Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6ᵗʰ & 7ᵗʰ March 2014

An Effective Energy and Latency Of Full Text Search Based On TWIG Pattern Queries Over Wireless XML

A.Rajathilak, A.Lalitha

PG Student, M.E(CSE), Valliammai engineering college, Chennai, Tamilnadu, India[1]

Asst prof , Dept of CSE, Valliammai engineering college, Chennai, Tamilnadu, India[2]

ABSTRACT: In wireless dissemination scheme, major issue is based on the retrieval of requested text search over internet and power is much requires to traversal and acquires it. Our previous work mainly concern about transmission of XML data to the client using lineage encoding and it must be decode in the client side then get the required and selected data from the broadcasting network. Performance evaluation can be attain by comparison of data set and stored. Major issue are communication is unstable in the wireless broadcasting environment hence it is difficult to identify package loss and tail drops. To resolve this conflict, index mechanism is to be use for acknowledgment for identify package loss in the wireless network. Depth-first traversal of elements improves the access time for the specific query. Introducing preprocessor in the server side that enhance the pattern mining by implementing stop word removal, stemming algorithm and part of speech tagger to improve mining in the document. We also suggest holistic twig and multipredicate merge join algorithm improves the XML parsing and mining.

KEYWORDS: stemming algorithm, tail drops, XML

I. INTRODUCTION

    With the rapid development of wireless network technologies, users with mobile devices can access a large amount of information at anytime from anywhere. Mobility and portability have created an entire new class of applications. For example, information services, including news, stock quotes, airline schedules, weather report and traffic information, are becoming more and more popular and helpful. [1]

Over the past decade, users communicate in the wireless mobile environment using their mobile devices such as smart phones and laptops while they are moving around various places. In broadcasting approach mobile user suffers utilization of bandwidth and energy consumption for various processes. In the same manner wireless communication is one of the most efficient schemes for the information exchange than any other recent approach and it also used in research, government, railways, telecommunications, etc.

Some of the major conflict arises in the wireless network approach are security-authenticate users can only access privilege data, reliable-handle multiple clients at a time. These problems can be minimizing by using evolving security algorithms and various mechanisms.

In our approach, XML repository collects the XML data and stores multiple XML documents. Using XML parser to read the XML value and make them to utilize by DOM parser and XPath. Server monitoring parsing function and XML value reader is done then the lineage encoding is introduced. Lineage encoding is one of the light weight efficient encoding scheme that convert XML into binary format (0,1) and transmit to the wireless dissemination scheme for clients. It improves authentication among the transmission and helpful to attain more secure data transfer.

II. RELATED WORKS

A. Energy and Latency Efficient Access of Wireless XML Stream

Distributed index structure and a clustering strategy for streaming XML data that enables energy and latency- efficient broadcasting of XML data. Initially, DIX node structure to implement a fully distributed index structure which contain the XML tag name, tag attributes, and its text content of an element, and also its corresponding indices. Exploit the index information in the DIX node stream, a client can access it with shorter latency. Implement a method of cluster the DIX nodes in the stream, which can provide enhance the performance of query processing in the mobile clients [2].
This approach have some of the disadvantages, indexing mechanism doesn't support larger data. Clustering of multiple nodes is helpful but it makes lesser efficiency.

B. A Novel Three-Phase XML Twig Pattern Matching Algorithm Based on Version Tree

Structure index method, node-based encoding method and sequence method are the common solution to querying and searching XML data that the execution time as well as the input size of algorithms grows rapidly as the size of XML document increases. To minimize this problem, a new 3-phase XML twig pattern matching algorithm called Twig3Version which executes holistic XML twig pattern matching algorithm on the structure index named Version Tree that compresses all repetitive structures in XML document, and returns sub trees of Version Tree that matching query twig in structure. Algorithm implements a trouble-free and efficient version filter module on the concise intermediate results to find matching versions. Then, it merges elements in the original document corresponding to these matching versions to generate final results [3].
This approach also have some of the disadvantages, the performance of Twig3Version may be poor in a special situation that the structure of XML documents is complex and query almost contain ancestor-descendant relationship and each query node has a high frequency in original document.

C. Xmill: an efficient compressor for xml data.

XMill compresses XML data by separating it into three components: The element and attribute names, the text values, and the tree structure of the XML document. The text values are grouped by parent element name and the three components are then compressed using standard text compression techniques. The disadvantage in this approach is the new algorithm outperforms our existing approach.

III. PRELIMINARIES

A. Twig Pattern Queries

It takes input as a set of XML documents with nodes annotated by the user. Decompose twig into linear patterns and find all matching of each linear pattern then merge them to produce the result. STEMMING ALGORITHM. A stemmer function is to identify the string as based on the root. Example: argues, argued, argue are reduced to the stem "argu".

IV. ORGANISATION OF THE PROJECT

The remainder of this paper is organized as follows: Chapter 5 presents the PREVIOUS WORK which provide the details of the existing system and its disadvantages and the system to be proposed in the earlier. In Chapter 6 represent the PROBLEM DEFINITION consists exact definition for problem and key idea to resolve it. CHAPTER 7 which holds the propose scheme and algorithm and CHAPTER 8 has System Architecture with their description.. Finally in Chapter 9 CONCLUDE the proposed work and discussing about the Future work.

