November 14, 2002

RT0493

Human-Computer Interaction 15 pages

Research Report

Layout Group Extraction from Web Content for Effective Adaptation

Kentarou Fukuda, Hironobu Takagi, Junji Maeda and Chieko Asakawa

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Layout Group Extraction from Web Content for Effective Adaptation

Kentarou FUKUDA   Hironobu TAKAGI   Junji MAEDA   Chieko ASAKAWA

Tokyo Research Laboratory, IBM Japan, Ltd.
1623–14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan
Tel: +81 46 215 4659   Fax: +81 46 274 4282
E-mail: kentarou@jp.ibm.com

Abstract

These days, people access the Web by using various devices and methods, such as PDAs, cellular phones, and voice-based browsers. However, most Web content is designed for desktop computers. Therefore, already-existing Web content should be transcoded to be suitable for each access device and method. For this purpose, some annotation-based transcoding systems have been developed. An annotation is additional information of Web content, and effective adaptation can be achieved by using it. One of the most difficult problems of annotation is the cost of annotating Web content. Many popular sites, such as news sites, have a large number of Web pages and add new content continually. Hence, it is almost impossible to annotate all of the content in these sites. To solve this problem, we introduce a method to extract common layouts from Web pages. We focus on the structure and characteristics of particular HTML tags that affect the layout of Web pages. Our method calculates the distance between Web pages using this method. When the distance is below the threshold, these pages can be considered as the same layout pages. By using this method, a certain annotation can be applied to any Web pages that have the same layout. Therefore, the cost of adaptation will be reduced.

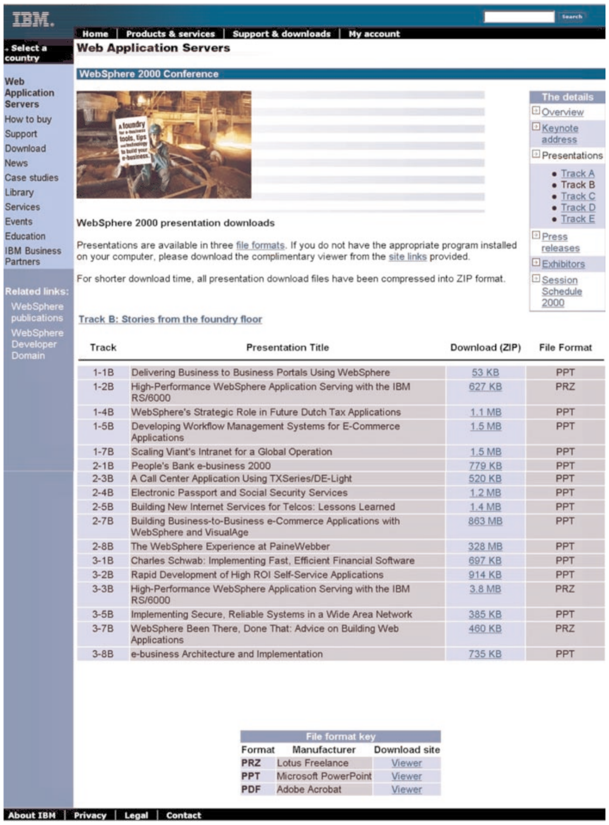
key words: HTML, layout group, transcoding, adaptation, annotation

1 Introduction

With dramatic improvements in computing power, network bandwidth, and data compression techniques, the role of World Wide Web has been diversified. The importance of the Web has increased not only in information dispatch and collection but in various fields, such as electronic commerce, remote education, and the formation of communities. Simultaneously, the tools to access the Web have also been diversified. In addition to conventional Web browser on desktop computers, people now access the Web by using various devices and methods, such as PDAs (Personal Digital Assistants), cellular phones, and voice-based browsers.

Because mobile devices such as PDAs and cellular phones offer portability and real-time connectivity, the population of mobile users is increasing explosively. Another aspect is that visually impaired people have difficulties accessing printed documents. Voice-based browsers [1, 2], which read the Web content aloud by using a synthetic voice, etc., are important information sources for such people [3-7]. However, most Web page is designed for conventional Web browsers on desktop computers. Therefore, already-existing Web content should be transcoded to be suitable for each access device and method. For this purpose, Web content transcoding systems have been developed [3-5, 8-11] .

Since most mobile devices have restrictions in screen size, adaptation of the layout, such as by splitting pages and reducing picture sizes, is required in order to display a Web page effectively [9, 12-14]. Next concern is the visual expression of Web content. To improve the density of information and support visual operations, many Web pages arrange information by using visual expressions [4, 6, 10]. For example, as shown in Fig. 1, the main content is placed in the center, and additional information such as a header, a footer, advertisements and an index surround it. This information is visually fragmented and organized into groups by using the different background colors, layout tables, spacing and so on. When a voice-based browser is used, it is very difficult to access this visual



Web Application Servers

WebSphere 2000 Conference

WebSphere 2000 presentation downloads

Presentations are available in three file formats. If you do not have the appropriate program installed on your computer, please download the complimentary viewer from the site links provided.

For shorter download time, all presentation download files have been compressed into ZIP format.

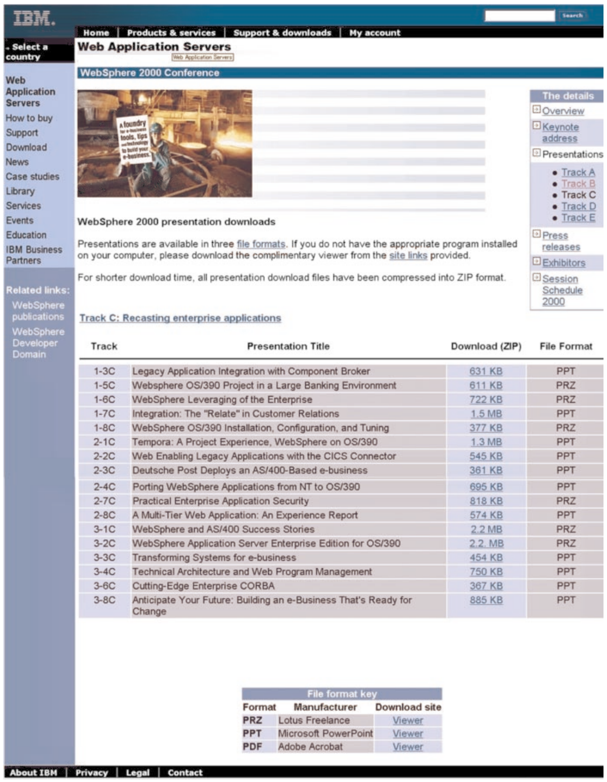
Track B: Stories from the foundry floor

Track	Presentation Title	Download (ZIP)	File Format
1-1B	Delivering Business to Business Portals Using WebSphere	53 KB	PPT
1-2B	High-Performance WebSphere Application Serving with the IBM RS/6000	627 KB	PRZ
1-4B	WebSphere's Strategic Role in Future Dutch Tax Applications	1.1 MB	PPT
1-5B	Developing Workflow Management Systems for E-Commerce Applications	1.5 MB	PPT
1-7B	Scaling Viant's Intranet for a Global Operation	1.5 MB	PPT
2-1B	People's Bank e-business 2000	779 KB	PPT
2-3B	A Call Center Application Using TXSeries/DE-Light	520 KB	PPT
2-4B	Electronic Passport and Social Security Services	1.2 MB	PPT
2-5B	Building New Internet Services for Telcos: Lessons Learned	1.4 MB	PPT
2-7B	Building Business-to-Business e-Commerce Applications with WebSphere and VisualAge	863 MB	PPT
2-8B	The WebSphere Experience at PaineWebber	328 MB	PPT
3-1B	Charles Schwab: Implementing Fast, Efficient Financial Software	697 KB	PPT
3-2B	Rapid Development of High ROI Self-Service Applications	914 KB	PPT
3-3B	High-Performance WebSphere Application Serving with the IBM RS/6000	3.8 MB	PRZ
3-5B	Implementing Secure, Reliable Systems in a Wide Area Network	385 KB	PPT
3-7B	WebSphere Been There, Done That: Advice on Building Web Applications	460 KB	PRZ
3-8B	e-business Architecture and Implementation	735 KB	PPT

File format key

Format	Manufacturer	Download site
PRZ	Lotus Freelance	Viewer
PPT	Microsoft PowerPoint	Viewer
PDF	Adobe Acrobat	Viewer

www.ibm.com/software/webservers/ws2000_present_b.html



Web Application Servers

WebSphere 2000 Conference

WebSphere 2000 presentation downloads

Presentations are available in three file formats. If you do not have the appropriate program installed on your computer, please download the complimentary viewer from the site links provided.

For shorter download time, all presentation download files have been compressed into ZIP format.

Track C: Recasting enterprise applications

Track	Presentation Title	Download (ZIP)	File Format
1-3C	Legacy Application Integration with Component Broker	631 KB	PPT
1-5C	Websphere OS/390 Project in a Large Banking Environment	611 KB	PRZ
1-6C	WebSphere Leveraging of the Enterprise	722 KB	PRZ
1-7C	Integration: The "Relate" in Customer Relations	1.5 MB	PPT
1-8C	WebSphere OS/390 Installation, Configuration, and Tuning	377 KB	PRZ
2-1C	Tempora: A Project Experience, WebSphere on OS/390	1.3 MB	PPT
2-2C	Web Enabling Legacy Applications with the CICS Connector	545 KB	PPT
2-3C	Deutsche Post Deploys an AS/400-Based e-business	361 KB	PPT
2-4C	Porting WebSphere Applications from NT to OS/390	695 KB	PPT
2-7C	Practical Enterprise Application Security	818 KB	PRZ
2-8C	A Multi-Tier Web Application: An Experience Report	574 KB	PPT
3-1C	WebSphere and AS/400 Success Stories	2.2 MB	PRZ
3-2C	WebSphere Application Server Enterprise Edition for OS/390	2.2. MB	PRZ
3-3C	Transforming Systems for e-business	454 KB	PPT
3-4C	Technical Architecture and Web Program Management	750 KB	PPT
3-6C	Cutting-Edge Enterprise CORBA	367 KB	PPT
3-8C	Anticipate Your Future: Building an e-Business That's Ready for Change	885 KB	PPT

File format key

Format	Manufacturer	Download site
PRZ	Lotus Freelance	Viewer
PPT	Microsoft PowerPoint	Viewer
PDF	Adobe Acrobat	Viewer

www.ibm.com/software/webservers/ws2000_present_c.html

Figure 1: Sample Web Pages

information [4, 6]. In addition, since voice-based browsers read out the Web content according to the order of the HTML [15] tags, users have to listen to this additional information before arriving at the content they are looking for [4, 5].

