Web Orchestration: Customization and Sharing Tool for Web Information

Lei Fu[1], Terunobu Kume[2], and Fumihito Nishino[2]

[1] Fujitsu R&D Center CO., LTD
13/F, Tower A, Ocean International Center,
No.56 Dong Si Huan Zhong Rd, Chaoyang District, Beijing, China
[2] FUJITSU LABORATORIES LTD
1-1 Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan
fulei@cn.fujitsu.com, {t-kume,nishino}@jp.fujitsu.com

Abstract. In this paper, we present a tool, Web Orchestration, which allows people to customize and share the web information in a simple way. Our work is based on the web annotation and web scraping technique. It adopts B/S architecture, and has a user-friendly interface. It can be used in many aspects, such as web information monitoring, web information sharing, web information integration , recombination and so on. As an application of web 2.0 technique, it's easy to use, simple but powerful; it can enhance collaboration of each other, and make web information sharing and personalized web information customization much easier to use.

Keywords: Web annotation, web scraping, information sharing, information customization.

1 Introduction

Currently, with the spread of the World Wide Web, diverse information floods every corner of the Internet. When we want to get some information which we are interested in, we may often feel overwhelmed by the information floods. Therefore, how to manage and share the information with others on the internet gain more and more attention.

Against such a background, W3C(The World Wide Web Consortium) launches a project called "Annotea Project[1]", which aims to enhance collaboration via shared metadata based web annotations, bookmarks, and their combinations. By annotations they mean comments, notes, explanations, or other types of external remarks that can be attached to any web document or a selected part of the document without actually needing to touch the document. When the user gets the document, he or she can also load the annotations attached to it from a selected annotation server or several servers and see what his peer group thinks. One part of our work is based on this project.

The other part of our work is based on the "web scraping" technique, web scraping (sometimes called harvesting) generically describes any of various means to extract content from a website over HTTP for the purpose of transforming that content into another format suitable for use in another context. Those who scrape websites may

M. Kurosu (Ed.): Human Centered Design, HCII 2009, LNCS 5619, pp. 689–696, 2009.

wish to store the information in their own databases or manipulate the data within a spreadsheet. Others may utilize data extraction techniques as means of obtaining the most recent data possible, particularly when working with information subject to frequent changes. Investors analyzing stock prices, realtors researching home listings, meteorologists studying weather, or insurance salespeople following insurance prices are a few individuals who might fit this category of users of frequently updated data. He or she can manipulate the frequently updated data conveniently with the web scraping technique.

In this paper, we present a browser-based tool: Web Orchestration, which adopts the ideas above and provides a user-friendly view of diverse information on the internet and by which you can also comment and share the information on the internet with others conveniently. It's a light-weighted realization of web annotation and web scraping, easy to realize and easy to use. It mainly has two functions: Web Information Customization Module (WICM, for short) and Annotation Posting & Sharing Module (APSM, for short). The first one is to get and manage the information from different web sites. The other is to comment and share the information on the internet with other persons.

2 Web Information Customization Module (WICM)

This module is based on the web scraping technique, it provides a simple but powerful way for getting your desired part of information on web pages, then reorganizing them, and displaying them according to your requirement. This can be looked upon as a personalized web information customization process, the users can remix all the content they want on any web pages. It's a light-weighted realization of web scraping technique, a little similar with mash-up application.

In our method, to complete this customization procedure, firstly, we should analyze and generate the HTML DOM (Document Object Model) tree structure of the web page (Fig.1 shows an example for HTML DOM tree).

