Temporal Structuring of Web Content for Adaptive Reduction in Obsolescence Delivery over Internet

(Global content delivery platform for knowledge creation and dissemination)

Mithilesh Kumar Mishra,
Research Scholar, UP Rajarshi Tandon Open University,
Allahabad, UP, INDIA-211013,
mmishra@iiita.ac.in

Dr. Anurika Vaish,
Assoc. Professor, Indian Institute of Information Technology,
Allahabad, UP, INDIA-211012,
anurika@iiita.ac.in

***Abstract*:-** With increased cyber penetration in the civil society and popularity of web services, web sites have now been seen as web based Information Systems [1]. The ever evolving web technology has increased flow of information on Internet many folds. Simultaneously flow of web obsolescence [2] has also increased with time. This is mainly due to more and more temporal contents are getting pumped into World Wide Web (WWW) and content authoring as well as content delivery platforms lack adequate mechanism to define, detect and filter obsolete contents. Since HTTP [4] transfer of web contents between two nodes on Internet takes place in the form of HTML documents [4], proper temporal structuring of HTML documents for defining age of web content may enable delivery platforms to detect and filter obsolescence in the under-delivery web content automatically. Presently this can be achieved using server-side scripts [5] by dynamically generated HTML documents, which is difficult for naïve users who can hardly work in HTML. Present paper addresses this problem of naïve user by proposing extended attribute sets of HTML TAGs [6] that can be used in web authoring using HTML. In this paper two new attributes 'pubDate' for date of publish of web content and 'expDate' for date of expiry of web content have been proposed with a simple syntax for defining age of web content i.e. life span, at the time of web authoring in HTML. Paper also demonstrates auto detection and filtering of web obsolescence in the delivery of static HTML documents by proposed design of content delivery platform. Paper highlights value additions that proposed concept offers for various existing delivery platforms.

Keywords: *Web Content; Web Obsolescence; Content Aging; Obsolescence Filtering; Web Authoring.*

*****

I. INTRODUCTION

Until the last decade of past century the structure of WWW was remarkably simple: it was described as a collection of web servers and browsers. The concept of multimedia content documents linked to each other over a network across the globe was so attractive, that since mid-ninety's, an unprecedented growth of WWW has been witnessed, both in terms of content and usage [7]. This led to enormous contentization over Internet, heavy transfer of web content over Internet and increase in web content obsolescence.

A study shows 13% of the total URL referenced, only 77% remained valid, 37% were found broken link, however, 29% were found inactive URLs i.e. invalid links or obsolete links just within 2 years as observed by Kitchens and Mosley (2000) study cited by Wales (2005) [8].

To control delivery of web obsolescence on Internet, we need to look into web authoring practices using HTML and existing mechanism of content delivery platforms. Hyper Text Markup Language, as we know, facilitates creation of static web pages using a set of commands known as HTML tags. These HTML tags mainly handle data presentation inside the web browser are classified in two categories

Unary Tags (only opening tag) and Binary Tags (a pair of opening and closing tags)

All HTML tags [6] possess a set of attributes for defining style of data presentation inside the web browser. Simple syntax of HTML tags for defining various aspects of content presentation on web pages is as below:

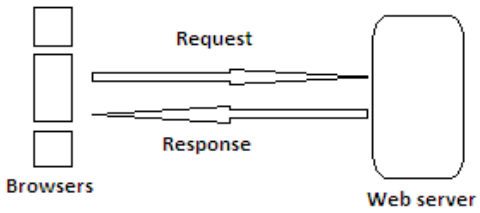
<**tag_name**
attribute_name="*attribute_value*"> Content

in case of unary TAG, however,

<**tag_name**
attribute_name="*attribute_value*">
Content
</**tag_name**>

in case of binary TAG.

