METHOD AND APPARATUS FOR ADAPTING WEB CONTENTS TO DIFFERENT DISPLAY AREA DIMENSIONS

Inventor: Vincent Wen-Jeng Lue, Sunnyvale, CA (US)

Correspondence Address:
Vincent Wen-Jeng Lue
1484 Owen Sound Drive
Sunnyvale, CA 94087 (US)

Appl. No.: 12/005,589
Filed: Dec. 27, 2007

Publication Classification
Int. Cl. G06F 7/30 (2006.01)
U.S. Cl. 707/102, 707/E17

ABSTRACT

A method is disclosed to generate, while preserving text, image, transactional and embedded presentation constraint information, a minimum set of simplified and navigable web contents from a single web document that is oversized for targeted smaller devices. The method includes a parser, a content tree builder, a document tree builder, a document simplifier, a virtual layout engine, a document partitioner, a content scalar and a markup generator. The parser generates markup and data tags from an HTML source document. The builder constructs a content tree. The simplifier transforms the document tree into an intermediate one defined by a subset of XHTML tags and attributes. Layout constraints, including size, area, placement order, and column/row relationships, are calculated for partitioning and scaling the document tree into sub documents trees with assigned navigation order and hierarchical hyperlinks. A simplified HTML document is then generated with the markup generator.
Figure 1
Figure 3

Diagram showing the flow of data through a system with components labeled as follows:
- Client
- Content Divide & Condense
- HTTP Handler
- Local Cache
- Web Server
- HTML Transcoder

The system involves HTTP communication between these components.
HTTP Proxy Handler

Content Divide & Condense

HTML Transcoder

Local Cache

Web Server

Figure 4
<html><head><title>Example</title></head><body bgcolor="blue" text="yellow"><center><h2>Colorful page with blue and yellow.</h2></center></body></html>

Figure 6
Figure 7
Figure 8
Figure 9b
Figure 10
Figure 13
Figure 14
Figure 15
Leaf Node T Associated with Start Text Style Tag

Find the First Ancestor TD, BODY or HTML Node A

Find the Last Leaf Node L in A Rooted Sub Tree

L == T ?

Create a New Node F with Paired Tags of T

Start from Next Leaf Node, Visit Node N Along Leaf Node List in Order

L == N ?

N Associated with Paired End Tag of T ?

N Associated with Same Start Tag of T ?

Find the Nearest Common Ancestor Node P of T and N

Visit Ancestor Node C of T in Order

Create a Copy of F as F'

Insert F' as Right of T Under C

Does F Have Any Child ?

No

Remove F'

C == P ?

Yes

Remove F

No

Remove T

Does F Have Any Child ?

No

Remove F

Yes

Remove T

Visit Ancestor Node C of N in Order

Create a Copy of F as F'

Insert F' as Left of N Under C

Does F Have Any Child ?

No

Remove F'

C == P ?

Yes

Insert F as between T and N Under P

Yes

Figure 16
(a) Insert F as right of B under P

(b) Insert F as left of D under P

Figure 17a
(c) Insert F as between A and E under P

(d) Insert F between B and D under P
Collect Node into List L

Assign Node as RightN

Is L Empty?

Yes

Assign Node as LeftN

No

Visiting Document Tree Following Depth First Order

Starting from Tag Node F Associated with Start Form Tag with An Empty List L

Form Element Node?

Yes

Form Node?

No

Leaf Node Associated with Form Tag?

Yes

Leaf Node Associated with End Tag?

Yes

Last Node?

Yes

Is L Empty?

Yes

No

Locate Least Common Ancestor P Among Nodes in L

Is P TABLE Tag Node?

No

Is P TR Tag Node?

No

Insert T between LeftN and RightN under P

Remove Node F

Done
Figure 20
Frameset Node F

Create TABLE Node T

Set R as Next TR Created

Create TR for Each Row as Child of T in Order

Initialize Number of Columns C = 1

ROWS Attribute?

Create TR as Child of T

COLS Attribute?

Create TD Node D

Update C

Set First TR as R

Visit Each Child Node N of F

Is N Frame Node?

Is N FrameSet Node?

Adopt Child Nodes from N to D

Add D as Next Child of R

Any More Child of F?

Reset N as Child of D

Replace F Rooted Tree with T Rooted Tree

(a) Frameset Simplifier

Figure 22a
Node 272
Add to Context Indexed by Value of USEMAP
Remove Node 1

(e) Img Simplifier

273
Iframe Node 1

Create TABLE Node T
Create TR Node R as Child of T
Create TD Node D as Child of R
Visit Each Child Node N of F
Replace I Rooted Tree with T Rooted Tree
Any More Child of F ?
Is N DIV Node ?

(f) Iframe Simplifier

Figure 23b
Figure 26

```
<table>
  <tr>
    <td>...</td>
    <td>...</td>
    <td>...</td>
  </tr>
  <tr>
    <td>...</td>
    <td>...</td>
    <td>...</td>
  </tr>
</table>
```
For Each Row Ri of Placement Constraint

For Each Node Ti of Row Ri

\[ W = 0; A = 0; N = 0; M = 0 \]

\[ M_r(i) = 0; W_r(i) = 0; \]

\[ M_r(i) = M_r(i) + M(T_i); W_r(i) = W_r(i) + W(T_i); \]

\[ A = A + W(T_i); N = N + N(T_i); \]

\[ (M_r(i) + W_r(i)) > (M + W) ? \]

\[ M = M_r(i); W = M + W - M_r(i); \]

\[ W_r(i); \]

\[ M = M_r(i); W = W_r(i); \]

\[ W = M_r(i) + W_r(i) - M; \]

Figure 29
Overly Wide Node: T

Data Node? → Y

Applicable Word Break? → Y

Add New Break on the Longest Word

Calculate W, M Considering Placement Constraint Up to the ith Column Ci

Adjust M to MWmax

Adjust M

Done

N

N

338

(W+M) > MWmax?

N

Calculate W, M Considering Placement Constraint Up to the ith Column Ci

Create DIV Node D

340

N

T' = Split(T, All Nodes in Columns C0, C1, ... C(i-1))

Y

342

Figure 31

Figure 32

Over Sized Node

Data Node?

Locate Cut Word Wc

Create Document Partition Based on Consecutive Nodes of Ri from Ti0 upto Tij

Create Document Partition Based on All Nodes of Consecutive Rows R0 to Ri

Visit Each Node Ti of Ri in Order

Nr = N(Tij); Ar = A(Tij)

Nr > Nmax or Ar > Amax?

j=j-1

Visit Each Row Ri Top Down

Visit Each Node Ti of Ri in Order

Nr = N(Tij); Ar = A(Tij)

Nr > Nmax or Ar > Amax?

j=j-1

Nr=0;Ar=0

Visit Each Row Ri Top Down

Visit Each Node Ti of Ri in Order

Nr = N(Tij); Ar = A(Tij)

Nr > Nmax or Ar > Amax?

j=j-1

Nr=0;Ar=0

Done

Done

done

done

Create Document Partition Based on Wc

Create Document Partition Based on All Nodes of Consecutive Rows R0 to Ri

Done
Steps to Split a T Rooted Tree with Selected Content Child Nodes N0, N1, ... Nk: Split(T, N0, N1, ... Nk)

(a) Steps to Split a T Rooted Tree with Selected Content Child Nodes N0, N1, ... Nk: Split(T, N0, N1, ... Nk)

(b) Figure 33
T: Current Node
P: Parent of T

Find the Closest Ancestor HTML Node R

Clone Tree Path R→P as R'→P'

Does R Have HEAD Node Child H?

Y: Clone H Rooted Document Subtree as H' Rooted Tree

N: T Data Node?

Y: Attach H' Rooted Tree by Adding H' as Child of R'

N: T = Split(T, Set of Descendant Content Nodes)

Remove Data from First Character upto the Cut Word out of T

Attach T' as Child of P'

Done

Figure 34
Figure 35
"... lots of text data and more data ..."

"... lots of text data"

"and more data ..."
Dw = Display Width

Scalable Attributes?

Calculate Scaling Factor S Based on (M,W,Dw)

Scale Node Tag Attributes by S

Leaf Node?

Assign Display Width for Each Column

For Each Descendant Content Node in Placement Constraint

Assign Dw as Sum of Column Widths Spanned

Update Sizing Information

Tree Optimization

Parent Node?

Back to Parent Node

Figure 38
For Each Column $C_i$

For Each Node $N_{ij}$ of Column $C_i$

Column Span $> 1$?

$D_{m}(C_{i}) = M(N_{ij}) + S^{*}W(N_{ij})$

$D_{m}(C_{i}) > D_{s} = D_{m}(C_{j}) + D_{m}(C_{j}+1) + D_{m}(C_{j}+2) + D_{m}(C_{j}+k-1) + \ldots + D_{m}(C_{n})$?

$D_{m}(C_{i}) = M(N_{ij}) + S^{*}W(N_{ij})$

$D_{s} < M(N_{ij}) + S^{*}W(N_{ij})$?

$RepeatFlag = 1$

$D(C_{i}) = (D(C_{i}) + D(C_{j}) + D(C_{j}+1) + D(C_{j}+2) + \ldots + D(C_{n})) / (k + 1)$

For $i = i, i+1, \ldots, i+k-1$

$D(C_{i}) = (D(C_{i}) + D(C_{j}) + D(C_{j}+1) + D(C_{j}+2) + \ldots + D(C_{n})) / (k + 1)$

For Each Columns $C_{j}$ where $(i < j)$ or $(j > (i+k-1))$
Document Root

Ax=0; Nx=0; Oa=1; On=1

Leaf Node?

Y → Set O(T) as Original Document Leaf Node Reverse Count Order

N → For Each Node T in Placement Constraint

Y → N(T) > Nx?

On = O(T)

N → A(T) > Ax?

Oa = O(T)

For Each Node T in Placement Constraint

Oa > 0?

