Chapter 5

Matching of Twig Pattern with AND/OR Predicates over XML Streams

Weidong Yang
Fudan University, China

Hao Zhu
Fudan University, China

ABSTRACT

Chapter 5 presents a novel approach for processing complex twig pattern with OR-predicates and AND-predicates over XML document streams which a twig pattern is represented as a query tree. Its OR-predicates and AND-predicates are represented as a separate abstract syntax tree associated with the branch node, and all the twig patterns are combined into a single prefix query tree that represents such queries by sharing their common prefixes. Consequently, all the twig patterns are evaluated in a single, document-order pass over the input document stream avoiding the translation of the set of twig patterns into a finite state automaton. Chapter 1 introduces the background of this issue. Chapter 2 discusses the representation of complex twig pattern as a query tree, how to combine a set of twig patterns into a single query tree, how to match multi twig patterns over the incoming XML document, and possible optimization of computing logical AND/OR predicates. In section 3, the architecture of a XML stream process system named LeoXSQ is given. Section 4 shows the conducted experiments. In section 5, the related work is discussed. Section 6 summarizes this chapter.

5.1 INTRODUCTION

Stream-based continuous query processing fits a large class of new applications, such as sensor networks, location tracking, network management and financial data analysis. Stream based applications usually involve handling the infinite size and unpredictability of the data streams and requiring timely response, traditional database processing techniques cannot be used directly, then there has been plentiful of researches about data stream management (for examples, Babcock, Babu, Datar, Motwani & Widom, 2002). As extensible markup language XML is a standard for information exchange, the problem of processing streaming XML data is gaining widespread attention from the
research community. Many researchers presented their work of processing XML stream based on automaton (Altinel & Franklin, 2000; Chan, Felber, Garofalakis & Rastogi, 2002; Green, Miklau, Onizuka & Suciu, 2003; Gupta & Suciu, 2003; Diao, Altinel, Franklin, Hao & Fischer, 2003). The XFilter system (Altinel & Franklin, 2000) used a separate FSM per path query to allow all of the FSMs to be executed simultaneously during the processing of a document. Building on the insights of the XFilter work, YFilter (Diao, et al., 2003) describes a new method, called “YFilter” that combines all of the path queries into a single Nondeterministic Finite Automaton (NFA).

XML stream systems aim to provide fast, on-the-fly matching of XML-encoded data to large numbers of query specifications containing constraints on both structure and content. An XML twig query is essentially a complex selection predicate on both structure and content of an XML document. Matching a set of twig patterns with the incoming XML stream is a core operation in XML stream system. Kwon, Rao, Moon & Lee (2005) recently presented a system called FiST, which performs holistic matching of twig patterns with incoming XML documents by transforming twig pattern into prufer sequences. But their work considers only twig patterns whose sibling edges are combined by AND logic.

Queries in real application, however, often contains logical-OR and logical AND in a single twig patter. For example, the following query

\[
Q = /\text{dblp/paper}[\text{title = 'XML Stream'} \text{ or (year = 2006 and conf = 'CAiSE')}]/\text{author}
\]

This twig pattern contains complex AND/ OR logical operations. Although, theoretically, some existing systems, for example, YFilter (OR predicate is not implemented in YFilter) can deal with this case, but their approaches are not very efficient. YFilter treats twig-patterns as nested path expressions, which are processed by using query decomposition. In their approach, the nested paths are extracted from the main path expressions and processed individually. A post-processing phase is used to link matched paths back together to determine if an entire query expression has been matched. This approach has some disadvantage: a) many intermediate results may be produced; b) the intermediate results should be buffered, and then are linked together.

This chapter presents an approach to match complex twig pattern with OR-predicates and AND-predicates over XML document streams in a way of combining top-down and bottom-up process. In the approach, a twig pattern is represented as a query tree. Its OR-predicates and AND-predicates of a branch node are represented as a separate abstract syntax tree associated with the branch node and processed in a bottom-up way. For the improvement of the processing performance of twig patterns, all the twig patterns are combined into a single prefix query tree that represents such queries by sharing their common prefixes. Consequently, all the twig patterns are evaluated in a single, document-order pass over the input document stream efficiently avoiding the translation of the set twig patterns into a finite state automaton.

5.2 MULTIPLE TWIG PATTERNS PROCESSING OVER XML STREAM

In this section, we address the central topic: processing strategies for multiple twig patterns over an XML stream. In 5.2.1, we describe a mechanism to represent the twig pattern with OR-predicates and AND-predicates as a query tree, and then they are combined into a single query tree with common prefix. In 5.2.2, we describe how to match the document stream with the twig patterns. In 5.2.3, we discuss the short-circuit evaluation for optimization.
Related Content

**Text Mining in Bioinformatics: Research and Application**
[www.irma-international.org/article/text-mining-in-bioinformatics/100039/](www.irma-international.org/article/text-mining-in-bioinformatics/100039/)

**Latent Topic Model for Indexing Arabic Documents**
[www.irma-international.org/article/latent-topic-model-for-indexing-arabic-documents/113331/](www.irma-international.org/article/latent-topic-model-for-indexing-arabic-documents/113331/)

**Pharaoh: Context-Based Structural Retrieval of Cognitive Scripts**
[www.irma-international.org/article/pharaoh-context-based-structural-retrieval/78314/](www.irma-international.org/article/pharaoh-context-based-structural-retrieval/78314/)

**Compressing and Vague Querying (XCVQ) Design**
[www.irma-international.org/chapter/compressing-vague-querying-xcvq-design/69133/](www.irma-international.org/chapter/compressing-vague-querying-xcvq-design/69133/)

**Technostress: Effects and Measures Among Librarians in University Libraries in Nigeria**
[www.irma-international.org/chapter/technostress-effects-measures-among-librarians/73779/](www.irma-international.org/chapter/technostress-effects-measures-among-librarians/73779/)