Normal sequence of flow of HTML content between server and client over Internet [9] can be depicted as



(Fig. 1) HTTP Request/Response

 ISSN: 2321-8169
470 - 475

HTTP Request of browser for web content
in the form of Universal Resource Locator (URL)

Parsing of URL by web server for requested document
Filename (HTML file)

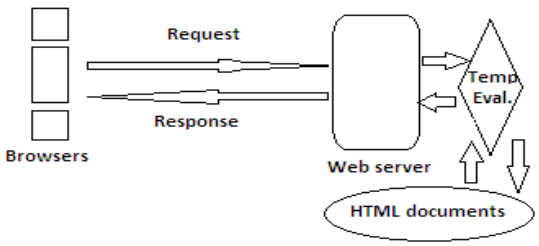
Locating document in the web root as per MIME type
Static files are stored in the web root of web server

Opening requested file if MIME type is TEXT/HTML
In the memory of web server

Reading contents of HTML file sequentially
Using file handler by the web server

Sending file as response to the web browser
Writing content sequentially through server socket on web server

Here it is evident that web server reads file content after opening it. Our proposed system integrates a Temporal Evaluation (TE) module at this point in the above sequence. The TE module parses web content, prior to delivery, for their validity based on the values of 'pubDate' or 'expDate' defined by the web author also know as web master. It is pertinent to mention that only web authors are the right persons who know the life span of web content presented on their web pages.



(Fig. 2) Proposed System

II. OVERVIEW OF THE SYSTEM

In the proposed framework, temporal attributes 'pubDate' and expDate' are used in the HTML tags for defining life span or age of web content enclosed by any HTML tag. For example, a paragraph text can be assigned age using <P> tag as below:

<P **pubDate**="value" **expDate**="value">
Content of paragraph goes here…
</P>

Here 'pubDate' defines date of publish of <P> tag and 'expDate' defines date of expiry of <P> tag. Date values can be a valid date using 24-hour standard format. In our demonstration we have used standard date format as

"DD-MMM-YYYY"

using a 24-hour clock with following notations:

DD – two digits number of day i.e. 00-31
MMM – three letter name of month i.e. Jan…Dec
YYYY – four digits number of year e.g. 2014, 2015

For defining content that does not expire, we can omit either 'pubDate' or 'expDate' attributes or simply we can use a null value for these dates as below:

<P **pubDate**="" **expDate**="">
Knowledge content of paragraph goes here…
</P>

Otherwise we can completely omit use of temporal attributes to define as same paragraph as below:

<P>Knowledge content of paragraph goes here…</P>

In our present work, contents without age or expiry are referred as Knowledge Content (KC) [10] or Perpetual Content (PC) [11]. However, contents with definite age or definite expiry are referred as Information Content (IC) [12].

Similarly proposed attributes can be applied to any HTML tag in its opening tag to define date of publish and date of expiry of piece of content enclosed with particular tag. For example

<A href="www.google.com" pubDate="01-Jan-2015" expDate="30-May-2015">Google Home</A>
Or

<DIV bgcolor="orange" pubDate="01-Jan-2015" expDate="30-May-2015">Content under this division goes here…</DIV>
Or

<SPAN pubDate="01-Jan-2015" expDate="30-May-2015">
Inline content under this tag goes here…
</SPAN>

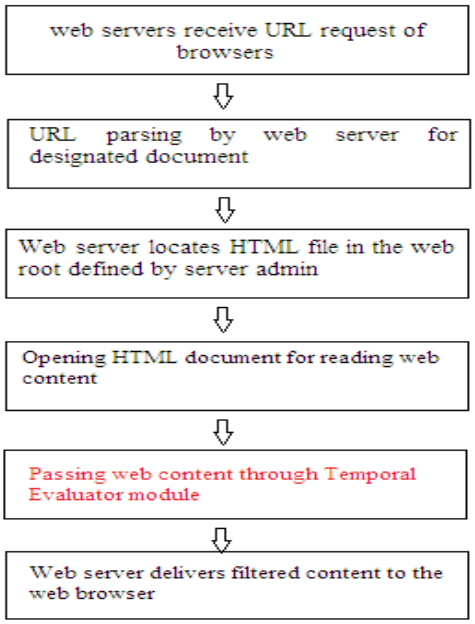
In the present work it is assumed that all the paired tags in a HTML document is properly closed as per standard of HTML 5 [13]. The improper pairing or incomplete pairing of tag leads to improper presentation of content inside web browser.