N → O(T) = On

Y → On > 0?

N → O(T) = Oa

Y → O(T) = max(Oa, On)

Back to Parent Node

Parent Node?

N → Done

Done

Figure 40
METHOD AND APPARATUS FOR ADAPTING WEB CONTENTS TO DIFFERENT DISPLAY AREA DIMENSIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a continuation of co-pending U.S. patent application Ser. No. 10/757,840, filed Jan. 14, 2004, which claims the benefit of U.S. Provisional Application No. 60/442,873, filed Jan. 27, 2003. The disclosure of the above identified applications is hereby incorporated herein in its entirety by reference.

BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention

[0003] The present invention relates generally to automatic markup language based digital content transcoding and, more specifically, it relates to a method to simplify, split, scale and hyperlink HTML web content for providing a new method to repurpose a large web content authored for desktop top viewing to support smaller devices using limited network bandwidth such as palmtops, PDAs and data-enabled cell phones wirelessly connected with small display areas and processing capacities.

[0004] 2. Description of the Related Art

[0005] With the popular use of Internet, vast and still growing amount of content have been made available through typical desktop browsers such as Internet Explorer (from Microsoft), Navigator (from AOL), and Opera (from Opera). They are coded in standard markup languages such as HTML and JavaScript. However, majority of them have been authored to fit regular desktop or notebook computers with large screen size, big processing capacity connected with high speed network.

[0006] As the web steadily increases its reach beyond the desktop to devices ranging from mobile phones, palmtops, PDAs and domestic appliances, problem in accessing legendarily content start to surface. Constraints from form factor and processing capacity render them practically useless on these devices. To solve this device dependency problem, one most cost effective approach is to provide intermediary adaptation in the content delivery chain.

[0007] Examples such as transcoding proxies can transform markup languages by removing HTML tags, reformating table cells as text, converting image file formats, reducing image size, reducing image color depths, and translating HTML into other markup languages, e.g., WML, CHTML, and HDML. More involved approaches extract subsets of original content, either automatically or manually, or employ text summarizing techniques to condense the target content. Even more elaborated systems include client components using proprietary protocols between intermediaries and corresponding programs running in client devices to emulate standard browser interfaces, such as ZFrame from ZFrame Inc.

[0008] The main problem with conventional markup content transcoding is its inability to handle the sheer volume of content, both text and images, etc. inside the document for small devices. Arbitrary linear approach to partition the content based on markup language codes often makes the results unorganized with the original presentation intent lost. Summarization techniques second guess the author’s intent and are not able to always satisfy user’s need.

[0009] Another problem with conventional markup content transcoding is its inability to handle common hidden semantics inside web documents such as HTML tables. However, authors are increasingly marking up content with presentation rather than semantic information and render the adapted content unusable.

[0010] Another problem with conventional markup content transcoding is its complexity in supporting new devices with different form factors. Instead of gracefully scaling the target transcoding result from small to large display devices, it relies on case-by-case settings requiring expensive development effort to support new devices.

[0011] Another problem with some conventional markup content transcoding is reliant on manual customizations to edit, select or annotate original content to assist adaptation process, which tends to be costly, error prone and not readily scalable.

[0012] Another problem with some conventional markup content transcoding is its dependency on specialized client software. Both deploying proprietary software to various client devices and administrating/configuring server adaptation engine increase cost significantly. This defies the original purpose of automatic content adaptation in place of adopting complete content re-authoring.

[0013] In these respects, Content Divide & Condense, the method to generate and scale document partitions with navigational links from single web content according to the present invention substantially departs from the conventional concepts and designs of the prior art, and in so doing provides an apparatus primarily developed for the purpose of providing a new method to transcoding web content authored for desktop top viewing into smaller ones to accommodate small display areas and capacities in mobile devices.

SUMMARY OF THE INVENTION

[0014] In view of the foregoing disadvantages inherent in the known types of markup content transcoding now present in the prior art, the present invention provides a new method, hereby named Content Divide & Condense, to simplify, partition, scale, and structure single content page onto hyperlinked and ordered set of content pages suitable for small device viewing before direct transcoding from HTML to the target markup language is applied, wherein the same can be utilized for providing a new method to transcoding web content authored for desktop top viewing into smaller ones to accommodate small display areas and capacities in mobile devices.

[0015] The general purpose of the present invention, which will be described subsequently in greater detail, is to provide a method to generate a minimum set of simplified and easily navigable web contents from a single web document, oversized for targeted small devices, while preserving all text, image, transactional as well as embedded presentation constraint information. Each of the simplified web content fits in display size and processing/networking capacity constraints of the target device. The whole set of generated pages are hyperlinked and ordered according to the intended two dimensional navigation semantics embedded inside the original content. A subset of XHTML is adopted to define the kind
of content to be extracted from the original document. With the reduced content complexity in each partitioned page and the preserved navigational organization from original content, final set of documents after applying direct transcoding from each HTML partition to target markup language represent a much more accurate presentation with respect to the original content yet suitable for small device viewing.

[0016] To attain this, the present invention, named as Content Divide & Condense, generally comprises HTML parser, content tree builder, document tree builder, document simplifier, virtual layout engine, document partitioner, content scalable, and markup generator. The parser generates a list of markup and data tags out of HTML source document. It handles script-generated content on the fly and redirects content fetch similar to how common web browsers behave. Based on a specific set of layout tags, the builder constructs a content tree out of the markup and data tags. It interprets loosely composed HTML document following a set of heuristic rules to be compatible with how standard browsers work. This builder completes document tree build from the rest of markup and data tags on top of content element tree. It also adjusts the tree structure to be in compliant with XML specification without changing rendering semantics of the source HTML document interpreted by common browsers. The simplifier transforms the document tree onto an intermediate one defined by a subset of XHTML tags and attributes through filtering and mapping operations on tree nodes. Spatial layout constraints are heuristically estimated and calculated for data and image content embedded inside the document tree according to the semantics of HTML tags. Layout constraints include size, area, placement order, and column/row relationships. Based on the display size and rendering/network capacity constraints, the document tree is partitioned into a set of sub document trees with added hyperlinks and order according to the layout order and content structure. With target device display size constraint, each sub document tree is scaled individually by adjusting height and width attributes through the scalar. Source image references are modified if needed to assure server side image transcoding capability is leveraged. Each document tree defines a simplified HTML document which is generated during the markup generation step. Navigation order and hierarchical hyperlinks are assigned at the same time. The original content is thus represented by the set of smaller documents with hyperlinks and order defined between each other. Additional files such as catalog file indicating network bandwidth required for each document or text only document partitions can be generated and hyperlinked together in the same manner. Each simplified document can be transcoded onto target markup languages such as WML and cached by applying available direct transcoding technique.

[0017] There has thus been outlined, rather broadly, the more important features of the invention in order that the detailed description thereof may be better understood, and in order that the present contribution to the art may be better appreciated. There are additional features of the invention that will be described hereinafter.

[0018] In this respect, before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and to the arrangements of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments and of being practiced and carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein are for the purpose of the description and should not be regarded as limiting.

[0019] A primary object of the present invention is to provide a method to simplify, split, scale, and structure web content for small devices that will overcome the shortcomings of the prior art devices.

[0020] An object of the present invention is to provide a method to simplify web content to contain only the most primitive parts such as texts, images, forms, hyperlinks, and layout presentation arrangements etc., supported by standard markup language browsers for small devices.

[0021] An object of the present invention is to provide a method to extract web content to contain only the selected parts, such as text only with images as text links, or forms only, while preserving layout presentation arrangements etc., supported by standard markup language browsers for small devices.

[0022] Another object is to provide a method to split two dimensional layout arrangement such as tables, framesets and alignment to fit content display to the screen width constraint of the target device.

[0023] Another object is to provide a method to partition web content along both logical and embedded layout structure according to display area and capacity constraints of the target client device.

[0024] Another object is to provide a method to apply minimal scaling to each document partition individually to fit in target device display width constraint.

[0025] Another object is to provide a method to present the original web content by a set of hyperlinked and ordered document partitions according to the two-dimensional navigation order embedded inside the original document.

[0026] Another object is to provide a method to utilize target device display size and resource capacities to partition the document by conducting virtual layout against the original content represented by a markup language.

[0027] Another object is to provide a method to present a hyperlinked catalog content indicating the required network bandwidth required for accessing each document partition from the target device.

[0028] Other objects and advantages of the present invention will become obvious to the reader and it is intended that these objects and advantages be within the scope of the present invention.