Recently, methods of extracting the updated portions of HTML documents are being studied [3, 5]. By finding differences among the target HTML documents, the old versions or the neighborhood documents, these methods make it possible for users to access the latest information easily. However, these methods have some problems. For instance, when the layouts among these HTML documents are different, these methods can't extract the updated portions. They also can't remove dynamically inserted parts of a document such as advertisements. Another serious problem is that the unchanged portion will be discarded even though it is an important part of the content. Hence, a more accurate adaptation method is required.

Annotation-based transcoding is expected to become a method for performing accurate and detailed content translation [4, 9-11] . Annotation is a kind of Metadata [9, 11, 16-20]. It has additional information about some Web content, and effective adaptation can be achieved by using it. In general, for each annotation related to URI, an XPath [21] or XPointer [22] is used to define where the annotation is attached in the document. Comments, role, importance, etc. of the designated part can be denoted as an annotation. In [4], the authors divide the Web page into semantic groups. The information about these groups, such as the role and the location (e.g., XPaths), is saved as annotations in XML [23] format. By using the annotations, the layout of the Web page can be changed without breaking up semantic groups or losing information. When semantic groups are rearranged in the order of importance, users can easily access the important information, even if they use a voice-based browser. An index of the content can be easily made from the annotation of semantic groups, and this helps mobile users understand an outline of the content in a small display. This is also useful for users using a voice-based browser.

In this way, by adding annotation to each bit of the content, accurate and detailed content translation becomes possible. One of the most difficult problems of annotation is the cost of annotating Web content. Many popular sites, such as news sites, have a great number of Web pages and add new content continually. Hence, it is almost impossible to annotate all of the content in these sites. Additionally, since Web content may be revised if time passes, an annotation has to be *robust* [14, 18]. This increases the difficulty of annotation.

In these sites, however, most of the content is created by using templates. Web page written with the same template generally has a similar layout. When using the same template, the location (e.g., XPaths) of the each semantic group and it's role are almost identical. Hence, if annotation can be shared among Web page with the same layout, the efficiency and the robustness of annotation can be improved drastically.

To solve this problem, we have developed a method for extracting the layout information from Web page. We focus on the structure and characteristics of particular HTML tags affecting the layout of Web pages. Our method calculates a distance between Web pages. If the distance is below a certain threshold, those pages can be considered as using the same layout. By using this method, a particular annotation can be applied to any Web pages that have the same layout. Therefore, the cost of annotation is reduced.

In this paper, we also explain the practical application of our layout group extracting method, that is insertion of a "Skip Navigation Link" into the Web page. Insertion of a "Skip Navigation Link" is an important item of Section 508 [24], and it improves the Web's accessibility for blind people [3, 25]. Our system can drastically reduce the Web authors' workload to improve Web accessibility.

This paper is organized as follows. In Section 2, we first introduce our layout group extracting algorithms. In Section 3, we evaluate the effectiveness of proposed algorithms through experiments. In Section 4, we will explain how our proposed algorithms can be applied to an actual system. We conclude our paper in Section 5.

2 Layout Group Extraction Algorithms

In this section, we introduce layout group extraction algorithms. To improve the density of information and visual accessibility for people with normal vision, many Web pages arrange information by using visual expressions [4, 6, 10]. In general, the main content of a Web document is placed in the center, and additional information such as a header, a footer, advertisements, and an index surround it (see Fig. 1). Most of these documents are created by inserting updated content (e.g., the latest news or search results) into templates. Hence, common layouts are used for many documents. For documents created from the same template, the location and the role of each semantic group will be similar. Therefore, if an annotation is shared among all of the content with the same layout, the efficiency of annotation can be drastically improved.

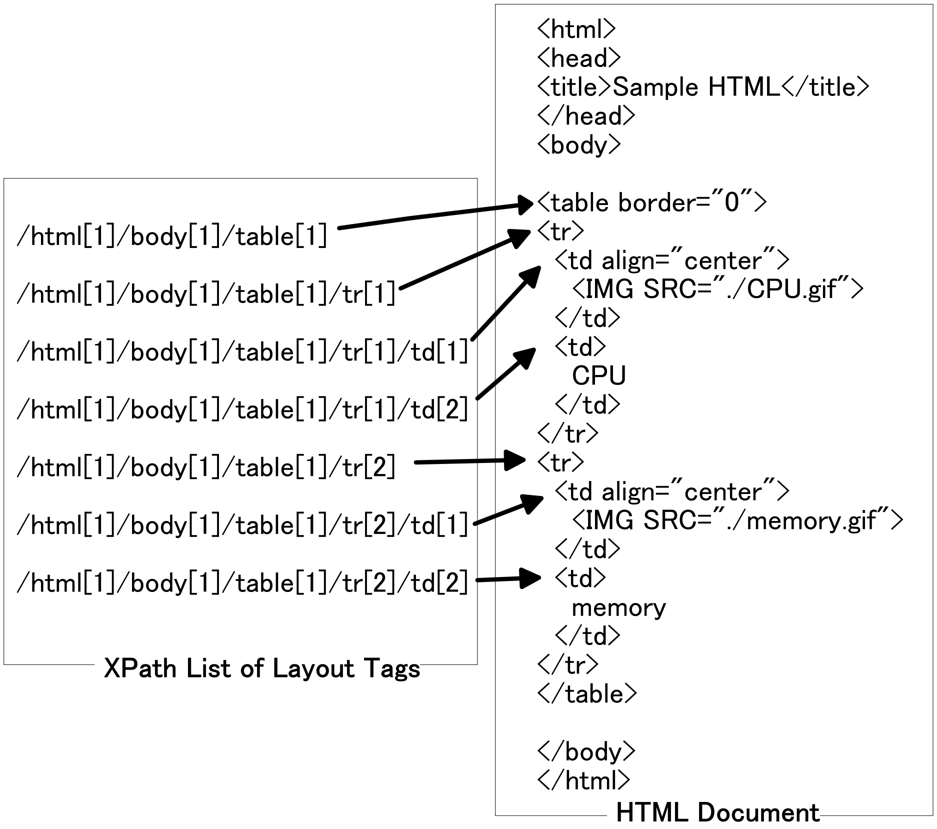
In this paper, we describe the algorithms for extracting the layout groups from Web documents. Our method focuses on the structure and characteristics of particular HTML tags affecting the layout of Web pages. Our method calculates a distance between Web pages based on the information. If the distance is below the specified threshold, those pages can be considered as being the same layout pages. By using this method, we can easily categorize Web documents from the point of view of the layout.

We first summarize HTML tags that affect the layout of Web document. Then, we will explain how to calculate a distance between Web documents and how to extract the layout group based on the calculated distances.

2.1 Layout-Related Tags and their Characteristic Uses

The structure of the block-level HTML tags is the main factor determining a Web page's layout. Some of the text-level (inline) tags also affect the layout of Web page. Examples of such tags are:

- Table (TABLE, THEAD, TBODY, TFOOT, TR, TH, TD)
- Form (BUTTON, FORM, INPUT, SELECT, TEXTAREA)
- Horizontal Line (HR)
- List (OL, UL, LI)
- Style Container (DIV, SPAN)
- Paragraph (P)



```
/html[1]/body[1]/table[1]

/html[1]/body[1]/table[1]/tr[1]

/html[1]/body[1]/table[1]/tr[1]/td[1]

/html[1]/body[1]/table[1]/tr[1]/td[2]

/html[1]/body[1]/table[1]/tr[2]

/html[1]/body[1]/table[1]/tr[2]/td[1]

/html[1]/body[1]/table[1]/tr[2]/td[2]
```

```
<html>
<head>
<title>Sample HTML</title>
</head>
<body>

<table border="0">
<tr>
 <td align="center">
  <IMG SRC="./CPU.gif">
 </td>
 <td>
  CPU
 </td>
</tr>
<tr>
 <td align="center">
  <IMG SRC="./memory.gif">
 </td>
 <td>
  memory
 </td>
</tr>
</table>

</body>
</html>
```

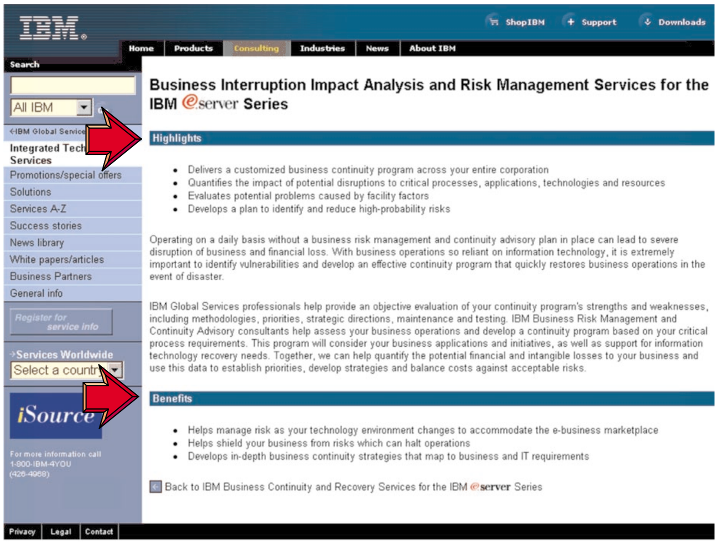
Figure 2: XPath List of Layout Tags

In this paper, we call these tags "*Layout Tags*". In many sites, the fragmentation of semantic groups and the page layout are controlled by the hierarchical structure of table elements. We can often recognize a layout group from observation of the *Layout Tag* structure. In our algorithm, we first analyze the HTML document and find the *Layout Tags*. For the purpose of dealing with the *Layout Tags* of the HTML documents, each tag is described as an XPath [21] expression. Fig. 2 shows an example of an HTML document and the extracted list of *Layout Tags.*

The enumerated *Layout Tags* captures the outline of a Web page. Fig. 1 shows Web pages having identical *Layout Tags*, and that these pages have visibly the same layout. However, not all HTML documents having an identical list of *Layout Tags* have the same layout. For example, Fig. 3 shows snapshots of two HTML documents whose lists of *Layout Tags* are identical. From this figure, it is obvious that layouts of these HTML documents are different. This mismatch is caused by the differences in how they use the *Layout Tags*. Arrows in the figure point to the corresponding "TD" tags in these documents. In the upper document, the "TD" has attributes such as "WIDTH" and "BGCOLOR", and it contains a "TEXT" element inside the block. On the other hand, in the lower document, the "TD" element has no attributes and it contains only an "IMG" element. These factors make the layouts of the Web pages different. Hence, to extract layout groups accurately, we have to consider how the *Layout Tags* are used.