III. IMPLEMENTATION OF THE SYSTEM

For the sake of demonstration of the proposed system, it is implemented using PHP [14] socket programming for running a sample web server. The Temporal Evaluation module is also implemented using PHP string processing library functions. A simple block diagram of proposed system is presented below that depicts integration of TE module in the sequence of normal web content delivery process. This server

 ISSN: 2321-8169
470 - 475

runs over system IP address [15] and port number 80 [16] to listen HTTP request of browser. It throws an error if requested HTML documents does not exist in the designated web path i.e. wwwroot [17].

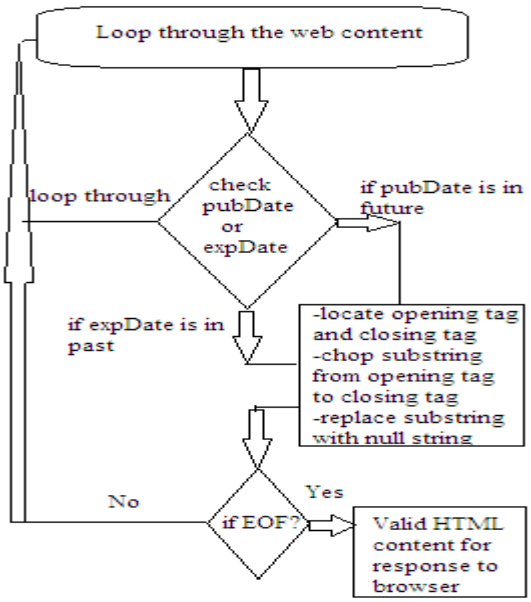


(Fig. 3) Process flow diagram of proposed system

The TE module scans through web content for 'pubDate' and 'expDate' specified with any paired tags in the HTML document. The corresponding date values of these two attributes are extracted from the tag and it is compared with the system date/time depending on the values. Once TE determines 'expDate' is from past, positions of opening and closing of container tag are identified. At the end substring from the position of opening tag till closing tag is replaced with null character. Similarly if value of 'pubDate' is from future, positions of opening and closing of container tag are identified to chop the substring from the delivery of web content. Here various cases may arise as below:

- Value of 'pubDate' is missing
- Value of 'expDate' is missing
- Value of 'expDate' may be specified less than the value of 'pubDate'
- Values are time only
- One value is date, other is time etc.

All such cases are covered while parsing HTML document and appropriate measures have been taken to ensure delivery of valid web content by the proposed framework. Flowchart of the framework is as below:

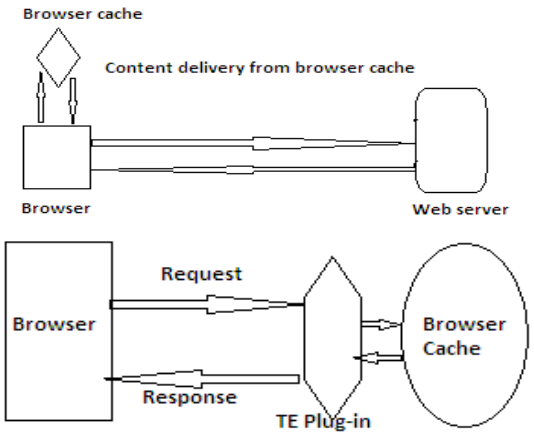


(Fig. 4) Flowchart of proposed system

The proposed framework integrates a Temporal Evaluator (TE) module in the delivery platform that parses an HTML documents for temporal attributes and compares the date/time values against system date/time.

Applications of Proposed Framework

Browser: Though web content delivery takes place mainly at the web servers; other web techniques also play important roles in delivery of web content over Internet. Browsers also use their local cache memory to act as a delivery platform for serving recently visited web pages [18] from the History.