[0029] To the accomplishment of the above and related objects, this invention may be embodied in the form illustrated in the accompanying drawings, attention being called to the fact, however, that the drawings are illustrative only, and that changes may be made in the specific construction illustrated.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] Various other objects, features and attendant advantages of the present invention will become fully appreciated as the same becomes better understood when considered in conjunction with the accompanying drawings, in which like
reference characters designate the same or similar parts throughout the several views, and wherein:

[0031] FIG. 1 is to illustrate the function of Content Divide & Condense;

[0032] FIG. 2 is to show Content Divide & Condense working as part of a transcoding server;

[0033] FIG. 3 is to show Content Divide & Condense working as part of a proxy server;

[0034] FIG. 4 is to show Content Divide & Condense working as part of a web server;

[0035] FIG. 5 is to show affects and steps of content reduction engine;

[0036] FIG. 6 is an example showing sample HTML code and the corresponding markup and data list;

[0037] FIG. 7 is to show steps to parse source HTML document into markup and data element list;

[0038] FIG. 8 is to show steps to build layout element tree from markup and data element list;

[0039] FIG. 9a and FIG. 9b are a set of handlers for inserting implicit tags;

[0040] FIG. 10 is an additional set of handlers for inserting implicit tags;

[0041] FIG. 11 is to show three types of relationships between tag node and associated markup or data element;

[0042] FIG. 12 is an example of HTML source code and its corresponding content element tree;

[0043] FIG. 13 is to show steps to build a document tree from the corresponding content element tree;

[0044] FIG. 14 is to show steps to build document sub-tree from markup and data element list between content nodes;

[0045] FIG. 15 is to show steps to rectify an HTML document tree to an XML compliant one;

[0046] FIG. 16 is to show steps to handle non-XML-compliant style tags;

[0047] FIG. 17a and FIG. 17b are to demonstrate node insertion operations;

[0048] FIG. 18 is a sample of HTML code, its corresponding preliminary document tree and XML compliant document tree;

[0049] FIG. 19 is an example to show handling of non-XML compliant style tags;

[0050] FIG. 20 is to show steps to handle non-XML compliant form tags;

[0051] FIG. 21 is an example of source HTML code and the corresponding document tree after form and style tag handling;

[0052] FIG. 22a and FIG. 22b show steps to map document tree onto a simplified one based on a subset of XHTML tags;

[0053] FIG. 23a and FIG. 23b show continuous steps to map document tree onto a simplified one based on a subset of XHTML tags according to FIG. 22a and FIG. 22b;

[0054] FIG. 24 is an example to demonstrate simplification of map tags onto list tags;

[0055] FIG. 25 is a frameset HTML code sample and its corresponding document tree;

[0056] FIG. 26 is a sample of document tree and its corresponding HTML code from frameset simplification;

[0057] FIG. 27 shows steps to apply layout and style constraints;

[0058] FIG. 28 shows virtual layout steps to assign sizing information to document node;

[0059] FIG. 29 shows steps to calculate layout sizing parameter through placement constraint;

[0060] FIG. 30 shows steps to split and partition document tree based on estimated layout width and content size;

[0061] FIG. 31 shows steps to split a document node with oversized layout width;

[0062] FIG. 32 shows steps to partition document against an oversized document node on sizing parameters A and N;

[0063] FIG. 33 shows steps to split a node and an example;

[0064] FIG. 34 shows steps to create document partition based on a set of descendant content nodes from placement constraint;

[0065] FIG. 35 is an example to show document tree change before and after node split without document partitioning;

[0066] FIG. 36 is a document partition example;

[0067] FIG. 37 is an example of document data node partition;

[0068] FIG. 38 shows steps to scale document partition;

[0069] FIG. 39 shows steps to calculate minimum width required for a column of document node;

[0070] FIG. 40 shows steps to obtain navigation order of a document tree; and

[0071] FIG. 41 shows a sample of hyperlinks and navigation order among document partitions.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0072] Turning now descriptively to the drawings, the attached figures illustrate a method to generate and scale document partitions with navigation links from single web content for small device viewing, as shown in FIG. 1 which comprises HTML parser 16, content tree builder 18, document tree builder 20, document simplifier 22, virtual layout engine 24, document partitioner 26, content scalar 28, and markup generator 30. The parser 16 generates a list of markup and data tags out of HTML source document 12. It handles script generated content on the fly and redirected content fetch similar to how common web browsers behave. Based on a specific set of layout tags, the builder 18 constructs a content tree out of the markup and data tags. It interprets loosely composed HTML document following a set of heuristic rules 34 to be compatible with the manner how standard browsers work. This builder 18 completes document tree build from the rest of markup and data tags on top of content element tree. It also adjusts the tree structure to be in compliant with XML.
specification without changing rendering semantics of the source HTML document interpreted by common browsers. The simplifier transforms the document tree onto an intermediate one defined by a subset of XHTML tags and attributes through filtering and mapping operations on tree nodes. Spatial layout constraints are heuristically estimated and calculated for data and image content embedded inside the document tree according to the semantics of HTML tags. Layout constraints include size, area, placement order, and column/row relationships. Based on the display size and rendering/network capacity constraints, the document tree is partitioned into a set of sub-document trees with added hyperlinks and order according to the layout order and content structure. With target device display size constraint, each sub-document tree is scaled individually by adjusting height and width attributes through the scalar. Source image references are modified if needed to assure that the server-side image transcoding capability is leveraged. Each document tree defines a simplified HTML document which is generated during the markup generation step. Navigation order and hierarchical hyperlinks are assigned at the same time. The original content is thus represented by the set of smaller documents with hyperlinks and order defined between each other. Additional files such as catalog files indicating network bandwidth required for each document or text only document partitions can be generated and hyperlinked together in the same manner. Each simplified document can be transcoded onto target markup languages such as WML and cached by applying an available direct transcoding technique.

Turning now to FIG. 2, overall effects of Content Divide & Condense 36 is illustrated. Input to the engine is a single web document such as HTML page 38. The engine then generates a set of simplified and small HTML documents 40 hyperlinked together. A linear navigation order is also assigned to each partition document.

The engine could process the same document with more than one settings at the same time. For example, it generates the partition both with and without images to allow the flexibility to turn on or off the image content while ensuring device capacity is fully utilized. The same text paragraph could appear in two partitions, one consists of only text data and the other contains also image links. Because image capacity is replaced by text data, these two partition documents can not be transcoded directly between each other by adding or removing image links. However, cross links can be inserted such that it is possible to access text data as preview and retrieve full image embedded one when interested.

Systems with Content Divide & Condense working together with client device and other servers are shown in FIG. 3, FIG. 4, and FIG. 5, as examples. Referring to FIG. 3, a transcoding server consists of HTTP handler component 46, a local cache 48, HTML transcoder 50 and Content Divide & Condense 42. The client 44 sends an HTTP request 52 along with client agent id to the transcoding server with an URL referencing a web content from the Web Server 56. The HTTP Handler 46 then sends another HTTP request 54 to the Web Server 56 for the document identified by the URL from the client request 52. Web Server 56 returns the document to the HTTP Handler 46, which passes the document along with a client agent id through Content Divide & Condense 42, which generates a set of hyperlinked and simplified document partitions along with pre-fetched and properly scaled images. Each partition is then passed through an HTML Transcoder 50 to map HTML onto target ML language for the client and stored in the local cache 48 along with scaled images. The HTTP Handler 46 selects the first page from the local cache 48 and returns to the client 44 as part of HTTP response 58.

Referring to FIG. 4, Content Divide & Condense 60 works as part of an HTTP Proxy Server. It works similarly as in FIG. 3. The differences are source link updates and HTTP request/response cache. There is no need to resolve absolute source links when working as the HTTP Proxy Server. By default, HTTP request/response pairs are cached and cache hit is always checked before making remote fetch.

Referring to FIG. 5, Content Divide & Condense 62 works as part of an HTTP Server. The client 64 sends an HTTP request 66 along with the client agent id to the server. The HTTP Handler 68 fetches the target HTML document 70 from local storage. Based on the client agent id, the HTTP Handler 68 determines whether to send the document back directly or pass the client agent id and document to Content Divide & Condense 62. If transcoding is needed, Content Divide & Condense 62 generates a set of hyperlinked and simplified document partitions along with properly scaled images. Each partition is then passed through an HTML Transcoder 72 to map HTML onto target ML language for the client 64 and stored in the local cache 74 along with scaled images. HTTP Handler 68 selects the first transcoded page from the local cache 74 and returns to the client 64 as part of HTTP response 76.

The HTML parser translates HTML document into a list of markup and tags similar to what common browsers do. Each element of the list is either a markup with its attributes or a block of raw data, such as text data or script codes. An example of HTML code sample 77a and its corresponding markup and data list 77b is shown in FIG. 6. The overall steps are shown in FIG. 7, including a syntactic parser 78, frame source handler 80, script source handler 82 and final tag list selector 84.

When the parser 78 encounters <FRAME> tag, source links inside the tag are resolved and corresponding document fetched 80a/parsed 78 on the fly. <FRAME> source is inserted into the original tag list right after the corresponding <FRAME> tag with an added <FRAME> tag at the end to enclose it. The process continues recursively as shown in FIG. 7.

When the parser 78 encounters <SCRIPT> tag, JavaScript source codes are executed by a JavaScript engine 86 with a simplified document object model 88. Source links are followed to fetch remote codes 82a, if there is any. The simplified document object model 88 supports both document.write and document.writeln functions and is capable of generating HTML content 86a on the fly. The in-line generated codes, if there is any, are parsed by the parser 78 and the resulting tag list is inserted right after the corresponding <SCRIPT> tag. This process runs recursively as shown in FIG. 7. The document object model could be expanded when needed by implementing additional objects and functions, including handling of specific client and/or user browser settings, cookies, etc.

After parser 78 exhausts all input sources, HTML tags requiring exclusive or selection or filtering are handled before final list 90 is generated. They include <SCRIPT> vs. <NOSCRIPT>, <FRAME> vs. <NOFRAME>, and
<EMBED> vs. <NOEMBED>, <LAYER> vs. <NOLAYER>. The parser tag selection 84 ignores <NOSCRIPT>, <NOFRAME>, and <EMBED> tags. These tags and all source markup data enclosed are left out from final tag list. Capability for the parser tag selection 84 to select an intended subset of tag list from the source document could readily be added. Depending on target client device context and document semantics, the parser might have an option to choose <NOFRAME> instead of <FRAME>, <NOSCRIPT> instead of <SCRIPT>, <EMBED> instead of <NOEMBED>, <LAYER> instead of <NOLAYER>, etc. Additional tags accepted as standards moving forward could also be supported in the similar manner.

The content tree builder constructs a tree out of the set of content markup elements based on the tag list generated by the parser. An HTML tag is considered content element if it designates directly an actual layout area when the content is rendered. The set of HTML tags considered content elements are listed in Table 1(a). These tags are different from those specifying mainly display styles, user interface context, or executable script codes such as those shown in Table 2(b). The set of content tags are focused first to simplify handling of many loosely composed HTML documents where style and context tags are not required to follow strict XML structures.

