Chapter 7
Introduction

Badya Al-Hamadani
University of Huddersfield, UK

Joan Lu
University of Huddersfield, UK

ABSTRACT
The eXtensible Markup Language (XML) is a World Wide Web Consortium (W3C) recommendation which has widely been used in both commerce and research. As the importance of XML documents increase, the need to deal with these documents increases as well. This chapter illustrates the methodology that has been used throughout the research, discussing all its parts and how these parts were adopted in the research.

1.1 INTRODUCTION
The eXtensible Markup Language (XML) is a World Wide Web Consortium (W3C) recommendation which has widely been used in both commerce and research. In recent years, we have witnessed a dramatic increase in the volume of XML digital information that is either created directly as an XML document or converted from another type of data representation. The importance of XML is mainly due to its ability to represent different data types within one document, solving the problem of long-term accessibility, and providing a solution to the problem of interoperability (Al-Hamadani et al., 2009).

Due to the replication of the XML schema in each record, the XML document is considered to be one of the self-describing data files, which means that these kinds of files have a lot of data redundancy in relation to both its tags and attributes (Ray, 2001). For the above reason the need to compress XML documents is becoming increasingly dramatic. Furthermore, what has evolved is the urgent need to retrieve information directly from the compressed documents and then decompress only the retrieved information (Ferragina et al., 2006).

DOI: 10.4018/978-1-4666-1975-3.ch007
Introduction

Because of the wide range of XML documents in use and the different kinds of users, being able to deal with all kinds of queries has become a key issue. Some of these queries may have imprecise constraints which cannot be processed directly due to the grammar restriction in the existing query languages. However, these types of queries, which are known as vague queries, appear to be common when the users of the XML documents have little knowledge about the document structure, or may lack the skills to write a precise and meaningful query. Another type of vague queries occurs when the query is presented without the presence of a Schema or the data type definition (DTD) of the document.

According to the relevant literature, there are a number of techniques that compress the XML documents and query the compressed version with no or partial decompression. These techniques process almost all types of queries but not the vague queries; admittedly, there are a number of researchers now trying to process vague queries on the original XML document.

The research carried out in this thesis primarily concerns designing and implementing a new technique called XML Compressing and Vague Querying (XCVQ) which consists of two stages. In the first stage, it separates the data part of the XML document into several containers according to the path of that data within the document. Then each of the containers is compressed separately using a back-end compressor. The second stage processes the vague queries by decomposing them into multiple sub-queries, retrieves information from the compressed XML document according to each sub-query, combines the retrieved information according to the given query, and finally decompresses only the most relevant information.

To eliminate the amount of technologies associated with the XML documents and to make the process of compressing and retrieving information easier for the inexperienced users, XCVQ is designed to be schema independent in both phases of the compressor and the query processor.

1.2 RESEARCH HYPOTHESIS AND RESEARCH METHODOLOGY

This thesis is based on the following hypotheses:

1. The existing XML compression techniques can be improved to construct a new schema independent XML compressor with a higher compression ratio.
2. The redundancy in the XML documents significantly affects the size of those documents and can be reduced to more than half of the original file size.
3. The compressed XML document can be retrieved according to vague queries. Vague queries are those queries which do not follow the semantic rules of current query languages. They occur when the exact matching user’s query does not retrieve the required information either because of the lack of experience in writing a query or the absence of the document’s schema.
4. The necessity of retrieving information from more than one XML documents without the need to specify an exact relative document.

The above hypotheses are tested throughout this research by using the System Development Methodology (SDM) (Nunamaker et al., 1991; Morrison and George, 1995; Hevner et al., 2004). This methodology has been widely used by software developers and information system specialists (Meersman et al., 2008; Yousof et al., 2011) As depicted in Figure 1, this methodology consists of four main stages:

1. **Identifying research problems:** This stage focuses on drawing up the research questions due in part to the lack of theories in the research field and/or build upon existing theories. In this thesis, the research questions are set from two XML fields, compressing the XML documents and querying them. As a result, a new XML compressor is introduced.
Related Content

Information Retrieval from Deep Web Based on Visual Query Interpretation
[www.irma-international.org/article/information-retrieval-from-deep-web-based-on-visual-query-interpretation/90441/](http://www.irma-international.org/article/information-retrieval-from-deep-web-based-on-visual-query-interpretation/90441/)

Domain Specific Custom Search for Quicker Information Retrieval
[www.irma-international.org/article/domain-specific-custom-search-for-quicker-information-retrieval/101358/](http://www.irma-international.org/article/domain-specific-custom-search-for-quicker-information-retrieval/101358/)

Adaptive Summarization of Digital Video Data
[www.irma-international.org/chapter/adaptive-summarization-digital-video-data/30762/](http://www.irma-international.org/chapter/adaptive-summarization-digital-video-data/30762/)

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

Studying Web Search Engines from a User Perspective: Key Concepts and Main Approaches
[www.irma-international.org/chapter/studying-web-search-engines-user/64435/](http://www.irma-international.org/chapter/studying-web-search-engines-user/64435/)