Recent Development in XML-IR

Baydaa T. Rashid[1], Raad F. Alwan[2], Joan Lu[1], Jim Yip[1]
[1] University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK.
[2] Philadelphia University, Amman (19392), Jordan.

Abstract

The Web is characterized by a huge amount of heterogeneous data sources, which have different media support and format representation. Because XML can represent files of different formats, it can play an important role in IR since it is becoming a standard form for data representation and exchange over the Web. Under this assumption, the problem of querying heterogeneous sources can be reduced to the problem of querying XML data sources. This paper shows the influence of XML on the IR techniques and methodologies during the last five years through serving over 400 papers published in different conferences and journals.

Keywords: XML, Information Retrieval, Query languages.

1. Introduction

Millions of people all around the world use the Web for their daily life needs and many companies invest a lot on developing a better system for information retrieval on the Web. Since the information on the Web is of different types, such as image, music, movies, text …etc, there was an urgent need to represent all these kinds of documents in a unified way. EXtensible Mark-up Language (XML) can be used to describe documents and data in a standardized, text-based format that can be easily transported via standard Internet protocols. It does a great job in describing structured data as text, and the format is open to extension. This means that any data that can be described as text and can be nested in XML tags will generally be accepted as XML. [7]

For all the previous reasons, the researchers tried to investigate the field of retrieving information from the XML documents (XML-IR). A review on the history of XML-IR shows that not only many developments have been made in the past but also there is an extensive desire to develop XML-IR techniques and their query languages.

2. XML-IR Overview:

This paragraph will give a brief background on XML, its importance and IR query languages on XML documents.

2.1 A Brief history of XML:

The XML was created in 1996 by World Wide Web Consortium (W3C) to improve Web page functionality beyond HTML and to simplify Standard Generalized Mark-up Language (SGML) [2, 8]. Although SGML was a good format for document sharing, and HTML was a good language for describing the layout of the documents in a standardized way, there was no standardized way to describe and share data that was stored in the document.

In 1998 the (W3C) met this need by combining the basic features that separate data from format in SGML with extension of the HTML tag formats that were adapted for the Web and came up with the first Extensible Mark-up Language Recommendation. [7]

2.2 Information retrieval (IR) system stages:

IR is the science of extracting relevant documents from a collection of documents [3]**,** or it is the process of extracting information from documents [4]. An IR system has two main objectives: [5]

1. Support of user search generation.
2. Present the search results in a format that helps the user in determining relevant items.

Each information retrieval system must pass in the following five stages: [3]

1. Before the information process starts, the original document must be transformed into a *logical* view. This stage done by the database manager to define the documents to be used and the operations to be performed.
2. Indexing: is the process of extracting index keywords from the document. The *inverted file* is one of the index structure used.
3. The user specifies his/her need, which is transformed by the same text operations applied to the text.
4. The *query* is generated, which will be used to obtain the *retrieved documents*.
5. Ranking: the process of sorting the retrieved documents according to its *likelihood* of relevance before they could be submitted to the user.

2.3 Importance of XML Retrieval

There are a lot of important reasons to retrieve information from XML documents. Some of them are:

1. XML is the future of the Internet. It becomes the standard form for representing and exchanging data over the web. Its syntax has already replaced the Hypertext Mark-up Language (HTML).
2. The structure of XML documents can answer the *content-and-structure (CAS)* queries which enable users to specify queries more precisely than the traditional *content-only (CO)* queries.
3. XML documents can be converted to a Tree-like structure which makes it easy to be searched for specific information.

2.4 XML Query Languages

Several query languages for XML data have been proposed so far. Most of them came from the database community and are greatly inspired by the standard database query languages (i.e., SQL and OQL) [1]. Figure-1 shows two main types of XML query languages can be recognized. The traditional type which lacks most IR-related features [6]. While the other type, the IR-XML type, depends on these features, which are weighting and ranking, relevance-oriented search, data types with vague predicates, and semantic relativism.

2.5 XML-IR Approaches:

At least two approaches can be devised for applying IR techniques to querying XML documents. [1, 8]

1. Disregard tag information to obtain a simple text document, and then apply the IR techniques. Although this is a simple approach but it has several disadvantages since, removing tags can lead to lose a lot of important information that can be used to improve the querying process.

2. Considering both textual and tag information. In this case, the IR models have to be extended to cope with both of this information. Several approaches have been defined to achieve this target, such as XXR, XIRQL, ELIXIR, ApproXQL, and XIRS [1].

3. The Survey

More than 300 papers presented in different international conferences over the last five years have been examined in order to trace the shift in research interests toward XML-IR domain. SIGIR, CIVR, HT, and JCDL/DL are the conferences that have been examined in this study.