### Table 1(a)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>abbr</td>
<td>acronym</td>
</tr>
<tr>
<td>acronym</td>
<td>address</td>
</tr>
<tr>
<td>address</td>
<td>applet</td>
</tr>
<tr>
<td>applet</td>
<td>area</td>
</tr>
<tr>
<td>area</td>
<td>base</td>
</tr>
<tr>
<td>base</td>
<td>blockquote</td>
</tr>
<tr>
<td>blockquote</td>
<td>dfn</td>
</tr>
<tr>
<td>dfn</td>
<td>frameborder</td>
</tr>
<tr>
<td>frameborder</td>
<td>iframe</td>
</tr>
<tr>
<td>iframe</td>
<td>img</td>
</tr>
<tr>
<td>img</td>
<td>input</td>
</tr>
<tr>
<td>input</td>
<td>isindex</td>
</tr>
<tr>
<td>isindex</td>
<td>label</td>
</tr>
<tr>
<td>label</td>
<td>legend</td>
</tr>
<tr>
<td>legend</td>
<td>li</td>
</tr>
<tr>
<td>li</td>
<td>link</td>
</tr>
<tr>
<td>link</td>
<td>marquee</td>
</tr>
<tr>
<td>marquee</td>
<td>menu</td>
</tr>
<tr>
<td>menu</td>
<td>meta</td>
</tr>
<tr>
<td>meta</td>
<td>noeml</td>
</tr>
<tr>
<td>noeml</td>
<td>noframes</td>
</tr>
<tr>
<td>noframes</td>
<td>nowrap</td>
</tr>
<tr>
<td>nowrap</td>
<td>p</td>
</tr>
<tr>
<td>p</td>
<td>pre</td>
</tr>
<tr>
<td>pre</td>
<td>td</td>
</tr>
<tr>
<td>td</td>
<td>th</td>
</tr>
<tr>
<td>th</td>
<td>thread</td>
</tr>
<tr>
<td>thread</td>
<td>title</td>
</tr>
<tr>
<td>title</td>
<td>tr</td>
</tr>
<tr>
<td>tr</td>
<td>ul</td>
</tr>
<tr>
<td>ul</td>
<td>xmp</td>
</tr>
<tr>
<td>xmp</td>
<td>wbr</td>
</tr>
</tbody>
</table>

The steps to build content tree 104 is shown in FIG. 8. Each element of markup and data list 92 generated by the parser is visited in order 92a. Those not belonging to content tags, including data elements, are ignored 94. Conditions are then checked for the need to insert implicit tag 96 into the list to ensure consistency between layout semantics and document tree structure. After new tag is inserted 96a, the process returns back to the list 92a and repeats from the new one on. If the current tag is a start tag 98a, it is pushed to the top of stack 98b while an end tag requires additional handling. Normally, the top of stack would match an end tag encountered. Otherwise 100, the whole stack is examined to see if there is matching one 101. If yes 102, the stack is popped 102a until the matching one is reached. An end tag without any matching tag in the stack is ignored.

Implicit tags are generated on the fly as shown in FIG. 9a, FIG. 9b and FIG. 10. A state refers to the name of the tag at the top of the stack. All the rules from (a) 106 to (n) 108 specify conditions when implicit end tags are detected. Rules (a) 106, (e) 110, (f) 112, and (g) 114 describe how <TR> tag is implied as well. Rule (k) 116 is needed because the parser fetches frame source and inserted the tag list after the associated <FRAME> tag followed by an added <FRAME> one. The last rule (n) 108 states that if </BODY> tag is encountered 108a without a matching state 108b, the end tag of the current state, if needed, is added automatically 108c. The list of HTML tags without end tags are <AREA>, <BASE>, <BR>, <COL>, <COLGROUP>, <FRAME>, <HR>, <IMG>, <INPUT>, <ISINDEX>, <LINK>, <PARAM>, and <BASEFONT>. These rules essentially implement what specified by HTML standard.

The document tree is built during popping tags from the stack. A tree node is defined after the top element is popped from the stack. There are three possible kinds of nodes, as shown in FIG. 11, depending on the relationships between a tag node and its associated markup elements. The most common one is formed by a paired start 120 and end 122 elements as in (a) 118. A degenerated one could be like (b)
Based on the content element tree 132, the remaining non-content markup and data tags are handled to complete the document tree 134. Firstly, these tags are visited following the steps shown in FIG. 13 together with FIG. 14 to complete a preliminary document tree 136. Then, a set of adjustments are applied to special set of non content tags according to the underlying HTML semantics to the final document tree 138 to be compliant with XML structure as shown in FIG. 15.

Based on the content element tree 132, each node is visited following a depth first order 132a and sub document trees, based on non-content tags, are built 142 and inserted 144 onto the content element tree 132 to form a preliminary document tree 134. The steps shown in FIG. 13 build sub document trees 142 based on segments of tag lists partitioned by content tags in the content element tree 132. A list of non-content tags is defined between the first child node and parent node, two neighboring sibling nodes, the last child node and parent node, or simply a single terminal node. A set of sub document trees 142 are constructed out of each such segment of tags and inserted as new child nodes 144 of the defining parent node in order.

The tree building steps are shown in FIG. 14, similar to FIGS. 9a and 9b but a bit simplified. The main difference lies in the handling of end tag without matching start tag in the stack. A tree node with this single end tag is created instead of being removed. This happens often because HTML does not require strict XML structure on style and context tags and parsing start/end tags might belong to two different segments of tag lists partitioned by content element tree nodes. This process on each segment of tag list results in an ordered list of sub trees 146 be inserted back to the content element tree 132. A sample HTML code 147, its corresponding preliminary document tree 147a and XML compliant document tree 147b are shown in FIG. 18.

Steps to rectify preliminary document tree 136 to be XML compliant are shown in FIG. 15. It iterates through the list of leaf nodes (node without any children) in order 148 and calls proper handlers for different types of nodes. Three handlers are considered here. If the leaf node is associated with an end tag 150, it is regarded as extra end tag and removed from the document tree 150a. If the leaf node is associated with a form tag 152, a form tag handler 152a is called. Otherwise, if it is a text style tag 154, a style tag handler 154a is called.

There are four tree insertion operations employed in the handler as shown in FIG. 17a and FIG. 17b. The insertion adds a new child node to a parent node but reset a set of original child nodes from the parent node to itself. Assuming a new node F, and a parent node P, FIG. 17a (a) shows the operation of inserting F as the right of B under P 156, where P is an ancestor node of B. FIG. 17a (b) shows the operation of inserting F as the left of D under P 158, where P is an ancestor node of D. FIG. 17b (c) shows the operation of inserting F as between A and E under P 160, where P is ancestor node of both A and E. And FIG. 17b (d) shows the operation of inserting F between B and D under P 162, where F becomes ancestor node of both B and D as a child of P.

Detailed steps of text style tag handler are shown in FIG. 16. It follows the rule that style specification does not pass <TD>, <BODY>, or <HTML>. In this implementation, multiple <BODY> or <HTML> nodes are possible because of pre fetched frames. An attempt is made to locate the matching leaf node among the rest of leaf nodes following this range rule first 164. If none is found, the style effect is assumed to cover the rest of document element under the first ancestor <TD>, <BODY>, or <HTML> node 166. Existence of leaf node of the same non-ending style tag is also considered an implicit matching end tag. Once matching node pairs are identified, the closest common parent node is located 166a and the new style nodes are inserted accordingly to cover all elements enclosed by these two nodes under this common ancestor node.

The form handler follows the steps shown in FIG. 19. When a leaf FORM tag node is encountered 168, it tries to search for all form element nodes belonging to this form. Form element nodes are those with tags such as <INPUT>, SELECT, FIELDSET, OPTION, OPT-GROUP, and TEXTAREA which are expected to be enclosed by a matching pair of <FORM> tags in the source document. These nodes are content elements and have already been built into the document tree. Depth first order searching following the leaf <FORM> node is required to collect them. The search ends either normally or with an error. If another <FORM> node is encountered 170 before a matching end <FORM> node is found, it is considered document error. Otherwise, the search continues and collects a list of <FORM> element nodes 172 until the matching end <FORM> leaf node is found 174. If the search exhausts the tree, it is assumed an implicit end <FORM> tag is intended right before </BODY> tag. When the search ends without error, it checks to see if the list of <FORM> element nodes is empty 174a. If so, no <FORM> node is needed 176. Otherwise, a new <FORM> node is created based on matching pair of <FORM> tags of the leaf <FORM> node 178. The least common ancestor node 180, which is neither TABLE nor TR nodes, among the collected list of <FORM> element nodes is then found. The new <FORM> node is then inserted between the first and last nodes of the collected node list under this least common ancestor found 182 as depicted in FIG. 17b (d) 162.

A sample document tree 183 built by the above stated steps from sample content element tree 131b in FIG. 12 is shown in FIG. 20. The handlers are provided for correcting loosely structured HTML document. Heuristic assumptions are made in these handlers with respect to when erroneous documents are encountered. More handlers could be added for other conditions not discussed above. In addition, assumptions on how browsers behave might also evolve as new versions and/or new kinds of browsers continue to be adopted in the market.

