Chapter 1

Foundation of Keyword Search in XML

Weidong Yang
Fudan University, China

Hao Zhu
Fudan University, China

ABSTRACT

It has become desirable to provide a way of keyword search for users to query structured information in an XML database (data-centric retrieval) by combining database and information retrieval techniques. Therefore, the key challenges of keyword search in the XML database are how to define appropriate result models meeting user’s search intents, how to search the results by using efficient algorithms, and how to ranking the results. In this chapter, on one hand, the authors present the foundational knowledge of XML keyword search such as XML data models, XML query languages, inverted index, and Dewey encoding. On the other hand, some existing typical researches of keyword search in XML are presented, including the results models such as Smallest Lowest Common Ancestor (SLCA), Exclusive Lowest Common Ancestor (ELCA), Meaningful Lowest Common Ancestor (MLCA), the related search algorithms, and the ranking approaches.

1.1 INTRODUCTION

As a standard for the representation and exchange of semi-structured data on the Internet, XML has attracted much research in XML retrieval, which enables information discovery in XML data.

With regard to the retrieval mode, traditional structured query languages, such as XPath and XQuery, are used to search XML data, which can convey complex semantic meanings and therefore retrieve precisely the desired results. Nevertheless, the syntax of such a language is often rather complicated which makes it not appropriate for a naive user. One still needs sufficient knowledge of the structure, role of the requested objects in
order to formulate such a meaningful query. In contrast, *keyword search* is a proven user-friendly way of querying XML data, since the user does not need to know either a query language or the structure of the underlying data. The main disadvantage lies in the lack of expressivity and inherent ambiguity, which also poses challenges in interpreting the semantics when performing keyword search on XML data.

Considering the organization of underlying XML data, *data-centric retrieval* and *text-centric (or document-centric) retrieval* are to be examined. While both text and structure are important, text-centric retrieval gives higher priority to text. The premise of this approach is that XML document retrieval is characterized by long text fields and inexact matching. In contrast, data-centric XML retrieval can execute exact match conditions upon mainly numerical and non-text attribute-value encoded data. Therefore, this puts the emphasis on the structural aspects of XML data and queries. In a word, text-centric approaches are appropriate for data that are essentially text documents, marked up as XML to capture document structure, while data-centric approaches are commonly used for data collections with complex structures that mainly contain non-text data. (Further comparisons can be referred in later sections).

In the chapter, some basic XML concepts will be described in Section 1.2. Then we will mainly discuss data-centric keyword search from several aspects, namely semantic model, search algorithms and relevance ranking, which be further presented during section 1.3-1.5. Other related issues can be found in Section 1.6.

### 1.2 BASIC XML CONCEPTS

#### 1.2.1 XML Documents

The Extensible Markup Language (XML) is a general-purpose specification for creating custom markup language. Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. XML is recommended by the *World Wide Web Consortium (W3C)*, a fee-free open standard, which also specifies XML lexical grammar and parsing requirements.

#### 1.2.1.1 XML Element and Attribute

An *XML element* is everything from (including) the element’s start tag to (including) the element’s end tag. An element can contain other elements, simple text or a mixture of both. Elements can also have attributes. *XML attributes* provide additional information about elements.

In the XML document showed in Figure 1, `<bookstore>` and `<book>` have element contents, because they contain other elements. `<title>` has text context since it only contains text. And `<book>` has an attribute (category="NOVEL").

#### 1.2.1.2 XML Schema

An *XML document* can optionally have a schema to describe the structure of the document. An *XML schema* defines elements and attributes that can appear in a document, the child elements of a particular element, the child elements’ order and number, data types of elements and attributes and their default or fixed values, etc. Figure 2 displays an example schema, which is used to describe the structure of the XML document showed in Figure 1.

#### 1.2.1.3 Data Model

As an XML document is formed by a sequence of elements that enclose text values and other elements, it is typically represented by a *tree* structure where the nodes correspond to elements, attributes or text values, and the edges represent immediate element-subelement or element-value relationships. Thus, an XML database is then...
Related Content

Advanced Branching and Synchronization Patterns Description Using Pi-Calculus
www.irma-international.org/chapter/advanced-branching-synchronization-patterns-description/69151/

Using Association Rules for Query Reformulation
www.irma-international.org/chapter/using-association-rules-query-reformulation/64430/

Semantic Role Labeling Approach for Evaluation of Text Coherence
www.irma-international.org/article/semantic-role-labeling-approach-for-evaluation-of-text-coherence/101360/

Social Coordination with Architecture for Ubiquitous Agents-CONSORTS
Koichi Jurumatani (2004). Intelligent Agents for Data Mining and Information Retrieval (pp. 154-163).
www.irma-international.org/chapter/social-coordination-architecture-ubiquitous-agents/24161/

Visualizing Information Science Knowledge by Modelling Domain Ontology (OIS)
Ahlam F. Sawsaa and Joan Lu (2013). International Journal of Information Retrieval Research (pp. 52-80).
www.irma-international.org/article/visualizing-information-science-knowledge-by-modelling-domain-ontology-ois/109662/