In this paper, we introduce a characteristic value of a *Layout Tag* to explain how it is used. Each tag's attributes and information about the elements within the subtree are stored as its characteristic value. For instance, the attributes of "TD" are "ALIGN", "BGCOLOR", "WIDTH", "HEIGHT"', "COLSPAN", "ROWSPAN", "CLASS", etc. Furthermore, the total length of the text, and the number of images and links in a "TD" cell are also recorded as characteristic values. By using a list of *Layout Tags* with these characteristic values, we can describe the layout of a Web page in detail.

In following subsection, we describe how to calculate the distance between two Web pages based on lists of *Layout Tags* and these characteristic values. We also explain the method for extracting the layout groups from Web documents.



Business Interruption Impact Analysis and Risk Management Services for the IBM @server Series

Highlights

- Delivers a customized business continuity program across your entire corporation
- Quantifies the impact of potential disruptions to critical processes, applications, technologies and resources
- Evaluates potential problems caused by facility factors
- Develops a plan to identify and reduce high-probability risks

Operating on a daily basis without a business risk management and continuity advisory plan in place can lead to severe disruption of business and financial loss. With business operations so reliant on information technology, it is extremely important to identify vulnerabilities and develop an effective continuity program that quickly restores business operations in the event of disaster.

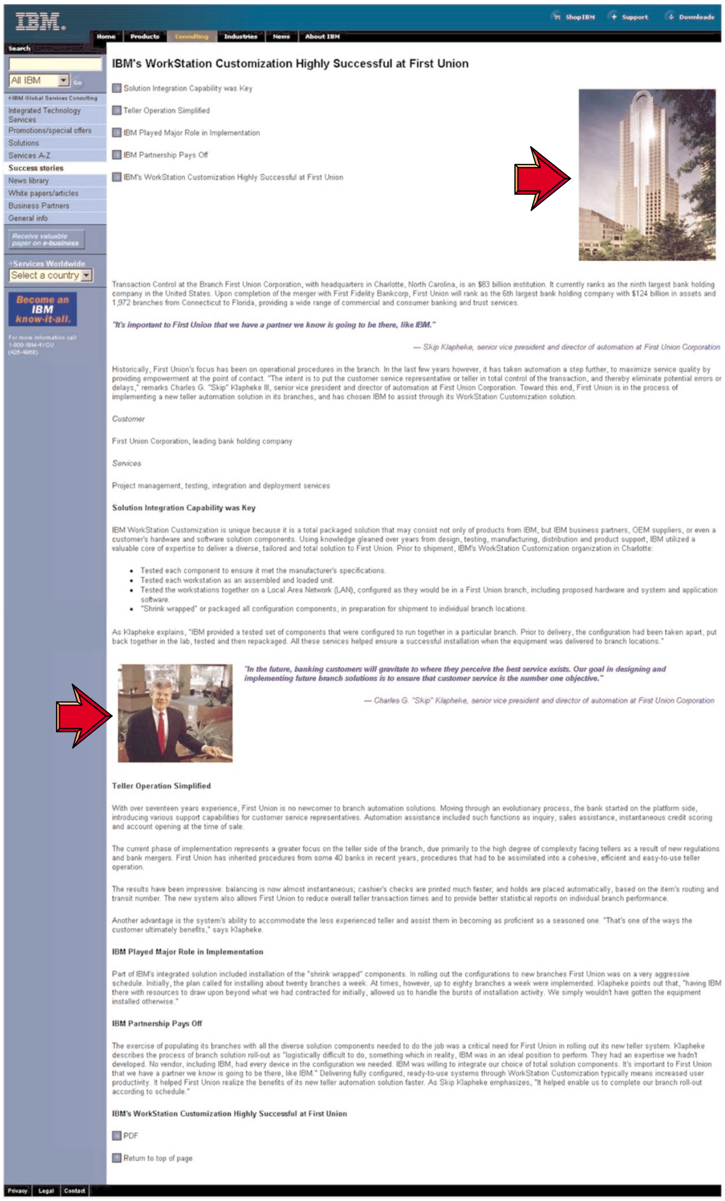
IBM Global Services professionals help provide an objective evaluation of your continuity program's strengths and weaknesses, including methodologies, priorities, strategic directions, maintenance and testing. IBM Business Risk Management and Continuity Advisory consultants help assess your business operations and develop a continuity program based on your critical process requirements. This program will consider your business applications and initiatives, as well as support for information technology recovery needs. Together, we can help quantify the potential financial and intangible losses to your business and use this data to establish priorities, develop strategies and balance costs against acceptable risks.

Benefits

- Helps manage risk as your technology environment changes to accommodate the e-business marketplace
- Helps shield your business from risks which can halt operations
- Develops in-depth business continuity strategies that map to business and IT requirements

Back to IBM Business Continuity and Recovery Services for the IBM @server Series

http://www.ibm.com/services/its/us/bcintalert.html



IBM's WorkStation Customization Highly Successful at First Union

- Solution Integration Capability was Key
- Teller Operation Simplified
- IBM Played Major Role in Implementation
- IBM Partnership Pays Off
- IBM's WorkStation Customization Highly Successful at First Union

Transaction Control at the Branch First Union Corporation, with headquarters in Charlotte, North Carolina, is an $83 billion institution. It currently ranks as the ninth largest bank holding company in the United States. Upon completion of the merger with First Fidelity Bankcorp, First Union will rank as the 6th largest bank holding company with $124 billion in assets and 1,972 branches from Connecticut to Florida, providing a wide range of commercial and consumer banking and trust services.

"It's important to First Union that we have a partner we know is going to be there, like IBM."

— Skip Klapheke, senior vice president and director of automation at First Union Corporation

Historically, First Union's focus has been on operational procedures in the branch. In the last few years however, it has taken automation a step further, to maximize service quality by providing empowerment at the point of contact. "The intent is to put the customer service representative or teller in total control of the transaction, and thereby eliminate potential errors or delays," remarks Charles G. "Skip" Klapheke III, senior vice president and director of automation at First Union Corporation. Toward this end, First Union is in the process of implementing a new teller automation solution in its branches, and has chosen IBM to assist through its WorkStation Customization solution.

Customer

First Union Corporation, leading bank holding company

Services

Project management, testing, integration and deployment services

Solution Integration Capability was Key

IBM WorkStation Customization is unique because it is a total packaged solution that may consist not only of products from IBM, but IBM business partners, OEM suppliers, or even a customer's hardware and software solution components. Using knowledge gleaned over years from design, testing, manufacturing, distribution and product support, IBM utilized a valuable core of expertise to deliver a diverse, tailored and total solution to First Union. Prior to shipment, IBM's WorkStation Customization organization in Charlotte:

- Tested each component to ensure it met the manufacturer's specifications.
- Tested each workstation as an assembled and loaded unit.
- Tested the workstations together on a Local Area Network (LAN), configured as they would be in a First Union branch, including proposed hardware and system and application software.
- "Shrink wrapped" or packaged all configuration components, in preparation for shipment to individual branch locations.

As Klapheke explains, "IBM provided a tested set of components that were configured to run together in a particular branch. Prior to delivery, the configuration had been taken apart, put back together in the lab, tested and then repackaged. All these services helped ensure a successful installation when the equipment was delivered to branch locations."

"In the future, banking customers will gravitate to where they perceive the best service exists. Our goal in designing and implementing future branch solutions is to ensure that customer service is the number one objective."

— Charles G. "Skip" Klapheke, senior vice president and director of automation at First Union Corporation

Teller Operation Simplified

With over seventeen years experience, First Union is no newcomer to branch automation solutions. Moving through an evolutionary process, the bank started on the platform side, introducing various support capabilities for customer service representatives. Automation assistance included such functions as inquiry, sales assistance, instantaneous credit scoring and account opening at the time of sale.

The current phase of implementation represents a greater focus on the teller side of the branch, due primarily to the high degree of complexity facing tellers as a result of new regulations and bank mergers. First Union has inherited procedures from some 40 banks in recent years, procedures that had to be assimilated into a cohesive, efficient and easy-to-use teller operation.

The results have been impressive: balancing is now almost instantaneous; cashier's checks are printed much faster; and holds are placed automatically, based on the item's routing and transit number. The new system also allows First Union to reduce overall teller transaction times and to provide better statistical reports on individual branch performance.

Another advantage is the system's ability to accommodate the less experienced teller and assist them in becoming as proficient as a seasoned one. "That's one of the ways the customer ultimately benefits," says Klapheke.

IBM Played Major Role in Implementation

Part of IBM's integrated solution included installation of the "shrink wrapped" components. In rolling out the configurations to new branches First Union was on a very aggressive schedule. Initially, the plan called for installing about twenty branches a week. At times, however, up to eighty branches a week were implemented. Klapheke points out that, "having IBM there with resources to draw upon beyond what we had contracted for initially, allowed us to handle the bursts of installation activity. We simply wouldn't have gotten the equipment installed otherwise."

IBM Partnership Pays Off

The exercise of populating its branches with all the diverse solution components needed to do the job was a critical need for First Union in rolling out its new teller system. Klapheke describes the process of branch solution roll-out as "logistically difficult to do, something which in reality, IBM was in an ideal position to perform. They had an expertise we hadn't developed. No vendor, including IBM, had every device in the configuration we needed. IBM was willing to integrate our choice of total solution components. It's important to First Union that we have a partner we know is going to be there, like IBM." Delivering fully configured, ready-to-use systems through WorkStation Customization typically means increased user productivity. It helped First Union realize the benefits of its new teller automation solution faster. As Skip Klapheke emphasizes, "It helped enable us to complete our branch roll-out according to schedule."

IBM's WorkStation Customization Highly Successful at First Union

PDF

Return to top of page

http://www.ibm.com/services/its/us/firstunion.html

Figure 3: Differently Laid Out Pages with Similar Structure of Tags

2.2 Layout Group Extraction

As mentioned in the previous subsection, the judgment whether Web documents utilize the same layout is done by comparing the lists of *Layout Tags* and their characteristic values. This is how we evaluate the differences between layouts as the distance between Web documents.