The simplifier transforms the document tree onto an intermediate one defined by a subset of XHTML tags and attributes through filtering and mapping operations on tree node. A document tree is condensed and simplified based on a subset of XHTML 1.0 markup tag list specified in Table 2. The main objective of this design is to render the content in terms of document tree while preserving as much as possible the intended content, style, hyperlinks and form interactions. Markup tag associated with original document tree node could belong to HTML, XHTML, or even generic XML. The simplification process goes through each node and performs transformation or filtering against a node or a sub tree. Semantics of HTML and XHTML tags are embodied in these transformation rules.
<table>
<thead>
<tr>
<th>Name</th>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>href = URL</td>
<td>anchor</td>
</tr>
<tr>
<td></td>
<td>name = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rel = Link Type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rev = Link Type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>type = Content Type</td>
<td></td>
</tr>
<tr>
<td>ABBR</td>
<td></td>
<td>abbreviation (e.g. WDVL)</td>
</tr>
<tr>
<td>ACRONYM</td>
<td></td>
<td>information on author</td>
</tr>
<tr>
<td>ADDRESS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BASEFONT</td>
<td>size = CDATA</td>
<td>base font size</td>
</tr>
<tr>
<td></td>
<td>color = Color Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>face = CDATA</td>
<td></td>
</tr>
<tr>
<td>BIG</td>
<td></td>
<td>large text size</td>
</tr>
<tr>
<td>BLOCKQUOTE</td>
<td>cite = URL</td>
<td>long quotation</td>
</tr>
<tr>
<td>BODY</td>
<td>align = Color Number</td>
<td>document body</td>
</tr>
<tr>
<td></td>
<td>background = URL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hbgcolor = Color Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>link = Color Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>text = Color Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>vlink = Color Number</td>
<td></td>
</tr>
<tr>
<td>BR</td>
<td></td>
<td>forced line break</td>
</tr>
<tr>
<td>BUTTON</td>
<td>disabled</td>
<td>push button</td>
</tr>
<tr>
<td></td>
<td>name = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>type = button</td>
<td></td>
</tr>
<tr>
<td></td>
<td>value = CDATA</td>
<td></td>
</tr>
<tr>
<td>CAPTION</td>
<td>align = calign</td>
<td>table caption</td>
</tr>
<tr>
<td>CENTER</td>
<td></td>
<td>shorthand for DIV align = center</td>
</tr>
<tr>
<td>CODE</td>
<td></td>
<td>citation</td>
</tr>
<tr>
<td>DO</td>
<td></td>
<td>computer code</td>
</tr>
<tr>
<td>DFN</td>
<td></td>
<td>fragment</td>
</tr>
<tr>
<td>DER</td>
<td>compact</td>
<td>definition description</td>
</tr>
<tr>
<td>DEV</td>
<td>align = left</td>
<td>directory list</td>
</tr>
<tr>
<td></td>
<td>center</td>
<td>generic language/style container</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>justify</td>
<td></td>
</tr>
<tr>
<td>DL</td>
<td>compact</td>
<td>definition list</td>
</tr>
<tr>
<td>DT</td>
<td></td>
<td>definition term</td>
</tr>
<tr>
<td>EM</td>
<td></td>
<td>emphasis</td>
</tr>
<tr>
<td>FIELDSET</td>
<td></td>
<td>form control group</td>
</tr>
<tr>
<td>FONT</td>
<td>color = Color Number</td>
<td>local change to font</td>
</tr>
<tr>
<td></td>
<td>face = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>size = CDATA</td>
<td></td>
</tr>
<tr>
<td>FORM</td>
<td>action = URL</td>
<td>interactive form</td>
</tr>
<tr>
<td></td>
<td>accept-charset = CharSet</td>
<td></td>
</tr>
<tr>
<td></td>
<td>enctype = ContentType</td>
<td></td>
</tr>
<tr>
<td></td>
<td>method = get</td>
<td></td>
</tr>
<tr>
<td>H1, H2, H3, H4, H5, H6</td>
<td>align = left</td>
<td>heading</td>
</tr>
<tr>
<td>HEAD</td>
<td>center</td>
<td>document head, contains BASE, LINK, META, SCRIPT, STYLE, TITLE</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>justify</td>
<td></td>
</tr>
<tr>
<td>HR</td>
<td>noshade</td>
<td>Horizontal rule</td>
</tr>
<tr>
<td>HTML</td>
<td>version = CDATA</td>
<td>document root</td>
</tr>
<tr>
<td></td>
<td>lang = Language Code</td>
<td>element</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>italic text style</td>
</tr>
<tr>
<td>IMG</td>
<td>alt = Text</td>
<td>Embedded image</td>
</tr>
<tr>
<td></td>
<td>src = URL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>height = Length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>longdesc = URL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>width = Length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>align = top</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bottom</td>
<td></td>
</tr>
<tr>
<td></td>
<td>middle</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
<td></td>
</tr>
<tr>
<td></td>
<td>right</td>
<td></td>
</tr>
<tr>
<td>INPUT</td>
<td>accept = ContentText</td>
<td>form control</td>
</tr>
<tr>
<td></td>
<td>alt = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>checked</td>
<td></td>
</tr>
<tr>
<td></td>
<td>disabled</td>
<td></td>
</tr>
<tr>
<td></td>
<td>maxlength = Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>name = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>readonly</td>
<td></td>
</tr>
<tr>
<td></td>
<td>size = CDATA</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 2-continued

<table>
<thead>
<tr>
<th>Name</th>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDB</td>
<td>type = Input Type</td>
<td>text to be entered by the user</td>
</tr>
<tr>
<td></td>
<td>value = CDATA</td>
<td></td>
</tr>
<tr>
<td>LABEL</td>
<td>align = halign</td>
<td>form field label text</td>
</tr>
<tr>
<td>LEGEND</td>
<td></td>
<td>fieldset legend</td>
</tr>
<tr>
<td>LI</td>
<td>type = li style</td>
<td>list item</td>
</tr>
<tr>
<td></td>
<td>value = number</td>
<td></td>
</tr>
<tr>
<td>MENU</td>
<td>compact</td>
<td>menu list</td>
</tr>
<tr>
<td>META</td>
<td>content = CDATA</td>
<td>generic meta</td>
</tr>
<tr>
<td></td>
<td>http-equiv = Name</td>
<td>information</td>
</tr>
<tr>
<td></td>
<td>scheme = CDATA</td>
<td></td>
</tr>
<tr>
<td>OBJECT</td>
<td></td>
<td>generic embedded object</td>
</tr>
<tr>
<td>OL</td>
<td>compact</td>
<td>ordered list</td>
</tr>
<tr>
<td></td>
<td>start = Number</td>
<td></td>
</tr>
<tr>
<td>OPTGROUP</td>
<td>label = Text</td>
<td>option group</td>
</tr>
<tr>
<td></td>
<td>type = ol Type</td>
<td></td>
</tr>
<tr>
<td>OPTION</td>
<td>disabled</td>
<td>Selectable choice</td>
</tr>
<tr>
<td></td>
<td>label = Text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>selected</td>
<td></td>
</tr>
<tr>
<td></td>
<td>value = CDATA</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>Paragraph</td>
</tr>
<tr>
<td>PRE</td>
<td></td>
<td>preformatted text</td>
</tr>
<tr>
<td>Q</td>
<td>cite = URL</td>
<td>short inline quotation</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>strike-through text style</td>
</tr>
<tr>
<td>SAMPlE</td>
<td></td>
<td>sample program output, scripts, etc.</td>
</tr>
<tr>
<td>SELECT</td>
<td>disabled</td>
<td>option selector</td>
</tr>
<tr>
<td></td>
<td>multiple</td>
<td></td>
</tr>
<tr>
<td></td>
<td>name = CDATA</td>
<td></td>
</tr>
<tr>
<td>SMALL</td>
<td></td>
<td>small text style</td>
</tr>
<tr>
<td>SPACER</td>
<td>generic language/style container</td>
<td></td>
</tr>
<tr>
<td>SPAN</td>
<td>generic language/style container</td>
<td></td>
</tr>
<tr>
<td>STRIKE</td>
<td></td>
<td>strike-through text</td>
</tr>
<tr>
<td>STRONG</td>
<td></td>
<td>strong emphasis</td>
</tr>
<tr>
<td>SUB</td>
<td></td>
<td>subscript</td>
</tr>
<tr>
<td>SUP</td>
<td></td>
<td>subscript</td>
</tr>
<tr>
<td>TABLE</td>
<td>bgcolor = Color Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>border = Pixels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>frame = ThFrame</td>
<td></td>
</tr>
<tr>
<td></td>
<td>summary = Text</td>
<td></td>
</tr>
<tr>
<td>TD</td>
<td>abbr = Text</td>
<td>table data cell</td>
</tr>
<tr>
<td></td>
<td>axis = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bgcolor = Color Number</td>
<td></td>
</tr>
<tr>
<td>TEXTAREA</td>
<td>cols = Number</td>
<td>multi-line text field</td>
</tr>
<tr>
<td></td>
<td>rows = Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>name = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>disabled</td>
<td></td>
</tr>
<tr>
<td></td>
<td>readonly</td>
<td></td>
</tr>
<tr>
<td>TH</td>
<td>abbr = Text</td>
<td>table header cell</td>
</tr>
<tr>
<td></td>
<td>axis = CDATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bgcolor = Color Number</td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td></td>
<td>Document title</td>
</tr>
<tr>
<td>TR</td>
<td></td>
<td>table row</td>
</tr>
<tr>
<td>TT</td>
<td></td>
<td>teletype or monospaced text style</td>
</tr>
<tr>
<td>U</td>
<td></td>
<td>underlined text style</td>
</tr>
<tr>
<td>UL</td>
<td>compact type = UI Style</td>
<td>Unordered list</td>
</tr>
<tr>
<td>VAR</td>
<td></td>
<td>instance of a variable or program argument</td>
</tr>
</tbody>
</table>
The simplification steps are shown in FIG. 21. It walks through the document tree following depth first order starting from the root node 184. If it is a data node 186, a simple filtering process is applied to remove consecutive space, carriage returns or line feeds 186a. Otherwise 188, the rooted sub tree is removed 188a if the associated tag belongs to the set of five types, <APPLET>, <SCRIPT>, <NOSCRIP>, <NOFRAME>, or <NOLAYER>. They are ignored because of 1. Java support: creating an applet is not allowed; 2. Script codes have either been executed to get dynamic client content or not yet supported; 3. <NOSCRIP>, <NOFRAME>, or <NOLAYER> do not usually contain useful information, but simply advisory or warning messages.