V. PREVIOUS WORK

In our previous work, XML data is taken to the repository that stores multiple XML contents and it can provide ease of access for further server requests. Server can request and acquire the respective XML data from the repository it act like as a database to the entire system. Generally, XML data to be parses for better understandable by using XML parser like SAX parser. Generating, a streaming unit of a wireless XML stream, called G-node structure which is eliminates structural overheads of XML documents and enables mobile clients to skip downloading of irrelevant data during query processing. Using XPATH as a query language that can be improve by implementing structure indexing and attribute summarization. Then wireless XML stream is generated from the server side.

It can be transfer to the broadcasting environment using efficient encoding scheme called Lineage Encoding. It consists of vertical code [V] and horizontal code [H], which transfer entire wireless streaming XML data into client side in the BINARY format. Streaming XML data from the broadcasting environment can be decoding using lineage code in the client side. Simple path query processing algorithm and Twig pattern query processing are used to construct a query tree its function is to identify relevant data from the corresponding wireless stream. It allows downloading with the interests of client and also provides skip downloading. Pattern matching to be carried out for better XML stream then the energy and latency efficiency attained and proved by performance evaluation with the help of real and synthetic data sets [4].

VI. PROBLEM DEFINITION

Depth-first traversal of elements increases the access time for specific queries. Communication is not stable in wireless broadcasting environment, so it is difficult to identify the packet loss and tail drops which can be overcome by using the indexing mechanism.

VII. PROPOSED SCHEME

A. Xml Value Reader

An XML document can be access from the XML repository and it can be represented as a rooted, ordered, and labeled tree which has elements, attributes, and texts are represented by nodes, and the parent-child relationships are represented by edges in the XML tree. It shows a simple XML document that will be used as a running example in the paper.

Here XPath is used as a query language and its results of XPath query are selected by a location path which consists of location steps. Dispensation each location step and selects a set of nodes in the document tree that meets axis, node test and predicates.

B. Preprocessor

Preprocessing consists of Stop word removal, Stemming (Porter Stemmer Algorithm) and Part of speech tagger. Here we are using affix removal stemmer.

1) Stemming algorithm:

A consonant will be denoted by c, a vowel by v. If a list contains ccc... of length greater than 0 will be represent as C, and a list vvv... of length greater than 0 will be represent as V. Therefore, any word should has one of the following four forms: CVCV ... C, CVCV ... V ,VCVC ... C ,VCVC ... V

These may all be represented by the single form [C]VCVC ... [V] where the square brackets denote arbitrary presence of their contents. Using $(VC)^m$ to denote VC repeated m times, this may again be written as $[C](VC)^m[V]$.m will be called the measure of any word or word part when represented in this form. The case m = 0 denotes the null word in the document. Here are some basic algorithm and steps:

m=0 TR,EE,TREE,BY.

m=1 TROUBLE,OATS,TREES,IVY.
M=2 PRIVATE,OATEN,ORRERY.

ALGORITHM

Step 1a
SSES-SS caresses-caress
IES-I ponies-poni
Step 1b
 $(m>0)$EED-EE feed>feed
Step 1c
$(*Y*)$-I happy-happi
Deals with plurals
BILITIES-BLE sensibilities-sensible

Step2: The test for the string S1 can be made fast by doing a program switch on the penultimate letter of the word being tested. This gives a reasonably even breakdown of the possible values of the string S1. It will be seen in fact that the S1-strings in step 2 are presented here in the alphabetical order of their second last letter. Similar techniques may be applied in the other steps.

Step 3
        $(m>0)$ICATE-IC triplicate-triplic
        $(m>0)$ALIZE-L formalize-formal
Step 4
        $(m>1)$IVE- effective-effect
        $(m>1)$IZE- bowlderize-bowlder
Step 5
        $(m>1$ and $*d$ and $*L)$ control-control

    Stop word removal and speech tagger are used in the preprocessor which are responsible for removing grammar errors and analysis part of speech [5].

C. Lineage Encoding
    Lineage encoding is a simple and efficient encoding scheme which is responsible for secure transmit and receive XML document over the network. In server   side output of stemmer function taken to the lineage code which have vertical code[V] and horizontal code[H] to convert XML document into binary format.

D.Client query tree formation and retrieval
    In client side frequent pattern mining and matching to be done by constructing query tree formation and retrieve the matching document.