Assuming that A and B denote lists of *Layout Tags* in two Web documents H_A, H_B, we define the inter-page distance D as:

$$D = \sum_i d_i(T_i) \tag{1}$$

where T_i is the ith *Layout Tag* that satisfies $A \cup B$. d_i denotes the distance function of the *Layout Tag* T_i and can be given as:

$$\begin{array}{rcll} d_i(T_i) & = & W_i * \sum_j \{W_{C_{ij}} * f_{C_{ij}}(C_{A_{ij}}, C_{B_{ij}})\} & if(T_i \in (A \cap B)) \\ & = & W_i * L_i & otherwise \end{array} \tag{2}$$

where W_i denotes a weighting parameter for T_i. The value C_{ij} denotes the jth characteristics of T_i and $W_{C_{ij}}$ is a weighting parameter for C_{ij}. $C_{A_{ij}}$ and $C_{B_{ij}}$ denote the characteristic values correspond to C_{ij} in H_A, H_B, respectively. The function $f_{C_{ij}}$ represents the distance between $C_{A_{ij}}$ and $C_{B_{ij}}$. The function $f_{C_{ij}}$ returns 0 when $C_{A_{ij}}$ and $C_{B_{ij}}$ are equivalent. Otherwise, $f_{C_{ij}}$ returns a positive value. For instance, when the characteristic C_{ij} is given as a numerical value, $f_{C_{ij}}$ can be defined as a monotonically increasing function of $|C_{A_{ij}} - C_{B_{ij}}|$. If a certain *Layout Tag* appears in either H_A or H_B, L_i is added to the distance. The value L_i is a constant that depends on the kind of *Layout Tag*.

By using the above equations, we can derive the distance between the Web documents H_A and H_B. If the calculated distance is below a certain threshold T, these pages can be considered as using the same layout. Further, based on the distance D, we can divide Web documents into layout groups by means of a clustering method (i.e., K-means clustering) [26].

In the next section, we'll apply our proposed method to existing Web sites and evaluate the effectiveness of the method.

3 Experiments and Results

In this section, we investigate the effectiveness of our layout group extraction algorithm. Five existing Web sites, i.e., LYCOS News (`http://news.lycos.com/`), ZDNet (`http://www.zdnet.com`), CNN (`http://www.cnn.com/`), MLB news on CNNSI (`http://sportsillustrated.cnn.com/baseball/mlb/`) and IBM (`http://www.ibm.com/`) were examined to evaluate our method.

In what follows, we'll first explain the parameter settings of the experiments. We then explain the results of layout group extraction on existing Web sites. From the results of experiments, we will discuss issues of our method and consider how to solve these problems.

3.1 Parameter Settings

In the experiments, we used nine HTML tags as *Layout Tag*. The parameter L in Eq. (2) and characteristics of each *Layout Tag* are shown in Table 1. In the table, characteristics of *Text*, *Link*, *Image* indicate whether that element exists in the block of the designated *Layout Tag*. We choose weighting parameters $W_i = 1, W_{C_{ij}} = 1$ for all i, j and determine the distance function $f_{C_{ij}}$ as follows:

$$\begin{array}{rcll} f_{C_{ij}}(C_{A_{ij}}, C_{B_{ij}}) & = & 0 & if(C_{A_{ij}} = C_{B_{ij}}) \\ & = & 1 & otherwise \end{array} \tag{3}$$

To observe the relationships among these parameters, the Web sites' characteristics and extracted layout group, we choose these simple parameters. Moreover, since the structures of the layout tables and forms deeply affect the layout of Web documents, we employ ∞ as L for TABLE, TR, TD, TH, FORM, BUTTON, INPUT, SELECT

Table 1: Parameter Settings

Layout Tag	L	characteristics
TABLE	∞	Attributes
THEAD, TBODY, TFOOT	1	Attributes
TR	∞	Attributes
TD, TH	∞	Attributes, *Text*, *Link*, *Image*
FORM	∞	Attributes
BUTTON	∞	Attributes
INPUT	∞	Attributes
SELECT	∞	Attributes
TEXTAREA	∞	Attributes
HR	1	Attributes

and TEXTAREA tags. $L = \infty$ means that Web documents with different structures of these tags are considered as different layout pages. As described below, these parameters have to be determined appropriately to meet the purpose of annotation and the characteristics of the Web sites.

3.2 Results

In this subsection, we have investigated the effectiveness of our proposed method by using the above parameters. Figs. 4 through 6 show the relationships between the threshold value T and the number of Web pages among layout groups. In the figures, the horizontal axis shows the layout group ID and the vertical axis shows the accumulated number of Web pages in the layout groups. For the purpose for comparison, we also show the results for $(T = 0, W_{C_{ij}} = 0)$. When $(T = 0, W_{C_{ij}} = 0)$, the layout group extraction depends only on the structure of the *Layout Tags*.

From the figures, we can observe that the result of $T = 10$ is similar to $(T = 0, W_{C_{ij}} = 0)$. As mentioned in Section 2, not all HTML documents with identical lists of *Layout Tag* have the same layout. The result for $T = 10$ shows that we can't effectively divide Web documents that have identical *Layout Tag* lists into semantically identical layout groups. As the threshold value T becomes smaller, the number of layout groups increases and the number of Web documents in each layout group decreases. This means that more detailed layout group extraction can be achieved by using smaller values of T. However, there exists a trade-off between the degree of detailed and the effectiveness of the layout group extraction. For instance, $T = 1$ is too sensitive to effectively divide Web documents into the layout groups.

In usual case of news sites, there are hyper-links to the "previous" and "next" articles (see Fig. 7). However, these hyper-links are often unused. Both sides of Fig. 7 show the same portion of two Web documents in LYCOS News, and arrows in the figure point at the corresponding "TD" tag in these documents. In the left-hand figure, the word "Next" is used as a hyper-link to the next article. On the other hand, the word "Next" is used as plain text in right-hand figure. As a result, the characteristic value of *Link* is different for these documents. For $T = 1$, these documents are divided into different layout groups even though their semantic layouts are identical. Noting that these kinds of differences are frequently generated in Web documents, we can conclude that a very small threshold value for T is not appropriate for layout group extraction. From our empirical investigation, the threshold value T chosen from the range of 5 to 8 is reasonable to divide semantically different Web documents.

Table 2 summarizes the results of layout group extraction when $T = 5$. From the table, we can observe that the cost to annotate the whole Web site can be dramatically reduced by using our method. For example, this table shows that annotating the recognized layout groups on Lycos News would handle over 95% of the site's pages. For LYCOS News, ZDNet, and CNN, most of the Web documents are match well with the templates. Since they rarely employ additional tables in the inserted parts (e.g., updated news, etc.), the structures of the *Layout Tags* in these documents mainly depends on the template. As a result, most of the Web documents in these sites belong to some layout group. Therefore, approximately 500 kinds of annotation can be applied to all of LYCOS News documents

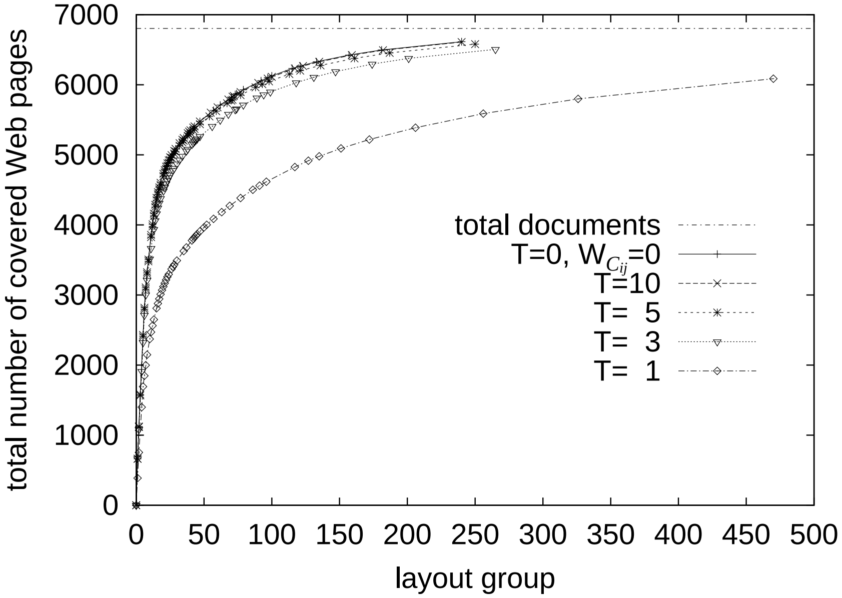


Figure 4: Layout Group and number of covered pages (news.lycos.com)

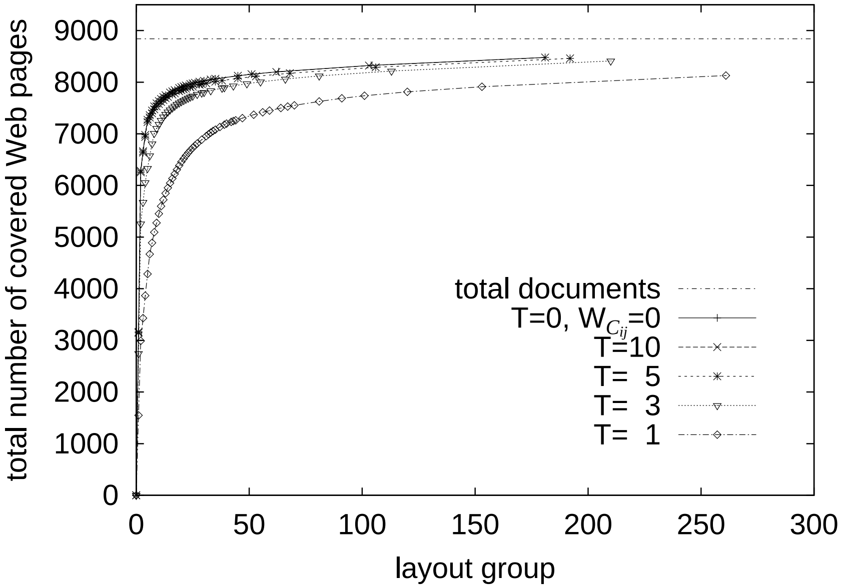


Figure 5: Layout Group and number of covered pages (www.zdnet.com)