For <META> tags 200, only those with the presence of HTTP-EQUIV attribute are retained 202. Other Meta tags used for naming, keywords or other purposes are removed 204, as they do not have significance either on content or how content is fetched. Response information is extracted from the HTTP attribute value pair denoted by the values of HTTP-EQUIV and CONTENT attributes, and stored as part of document context information 206, such as document encoding and language set specification.

Table simplifier 208 is applied for <TABLE> nodes as described in FIG. 22a (b) 209. It goes through its direct child nodes and ensures that: 1. <THEAD> is placed before <TR>, <TBODY>, and <TFOOT>; 2. <TFOOT> node is placed after <THEAD>, <TR>, and <TBODY> nodes; 3. order of <TR> nodes are kept the same, if there are <THEAD>, <TBODY>, or <TFOOT> nodes.

A node belonging to four types of tags are replaced by <DIV> node 210 to keep the structure in place while retaining the enclosed data. <ILAYER> 212 and <LAYER> 214 are used for positioning a block of content. This will not be proper after splitting and sealing the content.<MARQUEE> 216 is used for animating a block of content, not supported by most browsers. <OBJECT> 218 is to activate embedded client application and not handled by the simplification process. Alternate text enclosed by the <OBJECT> tags is preserved. The simplification process ignores presentation and functional controls intended by these tags and keeps only the content data as a division block.

When <BASE> node is encountered 220, the document context is updated 222 on the originating source URL. This node is removed afterwards 224, as the resulting content would be sent from servers of different URL.

An <INPUT> node with type attribute value FILE or IMAGE is removed 226. Image based input button might require client side image mapping capability which would be distorted during scaling.

<FRAMESET> node is handled by Frameset simplifier 228 as shown in FIG. 22a (a) 230. Contents enclosed by frameset and frame tags intended for separate client display windows are replaced by table structure preserving similar layout constraint. In essence, it removes intersections between frame windows on the client device while ensuring the original frame set content is properly displayed. A table node is created 234 for a <FRAMESET> node 232. Depending on ROWS attribute specification 236, one or multiple <TR> nodes are added to this table node. Single row table is assumed without the presence of ROWS attribute. Similarly, depending on COLS attribute specification 238, one or multiple <TD> nodes are added to each <TR> node. Again, one column row is assumed without COLS specification. Frame source content as children nodes of original <FRAME> node are reattached to the corresponding <TD> node as their parent node 240. However, when the child of <FRAMESET> node is also another <FRAMESET> node, it is reattached to the corresponding <TD> node as its child node. The resulting <TABLE> node rooted tree is connected to the document tree in place of the <FRAMESET> node 244. An example of mapping from <FRAMESET> node tree to <TABLE> node tree is demonstrated in FIG. 25 and FIG. 26.

<TR> node is handled by TR simplifier as shown in FIG. 22a (c) 246. If there is background color specified for the whole row, this attribute BGCOLOR is duplicated to each <TD> node 248 under this <TR> node, when no BGCOLOR is specified for the <TD> node. As table structure could be split and changed, this attribute will be honored at <TD> node but not <TR> node.

<MAP> node is handled by the map simplifier as shown in FIG. 23a (d) 250. The navigation links embedded inside a map are replaced by a newly created list of hyperlinks. In essence, a <MAP> node 251 is replaced by a <UL> node 252 and each <AREA> node under <MAP> node is replaced by a list <LI> node 254 with an anchor <A> node 256 presenting the reference specified in HREF attribute 258 inside <AREA> tag. Hyperlinked text for each <AREA> node is determined by its ALT attribute 260, if present. Otherwise, the file name of the URL specified by HREF attribute is used 262 instead. The resulting <UL> rooted document tree is then stored in the context 264 indexed by the name of <MAP> node through NAME attribute. <MAP> node rooted tree is then removed from the document 266. This is demonstrated in FIG. 24 with an example <MAP> rooted tree 266a and its corresponding <UL> rooted tree 266b.

<IMG> node 267 is handled by the img simplifier as shown in FIG. 23a (e) 268. If it is a server side image map, as specified by ISMAP attribute, this node is removed 270. Because of possible scaling, image map is not supported. If the node is a client side image map, this node is indexed in the content 272 for possible replacement later with corresponding <MAP> tree.

<IFRAME> node 273 is handled by Iframe simplifier as shown in FIG. 23b (f) 274. A newly created single cell table replaces the original <IFRAME> tag 276. To distinguish alternate text enclosed by <IFRAME> tags, the fetched frame source content is enclosed by the parser with <DIV> tags. The figure describes detailed steps on how the table tree and the frame source content are connected and inserted into the document tree while <IFRAME> and alternate text is removed.

If a node does not match any of tags considered above, it is checked against the list in Table 2. Those with tag names not preset in this table are removed from the document tree 278. Then its attributes are updated 280 as shown in FIG. 21. Those attributes not listed in Table 2 are removed. All relative URL, as in HREF or ACTION attributes are resolved according to document context with its absolute path. Actual
font size has to be used for SIZE attribute associated with FONT node as well. Because of the presence of `<BASE-FONT>`, relative font size can be resolved with the help of document context.

[0109] After walking through the whole document tree nodes, each `<IMG>` node with USEMAP attribute indexed by a map name, is further condensed 282 as shown in FIG. 21. The `<IMG>` node rooted tree in the document is replaced by the corresponding `<UL>` rooted tree created from original `<MAP>` node, if there is any, or removed if there is none. This is the last step to complete the simplification process.

[0110] Changes in the target tag and attribute list as well as how different types of document nodes are handled would result in variations of document tree reduction. For example, the data filter could employ a scheme to retain only content for hyperlinks or form interface but removing all others. Another example is the support of `<STYLE>` tags for getting more precise information and better control on how document would be rendered at client devices. Yet another example is support for international language attributes inside markup tags in addition to those from HTTP headers. As standards of markup language evolve, changes are expected to accommodate new developments.

[0111] Spatial layout constraints are heuristically estimated and calculated for text and image content embedded inside the document tree according to the semantics of HTML tags. Layout constraints include size, area, placement order, and column/row relationships. Display size and client capacity requirements are estimated for the simplified document tree through virtual layout on the underlying document tree. These parameters are used to determine how the document should be partitioned and scaled to accommodate a target client device. The process of virtual layout includes assigning placement constraints and calculating layout sizing information for each content node based on the constraints and a set of layout parameter settings.

[0112] Given a document tree, virtual layout determines the set of content children for each content node and assigns it placement constraint among these children nodes. A set of nodes C1, C2, ..., Cn form content children set S of a node N if 1. N is ancestor node of each node Ci in S and 2. each node Ci in S is either content node or data node and 3. for all leaf nodes under N rooted tree, there exists one and only one node in S as its ancestor node. By default, the content children set of a content node is defined as the collection of highest-level offspring content/data nodes. Virtual layout assigns placement constraint to document tree nodes such that 1. every leaf node of the document tree belongs to one and only one content children set and 2. each content node belongs to most one content children set.

[0113] To estimate the minimum display width needed for content rendering, placement constraint is designated to content nodes. Placement constraints adapted here are either table with rows/columns or simply a single column. Steps to assign placement constraint are illustrated in FIG. 27. It visits each node following depth first order 284 and adds row/column constraint to `<TABLE>` node on the associated `<TD>` node 286 accordingly, including row and column span. `<TR>` node is ignored 288 as its layout semantic has been considered when assigning row/column constraint for its parent `<TABLE>` node. All other content nodes, except for leaf ones 289, are assigned single column placement constraint 290 on its content children set.

[0114] Four sizing parameters could be derived from the document tree with placement constraints assigned and display font sizes selected for the target client device. They are scalable width (W) in pixel, minimum width (M) in pixel, image area (A) in square pixel, and total number of characters (N). W represents size required for scalable layout components such as `<IMG>` and `<TEXTAREA>`, for example. M characterizes the minimum fixed layout component needed. It is typically the width of the longest word in the document text. A is the total area of all images in the document. N is the number of all display characters inside the document, symbolizing the amount of text information carried. The minimum display width D required for rendering a document rooted at a node with W and M will be W + M.

[0115] Font size and language settings are needed to calculate layout sizing information. Character and word boundaries are determined by language encoding for the content text data. Average width of character is dependent on the specified font family and font size. To simplify the layout process, a single font family with minimum and default font size is indexed by the client agent and language code. For example, English content from IPAQ IE browser would use Times Roman font with minimum font size 2 and default font size 3. Selection of these parameters is to be as realistic as possible and depends on the settings of specific user agent.

[0116] A layout context is referenced and updated when visiting each node. Included in this context are current font size, layout sizing constraint (Nmax, Mmax, Amax), NoFlow flag, and Atomic flag, etc. Mmax is the maximum value of N allowed for the whole document. Mmax is the maximum (W + M) value for the whole document. Amax is the maximum image area allowed, NoFlow flag is used when text characters would be laid out in one line. And Atomic flag means no partition is allowed. Style nodes such as `<FONT>` node affect the font size. `<FORM>` node enables Atomic flag. Meaning elements of `<FORM>` tags should belong to the same document. `<SELECT>` node enables NoFlow flag to indicate text in a data node, mainly under `<OPTION>` node, should be shown in one single line.

