Web has made available a huge amount of text, image, audio and video data, collectively referred to as multimedia (MM) data. In recent years, many commercial MM tools have been developed (Flickner, 1995), among which are multimedia database systems (MMDBSs) whose aim is to provide unified frameworks for retrieving and integrating MM data (Subrahmanian, 1998). It is a fact that a key feature

of MMDBSs should be the capability of expressing highly powerful/complex queries by managing and interpreting the intrinsic imprecision of MM data and by exploiting classification processes that add semantic information to objects.

Many recent works have focused on the design of access methods suitable to index complex features (Ciaccia, 1997; Seidl, 1997), as well as on other performance-related issues, such as approximate queries (Shivakumar, 1998; Ciaccia, 2000a). On the other hand, more general issues relevant to MM query processing have only been partially addressed. These include models able to capture the "essential" aspects of MM objects needed by *content-based queries*, the impact of *user preferences* on query processing, the management of *imprecise* queries, new kinds of predicates and operators and so on. Indeed, contributions to the above issues (Fagin, 1996; Adali, 1998; Ciaccia, 1998; Montesi, 1999) often consider ad-hoc scenarios and/or completely ignore the other coordinates of the problem, thus resulting in a set of difficult to integrate recipes.

In this chapter we present an extended relational framework that takes into account the two major sources of imprecision arising from queries on MM databases (Soffer, 1998): 1) imprecision of *classification* of MM data and 2) imprecision in the *matching of features* that characterize the content of MM objects. As to the first point we rely on basic concepts from *fuzzy set theory*, and allow representation of vague classification both at tuple and at attribute level. We then introduce a *similarity algebra*, called SAME^w ("Similarity Algebra for Multimedia Extended with Weights"), that extends relational algebra in a conservative way and incorporates the use of *weights* in predicates and operators, in order to model user preferences. We show how complex queries can be easily expressed in SAME^w and sketch how equivalence rules can be exploited for the purpose of query rewriting and optimization.

RELATED WORK

Broadly speaking, there are two main approaches to deal with data imprecision. They are based on probability and fuzzy set theories, respectively. Probabilities are simple, yet intuitive to understand. However, the probability of complex (joint) events can be easily computed only under the hypothesis of events' independency. If this hypothesis does not hold (as it often happens), then an interval of probabilities needs to be considered for complex events, as the ProbView system does (Lakshmanan, 1997). Fuzzy concepts do not require such hypothesis since they do not rely on the notion of event. Due to their generality, they have been successfully applied to integrate information from different MM systems (Fagin, 1996).

As for query languages for MM data, specific functionalities have been investigated in recent years. For instance, the modeling of imprecisely classified data was thoroughly analyzed in Raju (1988). An interesting approach for "ranking" query results coming from different data sources is proposed in Gravano (1997), whereas Carey (1997) introduces an operator for selecting only the *n* "best" results of an SQL query. An algebraic setting more suitable for the integration of similarity measures coming from different sources rather than for similarity searches is proposed in Adali (1998).

In summary, there are no proposals considering a full-fledged query language expressing the basic MM functionalities within a uniform framework. This language must blend imprecision on data and their classification, specific operators needed to express complex queries on multimedia data and user preferences. This is the subject developed in the remaining of this chapter.