(Fig. 5) Browser cache service

Temporal Evaluator (TE) module has been implemented as a plug-in module of web browser application. This module scans web page on load and prior to rendering its contents, all HTML tags are evaluated for 'expDate' value against the system date browser runs on. If the content is expired, it is commented out so that expired contents are

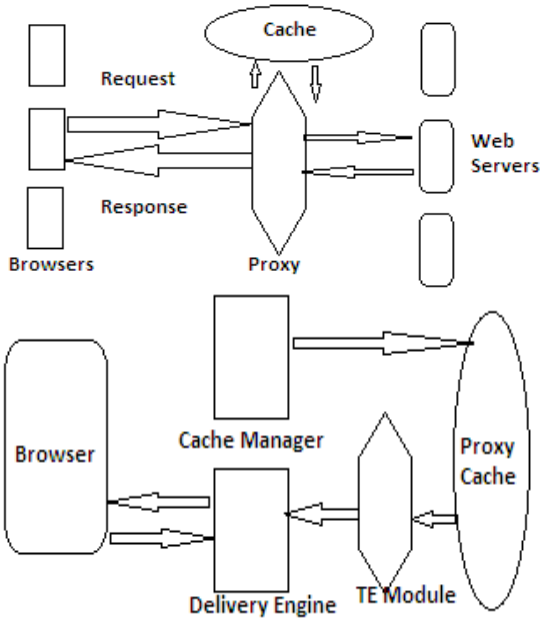
 ISSN: 2321-8169
470 - 475

invisible to the user. Standard HTML commenting methodology is used to implement the concept as below:

<!--P pubDate="01-Jan-2014" expDate="01-Jan-2015">
Paragraph content goes here...
</P-->

Proxy Server: Similarly, proxy servers also facilitate speedy delivery of web content over enterprise network through its cache management [19]. A proxy is a trusted agent that can access the Internet on behalf of its users with following main functions:

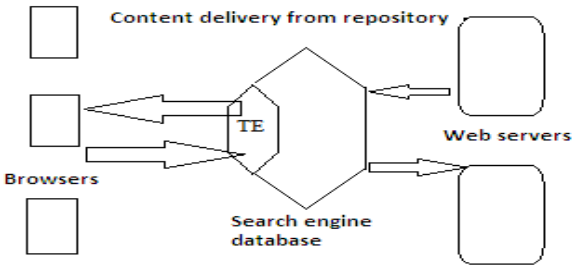
- providing access to Internet from within a local network;
- controlling the access to network recourses
- protecting individual users from network attack.
- reducing necessary bandwidth



(Fig. 6) Proxy service

Search Engine: Search engine crawls web content on Internet periodically and stores them in local database with adequate indexing [20]. This kind of intermediary short term storage facilitates cache based prompt delivery of web content; however, on the other hand it leads delivery of obsolescence from information point of view. Above figure illustrates TE module in search engine delivery mechanism for detection and filtering of obsolete content, similar to web server application.

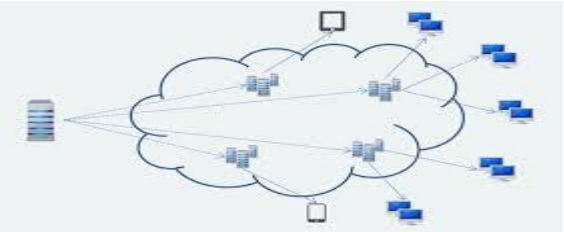
Prior to delivery of requested HTML document, TE scans whole document from <BODY> to </BODY> tag for 'expDate' attributes. If 'expDate' attribute value is specified in any HTML tag, this value is evaluated against system date, thus expired contents are filtered at the time of delivery.



(Fig. 7) Search engine service

This is useful in a situation where content is expired at source server however it is in delivery over Internet through various distributed cache servers.