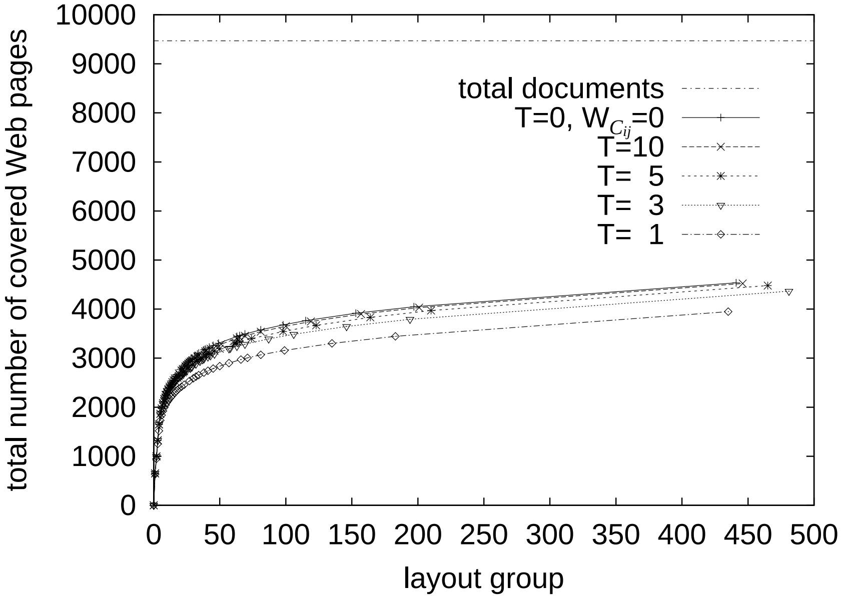


Figure 6: Layout Group and number of covered pages (www.ibm.com)

Table 2: Result of Layout Group Extraction ($T = 5$)

	# of Layout Group (coverage)	# of documents outside of layout group	total # of documents
`news.lycos.com`	250 (96.7%)	223	6803
`www.zdnet.com`	192 (95.7%)	381	8841
`www.cnn.com`	583 (95.9%)	1051	25792
`sportsillustrated.cnn.com/baseball/mlb`	411 (55.7%)	1179	2662
`www.ibm.com`	465 (47.3%)	4989	9469

even though there are 6,803 documents.

On the other hand, in the case of IBM and CNNSI, there are many Web documents outside of the layout groups, and the size of each layout group is generally smaller than that of LYCOS News or ZDNet (see Figs. 4 through 6 and Table 2). In these sites, main content with layout tables is inserted into a template. These additional layout tables mainly aims to align HTML elements (e.g., images, links, etc.) at the desired position in the document. Therefore, the documents have different *Layout Tag* structures even though they were created from the same template.

Moreover, there are many semantic tables in these sites such as the scoreboard in Fig. 8. Tables placed at the deeper parts of the hierarchy are often used as semantic table (e.g., lists of pictures, and so on). In the case of a semantic table, the whole table acts as the semantic group, and the structure of the "TR", "TH" and "TD" elements within the semantic table is frequently changed. For example, additional "TD" elements are inserted into the scoreboard in the case of an extra-inning game. Hence, the difference between the semantic table structures is often not important. In these sites, the effectiveness of our method is mainly degraded by additional layout table and semantic table issues.

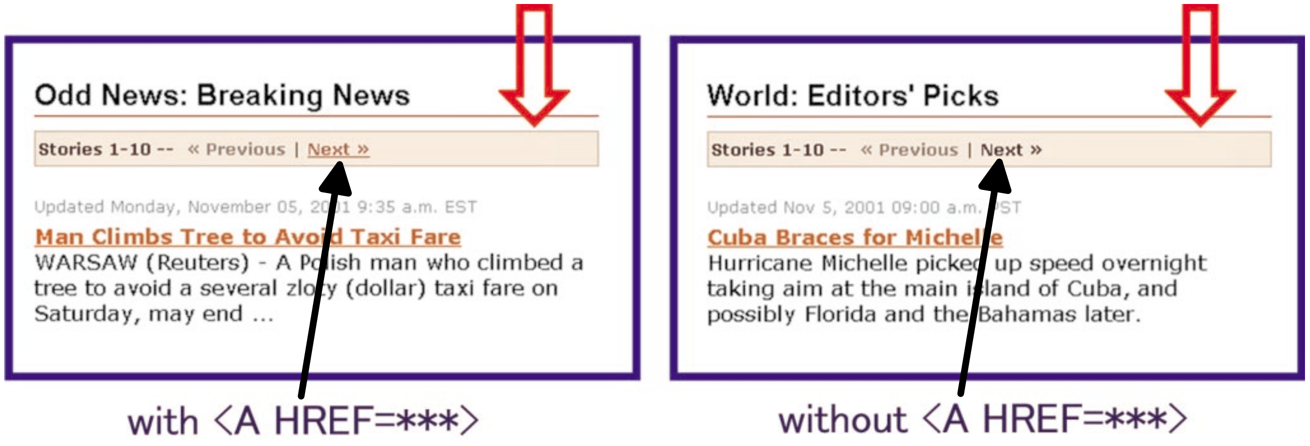
3.3 Discussion

From the observations in Subsection 3.2, our proposed method is generally effective to reduce the cost of annotation, even if the parameters are roughly selected. These parameters, however, may cause the unnecessary division of layout groups. For instance, the background color of *Layout Tags* may be changed periodically (e.g., seasonally). This would increase the distance between pages with the same layout but different colors. In this case, we can easily prevent the unnecessary division by decreasing the weighting of the parameter for "BGCOLOR". However, since some sites change the background color to reflect the different role of a tag, the reduction of the BGCOLOR's weight is not always valid. Further investigation is required to determine appropriate parameter settings to meet the purpose of annotation and reflect the characteristics of the Web sites.

To improve the effectiveness of layout group extraction, we have to solve the issues of additional layout tables and semantic tables. One of the realistic methods to solve these issues is that the weighting parameter W_i decreases according to the depth of the *Layout Tag* T_i. This method makes it possible to ignore the unimportant differences, such as an additional "TD" added to the semantic table. However, this method can't distinguish the layout table in the template from semantic tables and additional layout tables. As a result, some Web documents are considered as using the same layout, even though they employ different layouts. Therefore, we have to carefully select the combination of W_i and L_i to meet the characteristics of the Web site.

Another method to solve these problems is classification of the tables. We can estimate what role a table serve by using information such as the *Layout Tags* structure and characteristic values. For instance, a table whose "TD" elements have another table within a cell may be used as a layout table. On the other hand, a table whose "TD" elements only include short text, is probably a semantic table. We can expect a table located the deepest hierarchy which has only one "TD" cell with layout attributes (i.e., WIDTH, HEIGHT, ALIGN, etc.) to be additional layout table.

By decreasing the weighting parameter of tables classified as semantic tables or additional layout tables, more accurate layout group extraction becomes possible. Fig. 9 shows the effectiveness of the table classification. We


Odd News: Breaking News

Stories 1-10 -- « Previous | Next »

Updated Monday, November 05, 2001 9:35 a.m. EST

Man Climbs Tree to Avoid Taxi Fare

WARSAW (Reuters) - A Polish man who climbed a tree to avoid a several zloty (dollar) taxi fare on Saturday, may end ...

World: Editors' Picks

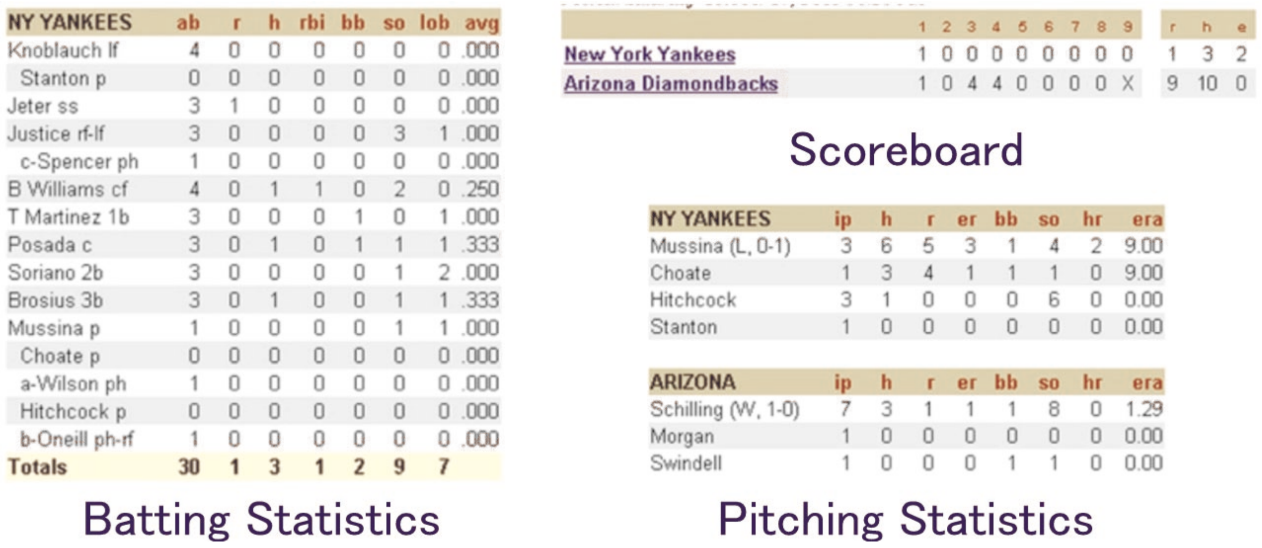
Stories 1-10 -- « Previous | Next »

Updated Nov 5, 2001 09:00 a.m. EST

Cuba Braces for Michelle

Hurricane Michelle picked up speed overnight taking aim at the main island of Cuba, and possibly Florida and the Bahamas later.

