ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ • ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING • UNIVERSITY OF IOANNINA

ΠΑΡΟΥΣΙΑΣΗ
ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΗΜΕΡΟΜΗΝΙΑ: Τετάρτη, 30 Αυγούστου 2023

ΩΡΑ: 12:00

ΑΙΘΟΥΣΑ: Αίθουσα Σεμιναρίων, ΤΜΗΥΠ

ΟΜΙΛΗΤΗΣ: **Γεώργιος Χριστοδούλου**

Θ έ μ α

« Interval Data Management in Main Memory »

Επταμελής Εξεταστική Επιτροπή:

1. **Νικόλαος Μαμουλής (Επιβλέπων),** Καθηγητής πρώτης βαθμίδας, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
2. **Παναγιώτης Βασιλειάδης,** Καθηγητής πρώτης βαθμίδας, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
3. **Παναγιώτης Τσαπάρας,** Αναπληρωτής Καθηγητής, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
4. **Ευαγγελία Πιτουρά,** Καθηγήτρια πρώτης βαθμίδας, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
5. **Παναγιώτης Μπούρος,** Assistant Professor, Institute of Computer Science of Johannes Gutenberg University Mainz (JGU), Germany
6. **Εμμανουήλ Κουμπαράκης,** Καθηγητής πρώτης βαθμίδας, Τμήμα Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Π.Α.
7. **Σπυρίδων Σκιαδόπουλος,** Καθηγητής πρώτης βαθμίδας, Τμήμα Πληροφορικής & Τηλεπικοινωνιών, Πανεπιστήμιο Πελοποννήσου

Π ε ρ ί λ η ψ η

The management of intervals has been an active research area since databases were invented. A popular direction of research is the indexing and retrieval of intervals, finding a wide range of applications. Emerging and widely used systems are built dependent on temporal and uncertain data. Many algorithms and indices have been proposed, concentrated on a variety of queries. Most algorithms are either suboptimal in space consumption or perform well only for specific query types. We need novel and

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τ.Θ. 1186, ΙΩΑΝΝΙΝΑ 45110
T: 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/

DEPT. OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA
P.O. BOX 1186, IOANNINA
GR - 45110 GREECE
T: +30 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/

efficient in-memory indices for intervals, which can evaluate queries with high performance.

In statistical and probabilistic databases, uncertain values are often approximated by confidence intervals. Real-world examples of uncertain values include temperature values obtained from IoT devices or time-series. For such cases, it would be more appropriate to record an observation using an interval range rather than a single value. In data anonymization attributes can be generalized to intervals. Stored values can be replaced with semantically consistent but less precise alternatives in the form of intervals. In this way, information from a private table, like the identity of any individual to whom the released data refer cannot be recognized. In XML data indexing techniques, the scope of an XML element can be modeled as an interval defined by the positions of the starting and closing tag of the element.

Intervals are representations of value ranges. Quite often, these ranges represent periods of time described as a tuple [start, end]. In a temporal database, an interval-based data model can timestamp each tuple or attribute value with a validity time interval. Along with valid time, an interval-based model can timestamp transaction time, which captures when a tuple is inserted and deleted from the database. We index intervals in data structures so that we can efficiently evaluate different types of queries. There are several query types over intervals, with the differentiation lying on the specifications that shape the resulting set of intervals, or the context in which we model data with an interval representation.

The topic of this dissertation is to study the problem of indexing and querying a large collection of records, based on an interval attribute that characterizes each object. We focus on the different aspects of temporal databases, as they form the most significant application of interval data. The collection can be known before indexing or evolve over time, which is common in temporal databases or streaming data. The challenge is to find solutions which, can take advantage of modern hardware such as large main memories, can handle traditional and on demand indexing of intervals, and provide high performance for a wide variety of query types and predicates. In this thesis, we study numerous problems and different scenarios which come down to indexing and querying interval data.

In the first part, we propose HINT, a novel and efficient in-memory index for large known collections of intervals, with a focus on range queries, which are a basic component of many search and analysis tasks. Our index is suitable for valid-time indexing in the context of temporal databases. HINT applies a hierarchical partitioning approach, which assigns each interval to at most two partitions per level and has controlled space requirements. We reduce the information stored at each partition to the absolutely necessary by dividing the intervals in groups, based on whether they begin inside or before the partition boundaries. In addition, our index includes storage optimization techniques for the effective handling of data sparsity and skewness.

The second problem we study, is a more general version of HINT, so that with the best trade-off in information storage, it will be able to handle queries with different predicates. Intervals may satisfy more sophisticated relations than intersections, which are based on Allen's relationships (e.g., find all intervals that are covered by the query interval). The principles of HINT are useful for the retrieval of data intervals based on Allen's relationships, because the hierarchical partitioning applies independently of the

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τ.Θ. 1186, ΙΩΑΝΝΙΝΑ 45110
T: 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/

DEPT. OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA
P.O. BOX 1186, IOANNINA
GR - 45110 GREECE
T: +30 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/

query type. We show how HINT can be tuned depending on the data and can be efficiently used to process joins and queries based on Allen's relationships.

In the last part of this dissertation, we study the problem of transaction-time indexing in the context of temporal databases, i.e., indexing versions of data in an evolving database. Given the fact that the main memories of modern commodity are large and cheap, we can afford to keep track of all versions of an evolving table in memory. This raises the question of how to index such a table effectively. We depart from the classic indexing approach, where both current (i.e., live) and past (i.e., dead) data versions are indexed in the same data structure, and propose LIT, a hybrid index, which decouples the management of the current and past states of the indexed column. LIT includes optimized indexing modules for dead and live records, which support efficient queries and updates, and gracefully combines them.

For the evaluation of our methods, we used multiple real and synthetic datasets with different characteristics so that we can safely conclude the robustness of our algorithms. The experiments showed that our algorithms are typically one order of magnitude faster than existing methods on static or evolving data collections and with multiple types of queries.

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

Τ.Θ. 1186, ΙΩΑΝΝΙΝΑ 45110
T: 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/

DEPT. OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

P.O. BOX 1186, IOANNINA
GR - 45110 GREECE
T: +30 265100 8817 - 8813 - 7196
http:/www.cse.uoi.gr/