A novel web application called "My Portal Viewer (MPV)" has been developed to provide web users with higher quality content, which is needed due to a rapidly growing amount of content on the web. It provides fused news to the user based on two viewpoints through a user friendly interface and the user's preferences. MPV automatically selects and merges content from many news pages based on the user's interest and knowledge after gathering these pages from various web sites. Our unique approach is that the layout of the MPV page is applied to the users' favorite news portal page, and a part of the original content is replaced by the fused content. Whenever a user accesses an MPV page after browsing other news pages, he/she can acquire the desired content efficiently because MPV presents a refreshed page based on the user's behavior, which reflects his/her interests and knowledge. In addition to the MPV framework, methods that are based on user reference for replacing and selecting have been developed using an HTML table model.

1. INTRODUCTION

As the amount of web content increases, novel applications that provide higher quality content are becoming more desired by users. In particular, applications that select information based on user preferences from gathering information on the web are needed for information retrieval. Also needed is a system that can efficiently merge high-quality content gathered from the web for each user. We have developed an application we call "My Portal Viewer (MPV)" that merges information gathered from many web pages using usage history based on the user's preferences, interests, and knowledge.

Existing fusion systems provide a summary of content distributed over various web sites [1, 2, 3]. The summarized content is based on the structure and sentences of the collected pages. Users can see the content on one screen without having to search for and access many web sites (Fig.1(a)). However, the user has to look for the target information in many lists for various categories because the fusion system has its own method of categorization, and the categories may not correspond to the user's viewpoint. Furthermore, novice users are not familiar with the various web interfaces and thus always have trouble finding the target information.

The proposed MPV application can provide targets efficiently to each user without the confusion of retrieval and operation (Fig.1(b)). MPV collects and stores pages from various web sites by crawling through web sites, and it provides an MPV page that fuses the stored pages. Our unique approach is that the layout of the MPV page and part of the original content is replaced by the fused content. MPV uses the layout of the users' favorite web page as the MPV page. Specifically, if the user frequently accesses a particular page, he/she can view the fused information through it.

The content fusion is based on the user's preferences related to his/her interests and knowledge. The gathered information is categorized based on these preferences and by specifying what is desired. The categories are defined based on the user's history of operations. As the history changes over time, the content in the MPV page is dynamically altered, reflecting the changes in the user's interest and knowledge. By layout and category, the user can easily locate desired information from the pile gathered.

In the next section, we describe related work on web pages fusion, and in Section 3 we describe the concept of MPV. Then in Sections 4 and 5, we describe the method used to construct a user-friendly layout and the adaptation of the user's interests and knowledge to fuse news pages into a page that matches the user's preference. In Section 6, we conclude the paper with a short summary and a look at future work.

2. RELATED WORK

Many studies have investigated summarization technology, including collecting, clustering, and text categorization. Columbia's Newsblaster [1] is an online news summarization system in which collected news is categorized by event using a topic detection and tracking (TDT) style and TF-IDF. When each news item has been assigned into six categories, each category is summarized using language technology. The user can then read a brief summary of
an event based on sources. However, the user always has to look for the summary of a particular item among various others because categorization and summarization are not based on the user’s preferences.

RSS reader [3, 4] provides titles of collected news content, and the user can browse the titles based on date, subject, or class and read the original page content after clicking on the link for the desired title. However, these applications also do not make it easy to find a particular news item, and the interface is not user friendly because he/she cannot modify the layout to match his/her preferences.

MyYahoo! [5] provides selecting categories based on a user’s selection from various content. However, these categories are defined by MyYahoo! system, and the user cannot select categories based on his/her interest and knowledge. Furthermore, the layout can be modified, but it is hard to reconstruct a more user-friendly layout.

3. CONCEPT OF MPV

As shown in figure 2, the MPV interface appears as a toolbar in the web browser. First, the user inputs the URL of a desired web page into the blank box in the toolbar of the MPV interface; in the example shown, the URL entered was that of the CNN home page. When the user hits the enter key, the MPV interface sends the entered URL to the MPV site. MPV extracts the layout of the entered URL page and the content on the layout. If MPV has not stored the specified page, MPV gets this page. MPV then creates an MPV page that changed the content on part of the specified page to fused content of many stored pages. The fused content is based on user’s preferences. The result is that the user can easily access desired content through a preferred layout.

In the figure 2, after user entered the URL, a MPV page appears which is based on the CNN portal page layout as transformed by MPV site. The layout is not converted, only some of the content: the categories, the top news item with an image, and the news headlines in each category. In this example, the original categories "World", "World Business", and "Technology" were changed to an Iraq-related news category, a baseball player-related news category, and a movie-related category because the user was interested in stories related to these categories. The top news item with an image and the news headlines in each category were also changed to match the user’s preferences.

The categories are modified dynamically based on the user’s history of browsing. If the user starts reading stories related to "football more often", this category is made into a new category, and a category that is no longer read as much is demoted. Once the user has read a story, it is deleted from the list. Stories are also deleted after a set period of time (hourly, daily, monthly, etc.) by the user. An user can thus quickly access current stories that have not been read yet.

On the MPV site, the system collects news pages by crawling through web sites and archives titles and meta data of all news pages and links of the titles. When the MPV receives a URL from a user, it analyzes the layout of the specified page and changes the three content items specified above to the specified content.

In the next section, we describe the method used to extract content from the original web pages and the one used to select a modified item based on the user’s preferences.

4. CONTENT EXTRACTION

MPV changes only three content items on the original page through the original layout. It does this by analyzing the original page and detecting the layout, original content, and new content.
4.1. Original Layout Extraction

We analyzed the portal page of six example major news sites. The basic layout for these pages was used to construct an HTML table model. All the pages had five content items:

- each site logo image
- category keyword
- a top news item with an image
- a list of news headlines in each category
- and advertisements.