Figure 7: Semantically identical but characteristic value is different


NY YANKEES	ab	r	h	rbi	bb	so	lob	avg
Knoblauch lf	4	0	0	0	0	0	0	.000
Stanton p	0	0	0	0	0	0	0	.000
Jeter ss	3	1	0	0	0	0	0	.000
Justice rf-lf	3	0	0	0	0	3	1	.000
c-Spencer ph	1	0	0	0	0	0	0	.000
B Williams cf	4	0	1	1	0	2	0	.250
T Martinez 1b	3	0	0	0	1	0	1	.000
Posada c	3	0	1	0	1	1	1	.333
Soriano 2b	3	0	0	0	0	1	2	.000
Brosius 3b	3	0	1	0	0	1	1	.333
Mussina p	1	0	0	0	0	1	1	.000
Choate p	0	0	0	0	0	0	0	.000
a-Wilson ph	1	0	0	0	0	0	0	.000
Hitchcock p	0	0	0	0	0	0	0	.000
b-Oneill ph-rf	1	0	0	0	0	0	0	.000
Totals	30	1	3	1	2	9	7	

	1	2	3	4	5	6	7	8	9	r	h	e
New York Yankees	1	0	0	0	0	0	0	0	0	1	3	2
Arizona Diamondbacks	1	0	4	4	0	0	0	0	X	9	10	0

NY YANKEES	ip	h	r	er	bb	so	hr	era
Mussina (L, 0-1)	3	6	5	3	1	4	2	9.00
Choate	1	3	4	1	1	1	0	9.00
Hitchcock	3	1	0	0	0	6	0	0.00
Stanton	1	0	0	0	0	0	0	0.00

ARIZONA	ip	h	r	er	bb	so	hr	era
Schilling (W, 1-0)	7	3	1	1	1	8	0	1.29
Morgan	1	0	0	0	0	0	0	0.00
Swindell	1	0	0	0	1	1	0	0.00

Figure 8: Sample of a semantic table

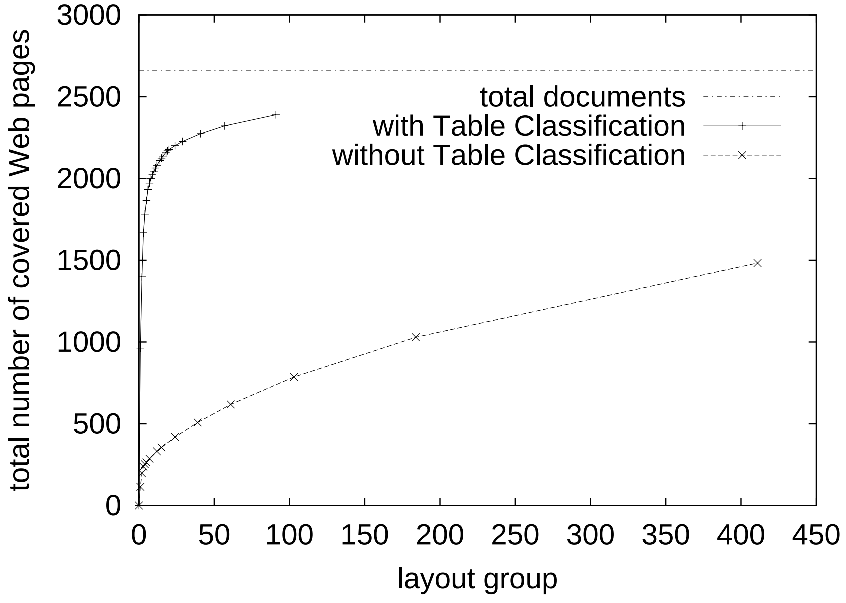


Figure 9: Effectiveness of table classification

employ simple heuristics to find a semantic table in the Web documents in CNNSI. The heuristics can be expressed as follows:

- If all of following conditions are satisfied, the table can be considered to be a semantic table.
  - TD, TH elements within the table don't contain another table.
  - Total number of TD and TH elements within each TR is identical.
  - The length of the strings within each TD, TH element doesn't exceed 20 characters.

For a semantic table, we decrease all weighting parameters and the constant value L to 0.5. As shown in Fig. 9, we can see that the efficiency of the layout group extraction is much improved by using such table classification.

To achieve effective classification, we have to make some heuristics for each kind of table, e.g., image list, link list or scoreboard, etc. The existing pattern discovery techniques, such as IEPAD [27], are helpful to discover the Web site specific classification heuristics.

The classification is a good approach to describe the characteristics of *Layout Tags*. For example, a "TD" element which include some hyper-links and where most of the "TD" cell is occupied by "A" elements is described as a "list of hyper-links". They are also subdivided into "lists of hyper-links to inside the site" and "lists of hyper-links to outside of the site". Other classifications, such as "long text", "short text", "numerical data", "image", etc., also improve the accuracy of the characteristics of *Layout Tags*.

For practical use, an annotator sometimes requires more detailed groupings. Since our method depends mainly on the templates of Web content, various kinds of content such as sports news and business news are in the same layout group. To annotate the main content as "business news" or "sports news", we have to divide the extracted layout group again. In these case, we can employ the existing methods such as filtering based on the URL by using regular expressions. An XPath expression, such as "//table[1]/td[1][(contains(string(),'business news')]", can be used to determine the detailed grouping. Since our method collects Web documents with same layout in advance, we can easily make XPath expressions without considering the robustness issues [14, 18].

4 Example Application

In what follows, we will explain how our results presented in Sections 2 and 3 can be applied to an actual application. In this paper, we use the insertion of the "Skip Navigation Link" into the Web content as a sample application.

Insertion of a "Skip Navigation Link" is an important item of Section 508 [24]. The "Skip Navigation Link" is very useful, since it allows people using voice-based browsers to skip the unrelated information to the main content at the top of the page and to jump directly to the main content. These links, however, are still rare on actual Web sites. Therefore, we have been developing a function that inserts a "Skip Navigation Link" into existing Web content, and this function is integrated into the Web content accessibility compliance tool [25].

An abstract model of the system organization is illustrated in Figure 10. First, user inputs target Web pages into the compliance tool. Then, the site-wide analyzing engine in the compliance tool starts checking the accessibility of the target content. At the same time, the Web pages are divided into some layout groups by using our layout group extracting algorithms. Each layout group is assigned a "Layout ID". Hence, Web pages which have similar layouts to a certain page can be easily detected by their Layout IDs.

User can easily check and correct non-compliant Web pages by using a GUI in our system. Snapshots of the GUI for insertion of a "Skip Navigation Link" are shown in Figs. 11 and 12. User first selects the target Web page from the list of not-compliant URLs, and then specifies the position of the main content in the browser view by using a mouse operation (see Figure 11). The specified position of the main content is stored as an XPath expression. If the site owner wants to directly repair his documents, the system inserts the "Skip Navigation Link", which jumps to the specified main content, at the top of the target Web page. When a user can't change the original documents, this tool can still make annotations. This kind of annotation contains information about the target Layout ID and an XPath for the main content. By using this annotation, a "Skip Navigation Link" is inserted by a the transcoder.