[0117] Steps to calculate sizing parameter values for a document node associated with placement constraint are shown in FIG. 28. Initial values of W, A, N, and M are set to zero for all nodes. A bottom up process is employed to propagate layout sizing information from leaf nodes to the root through these constraints. Starting from the root, it walks down the tree in a depth first order to size each node. For non-leaf node 292, it is checked if layout context, including font size, and character flow control, needs to be updated. A leaf node 294, which is either data node, containing only character text, or image node, linking an image source, provides the basic layout dimension data. An `<IMG>` node 296 with width w and height h would have size W = w, A = w*h, N = 0, and M = 0. A data node 298 without NoFlow flag on text layout context total number of characters t and the longest word in terms of characters l would have size N = t, M = l*F, W = 0, and A = 0, where F is the average character width under the current font setting in the text layout context. If NoFlow flag is set for text layout context, as in the case under `<SELECT>` node rooted tree, M is assigned as 1*F instead of 1*F. More precise calculation is possible when equipped with detailed display size information for each character instead of using average with.
After all children of a content node have been sized, the associated placement constraint is applied to obtain sizing information for this node. Generic steps to calculate these parameter values according to the constraint are shown in FIG. 29. A slight variation is applied for each node. (W, A, N, M) is initially set to (0, 0, 0, 0) 302 during the calculation. Each row is iterated through 304 to update these four values. The number of nodes in a row could be smaller than the number of columns in the constraint because of column span constraints. Area A and total number of characters N are additive but considered only once when spanning multiple rows. M and W are assigned such that both (M + W) and M should both be maximum among all rows.

Propagation function could be node specific. For the minimum of all M values among all its child nodes is assigned as node's M value. Based on a simplified document tree, the virtual layout engine derives document layout parameters without conducting actual document rendering. Final result depends on the set of sizing parameters used, placement constraints applied to each node, constraint propagation functions adopted, text layout style context employed, and the global display size setting including language encoding and user agent font families. Variation of these parameters is expected as additional aspects of document layout are considered.

Based on the display size and rendering/network capacity constraints, the document tree is partitioned into a set of sub-document trees with added hyperlinks according to the layout order and content structure. Based on the sizing estimation from virtual layout, a document is partitioned and/or split according to user agent size constraints. Partitioning applies to a document and creates new documents while splitting operates on a document node, generating new nodes but not additional document. Partitioning and split operations are applied in accordance with the document tree to preserve the original content structure as much as possible.

Virtual layout and document partitioning are interwoven together in a bottom up process from leaf tree nodes to arrive at a set of document nodes where each one satisfies the user agent constraint. The steps of this process are shown in FIG. 30. Starting with the root node, it traverses down the tree in a depth-first order and accumulates document elements, calculating layout sizing information, and performing partitioning or splitting to ensure sizing constraints are satisfied for each document node collected. Sizing parameters (W, A, M, N) for each node T shall be partitioned or split such that W + M ≤ MWmax; A ≤ AMax; N ≤ NMax.

Leaf node considered for sizing is either an <IMG> node or a data node. A data node cannot be split or partitioned but a scaling factor could always be found to satisfy the sizing constraint. With NoFlow flag on in the associated layout context, a data node cannot be split nor partitioned. Its sizing parameters are adjusted artificially to satisfy the layout constraint with an assumption that the user agent would be able to make proper adjustment on the client side.

(W, N, M, A) adjustment makes updates directly on the sizing parameter values without changing the document tree. If an <IMG> node with original sizing data as (W, A, 0, 0) where W > MWmax or A > AMax, the sizing parameters are adjusted through a scaling factor r = min(W/MWmax, sqrt(A/AMax)). The adjusted set of sizing parameters would be (rW, rA, 0, 0). A data node under NoFlow flag with original sizing parameter (0, 0, M, N) exceeding sizing constraints would be adjusted to (0, 0, min(M, MWmax), min(N, Nmax)).

Once sizing parameters (W, A, M, N) of a node is obtained 316, the constraint MWmax is checked and split operation 318 applied if (W + M) > MWmax until the constraint is satisfied, then both AMax and Nmax constraint are checked 320 and partition operation applied if (N > Nmax) or (A > AMax) until both are satisfied. Document partition 322 is based on node split but creating a new document tree.

To split a data node, an attempt is made to insert breaks in the longest word to bring the width requirement under the MWmax constraint. This is an update of the node without adding new ones. In the case no such break is possible, M value is artificially adjusted to MWmax with an intent for user client to handle and leave the node unchanged.

Split of multi-data node T separates the original T rooted sub tree into two separate ones. This operation, denoted as split (T, N0, N1, . . . Nk), requires the target node T and a set of descendant content nodes, N0, N1, . . . , Nk, from its associated placement constraint. The steps are shown in FIG. 33 (a) 324. A clone of T is created as T' 326 to be the root node of the new split out sub tree. All paths between Ni and T are cloned in the T' rooted sub tree. Each Ni rooted sub tree is removed from the original document and inserted into T' rooted one under the same path copied 328. A copy of placement constraint associated with T is attached to T' governing the node set N0, N1, . . . , Nk. A sample of split operation is shown in FIG. 33 (b) 332, where split (T, N2, N3, N4, N5) results into two trees rooted by T' 334 and T' 336 respectively. Note that clones of T, T' 1, T' 2 are created as T', T' 1 and T' 2 in this case.

A non-data node T with (W + M) > MWmax needs to be split based on columns in the associated placement constraint, as shown in FIG. 31. It is not possible for a node with placement constraint with single column to be with (W + M) > MWmax. Every one of the descendant content node of a column will have (W + M) ≤ MWmax before the current node is considered. A column C (i − 1) is selected 338 such that MWmax constraint is satisfied considering the partial placement constraints including all nodes belonging to columns from the first one up to (i − 1), but not when adding nodes from the next one column Ci. A new T' rooted sub tree is created 340 by node split based on nodes from maximum consecutive columns C0 up to Ci (i − 1). In addition, a dummy <DIV> node D is also created 342 with the T' rooted tree removed from the document tree. T' and T are attached to D 344 as its children with T as the next sibling of T'. D is then added back to the document tree 346 in the original place of T. Default single column placement constraint on T and T' is assigned to D 348. A sample of this split operation is shown in FIG. 35.

After MWmax constraint is handled, AMax and NMax are considered as shown in FIG. 30. If Atomic layout context flag is on 350, such as the case for <FORM> related nodes, no document partition is allowed, and the process proceeds by updating N and A values such that N ≤ NMax and A ≤ AMax without performing any partition operation. Otherwise document is partitioned on the current node.
Steps to partition a document are shown in FIG. 32. For a data node, a cut word is located 354 from the associated data such that number of characters before and including the cut word constitute the maximum number of consecutive words that would satisfy $N_{\text{max}}$ constraint. For a non-data node, a set of descendant content nodes is selected from rows of its placement constraint to split off a document partition. Selecting which nodes for partition follows rows and columns/row span specifications in the placement constraints. If ON or A of any of these elements is set, a set of consecutive nodes from this row are selected to be partitioned 358. Afterwards, it continues to examine row by row from top down 358 and finds the maximum number of consecutive rows to split off.

A document partition on a node is accomplished by cloning its ancestor nodes and a node split on itself, as shown in FIG. 34. The ancestor tree path includes the closest ancestor <HTML> node 360 to its parent node. It is possible for a document node to have multiple <HTML> ancestor nodes because of expanded frame sources. If the selected <HTML> node has child <HEAD> node, the <HEAD> node rooted sub tree is also cloned 362 and attached to the new <HTML> node as descendants 364.

Data node and non-data node are handled differently. For a data node 366, its clone $T'$ is created 366a and the set of data from first characters up to the cut word is identified is moved from the original node to the cloned one 366b. An example is shown in FIG. 37. For a non-data node 368, a subtree rooted by $T'$ is created 370 by node split based on the selected descendant content nodes. $T'$ is then attached to the cloned tree 372 path to form a document partition, rooted by an <HTML> node $R$. Document changes based on tree operations according to layout sizing constraints depend on what constraints to use, how they are used and which operations to apply. Variations are possible for different considerations. For example, a new constraint $N_{\text{min}}$ and $A_{\text{min}}$ could be introduced to ensure each document partition would have $N_{\text{min}}$ and $A_{\text{min}}$ satisfying $N_{\text{min}} > N_{\text{min}}$ or $A_{\text{min}} > A_{\text{min}}$ with split and partition decisions updated accordingly. Partition could be applied accordingly with slightly different results. FIG. 36 is a document partition example.

Based on target device display size constraint, each sub document tree is scaled individually by adjusting height and width attributes through the scalar. Source image references are modified, if needed, to assure server side image transcoding capabilities, including, for example, image format change, color depth adjustment, and width/height scaling, are leveraged. Scaling process is applied to each partitioned document as well as the updated original one to change tag attributes and perform tree optimization at the same time. Scaling factor is calculated according to estimated document size information and the target client display width available.

Overall steps for scaling are shown in FIG. 38. Starting from the document root 374, it assigns maximum display width available for each document node. Initially, maximum width available for the root node is assigned as the display width 376 of target client device. If scalable attributes are present 378, scaling factor is calculated 380 and applied 382. Scalable attributes are also applicable for text related nodes, in addition to image ones, such as WIDTH, HEIGHT, and SIZE for <INPUT> tag and COLS for <TEXTAREA> tag. SRC attribute for scaled <IMG> tags should be updated to embed scaling information with redirecting path for special image processing server when needed.

Given $M$, $W$, and $D_w$, scaling factor $S$ is calculated as $(D_w - M)/W$ if $(D_w - (W + M))$ and $(W > 0)$. Otherwise, $S$ is set to 1, i.e. the content fits the screen without the need for scaling. As $M$ represents non-scalable sizing information such as minimum word length, only $W$, usually minimum image width, could be scaled.

Sizing information for a document node is updated 384 and optimized 386 after scaling operations have been performed on all descendant content child nodes according to its placement constraint. The optimization removes content nodes with empty A and N. Non-content nodes without any content offspring nodes are also deleted. Placement constraint is also simplified by removing rows and columns without any descendant content nodes. Column and row span values are updated accordingly. Whether to allow ALIGN right or left for a child <IMG> node, when present, can be determined by the available display width for the current node and the minimum width needed for the rest of child nodes. Additional constraint could be employed to eliminate document nodes that don’t satisfy minimum height, width or maximum scaling factor values, for example.

