4T−10 WebBeholder: Filtering Issue for HTML Document's Difference Stream Based on User's Interest

Saeyor SANTI Mitsuru ISHIZUKA
東京大学 工学部 電子情報工学科
e-mail: {santi,ishizuka}@miv.t.u-tokyo.ac.jp

1 Introduction

The WebBeholder is a cooperative agent community framework that provides open services on tracking and displaying changes on the World Wide Web. The framework is contructed by tracking service provider, mediator sites, and user's personal agents. The service provider agent is responsible for tracking changes found in assigned web pages. The infomation of changes need to be filtered in order to eliminate insignificant changes. A personal agent in the community notifies its user once it found any significant changes on assigned web pages. Moreover, the HTML differentiating module in the framework generates a summary page of changes which coexist with the original web page. The way of presentation enables the user to glance the overall changes and skip from one change to another at ease. Behind this convenient way of presentation and reviewing, the LOGTAGS algorithm plays an important role [2].

This paper focuses on the explanation of the mechanism we use in our innovative filter for classifying changes in HTML document's difference stream. The criteria of the filter are based on user's interest. The filter consults user's interest and assigns score to each change in the stream. The scores are summed up to indicate the level of interest related to user's preference. The framework uses this indicator to decide that whether it should notify the user of the detected changes.

With great help of this framework, it saves the user a lot of time for having no need to desperately check any web page from time to time in order to find whether it has been changed in meaningful way. The framework carries out those tedious and time consuming works on behalf of the user. Moreover the user can be sure that the notifications from the personal agent are trustworthy and basicly based on the user's interest.

Saeyor SANTI, Mitsuru ISHIZUKA
Dept. of Information and Communication Engineering,
Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN

2 User's Interests

Users of the framework is served by personal agents based on their interests. The personal agents represent their users in the WWW and work as if the users perform the tasks themselves. In order to deliver such performance, the personal agents need to know their users well enough to give the result that satisfy their needs. A personal agent places the orders received from its user to the service provider agent together the preference and interests of the user. This section discusses about the information that we incorporate into our filtering process. This research divides the interests of the users into following categories:

- **Existence of contents:** Many users are curious to know whether the contents they interested in are exist in the document. It is inevitable to check the existence of components again once the document was reviewed. The filter used in our service provider agent deals with this demand by checking the existence of links, images, java applet, etc. that appear in new version of document. The filter provides a report of the change in existence of component by comparing with the old version. The personal agent notify the user via email with this information in order to roughly tell the user how much the document has been changed.

- **Appearance of document:** HTML specification includes many tags that control the flow and structure of HTML documents. In some cases, many users are curious about the change in the appearance of the documents. Unfortunately, the presentation of change in appearance is not straightforward. The appearance of HTML document must be either the old appearance or the new one. This issue is worth taking into account. We deal with this issue by selecting the new appearance for our summary document but counting the change of appearance in the filtering process. The information of change of appearance participates in the decision making process which results in whether the overall changes are significant enough to notify the user.

- **Topics in interest:** This category is the most important issue when we deal with the user's interests. In real world, people evaluate whether the content they are reading fall in the area of their interest by the context. If we insist to deal with this issue rigorously, we have an expensive processing to pay. The evaluation of relevance to user's interest areas is considered to be heuristic. On the other hand, the process need natural language processing techniques. We compromised the correctness with computational costs. The filter deals with this issue via the hint of words that relevant to the user's area of interests.

It is obvious that we have many categories of user interests to deal with in our HTML document's difference stream filter. The details of incorporating these interest criteria into the filtering process will be discuss in next sections.

3 HTML Document Streaming

HTML Document is a mark up language which can be regarded as a sequence of tags and context even though those tags control the structure and appearance of the document. In this research we assume that all HTML documents are generated perfectly correct to HTML grammar. We have to assume this way because we consider all kinds of tags as one kind no matter they are beginning tags or ending tags. By considering this way we can parse the whole HTML document into a stream of tags with context. In Fig.[1], the tag parser parses both the old HTML document and the new HTML document into two streams for the Longest Common Tag Sequence Detector. The details of how the detector works is described in [2].

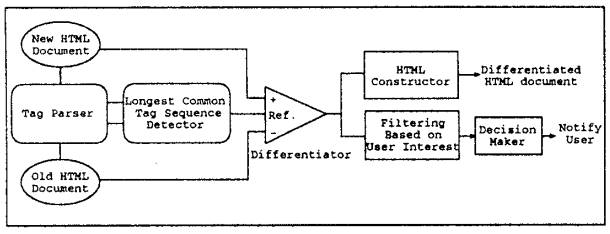

Figure 1: *HTML Difference Engine with User's interest based Filter*

On the other hand, the tags in HTML documents can be divided into two categories. The first category is the tags that only control the format or appearance of the document. Another category is the tags that define some information for the document such as links, images, email links, etc. The latter category is treated separated from the first one since we have more information to deal with. The filter implement its configuration upon user's interest with some information in those tags. From this reason, we need to combine the information within those tags into the stream of HTML document.

Due to the mechanism of the LOCTAGS detector, it is inevitable to combine the data into the prior tag. The appearance of the concatenated tag in the stream seems a bit strange but that serves the propose of tag positioning well. The differentiator is able to deal with the tag and efficiently find the difference between the old and new documents.

4 HTML Document Stream Filtering

In Fig.[1], the result from the differentiator is fed into the HTML constructor and the user's interest based filter. The filtering process is perform right in this filter. The filter implements the three categories of user interest as described above. The existence of contents can be checked by finding whether the old contents are still in the document. At the same time, the filter scans the document whether is there any new content inserted to the document. In the same manner, the filter check whether is there anything changed with the tags that control the appearance of the document. Once the filter found any changes that fall into these two categories, it evaluates a score those changes. The topics in user's interests can be found by checking the key words that user specified. The filter checks the existence of those key word in the context of document then evaluates another score for this kind of user interest. The total score is then determined by the decision maker. If the score is above the threshold, the personal agent will be informed to notify its user of the changes.

5 Conclusion

The WebBeholder was proposed as another approach to finding and displaying changes on the World Wide Web. The personal agents in the WebBeholder community work on the World Wide Web on behalf of their users. A personal agent carries its user's interests and roam over the Internet. The user's interests are submitted together with the request for page tracking service. The user's interest based filter carry out the information for decision making process. It's still difficult to incorporate all categories of user interests into one criterion. The scoring method need improvement.

References

1) Santi Saeyor, Ishizuka Mitsuru *WebBeholder: A Revolution in Tracking and Viewing Changes on the Web by Agent Community* Webnet98 3rd World Conference on Internet Technology, Nov. 7-12, 1998, Orlando, Florida, USA

2) Santi Saeyor, Ishizuka Mitsuru *Longest Common Tag Sequence Algorithm for precise reviewing of changes in the WWW* The 56th Annual Convention, Information Processing Society of Japan