As shown in figure-2, the great portion of the examined papers had focused on XML retrieving techniques and methodologies, such as: improving ranking and indexing in order to answer vague user queries, solve the overlapping problem, expanding the existing IR models to deal with (CAS) queries, etc. Over 28 percent of the examined papers had focused on text retrieval. This includes electronic libraries and journals, geographic information, health care information, and learning.

Since multimedia files can be represented as XML documents, 20 percents of the papers had focused on representing, searching and retrieving information from multimedia files such as video, image, sound…etc. The rest of these papers were talking about XML query languages either by analysing and improving some existing languages or by creating new ones.

Figure-3 illustrates a clear growing trend has occurred in query languages during the last five years. While a gradual growth is shown in the areas of XML-IR techniques an methodologies, and multimedia retrieval since the year 2007. A decline is shown clearly in the area of text retrieval during period 2007-2008.

4. Conclusion

There is a growing momentum behind XML starting from its beginning until now. It is now used to tag all kinds of documents including images, video, audio, and multimedia. XML is also used for music, maths, and e-business documents. While the first XML-IR systems were established base on retrieving text only, many other algorithms and methods have since been added to them to improve their results.

Our study shows that about half the examined papers during the last five years are related to XML-IR techniques and methodologies. We can conclude that the major interest of the researchers is in this area. Another finding is the trend of interest in the area of XML-IR techniques and methodologies is fluctuated since the peak point is occurred during 2006, followed by a clear decline in 2007 and a promising boost is appearing in 2008. This shows that the XML-IR area is regaining the interest again.

The study shows that XML-IR topics such as ranking, filtering and query formulation are still hot topics. This reveals that XML-IR have many unsolved and research-interesting areas.

References

[1] Ajantha Dahanayake, Waltraud Gerhardt, Web-*Enabled System Integration: Practices and Challenges*, IDEA Group Publishing, 2002.

[2] Tittel, Ed. *Schaum's Outline of Theory and Problems of XML.* Blacklick, OH, USA: McGraw-Hill Trade, 2002.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. PEARSON Addison Wisley, 1999.

[4] Van Rijsbergen and C. J. *Geometry of Information Retrieval*. West Nyack, NY, USA: Cambridge University Press, 2004.

[5] Kowalski, Gerald. *Information Storage and Retrieval Systems: Theory and Implementation.* Hingham, MA, USA: Kluwer Academic Publishers, 2000.

[6] Norbert Fuhr and Kai Grobjohann, *XIRQL: A Query Language for Information Retrieval in XML Documents*, SIGIR'01*,* New Orleans, Louisiana, USA 2001.

[7] Benz and Brian. *XML Programming Bible*. Hoboken, NJ, USA: John Wiley & Sons, Incorporated, 2004.

[8] Simon, Solomon H. *XML Business Applications*. Blacklick, OH, USA: McGraw-Hill Professional Book Group, 2001.

