Chapter 8
Research Background

Badya Al-Hamadani
University of Huddersfield, UK

Joan Lu
University of Huddersfield, UK

ABSTRACT

XML documents are increasing in usage during the last years since the structure of these documents has lots of important specifications. This chapter explains the importance of XML documents and their structure. It spotlights the difference between the traditional text retrieval and retrieving information from XML documents as well as the query languages used to access parts of these documents.

1.1 XML COMMENCEMENTS AND IMPORTANCE

Before the rise of the internet, 1980s witnessed the invention of Standard Generalized Markup Language (SGML) as a way to display information dynamically. Later, in 1995, W3C recommended SGML to be used for the internet. Problems occurred when using SGML included the lack of widely supported style sheets, complexity and instability in the software that were using it, and the difficulties in interchanging SGML data due to its varying levels among SGML software packages.

In 1996, the first XML working draft was intended to be a powerful substitute to SGML. It was first recommended by the World Wide Web Consortium (W3C) in 1998 to be used as a mark-up language for storing and exchanging data through the web. The most recent recommendation was published in 2008, which is the fifth edition of the XML (W3C, 2008). In a very short period of time, XML has become the basis for data exchange through the Internet. This is due to its several features such as the following (NG et al., 2006; Gerlicher, 2007; Groppe, 2008):

- **Readability**: XML is readable by both human and machine. This means that the data represented by XML can be used by different users and by different parsing code.
**Research Background**

- **Interoperability:** This is the ability of the hardware and software to use XML documents without the need to make any changes to the software or the data itself. This means that XML data is stripped of any dependency on software and machine.

- **Long term usability:** Since XML documents are represented using the Unicode; these documents are expected to stay in secure storage and usage for years (Augeri et al., 2007; De Meo et al., 2007).

- **Extensibility:** This means that there are no fixed set of tags that should be used to represent data.

- **Generality:** XML documents have the ability to represent different kinds of data representation such as images, sounds, videos, texts, etc.

- **Internationality:** Almost all written languages can be represented in XML documents since they support Unicode (Norbert and Kai, 2004).

In spite of all these advantages, XML has also some weaknesses:

- They have a huge amount of redundancy which makes these documents demand high storage memory to be archives, high bandwidth to be transmitted, and high cost to be processed.

- The huge amount of technologies surrounding it complicates the use of these documents such as schema, DTD, XSLT, SAX, DOM, XPath, XQuery. These technologies render the use of these documents somewhat difficult especially with naive users or in cases where these technologies are absent, it would be just as difficult as they are considered necessary for dealing with XML documents.

- The problems that can occur when dealing with the document namespace should be carefully sorted out otherwise other problems and complications could occur during the processing of these XML documents.

**1.2 XML DOCUMENT TYPES**

The main building blocks of any well-formed XML document are nested open tags and their equivalent close tags. These tags can be formed as follows (Hunter, 2000; Anders, 2009; Goldberg, 2009):

1. **Elements:** each element starts with an open tag (<p>) and ends with an end tag (</p>). Everything between and including these tags is an element. The general structure of an element is as follows: 
   \[
   \text{<e at}_1="v_1" \text{at}_2="v_2" \text{at}_n="v_n">d_1d_2d_3\ldots d_m</e>
   \]
   such that \(n\geq0, \text{ and } m\geq0\) (1)

2. **Such that** \(n\geq0, \text{ and } m\geq0\) (1)

   Each element has an element-name (e) which should follow the following rules:

   - Case sensitive names.
   - Consist of characters, numerals, underscores and tabs.
   - Start with a character or an underscore.
   - Should not start with xml or XML.

   Elements can have optional element-value \((\{d_1d_2d_3\ldots d_m\} \text{ in } (1))\) which represent the actual data values for the XML document.

3. **Attributes:** attributes (if any) appear within an element and they provide more information about that element. Each attribute has an attribute-name \((\{\text{at}_1, \text{at}_2, \text{at}_n\} \text{ in } (1))\) which should follow the same rules for an element-name, and an attribute-value \((\{v_1, v_2, \ldots v_n\} \text{ in } (1))\) which can be any printable character between a pair of quotations.

4. **Data text:** the data in a XML document could either be attribute-values or element-
Related Content

A Knowledge Based Approach for Accident Analysis: Application to the Safety of Automated Rail Transport Systems
[www.igi-global.com/article/a-knowledge-based-approach-for-accident-analysis/210062?camid=4v1a](www.igi-global.com/article/a-knowledge-based-approach-for-accident-analysis/210062?camid=4v1a)

The Mutual Influence of Technology and Leadership Behaviors
[www.igi-global.com/chapter/mutual-influence-technology-leadership-behaviors/69747?camid=4v1a](www.igi-global.com/chapter/mutual-influence-technology-leadership-behaviors/69747?camid=4v1a)

Segmentation Free Word Spotting for Handwritten Documents Using Bag of Visual Words Based on Co-HOG Descriptor
[www.igi-global.com/article/segmentation-free-word-spotting-for-handwritten-documents-using-bag-of-visual-words-based-on-co-hog-descriptor/222768?camid=4v1a](www.igi-global.com/article/segmentation-free-word-spotting-for-handwritten-documents-using-bag-of-visual-words-based-on-co-hog-descriptor/222768?camid=4v1a)

Krikelas’ Model of Information Seeking Behavior (1983)
[www.igi-global.com/chapter/krikelas-model-of-information-seeking-behavior-1983/127124?camid=4v1a](www.igi-global.com/chapter/krikelas-model-of-information-seeking-behavior-1983/127124?camid=4v1a)