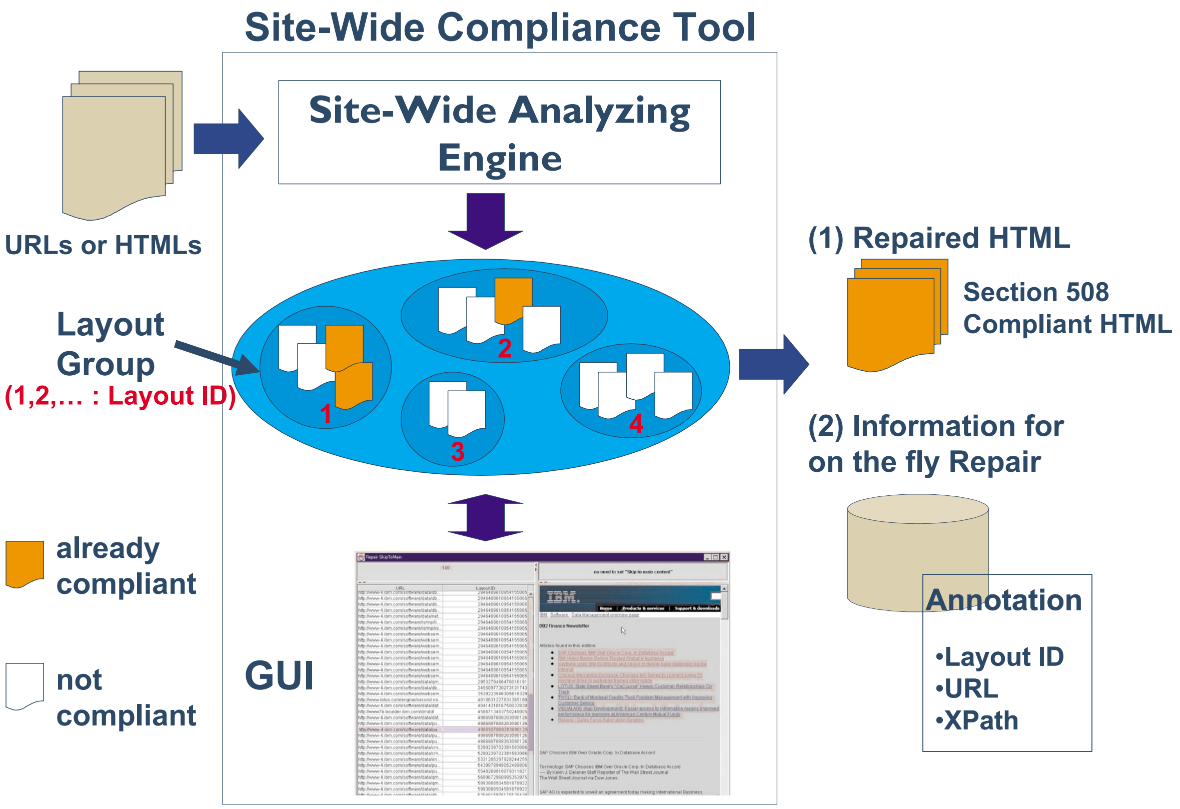


Figure 10: Compliance Flow

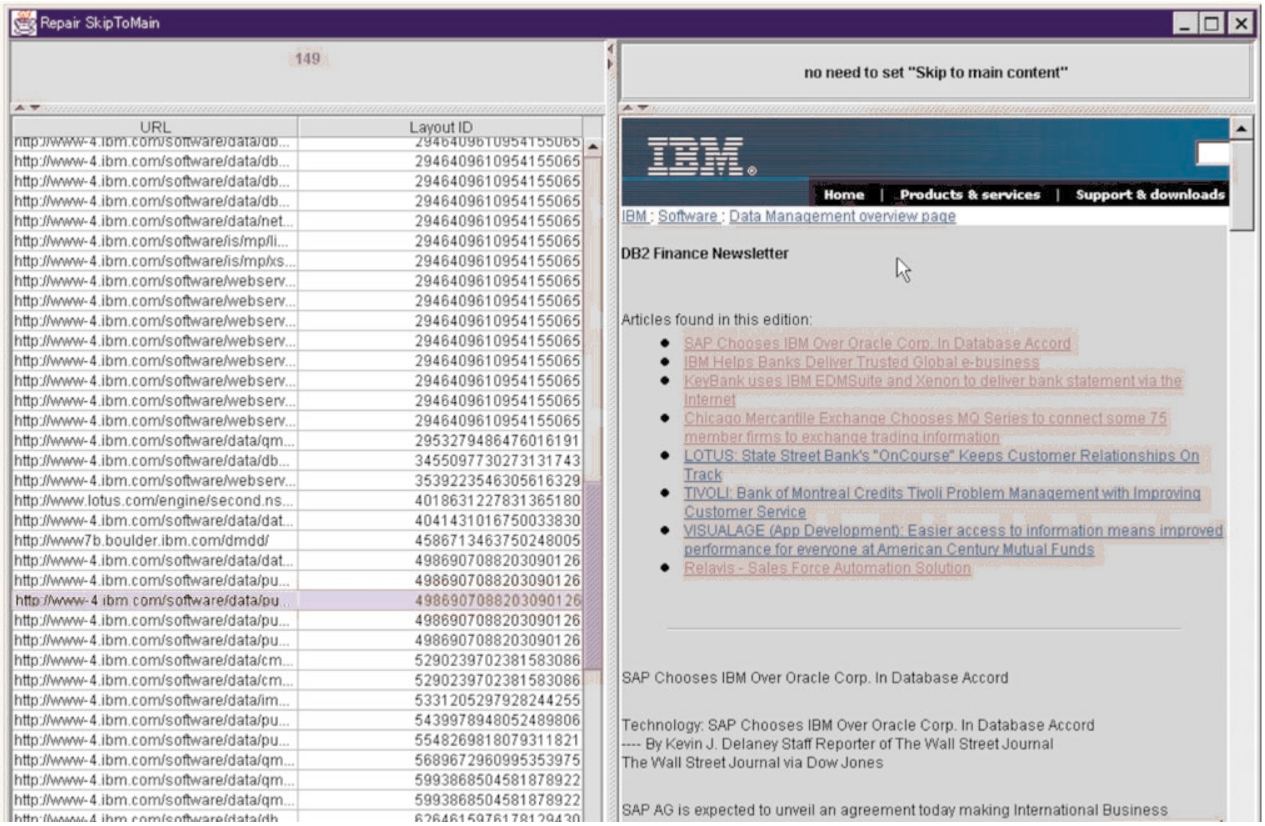
If there are non-compliant pages whose Layout ID is identical to the target page, a confirmation window pops up as shown in Figure 12. In the browser view of the confirmation window, the probable main content of the similarly laid out page is highlighted. In most cases, the main content of similarly laid out pages are located in the same position in the Web page. Therefore, we can easily estimate the position of the main content by using the XPath derived from the first target page. In almost all cases, the user only has to push a confirmation button. When the user notices that it is pointing to the wrong place, he can specify another position using the browser view.

With these methods, the user can insert "Skip Navigation Links" into similarly laid out pages at the same time. Figure 13 shows an example of a propagation of "Skip Navigation Link". From the observations in Section 3, a few kinds of layout group can cover almost all of the documents in a major site. Furthermore, even if a new page is created, we can apply the existing annotation to the page, as long as the same layout is being used in the Web site. Hence, we can conclude that the cost of "Skip Navigation Link" insertion for the whole Web site can be dramatically reduced.

5 Conclusion

In this paper, we have investigated the applicability of our proposed Layout Group Extraction algorithm. Our method calculates the distances between Web pages based on the structure and characteristics of the *Layout Tags*. Based on the calculated distance, we can extract Web documents with same layout as a Layout Group. Through experiments, we have shown that with our proposed algorithms, the required cost of annotating the whole Web site can be drastically reduced. Our method is also useful to improve the robustness and accuracy of annotations.

To obtain more detailed layout group extraction, several topics remains to be investigated. How to make effective heuristics for classification of the *layout tags* is our future research topic. We also have to investigate the appropriate parameter settings to match the purpose of the annotations and the characteristics of the Web sites.



149

no need to set "Skip to main content"

URL	Layout ID
http://www-4.ibm.com/software/data/db...	2946409610954155065
http://www-4.ibm.com/software/data/db...	2946409610954155065
http://www-4.ibm.com/software/data/db...	2946409610954155065
http://www-4.ibm.com/software/data/db...	2946409610954155065
http://www-4.ibm.com/software/data/net...	2946409610954155065
http://www-4.ibm.com/software/is/mp/li...	2946409610954155065
http://www-4.ibm.com/software/is/mp/xs...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/data/qm...	2953279486476016191
http://www-4.ibm.com/software/data/db...	3455097730273131743
http://www-4.ibm.com/software/webserv...	3539223546305616329
http://www.lotus.com/engine/second.ns...	4018631227831365180
http://www-4.ibm.com/software/data/dat...	4041431016750033830
http://www7b.boulder.ibm.com/dmdd/	4586713463750248005
http://www-4.ibm.com/software/data/dat...	4986907088203090126
http://www-4.ibm.com/software/data/pu...	4986907088203090126
http://www-4.ibm.com/software/data/pu...	4986907088203090126
http://www-4.ibm.com/software/data/pu...	4986907088203090126
http://www-4.ibm.com/software/data/pu...	4986907088203090126
http://www-4.ibm.com/software/data/cm...	5290239702381583086
http://www-4.ibm.com/software/data/cm...	5290239702381583086
http://www-4.ibm.com/software/data/im...	5331205297928244255
http://www-4.ibm.com/software/data/pu...	5439978948052489806
http://www-4.ibm.com/software/data/pu...	5548269818079311821
http://www-4.ibm.com/software/data/qm...	5689672960995353975
http://www-4.ibm.com/software/data/qm...	5993868504581878922
http://www-4.ibm.com/software/data/qm...	5993868504581878922
http://www-4.ibm.com/software/data/db...	6264615976178129430

IBM : Software : Data Management overview page

DB2 Finance Newsletter

Articles found in this edition:

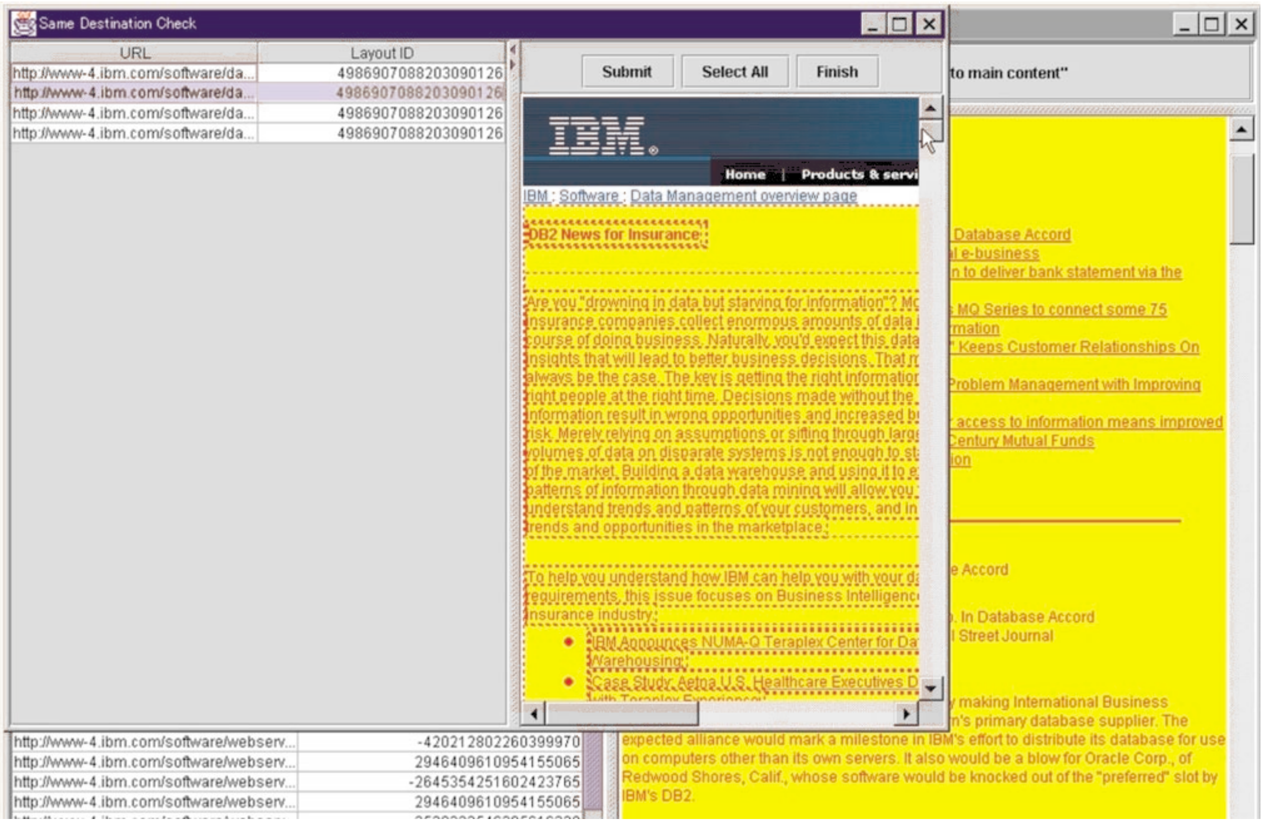
- SAP Chooses IBM Over Oracle Corp. In Database Accord
- IBM Helps Banks Deliver Trusted Global e-business
- KeyBank uses IBM EDMSuite and Xenon to deliver bank statement via the Internet
- Chicago Mercantile Exchange Chooses MQ Series to connect some 75 member firms to exchange trading information
- LOTUS: State Street Bank's "OnCourse" Keeps Customer Relationships On Track
- TIVOLI: Bank of Montreal Credits Tivoli Problem Management with Improving Customer Service
- VISUALAGE (App Development): Easier access to information means improved performance for everyone at American Century Mutual Funds
- Relavis - Sales Force Automation Solution

SAP Chooses IBM Over Oracle Corp. In Database Accord

Technology: SAP Chooses IBM Over Oracle Corp. In Database Accord
---- By Kevin J. Delaney Staff Reporter of The Wall Street Journal
The Wall Street Journal via Dow Jones

SAP AG is expected to unveil an agreement today making International Business

Figure 11: Insertion of "Skip Navigation Link"



URL	Layout ID
http://www-4.ibm.com/software/da...	4986907088203090126
http://www-4.ibm.com/software/da...	4986907088203090126
http://www-4.ibm.com/software/da...	4986907088203090126
http://www-4.ibm.com/software/da...	4986907088203090126

IBM : Software : Data Management overview page

DB2 News for Insurance

Are you "drowning in data but starving for information"? Mo
insurance companies collect enormous amounts of data i
course of doing business. Naturally, you'd expect this data
insights that will lead to better business decisions. That m
always be the case. The key is getting the right informatior
right people at the right time. Decisions made without the
information result in wrong opportunities and increased bu
risk. Merely relying on assumptions or sifting through large
volumes of data on disparate systems is not enough to st
of the market. Building a data warehouse and using it to e
patterns of information through data mining will allow you
understand trends and patterns of your customers, and in
trends and opportunities in the marketplace.