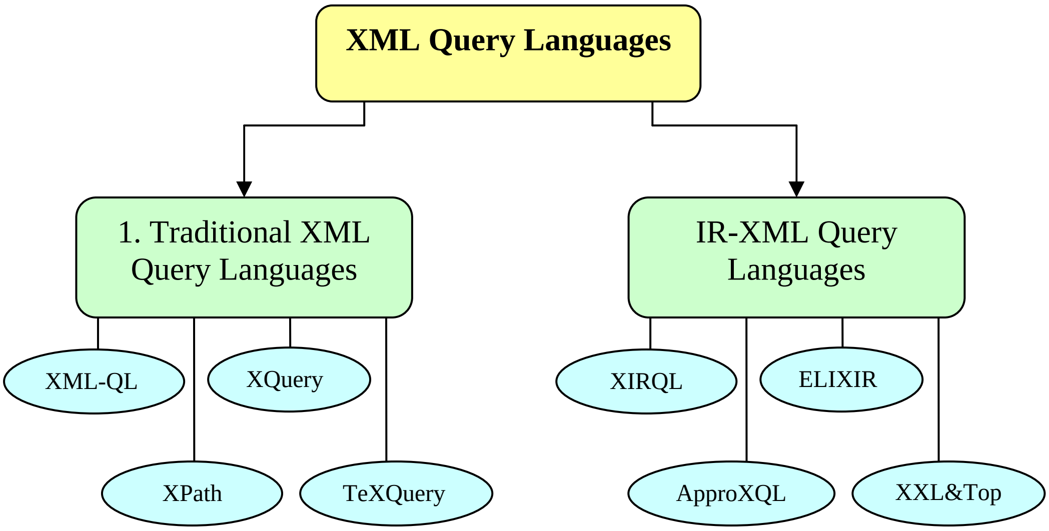

Figure-1 Taxonomy of XML Query languages

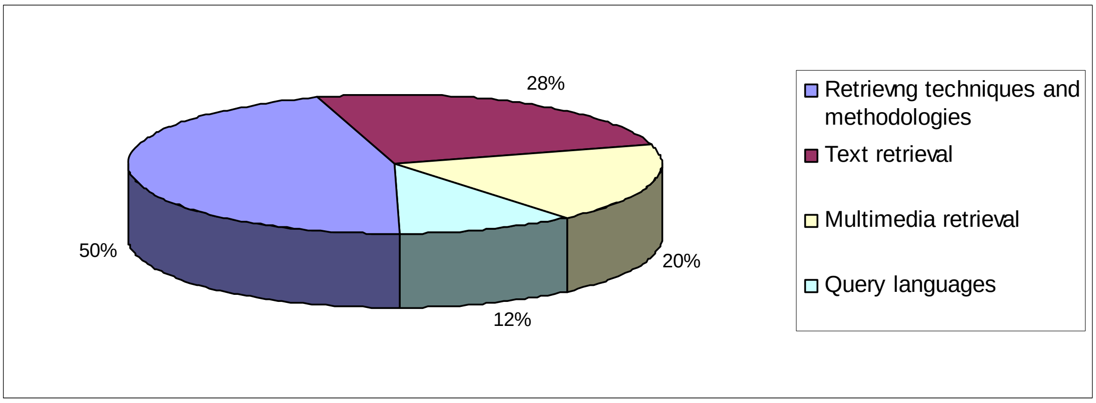

Figure-2 Major Topics Published in SIGER, CIVR, HT, and JCDL/DL from 2004 to 2008.

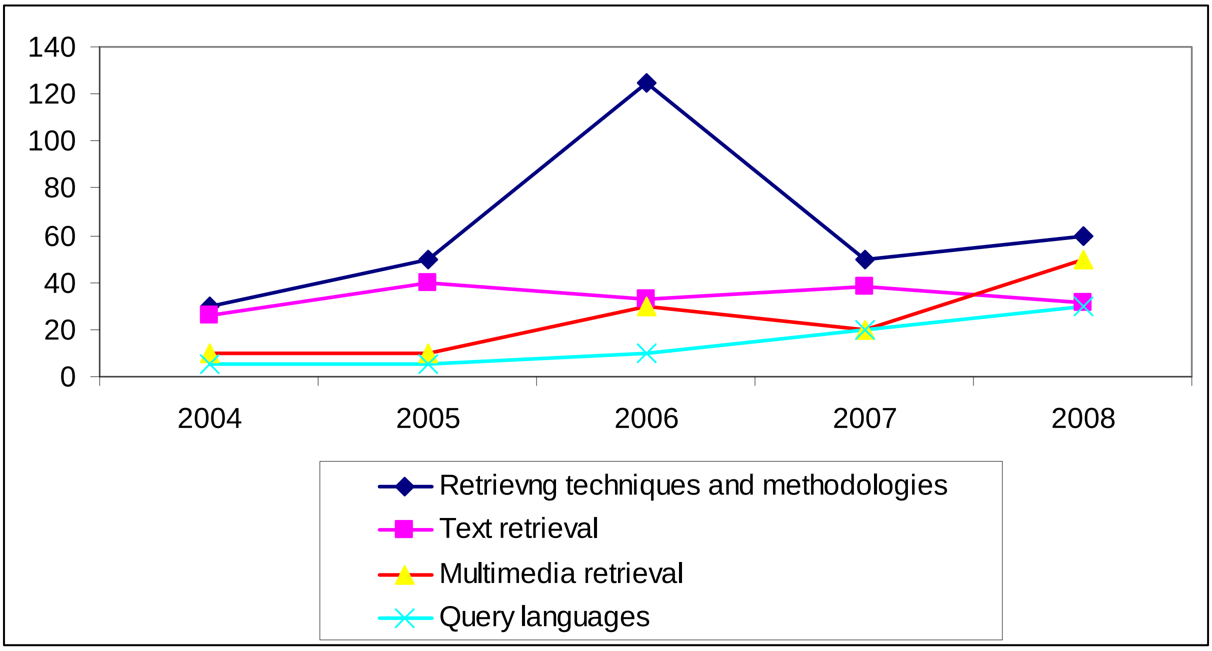

Figure-3 Subjects of Research Papers Published in Different Years