1)Algorithm 2.
    Twig Path stack Query Processing algorithm implemented as,
    Input: Wireless XML Stream DS, a twig pattern query Q
     Output: Result set R satisfying Q
    begin 01: initialization

02: Initialize the selection bit string SB as 1;

03: Initialize Lineage Code of the root G-node as (1, (1));

04: Initialize nextNode as the address of the root G-node in DS;

05: Tree traversal phase

06: Construct a query tree T for Q;

07: Tune a group descriptor GD of the G-node indicated by next Node; IF (current node CN is the leaf node in TÞ THEN

08 Store AVL and TL the node in T; ELSE IF (CN contains predicate conditions PÞ THEN Tune the relevant attribute values and/or text using AI   and/or TI; Store the relevant attribute values and/or text into the node in T;

09:Assing the address of the next node in CI to next Node; END IF Let N be the highest branching node inT;

10: Let C be the child node of P in MP;

    SelectChildrenðC, C is the leaf node in set R of elements   in C using the selection bit string

Return R; end

Twig query processing improving the client side query tree formation and retrieval the selected text from the broadcasted environment. It also provide content search in the disseminated scheme by using pattern matching technique [5].
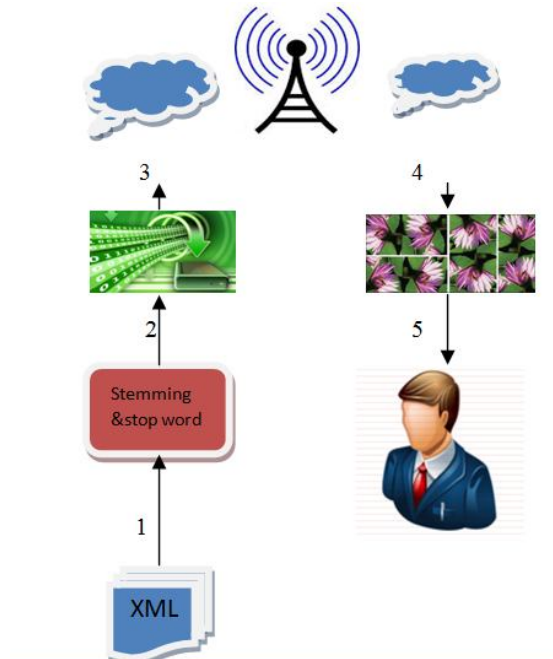
VIII. ARCHITECTURE DIAGRAM

Fig8. 1 architecture diagram

In fig8.1, it shows main architecture for this project which consist of following process, (1) where the XML document can be store in to repository. It can be parses using SAX parser that fed into preprocessor.

Preprocessor (2) has three main function stop word removal, part of speech and stemming algorithm which are responsible mining XML document depends upon the affix removal stemmer to refine the input document. Lineage code (3) is used to encode the wireless XML stream into binary format that can be disseminated to the network. Frequent pattern

matching (4) and client tree retrieval that retrieves the requested text-tree from the network. Client (5) can receive the search text.

IX. EXPERIMENTAL EVALUATION

By implementing stemming algorithm that provide better evaluation then previous work. It improves the XML mining and provides better results.

Table 8.1   Stemming Retrieval Effectiveness Studies Summary

Study	Question	Test Collection	Dependent Vars	Results
Salton	Full stemmer vs. suffix s stemmer	IRE-3	14 DV's	Full > s
		ADI		Full > s
		Cranfield-1		s > Full
Hafer and Weiss	HW stemmer vs. SMART stemmer	ADI	recall(R), precision (P)	HW > SMART
	HW stemmer vs. manual stemming	CRC		HW = Manual
	stemmed vs. unstemmed	ADI CRC		Stemmed > Unstemmed
VanRijsbergen et al.	Porter vs. Dawson stemmer	Cranfield-1	R,P	Porter = Dawson
Katzer et al.	stemmed vs. unstemmed	CCA 12,000	R,P(highest) R,P(all)	stemmed = unstemmed
	stemmed vs. other reps.			stemmed > title, descriptors

Stemmers are used to conflate terms to improve retrieval effectiveness and /or to reduce the size of indexing file. It will increase recall at the cost of decreased precision. Stemming can have marked effect on the size of indexing files, sometimes decreasing the size of file as much as 50 percent

X. CONCLUSION

In this paper, we consider the XML data broadcast problem. An affix removal stemmer algorithm is proposed to pre-process the original XML documents before broadcast scheduling, and it can mine about 50% element node on average. Because the stemmer algorithm processing is transparent to mobile clients, no modification is needed for the data access protocol on the client end. Air indexing is used to improve the tuning time.

REFERENCES

[1]   Yongrui Qin, Hua Wang and Lili Sun "Cluster-Based Scheduling Algorithm for Periodic XML Data Broadcast in Wireless Environments" IEEE INTL Conference on Advanced Information Networking and Applications, vol.21, pp855-860,June 2011.
[2]   Y.D. Chung, S. Yoo, and M.H. Kim, "Energy- and Latency-Efficient Processing of Full- Text Searches on a Wireless BroadcastStream," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 2,pp. 207-218, Feb. 2010
[3]   Guiquan Liu,Meiling Yao,Desheng Wang and Enhong Chen, " A Novel Three-Phase XML Twig Pattern Matching Algorithm Based on Version Tree" IEEE INTL Conference on Fuzzy Systems and Knowledge Discovery, vol.21, pp1678-1688,Jan 2011.
[4]   Jun Pyo Park, Chang-Sup Park, and Yon Dohn Chung, " Lineage encoding- an efficient wireless XML streaming support Twig pattern queries" IEEE transactions on knowledge and data engineering, vol. 25, no. 7, july 2013.
[5]   www.wikipedia.com