論 文 内 容 の 要 旨

博士論文題目　Efficient Indexing Techniques for XML Data

氏　　名　Dao Dinh Kha

（論文内容の要旨）

In today s just-in-time information paradigm, the ability to communicate efficiently is vital. In this context, Extensible Markup Language (XML) has been invented as a standard for data exchange among a variety of data sources and applications. Since its invention, the popularity of the XML has been dramatically increased because of its ability to provide a simple but standardized, extensible mean to express semantic information within documents. This property makes XML possible to address the shortcomings of existing markup languages and become a key technology that facilitates information exchange for science, technology, and industries, as well as many aspects of society. XML also has been selected as the data exchange standard in newly arising business domains, such as e-business. The semantics of XML data is richer than the one of relational data by including both of its content and metadata, the information that describes the structure of the data. Therefore, XML requires new techniques that are different from the relational databases technologies for data management.

The aim of this thesis is to design efficient indexing techniques for XML data. The structural characteristics of XML data raises several challenges to design of indexing structures. In this thesis, we investigate three critical issues of XML data management as follows.

The first issue is to cope with intensively content-updated XML data. Among several methods of storing XML documents, a straightforward yet efficient method is to store a string representation of the XML document. An XML node is usually represented by a region coordinate, which is a pair of integers expressing the start and end positions of the substring corresponding to the node. This approach, however, has the drawback that a change of a node┣ s region coordinate causes change of the region coordinates of many other elements that normally degrades the performance of XML applications. To deal with the issue, we propose the Relative Region Coordinate (RRC)

technique to effectively reduce the cost of re-computation by expressing the coordinate of an XML element in the region of its parent element. We present a method to integrate the RRC information into XML systems and provide experimental results.

The second issue is to cope with structure-update of XML data. XML queries involve both content search and structure search. For structure search, the structural information of XML data is essential to determine the structural relationships in XML documents. Several numbering schemes have been proposed so far to express the structural information using the identifiers of XML nodes. However, since the structure of XML documents can be changed, the robustness of these numbering schemes is vital for the whole indexing structure. For this purpose, we introduce a new numbering scheme called recursive UID (rUID) that has been designed to be robust in structural update and applicable to arbitrarily large XML documents. We investigate the applications of rUID to XML query processing in a system called SKEYRUS, which enables the integrated structure-keyword searches on XML data.

The third issue considered in this thesis is to exploit the schematic information to improve XML query processing efficiency. Although most of XML documents have associated DTD or XML schema, the prior query processing techniques have not utilized the document structure information efficiently. We propose a novel XML query processing method that uses DTD or XML schema to improve the I/O complexity of XML query processing. We design a Structure-based Coding for XML data (SCX) that incorporates both structure and tag name information extracted from the document structure descriptions. This property of SCX provides a Virtual Join mechanism that greatly reduces I/O workload for processing XML queries. Our experimental results indicate that SCX accelerates the processing of XML queries significantly.

XML is a new technology and its invention follows an industrial concept, where societies and industries are creating artifacts for researchers to study. Therefore, the XML-related techniques are subjects of a long development and improvement. The study presented in this thesis is an effort toward efficient XML data management.

氏　名	Dao Dinh Kha

（論文審査結果の要旨）

2003 年 12 月 26 日に開催した公聴会の結果を参考に、2004 年 1 月 16 日に本博士論文の審査を行った．次のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

Dao Dinh Kha は本論文において、XML 文書の索引付けを効率よく行う技術について研究した。XML による文書のタグ付けは、情報交換の基盤技術として、広く利用されつつあるが、XML データは構造をもっているので、索引付けに関しても新しい技術を要請している。本論文では、XML データ管理における次の三つの課題を研究し、実験により評価して、成果を得た。

第 1 の研究課題は、内容がひんぱんに更新される場合の XML データの処理である。ここでは、Relative Region Coordinate (RRC) と呼ぶ技術を開発し、XML 要素の座標をその親要素の範囲で相対的に表現する方式により、更新がもたらす座標の再計算のコストを軽減することに成功した。

第 2 の研究課題は、XML データの構造変化である。XML データの構造変化に強い番号付け方式を開発した。この番号付け方式を再帰的な UID (rUID)と呼ぶ。rUID は、XML 文書の構造変化に強く、任意の大きい XML 文書にも応用可能である。実際の問合せ処理に統合して、SKEYRUS システムを実験した。

第 3 の研究課題は、XML 問合せ処理におけるスキーマ情報の活用法である。XML 文書は DTD や XML スキーマと結びついているが、問合せ処理に文書の構造を活用するには至っていなかった。「構造を活用した XML データの符号化法（Structure-based Coding for XML data, SCX) 」では、文書構造の記述から構造情報やタグ名情報を得て、それを問合せの効率向上に活用する。実験により、SCX は XML 問合せ処理を大幅に高速にすることを示した。

本研究は，構造を持った文書を効率よく処理する技術の提案であり、独創的，かつ実用的で，研究開発の現場において高い貢献があるものと評価する．よって、本論文は博士（工学）の学位論文として価値あるものと認める．