1 F − 1

Longest Common Tag Sequence Algorithm for precise reviewing of changes in the WWW

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 World Wide Web (WWW) is changing in the ways in which the users obtain various information. Both invaluable and garbage information have been growing at exponential rates in the WWW. This is overwhelming the ability of its users to stay in tune with the changes. In the same time, the mode of data transfer becomes bimodal called "push-pull" model which arbitrary data is pushed to the users and pulled by them upon request. We may have heard about push-technology for a while. Actually, the push technology suites for the data which is temporarily exposed to the users. The technology brings remarkable ease to the users like many other mass-media broadcasting styles.

At one extreme, we can view the changes in specific WWW page as temporary information since the page keeps changing upon the time. The information of updated parts should be informed to the users who keep watching the growth of the page. At this point, we may see that push-technology should play an important role to carry this task out. Unfortunately, there is still large percentage of pull mode WWW pages out there in the WWW. A number of utilities have been devoloped to assist the users in tracking changes in WWW pages of interest. Those tools implement some mechanisms that work like a robot. The robot keeps watching the WWW pages according to the schedule issued by the users. Changes in each page are supposed to be trivial or interesting ones. That will be the responsibility of each robot to consider and we will not go further in detail on this topic. Instead, we will examine the way to express and display the updated part in specific WWW page.

While the WWW pages are changed dynamically, the effort for tracking and reviewing the changes has come into account. This paper details a method for checking and displaying changes in arbitrary two revisions of a specific WWW page. We developed an algorithm called Longest Common Tag Sequence to match tag sequences in old and new version of WWW page. The algorithm was applied to help finding the right places for context comparison within a pair of WWW pages. The differences are justified in a new HTML document conformes to its updated version's outlook so the user can identify the differences at ease.

2 Longest Common Tag Sequence

HTML document consists of markup tags and context. Thus the HTML document processing always deals with markup tags and context processing. Among markup tags used in each document, we divided them into two types [2]. Firstly, the content-defining markup tags (such as image < IMG src=... > and hypertext references < A href=... >) provide information of contend within the tags. Secondly, we consider the rest as the tags that control the format of context such as < P >, < I >, < HR >. Since the tags are divided in two categories, the preprocessor of the difference engine are designed to parse the HTML document based upon the basic that each tag is followed by context. In other words, the process flow comes as a sequence of tag and its context. The merit of this method is that the HTML parser does not have to understand all the given tags described in HTML specification. In long run, the method is still valid for new markup tags introduced in later version of HTML specification (to some extent, since the method is blind to the meaning of content-defining markup tags). However, the concept simplifies the parsing process remarkably and suits for processing large scale WWW pages comparison. The next problem is how to present the changes in each WWW page meaningfully and smartly. In common, the changes can be addition or insertion, deletion, or changing of markup tags and context. If we just parse and differenciate the HTML document in sequence, the result would be presented in nonsense format because the first change (especially, the addition or deletion) introduces sequential changes for the the markup tags and context after it. As the result, some parts of tags and context which are identical to the original page will be ignored.

In order to solve described problems, we applied the same concept of text comparing algorithm that implement the Longest Common Sequence (LCS) of characters in string. We view the HTML document as a string of markup tags and context. The algorithm treats the context and tag sequence separately but keeps the processing order in correct sequence. The algorithm can compare the context to its pair in right

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, JAPAN

position because the sequence of markup tags are checked and recognized.

So far, the idea seems to work well but there exist yet another problem. In general, when the identical tags come in a row (especially, the markup tag used in structured text such as < LI > and those of table display), the algorithm that relys merely on common tag sequence fails to deliver the right comparison result. This happened because the state of problem returns to the same as described in direct one-to-one comparing method. In such a situation, the algorithm cannot tell one tag from others. Fortunately, we can solve this problem easily by appending the context to the markup tag in order to make it difference from others.

The same problem seems to be a recursive one in case that the context coming up next to the tag is still identical to the next set of sequence. In the case, the algorithm must be designed to look ahead to the HTML document in order to completely solve the problem. In our work, the consideration on this problem is implemented up to few levels of markup tags looking ahead since such a situation is considered scarcely happens.