CDN Server: A content delivery network (CDN) is a system of distributed servers that delivers Web Content to a user based on the geographic locations of the user, the origin of the web page and a content delivery server [22] [23]. Content delivery network comprises of several web servers for serving the same set of web content to large number of users [24] [25]. Here also proposed framework plays a significant role in reducing replication of obsolete content on both, source server as well as mirror servers.



(Fig. 8) Content Delivery Network [24]

Integrating TE module on main web server will enable obsolescence detection and filtering prior to replicating web content on mirror servers. TE module on mirror servers ensures further delivery of obsolescence free web content to the end users.

The present work also highlights several other benefits that can be achieved using proposed framework such as demand based content authoring, time line navigation, synching of content delivery network for valid information, time specific content delivery etc.

IV. ANALYSIS

For analyzing the performance of proposed content delivery platform a HTML MIME type document "test1.html" was used. On Internet, delivery of 84 KB test document took .866s with 397ms CPU utilization when obsolete content was unfiltered. However, with filtering of obsolete content of 1 KB,

 ISSN: 2321-8169
470 - 475

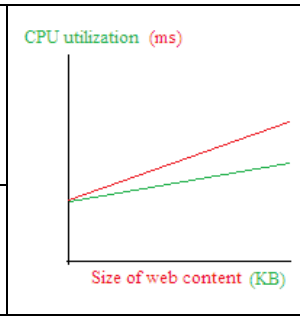
delivery of 83 KB test document took .857s with 386ms CPU utilization.

(Fig. 9) CPU utilization with & without obsolescence

It is evident that delivery of obsolescence free web document is not only .009s faster than the delivery of original document that contains obsolete content, but also CPU utilization on web server reduces by .011ms in delivering obsolescence free web content [21] [26].

Performance of the proposed delivery platform is more demonstrative in the following graph between size of obsolescence and time in delivery. CPU utilization and delivery time of different size of obsolete content have been recorded. This clearly shows that with increase in obsolescence in sample web page, CPU utilization reduces significantly.

WEB CONTENT SIZE IN KB	83	84	85	86
TE - CPU UTILIZATION IN MS	386	397	407	416



(Fig. 11) CPU utilization vs. Size of web content

CPU utilization depends on various factors related to network conditions, server hardware, running applications on web server etc. and it various in different environments [27] [28].

V. FUTURE SCOPE

1. Date time format such as "DD-MMM-YYYY HH:MM:SS" in a 24-hour clock can be used to define content age in terms of minutes, hours, days, months or years.
2. TE module can scan HTML documents of web server for periodically to regenerate temporary HTML document free of obsolete content.
3. Meta tag can be used to indicate temporal structuring of HTML documents. Accordingly TE module of content delivery platform can enable or disable for its functioning.
4. Since the obsolete content remains in the static HTML document web browsers can facilitate users a timeline navigation mechanism [18] using 'pubDate', 'expDate' attributes with respect to user's Computer Date as below:

   pubDate <= sysDate <= expdate

5. Presently 'pubDate' and 'expDate' are defined with date values, however, time values can also be used for defining age of fast moving contents.
6. With evaluation of system date/time against 'pubDate' and 'expDate' dynamic web pages can be designed even without use of client-side or server-side scripting. For example, to greet visitor, following two paragraphs can server the purpose:

   <P pubDate="00:00:00" expDate="12:00:00">
   Hello! Good Morning
   </P>

   and

   <P pubDate="12:00:01" expDate="24:00:00">
   Hello! Good Evening
   </P>

7. Browser-side plug-in module [19] can be made available in open source for popular web browsers.
8. Cache management features of proxy servers and browsers can be optimized by periodically purging the obsolete content stored in the cache memory [20] of these web tools.
9. Web authors can be acknowledged in advance by web server through mail or SMS for content obsolescence.
10. Auto feed to registered subscribers can be generated by web server for newly published web content on the web site.

VI. ACKNOWLEDGEMENT

The implementation of this framework was assisted by our two colleagues Mr. Prashant Kr. Sirvastava and Mr. Ajay Kr. Tiwari. However, other two colleagues Mr. Durgesh Kumar and Mr. Santosh Kr. Yadav played key roles in testing various cases. Their effort is heartily appreciated by authors of this paper.