To help you understand how IBM can help you with your da
requirements, this issue focuses on Business Intelligence
insurance industry.

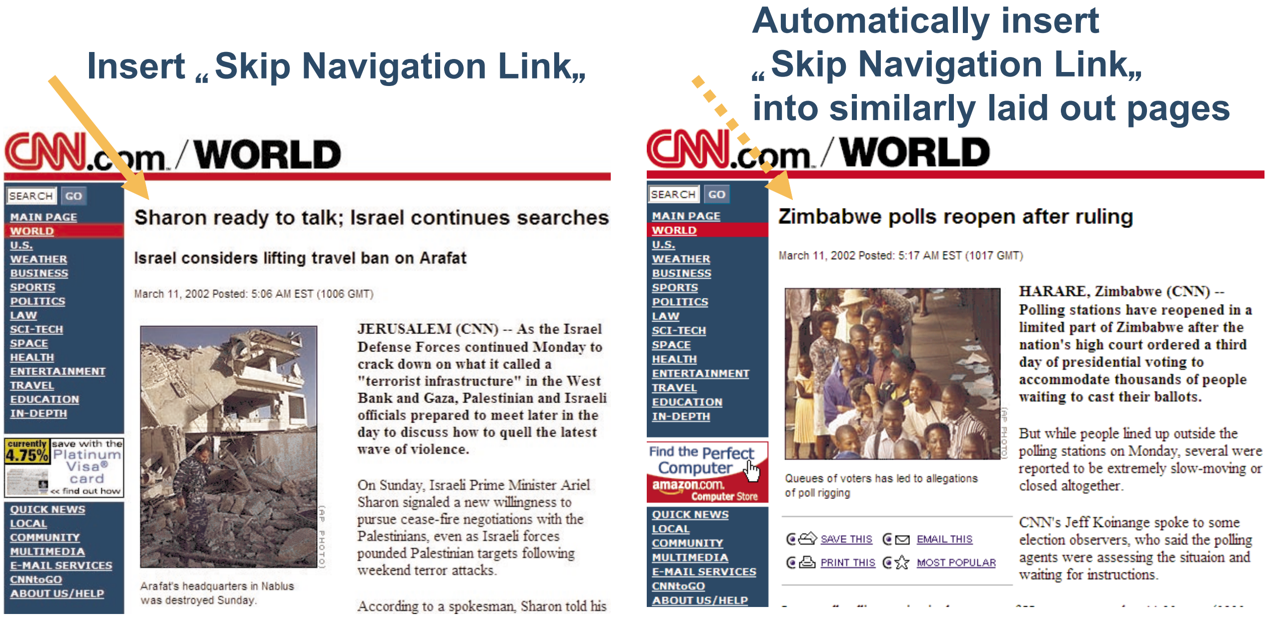
- IBM Announces NUMA-Q Teraplex Center for Da Warehousing
- Case Study: Aetna U.S. Healthcare Executives D

to main content"

URL	Layout ID
http://www-4.ibm.com/software/webserv...	-420212802260399970
http://www-4.ibm.com/software/webserv...	2946409610954155065
http://www-4.ibm.com/software/webserv...	-2645354251602423765
http://www-4.ibm.com/software/webserv...	2946409610954155065

y making International Business
n's primary database supplier. The
expected alliance would mark a milestone in IBM's effort to distribute its database for use
on computers other than its own servers. It also would be a blow for Oracle Corp., of
Redwood Shores, Calif., whose software would be knocked out of the "preferred" slot by
IBM's DB2.

Figure 12: Insertion of "Skip Navigation Link" to similar layout pages


Insert „Skip Navigation Link„

Automatically insert „Skip Navigation Link„ into similarly laid out pages

Sharon ready to talk; Israel continues searches

Israel considers lifting travel ban on Arafat

March 11, 2002 Posted: 5:06 AM EST (1006 GMT)

Arafat's headquarters in Nablus was destroyed Sunday.

JERUSALEM (CNN) -- As the Israel Defense Forces continued Monday to crack down on what it called a "terrorist infrastructure" in the West Bank and Gaza, Palestinian and Israeli officials prepared to meet later in the day to discuss how to quell the latest wave of violence.

On Sunday, Israeli Prime Minister Ariel Sharon signaled a new willingness to pursue cease-fire negotiations with the Palestinians, even as Israeli forces pounded Palestinian targets following weekend terror attacks.

According to a spokesman, Sharon told his

Zimbabwe polls reopen after ruling

March 11, 2002 Posted: 5:17 AM EST (1017 GMT)

Queues of voters has led to allegations of poll rigging

HARARE, Zimbabwe (CNN) -- Polling stations have reopened in a limited part of Zimbabwe after the nation's high court ordered a third day of presidential voting to accommodate thousands of people waiting to cast their ballots.

But while people lined up outside the polling stations on Monday, several were reported to be extremely slow-moving or closed altogether.

CNN's Jeff Koinange spoke to some election observers, who said the polling agents were assessing the situaion and waiting for instructions.

Figure 13: propagation of "Skip Navigation Link"

References

[1] C. Asakawa and T. Itoh, "User Interface of a Home Page Reader," *Proceedings of ACM ASSETS '98*, pp. 149–156, April 1998.

[2] "pwWebSpeak." The Productivity Works, `http://www.prodworks.com/`.

[3] H. Takagi and C. Asakawa, "Transcoding Proxy for Nonvisual Web Access," *Proceedings of ACM ASSETS 2000*, pp. 164–171, November 2000.

[4] C. Asakawa and H. Takagi, "Annotation-Based Transcoding for Nonvisual Web Access," *Proceedings of ACM ASSETS 2000*, pp. 172–179, November 2000.

[5] T. Ebina, S. Igi, and T. Miyake, "Fast Web by Using Updated Content Extraction and a Bookmark Facility," *Proceedings of ACM ASSETS 2000*, pp. 64–71, November 2000.

[6] T. Sullivan and R. Matson, "Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites," *Proceedings of ACM CUU 2000*, pp. 139–144, November 2000.

[7] D. Sloan, P. Gregor, M. Rowan, and P. Booth, "Accessible Accessibility," *Proceedings of ACM CUU 2000*, pp. 96–101, November 2000.

[8] R. Barrett and P. P. Maglio, "Intermediaries: new places for producing and manipulating Web content," *Proceedings of the 7th International World Wide Web Conference*, pp. 509–518, April 1998.

[9] M. Hori, G. Kondoh, K. Ono, S. Hirose, and S. Singhal, "Annotation-Based Web content transcoding," *Proceedings of the 9th International World Wide Web Conference*, pp. 197–211, May 2000.

[10] A. W. Huang and N. Sundaresan, "A Semantic Transcoding System to Adapt Web Services for Users with Disabilities," *Proceedings of ACM ASSETS 2000*, pp. 156–163, November 2000.

[11] K. Nagao, Y. Shirai, and K. Squire, "Semantic Annotation and Transcoding: Making Web Content More Accessible," *IEEE MultiMedia*, vol. 8, pp. 69–81, April 2001.

[12] M. Jones, G. Marsden, N. Mohd-Nasir, K. Boone, and G. Buchanan, "Improving Web interaction on samll displays," *Proceedings of the 8th International World Wide Web Conference*, pp. 1129–1137, May 1999.

[13] C. R. Anderson, P. Domingos, and D. S. Weld, "Personalizing Web Sites for Mobile Users," *Proceedings of the 10th International World Wide Web Conference*, pp. 565–575, May 2001.

[14] J. Freire, B. Kumar, and D. Lieuwen, "WebViews: Accessing Personalized Web Content and Services," *Proceedings of the 10th International World Wide Web Conference*, pp. 576–586, May 2001.

[15] "HTML 4.01 Specification." W3C Recommendation, `http://www.w3.org/TR/1999/REC-html401-19991224`, December 1999.

[16] "Resource Description Framework (RDF) Model and Syntax Specification." W3C Note, `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222`, February 1999.

[17] "Annotation of Web Content for Transcoding." W3C Note, `http://www.w3.org/1999/07/NOTE-annot-19990710`, July 1999.

[18] J. Kahan and M.-R. Koivunen, "Annotea: An Open RDF Infrastructure for Shared Web Annotations," *Proceedings of the 10th International World Wide Web Conference*, pp. 623–632, May 2001.

[19] "RDF Site Summary (RSS) 1.0." RSS-DEV Working Group, `http://web.resource.org/rss/1.0/spec`, December 2000.

[20] J. Kunze, "Encoding dublin core metadeta in html," *IETF RFC2731*, December 1999.

[21] "XML Path Language (XPath) Version 1.0." W3C Recommendation, `http://www.w3.org/TR/1999/REC-xpath-19991116`, November 1999.

[22] "XML Pointer Language (XPointer) Version 1.0." W3C Candidate Recommendation, `http://www.w3.org/TR/2001/CR-xptr-20010911`, September 2001.

[23] "Extensible Markup Language (XML) 1.0 (Second Edition)." W3C Recommendation, `http://www.w3.org/TR/2000/REC-xml-20001006`, October 2000.

[24] "Electronic and Information Technology Accessibility Standards." the Federal Register, `http://www.access-board.gov/sec508/508standards.htm`, December 2000.

[25] C. Asakawa, K. Fukuda, H. Takagi, and J. Maeda, "An Automatic Web Content Accessibility Compliance Tool for Section 508," *Proceedings of CSUN 2002*, March 2002.

[26] J. A. Hartigan, *Clustering algorithms*. John Wiley & Sons, 1975.

[27] C.-H. Chang and S.-C. Lui, "IEPAD: Information Extraction Based on Pettern Discovery," *Proceedings of the 10th International World Wide Web Conference*, pp. 681–688, May 2001.