3 HTML Diffference Engine

The HTML Difference Engine was constructed to compare a pair of HTML page. The output of the Difference Engine can be separated into two categories. The first one is the information of changes found when comparing. The second one is the HTML document that presents the changes. The code base of some links, images and JAVA applets are modified on the fly, so that we have no need to hold all images or applets' byte code locally. The longest common tag sequence algorithm is applied to construct the HTML Difference Engine which is able to create smart presentation of changes detected.

The architecture of the Difference Engine is shown in Fig. 1. Old and new version of pages are fed to the tag parser module in order to generate the tag sequence for Longest Common Tag Sequence Detector. All tag streams are compared by Differenciator and its information will be used to generate the final HTML document that indicates the changes detected.

4 Experiment

The Difference Engine has been used to compare HTML documents. The presentation capability has been tested over some listings and tables. The algorithm manifests its success in detecting and displaying changes in structured text and table. Fig. 2 shows addition (underlined text), deletion (stroked text) and changed parts in a table.

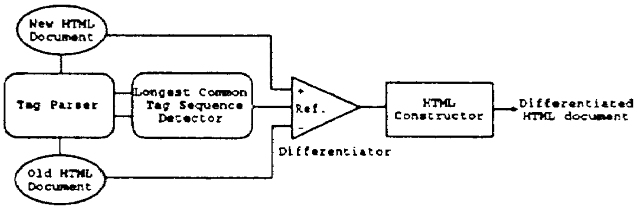
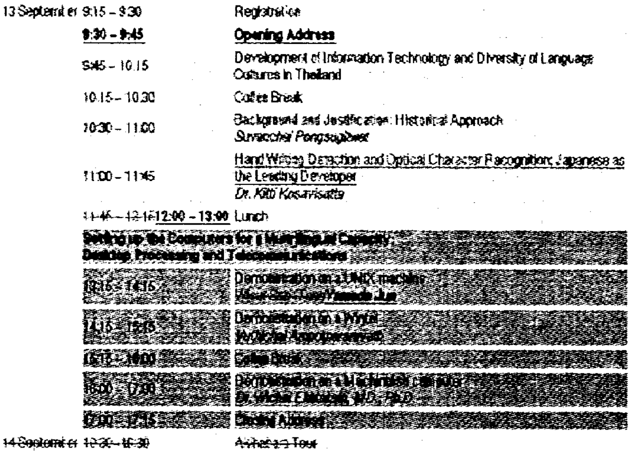


Figure 1: *Building blocks of Longest Common Tag Sequence HTML Difference Engine*



13 September	9:15 – 9:30	Registration
	9:30 – 9:45	Opening Address
	9:45 – 10:15	Development of Information Technology and Diversity of Language Cultures in Thailand
	10:15 – 10:30	Coffee Break
	10:30 – 11:00	Background and Justification: Historical Approach *Suracchai Pongsupawat*
	11:00 – 11:45	Hand Writing Detection and Optical Character Recognition: Japanese as the Leading Developer *Dr. Kitti Kosavisutte*
	~~11:45 – 12:15~~12:00 – 13:00	Lunch
	Setting up the Computers for a Multilingual Capacity: Desktop, Processing and Telecommunications	
	13:15 – 14:15	Demonstration on a UNIX machine [illegible]
	14:15 – 15:15	Demonstration on a Wintel [illegible]
	15:15 – 16:00	Coffee Break
	16:00 – 17:00	Demonstration on a Macintosh computer *Dr. Wichai [illegible], M.D., Ph.D*
	17:00 – 17:15	Closing Address
~~14 September~~	~~10:30 – 16:30~~	~~Ayutthaya Tour~~

Figure 2: *A sample of changes presentation in some elements of table*

5 Conclusions

We have seen that the Longest Common Tag Sequence algorithm can position the right portions of context that should be compared in a pair of HTML documents. The algorithm is suitable to be used in presentation part of HTML Difference Engine since the comparison does not concern the meaning of the changes detected.

References

1) Fred Douglis, Thomas Ball, Yih-Farn Chen, Eleftherios Koutsofios *WebGUIDE: Querying and Navigating Changes in Web Repositories* Fifth International World Wide Web Conference, May 6-10, 1996, Paris, France

2) Fred Douglis, Thomas Ball, Yih-Farn Chen, Eleftherios Koutsofios *The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web* AT&T Labs - Research Technical Report, April 1997

3) Oren Etzioni, Daniel Weld *A Softbot-Based Interface to the Internet* Comm. of ACM, July '94.