We used each characteristic in the table model and the second, third, and fourth items to change the content.

We used the HTML table model to acquire the x-y coordinates for the portal page layout.

Using the HTML table model enables authors to arrange data that consists of text, images, links, other tables, etc., into rows and columns of cells [6]. The directionality of a table by default is left-to-right (column 0 is on the left and row 0 is on top). The TABLE element contains the formatting that specifies the rows, columns, and other content. The number of rows in a table is equal to the number of TR tags contained in the TABLE element, and the rows are grouped according to the ROWSPAN values. The width of the table is equal to the WIDTH elements, and the number of columns is equal to the total of the number of TH or TD tags and COLSPAN values.

We derive the x-y coordinate in each table from the above definitions and calculate the table area based on the x-y coordinates. Figure 3 shows the layout rendered based on the following HTML table model.

```html
<TABLE width=100>
  <TR>
    <TH rowspan="4">category</TH>
    <TH>logo image</TH>
    <TH>advertisments</TH>
  </TR>
  <TR>
    <TH colspan="2">top news item with an image</TH>
    <TH>advertisments</TH>
  </TR>
  <TR><TH>advertisement</TH></TR>
  <TR><TH>advertisement</TH></TR>
</TABLE>
```

4.2. Three Content Extraction Items

First, we describe the method for extracting each table area in a news page; then, we describe the method for extracting the content (the three content items mentioned above). Each item is extracted using the characteristics of each table area and content. The characteristics of the content are as follows.

**Category keyword:** each category is a specific keyword such as world, business, sports, or weather. Each of the keywords has a regular array that has a similar tag structure with repeating rows or columns, and it has links to the content of other pages.

**Top news item with an image:** it has a title and an image. This area is arranged behind the category keyword area, and the title and the image link to the news content on the same page.

**List of news headlines for each category:** each list has a category keyword and titles for the news content. The keyword has the same link with the category keyword, and the titles have a link to the news content. The keyword and titles have a regular array that has a similar tag-structure with a repeating row or column in the table area.

5. CONTENT SELECTION

MPV changes the extracted original content to the new content based on the user’s interests and knowledge. It detects the user’s viewpoints and selects the page based on the user’s preferences. We propose a detection method of this user's viewpoints and the adaptation method to the viewpoints and each page content.

5.1. Viewpoints Detection

The viewpoints of the user’s preferences are represented by keywords extracted based on the user’s access history.

Figure 4 shows a flow chart of a user’s behavior through the MPV page. When the user accesses a news item of an original page, the keywords of the meta data in the original are extracted by the MPV. The meta data is defined as `<meta name="name", content="content">` by a META tag in the original page, and the MPV extracts the content of the name using "name=Description"
and the "name=Date". Keywords of the extracted description are detected, and the MPV adapts the TF algorithm to get the weight of the extracted keyword. The set of the extracted date, keywords, and weights are stored in the users machine using cookies. The date, keywords, and weights are changed by the user's behavior. If the same keyword as the stored one is extracted when the user accesses the other original page, the weight of the keyword is added to the weight of the stored keyword. The stored date is also modified when the weight value is changed.

The keyword of the users is interested in represented by the larger of the weight values. The original category keyword on the MPV page is changed to the larger weight keyword.

5.2. News Item Selection

Viewpoints of the keyword were extracted based on the user’s behavior, then, MPV selects the most interesting news item using the keywords of the viewpoints.

The method for item selection uses the weight of the keyword and the date of browsing. Figure 5 shows the tables of a user and pages. To select the top of a news item with an image, first, the URLs of the item in the page table are selected by the keyword of the largest weight in the user table, MPV then selects one that has the most recent date and has not yet been read by the user. In figure 5, the URL of "www.1.3" is selected as the top news item because the keyword of the largest weight is "A" in the user table, and the date of news item "www.1.2" is older than the date of A. As a result, the title and the image of the selected URL are replaced.

To select the list of the news headlines for each category, first, the URLs of the item are selected by the category keyword MPV then selects one that is more similar to the page graph. We describe the graph for selection in the next section. The titles of the headlines of the selected URLs appear on the MPV page.

5.3. Page and User Graph

To select the news item, we adapted a semantic web technique called an RDF graph. The RDF graphs of a user and pages are rebuilt based on the user’s behavior.

Figure 6 shows the procedure for constructing RDF graphs of the user and pages. The MPV constructs the user graphs reflecting the user’s interest based on the stored keywords, and each graph is built by using the RDF schema defined by classifications such as person, place, and date. The user graphs are restructured whenever the user accesses the page by using the page graph. This restructured graph represents a model of the users’ knowledge.

Each page graph is built by each user. First, the keywords of the meta data are extracted from the collected pages, then, and the page graph is built by using an RDF schema that is defined by classification after the user accesses the MPV page. The page graph is built not beforehand but after each unit of behavior. In fact, the graph of the page, as well as the graph of the viewpoint, can change dynamically based on changes in the user’s preferences.

Each constructed page graph matches the user graph. The title of the page of the selected graph is represented by the MPV page.

6. CONCLUSION

We described My Portal Viewer (MPV), which is an application for efficiently providing high-quality content from the web that is user specific. Many web pages are merged based on the user’s preferences, and the MPV can provide a user-friendly web interface based on these preferences from two viewpoints; the layout and the user’s interests and knowledge.

We presented a method for constructing the layout and an adaptation of the users’ interest and knowledge using an HTML table model and semantic web technology. We will demonstrate and evaluate the effectiveness of the MPV.

7. REFERENCES