XML Representation of Web Document used by Search Engine

Payal Kansal1, Mukesh Rawat2

1 M.tech student at M.I.E.T College, Dr. A.P.J. Abdul Kalam Technical University (APJAKTU) Uttar Pradesh, India
2 Assistant Professor at M.I.E.T, APJAKT University, India

Abstract- Information retrieval is the part of Computer Science that studies the recovery of information from a gathering of written documents. Searches can be based on either full-text or other content-based indexing. The retrieved documents aim at satisfying a user information requirement usually expressed in natural language. XML is a highly structured language and used to represent the logical structure of a document. This structural nature of XML gives the user of a XML retrieval system the ability to issue more complex and precise queries than those used in traditional flat (unstructured) document retrieval. Users can make use of the structural nature of XML documents to restrict their search to specific structural elements within the XML document collection.

Keywords- Information Retrieval, Structured Data, Semi-Structured Data, Data Centric, Indexing.

I. INTRODUCTION
Information retrieval can be characterized as strategy and procedure of looking, recouping and translating pertinent data from huge accumulation of information [2]. The World Wide Web is a huge dispersed computerized data space. The capacity to look and recover data from the Web productively and adequately is an empowering innovation for understanding its maximum capacity. This can be done by using information retrieval system [1].

In today's era of Internet and Web, immense measure of data in different structures and distinctive subjects are accessible to client. The accessible data can be separated into three classifications: structured, unstructured and semi-structured. Information retrieval system customarily recovers data from unstructured content which has no identifiable structure [2,3]. For structured content, i.e., information having identifiable structure, XML retrieval is utilized.

XML recovery is content-based recovery of structured documents with XML. Xml information retrieval system intends to execute a more engaged recovery technique and return document parts, so-called XML components [5], rather than complete document, in light of a client's query.

II. STRUCTURED RETRIEVAL
Records can be structured or unstructured. Unstructured reports have no settled pre-characterized format, while organized archives are generally sorted out as per an altered pre-characterized structure [5]. A case of structured archive is a book composed into parts, each with section made of passages, paragraphs, etc. The term Structured information refers to data arranged in an all-around justifiable way [4]. These days, the most widely recognized approach to organize structured content is with the W3C standard for data vaults and trades, the eXtensible Mark-up Language (XML).

Utilizations of structured retrieval incorporate digital libraries, patent databases, web journals, text in which entities like people and areas have been labeled and the output from office suites like OpenOffice that save documents as marked up text [6]. In these applications, we need to have the capacity to run queries that combine textual criteria with structural criteria. Besides, structured text retrieval plans to exploit the accessible structural information to return content sections (e.g. XML components) instead of whole content archives.

III. DATA CENTRIC XML
Based on the content of XML records, they are arranged into two classifications: data centric and document centric [7]. Data centric XML archives are highly structured though document centric XML are inexact structures and contain extensive content segments.
In object oriented databases, one regularly discovers semi-structured data.

V. XML CONCEPTS

Extensible Markup Language (XML) is utilized to depict information. The XML standard is an adaptable approach to make data designs and electronically share organized information by means of the general population Internet, and in addition by means of corporate systems [7].

XML information is known as self-portraying or self-characterizing, implying that the structure of the information is implanted with the information, in this way when the information arrives there is no compelling reason to pre-construct the structure to store the information; it is progressively comprehended inside the XML [8].

The fundamental building piece of a XML archive is an element, characterized by tags. An element has a start and a closure tag. All elements in a XML archive are contained in an outermost element known as the root element. XML can likewise support nested elements, or elements inside elements. This capacity permits XML to support hierarchical structures [1]. Element names depict the content of the element, and the structure portrays the relationship between the elements. XML likewise supports the capacity to characterize attributes for elements and portray characteristics of the elements in the first place tag of an element [1].

XML’s energy dwells in its straightforwardness. It can take expansive lumps of data and merge them into a XML record- important pieces that give structure and association to the data [7].

A. Tree representation of XML-

A XML report has a self-descriptive structure. It shapes a tree structure which is alluded as a XML tree. The tree structure makes it simple to depict a XML archive [5]. A tree structure contains root component (as parent), child component, etc. It is anything but difficult to navigate all succeeding branches and sub-branches and leaf nodes beginning from the root [5,6].

<?xml version="1.0"?>

<addresses>

<person>

<lastname>Sharma</lastname>

<firstname>Neha</firstname>

<phone>09484295133</phone>

</person>

</addresses>

Example of data centric xml document

Data-centric XML is XML in which occasions display a uniform distribution of element types. With regards to information retrieval, we are just inspired by XML as a language for encoding content and archives. The widespread utilization of XML is to encode non-text data [8]. For instance, we might need to export information in XML design from an enterprise resource planning system and afterward read them into an examination program to create diagrams for a presentation. This kind of use of XML is called data centric [6].

IV. SEMI-STRUCTURED RETRIEVAL

Semi-structured data can be characterized as information that might be sporadic or fragmented, and whose structure may change quickly and unusually.

Semi-structured data is a type of structured data that does not accommodate with the formal structure of data models connected with relational databases or different types of data tables, yet regardless contains labels or different markers to isolate semantic components and uphold hierarchies of records and fields inside the information. In this manner, it is otherwise called self-describing structure [3,4].

In semi-structured data, the elements having a place with a similar class may have diverse attributes despite the fact that they are gathered together [7]. Semi-structured data is progressively happening since the approach of the Internet where full-content reports and databases are not by any means the only types of information any longer and diverse applications require a medium for trading data [9].
Structured document retrieval principle: A framework ought to dependably recover the most particular part of a report as the answer to the query [2].

Second challenge in structured retrieval is requirement for indexing unit. Parallel to the issue of which parts of an archive to come back to the client is the issue of which parts of a report to index [4]. In structured retrieval, there are various diverse ways to deal with characterizing the indexing unit.

A few indexing techniques for XML retrieval have been developed as follows:

A. Group hubs into non-overlapping pseudo reports.

In the beneath case, books, chapters and sections have been assigned to be indexing units, but without overlap. For instance, the furthest left dashed indexing unit contains just those parts of the tree ruled by book that are not already part of other indexing units [4,6]. The disadvantage of this approach is that pseudo-documents may not sound good to the client since they are not coherent units.

B. Top-down indexing-

We can likewise utilize one of the largest elements as the indexing unit, for instance, the book element in a gathering of books or the play element for Shakespeare's works [7]. We can then post process indexed lists to discover for every book or play the sub element that is the best hit. Lamentably, this process neglects to give back the best sub element for some questions in light of the fact that the pertinence of an entire book is frequently not a decent indicator of the pertinence of small sub elements inside it.

C. Bottom-up indexing-

Rather than recovering substantial units and recognizing sub elements(top down), we can likewise search all the leaves, select the most
significant ones and afterward stretch out them to bigger units in post processing (bottom up) [3]. This approach has a comparable issue as the last one, i.e., the importance of a leaf element is regularly not a decent indicator of the pertinence of element it is contained in.

D. Index all elements-

The minimum prohibitive approach is to index all elements. This is additionally risky. Numerous XML components are not relevant query results [7]. Likewise, indexing all elements implies that query items will be exceedingly repetitive.

VII. CONCLUSIONS

It is most effortless to seek a keyword in a XML representation of an archive [4]. By this, we can retrieve the data as per our need. For instance, if the client needs to bring the address data, he/she will recover the content within the addressed tag. For recovering the data inside the tags, different XML parsers are available.

REFERENCES