

FIELD OF THE INVENTION

This invention relates to a software-based tool for tracking of new content on web pages selected by the users.

BACKGROUND OF THE INVENTION

The World-Wide Web (WWW) continues to grow at an astounding rate. Its popularity is partially encouraged by the fact that anyone can publish content on the web in the form of web pages, unlike the broadcast media of old. As such, the information society is truly upon us, and increasingly we find ourselves having to deal with the problem of information overload.

In general, we deal with web pages on the WWW in one of three broad ways: search, recall and tracking. *Search* refers to the exploratory mode of working, where we know what we want to look for, but do not know where the relevant web pages are located. *Recall* refers to situations where you already know where the web page is, you just need to retrieve its location from your information bank. Finally, *tracking* means you already know where the web page is, and you want to stay informed when there is new content on that page.

It is desirable to have an automatic method for tracking of new content on selected web pages. This method should focus on relevancy, whereby only new content which are relevant are detected. "Relevancy" is a concept largely defined by the user. For example, it could include only text and attachments, and exclude advertisements, graphics, and date/time stamps. By detecting only new content which are relevant, it could potentially save users a great deal of time in keeping themselves up-to-date with interested web pages.

This Sheet Has Been Replaced



SUMMARY OF THE INVENTION

The web page which the user wishes to track is retrieved from the remote web server. This web page becomes the control web page. A normalized template is obtained from the control web page by first dividing all the text content into sections using the non-formatting HTML tags. Each section should then consist only of formatted text without any advertisements or graphics. Each section is then normalized by removing the formatting HTML tags, meta characters and repetitive white spaces. The normalized sections are inserted into a template sorted in alphabetical order.

Any hyperlink tags are also extracted from the web page, treated as attachments and stored in the normalized template along with the last modified date and size information which is retrieved from the remote web server. So, each record in the normalized template can be either "text" that is stripped of all HTML tags, or "attachment" which is described by its location, last modified date and size of attachment.

Some time later, the target web page is retrieved from the same remote web server and a normalized template is generated from it using the same procedure as for the control web page. For each section in the target normalized template, the control normalized template is searched for a similar section. If none can be found, then that section is marked as "new". Certain rules can be applied to look only for relevant differences, such as ignoring sections which contain only date/time, and performing filtering on sections based on certain keywords.

All the "new" sections that have been identified are presented to the user in the form of a summary. The current control normalized template is discarded and replaced by the target normalized template, which becomes the new control normalized template, to compare with the next version of the target web page.



DETAILED DESCRIPTION OF AN EMBODIMENT OF THE INVENTION

A method called *normalized template matching* is described which detects new content in any given target web page with reference to a control web page. Typically, the control page is an older version of the target page. By matching the target page against an older version of itself, it is possible to detect the new content on the more recent target page.

i. Creation of normalized template from the control web page

Initially, an HTML web page is selected for tracking by retrieving it from a remote web server. This becomes the initial control web page. A normalized template of the control web page is obtained by first dividing it into *sections* by using non-formatting HTML tags. Formatting HTML tags typically refer to those tags whose primary purpose is to affect the size, color, style and orientation of text eg. bold, italics, monospace etc. They are normally embedded within sections of text to influence their presentation in the web browser. Non-formatting HTML tags refers to all other tags. These includes meta (non-display) tags (eg. <HEAD>), layout tags (eg. <TABLE>), comments and JavaScript programs.

FIG. 1 gives an example of formatting HTML tags versus non-formatting ones.

Each section obtained by the division process should then consists only of text (including formatting) information. All advertisements, graphics etc. should have been filter out. Each section is then individually normalized by removing formatting HTML tags, meta characters (eg. non-breaking white space) and repetitive white spaces. The normalized sections are inserted into a template sorted in alphabetical order. This is important because it enables the sections to be searched rapidly later in the process.

In addition, hyperlink tags in the control page are also processed. Each hyperlink tag consist of a link to another computer file. This could be another web page, a CGI script, a graphics file, etc. If the extension of this file matches a list of extensions



given by the user, it is considered to be an attachment. Its last modified date and size are retrieved from the WWW, and a record containing the location, last modified date and size information is entered into the template.

When the entire control page is processed, a *normalized template* consisting of one or more records is obtained. Each record consists of two fields: type and content. There are two types: text and attachment. For "text" type, the content will be the normalized section consisting of text minus formatting HTML tags, meta characters (eg. non-breaking white space) and repetitive white spaces. For "attachment" type, the content will be the location, last modified date and size of the attachment.

FIG. 2 gives an example of a normalized template.

ii. Detection of new content on the target web page

The target web page refers to a subsequent, newer version of the control web page which is retrieved from the remote web server. This target web page is similarly divided and processed, after which the normalized template of the target page is matched against that of the control page. For each section in the target normalized template, a search (using content as the search key) is performed in the control normalized template for a similar section. Since the control normalized template is sorted by its content, this search can be effected rapidly and efficiently. If a particular section in the target normalized template cannot be found in the control normalized template, that section will be marked as "new".

In addition, certain rules can be applied to enforce relevancy. For example, a section can be tested to see if it contains only a date, or date/time. Such information always changes, and it is possible to skip this section when testing for "newness". Similarly, the user can provide certain "include" or "exclude" keywords. "Include keywords" means only sections containing particular keywords will be tested for "newness". "Exclude keywords" means sections containing particular keywords will immediately disqualify for the test of "newness".



FIG. 3 gives an example of the result of matching a target normalized template against a control normalized template.

iii. Presentation of new content detected on the target web page

The sections which are marked as new are presented to the user by creating a summary web page which lists all those sections which are new. The summary web page can be stored on a server and delivered to the user as a hyperlink, or it can be delivered in its entirety to the user as an email attachment.

Finally, the control normalized template is discarded and replaced by the target normalized template, which becomes the new control normalized template, to compare with the next version of the target web page.

This Sheet Has Been Replaced



THE CLAIMS



1. A method for detecting new content in any given target web page with reference to a older, control web page, consisting of :

creating separate normalized templates for the target and control web pages;

performing a comparison between the target normalized template and the control normalized template;

deciding whether the target web page differs sufficiently from the control web page to indicate that there is new content.
2. A method as claimed in claim 1, wherein each normalized template is made up of one or more normalized sections created by first dividing the web page into sections using the non-formatting HTML tags and then removing all the formatting HTML tags from each section.
3. A method as claimed in claim 2, wherein the normalized sections that make up the control normalized template are arranged in alphabetical order to facilitate searching.
4. A method as claimed in claims 2 and 3, wherein each normalized section is classified according to its content. The content type for each normalized section is also stored in the normalized template.
5. A method as claimed in claim 1, wherein the comparison is achieved by taking each section in the target normalized template and searching the control normalized template for a similar section, using content as the search key.
6. A method as claimed in claim 1, wherein the comparison is controlled through system-supplied or user-defined rules.



7. A method as claimed in claim 6, wherein one of the rules is to check the content type of each normalized section to determine if the section should be skipped in the matching process.
8. A method as claimed in claim 6, wherein one of the rules is to search the content for certain "include" or "exclude" keywords to determine if the section should be skipped in the matching process.
9. A method as claimed in claim 1, wherein a web page is created containing the results of the comparison in the form of a summary.



This Sheet Has Been Replaced