Steps to assign $D_w$ to each descendant content child node of a placement constraint are shown in FIG. 39. The display width available for the current node, $D_w$, its scaling factor calculated, $S$, and a parameter to control the maximum number of iterations employed inside the steps. Maxiterate $S$, are required to proceed. For a single column placement constraint, each node of the column is assigned the same width 390 as $D_w$. Otherwise, an algorithm is used 392 to distribute $D_w$ to each column, hence each node.

The objective of this algorithm is to find a set of values for all column width such that each node in the placement constraint can be accommodated and the sum of all column width equals $D_w$. Because of the way $D_w$ is calculated, there always exists a set of values. This algorithm considers first the subset of nodes with single column span. It establishes minimum column width $D_i(C_i)$ for each column. $C_i$, by definition, is no less than sum of these minimum widths. The difference, if there is, is distributed among each column $C_i$ as $D(C)$. When it iterates through all other nodes with multiple column span and makes adjustment of column width accordingly for the new node constraint while maintaining the original minimum column width assigned. Because of convergence nature of this assignment, it is expected to settle down to a solution after certain steps. However, maximum number of iteration cycles along the nodes is set 394 to arrive at an acceptable solution without much cost.

Several additional notations used in FIG. 39 warrant explanations. $T$ is the minimum total column width based on nodes with single column span. $C_n$ stores then number of columns. RepeatFlag indicates whether a satisfactory set of column width values have been assigned after a loop considering all nodes with multiple column span. Iterate counts the number of iterations. $C_i$ stands for $i$th column and $R_i$ for $i$th row. $N_i$ is the node on $i$th row and $j$th column. With multiple column and row spans, $N_i$ and $N_j$ could stand for the same node although $i = j$ and or $j = i$. $D_s$ represents the cumulative column width allocated defined by column span of a...
node. DDs is the sum of all width in addition to the minimum one for each column not covered by the column span of a node.

[0140] After a node is scaled, optimization rules are applied to either remove the node or the whole rooted tree. A content document node which doesn’t have any content size, i.e. A = 0 and N = 0, would be removed together with its rooted tree. In addition, the HTML node which is not document root, created because of a method to determine how sizing specification in content elements should be updated so the resulting content would fit the display screen properly. This method considers available display window width, content sizing information, and the placement constraint embedded with an algorithm to calculate the minimum width and a method to scale the applicable sizing attributes. Variations are possible for the set of content attributes to size, how new sizes are calculated, handling of boundary conditions such as when an element becomes too small to be significant, and how the algorithm is designed to arrive at a solution.

[0142] Based on the set of partitioned and scaled document trees, corresponding markup files are generated according to the subset of XHTML specs defined in Table 2, along with navigational relationship among each other. Document partition operation defines a hyperlinked information estimated between the document tree with the split node and the one partitioned out. Additional ordering relationships are established for accessing one document from another in a linear manner based on the original document source text order.

[0143] Steps to calculate order for each document are shown in FIG. 40. The order of a node Ti, denoted as O(Ti), is obtained by the sequence count of content leaf nodes appearing in the original document tree in reverse. For example, the first content leaf node is designated with order 0, the next one is 1, etc. The order is propagated bottom up from leaf nodes of a document tree. The order of a node is determined by two descendant content nodes, if different, with the maximum A and maximum N sizing. If these two nodes are different, the one with the larger order is selected during propagation. Many other approaches could be adopted to create hierarchical links and linear order among document trees. They could be randomized, based solely on particular type of content nodes, or using the first, last, etc. leaf node order for navigational relationship.

[0144] Sample hyperlinks and navigation order so constructed are illustrated in FIG. 41. Six document partitions are ordered D0398, D1400, ..., D5402. Hyperlink [+] points to the previous document, [+] to the next, and [’] to its parent. If the first page selected to send back to the client is based on navigation order only, the client receives document D0398. The user could either click on [+] to go to the next page, D1400, or back to its hierarchical parent page, D5402, through [’]. From page D5406, the root page, four partitions D0398, D1400, D3408, and D4410 are directly linked as its child document pages. Although D2412 follows D1400 in order, it is also linked under D3408 hierarchically. Such hierarchy has been built during document partitions reflecting the original document layout semantics.

[0145] The first page returning to the client after partitioning varies depending on the need. It could be the first one based on navigation order, the root page along the partition hierarchy, or a separate page built from these partitions for special purpose. One such example is a catalog page with simple summary information on bandwidth requirement and navigation as well as hierarchy relationships among the pages, connected with hyperlinks. This will give user an overview of the target document without costing too much bandwidth resource before proceeding further.

[0146] As to a further discussion of the manner of usage and operation of the present invention, the same should be apparent from the above description. Accordingly, no further discussion relating to the manner of usage and operation will be provided.

[0147] With respect to the above description then, it is to be realized that variations and extensions of the embodiment are deemed readily apparent and obvious to one skilled in the art, and all equivalent relationships to those illustrated in the drawings and described in the specification are intended to be encompassed by the present invention.

[0148] Therefore, the foregoing is considered as illustrative only of the principles of the invention. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the invention to the exact construction and operation shown and described, and accordingly, all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.

What is claimed is:

1. A method for rectifying a markup document having one or more elements comprising:

   selecting a first portion of the markup document;

   partitioning a second portion of the markup document into one or more parts according to the first portion, the first portion being separate from the second portion;

   building a document tree according to a content tree and one or more element trees, wherein the content tree corresponds to the first portion, and wherein each of the one or more element trees corresponds to one of the one or more parts; and

   generating a rectified markup document in compliance with a markup language based on the document tree for a user interface presentation.

2. The method of claim 1, wherein the one or more elements include data elements, wherein each of the data elements corresponds to a physical partition of the presentation in the user interface.

3. The method of claim 2, wherein the selection comprises:

   determining if an element of the one or more elements is a data element, wherein each element of the first portion of the markup document is a data element.

4. The method of claim 2, wherein the data elements include a table element corresponding to a table in the user interface presentation.

5. The method of claim 2, wherein the first portion of the markup document includes a first data element associated with a tag name, wherein building the document tree comprises building the content tree based on the first portion of the markup document, wherein the content tree includes a first content tree node corresponding to the first data element and a first matching data element associated with the tag name, and wherein building the content tree comprises:
comparing the first data element with one or more of the data elements in the first portion of the markup document; and

determining the first matching data element based on the comparison.
6. The method of claim 5, wherein the first portion of the markup document includes a second data element, wherein the content tree includes a second content tree node corresponding to the second data element, wherein one of the one or more parts is partitioned from the one or more elements of the markup document according to the first matching data element and the second data element.
7. The method of claim 6, wherein the one or more elements are associated with an order, wherein each of the one or more elements is associated with a position in the order, wherein the first matching data element is associated with a first position in the order, wherein the second data element is associated with a second position in the order, and wherein each element in the one or more parts is positioned between the first position and the second position in the order.
8. The method of claim 7, wherein building the document tree further comprises:

building one of the element trees according to elements positioned between the first position and the second position in the order in the one or more elements.
9. The method of claim 5, wherein the one or more elements include the first matching data element.
10. The method of claim 5, wherein the one or more elements do not include the first matching data element and wherein the determination comprises:

generating the first matching data element; and

positioning the first matching data element among the one or more elements according to the order.
11. A machine-readable storage medium having instructions therein, which when executed by a machine, causes the machine to perform a method, the method comprising:

selecting a first portion of the markup document;

partitioning a second portion of the markup document into one or more parts according to the first portion, the first portion being separate from the second portion;

building a document tree according to a content tree and one or more element trees, wherein the content tree corresponds to the first portion, and wherein each of the one or more element trees corresponds to one of the one or more parts; and

generating a rectified markup document in compliance with a markup language based on the document tree for a user interface presentation.
12. The machine-readable storage medium of claim 11, wherein the one or more elements include data elements, wherein each of the data elements corresponds to a physical partition of the presentation in the user interface.
13. The machine-readable storage medium of claim 12, wherein the selection comprises:

determining if an element of the one or more elements is a data element, wherein each element of the first portion of the markup document is a data element.
14. The machine-readable storage medium of claim 13, wherein the data elements include an image element corresponding to an image in the user interface presentation.
15. The machine-readable storage medium of claim 12, wherein the first portion of the markup document includes a first data element associated with a tag name, wherein building the document tree comprises building the content tree based on the first portion of the markup document, wherein the content tree includes a first content tree node corresponding to the first data element and a first matching data element associated with the tag name, and wherein building the content tree comprises:

comparing the first data element with one or more of the data elements in the first portion of the markup document; and

determining the first matching data element based on the comparison.
16. The machine-readable storage medium of claim 15, wherein the first portion of the markup document includes a second data element, wherein the content tree includes a second content tree node corresponding to the second data element, wherein one of the one or more parts is partitioned from the one or more elements of the markup document according to the first matching data element and the second data element.
17. The machine-readable storage medium of claim 15, wherein the one or more elements are associated with an order, wherein each of the one or more elements is associated with a position in the order, wherein the first matching data element is associated with a first position in the order, wherein the second data element is associated with a second position in the order, and wherein each element of the one or more parts is positioned between the first position and the second position in the order.
18. The machine-readable storage medium of claim 15, wherein the one or more elements include the first matching data element.
19. The machine-readable storage medium of claim 15, wherein the one or more elements do not include the first matching data element and wherein the determination comprises:

generating the first matching data element; and

positioning the first matching data element among the one or more elements according to the order.
20. An apparatus for rectifying a markup document having one or more elements, the apparatus comprising:

means for selecting a first portion of the markup document;

means for partitioning a second portion of the markup document into one or more parts according to the first portion, the first portion being separate from the second portion;

means for building a document tree according to a content tree and one or more element trees, wherein the content tree corresponds to the first portion, and wherein each of the one or more element trees corresponds to one of the one or more parts; and

means for generating a rectified markup document in compliance with a markup language based on the document tree for a user interface presentation.