VII. CONCLUSION

It can be concluded that the proposed framework leads new way of web authoring with two new attributes and new technique of content delivery by web servers that use new attributes in sensing and filtering obsolete content of the web page. Temporal restructuring of web content achieved using proposed two new attributes “pubDate” and “expDate” leads new mechanism of cache management, content updation, auto subscription feeds etc. These attributes can be applied on any HTML tag that <opens> and </closes> for defining age of web content, thus, time specific web authoring is possible under proposed framework. The concept of green web can be assured by ensuring obsolescence free web content delivery over Internet. The performance of web server with obsolescence detection and filtering feature improves with size of HTML documents.

REFERENCES

[1] Beyond HTML: Web based Information System by Chandrinos, K.V.; Trahanias, P.E.
[2] Web content aging and filtering of static HTML obsolescence by Mishra, MK; Dr. Vaish, Anurika
[3] http://en.wikipedia.org/wiki/HTML
[4] The Evolution of Video Streaming and Digital Content Delivery by M. West, Darrell
[5] http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Server-side_scripting.html
[6] http://www.w3schools.com/tags/tag_html.asp
[7] The impact of Proxy caches on Browser Latency By Andrzej Sieminski
[8] The Lifecycle of Knowledge Management System for Organizational Learning by Jennex, Murray E. (Page 282-283)
[9] Relational Flow and the World Wide Web: Conceptualising the Future of Web Content By Scott Shaner
[10] Knowledge Management through Content Interpretation by R. Jones, Richard; A. Bremdal, Bernt; S. Paggiari, Christophe; Johansen, Fred; Engles, Robert
[11] Web content in perpetual motion By Richard
[12] How is information content measured? By Batten, Don
[13] http://www.html5rocks.com/en/
[14] http://www.php.net
[15] Mining the Web and the Internet for Accurate IP Address Geolocations by Guo, Chuanxiong; Liu, Yunxin; Shen, Wenchao; J. Wang, Helen; Yu, Qing; Zhan, Yongguang
[16] http://www.papercut.com/products/ng/manual/ch-customization-enable-additional-ports.html
[17] http://www.prophoto.com/support/find-web-root-folder/
[18] Optimization in Web Caching: Cache Management, Capacity Planning and Content Naming by P. Kelly, Terence
[19] Evaluating Content Management Techniques for Web Proxy Caches by Arlitt, Martin; Cherkasova, Ludmila; Dilley, John; Friedrich, Richard; Jin, Tai
[20] Enhancing Collaborative Web Search with Personalization: Groupization, Smart Splitting, and Group Hit – Highlighting by Morris, Meredith Ringel; Teevan, Jaime; Bush, Steve
[21] http://www.webpagetest.org/result/150115_4C_JVY/1/details/
[22] http://www.webopedia.com/TERM/C/CDN.html
[23] A Case for Peering of Content Delivery Networks by Buyya, Rajkumar; Pathan, Al-Mukaddim Khan; Broberg, James; Tari, Zahir
[24] http://www.bitrixsoft.com/products/cms/features/cdn.php
[25] Active cache: caching dynamic contents on the Web. In: Proceedings of IFIP international conference on distributed systems platforms and open distributed processing, 1998. p. 373-88 [http://www.cs.wisc.edu/~cao/papers/active-cache.ps] by Cao, P.; Zhang, J; Beach, K
[26] Eureka: A methodology for measuring bandwidth usage of networked applications by Vaishnavi,I.; Centrum Wiskundeen Inf.; Amsterdam, Netherlands
[27] Delay Based Network Utility Maximization by Neely, MJ, University of Southern California,Los Angeles, CA, USA
[28] Throughput analysis of IEEE 802.11 wireless networks with network coding by Jang, Bo Kyung, Dept of Electrical Engineering, Kyung Hee University, Youngin, South Korea