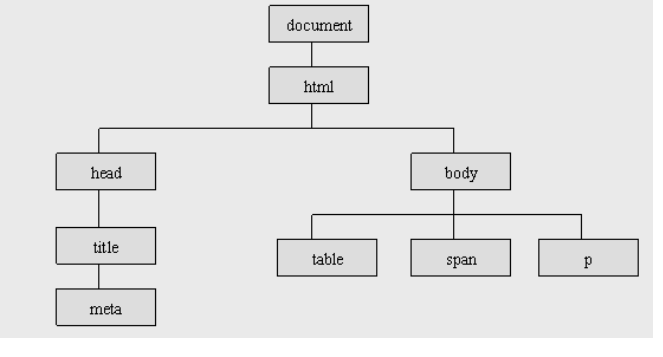

Fig. 1. HTML DOM Tree

In the DOM tree, each node corresponds to each block of the original web page seen by the terminal users. Then according to this, we can divide the DOM tree into some blocks automatically, for example, each <table>&</table>, <tr>&</tr>, <td>&</td>, & node or the whole <body>&</body> node in DOM tree, all these pair nodes can be a block.

Secondly, what you need to do is to choose the block you are interested in. WICM will record the path of the selected block and the URL of the web page, by the path, it's similar with XPath[2], and it's defined as the set of nodes which you have to pass if you want to reach the target node in the DOM tree. For example, if you want to reach the "table" node in Fig1. The path is "//body/table", "//" is the root of the tree. This path and the url of the page will be stored in the server.

Finally, you should choose a target page which you want to insert the selected block in. For the target page, WICM still firstly analyzes the DOM tree structure of the page, and then you can choose where to insert your interested block. The target page can be an existed page or a new page. WICM will record the insertion position, the path in the DOM tree of the target page. All the path information will be stored in the server, so when you load the page, WICM will load the insertion information of the page synchronically.

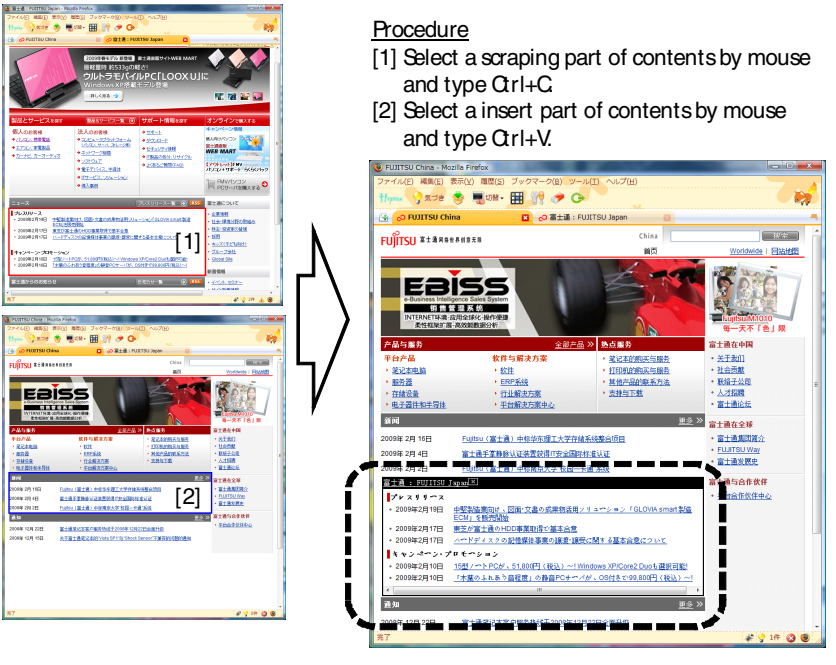

Fig. 2. WICM

Compared with the common method, WICM, as a light-weighted application of web scraping, has many advantages. Firstly, it's a dynamic process, not a simple copy & paste operation, you can choose any information you are interested in from any

websites on the internet, you will not be limited to a fixed frame, you can fully express your initiative and customize the information which interests you; Secondly, all the inserted information on the target page can change with their original pages, so you can also make use of it for monitoring some pages, such as that you can follow somebody's schedule or work progress, or you can monitor the price of all the branches of a shop and so on; Thirdly, by this module, you can avoid unnecessary switches between several web sites, and read the diverse information on one page; The last one, WICM can mash up important contents on personal page. It can recombine diverse information into one existed website, with it, you can establish a purely personal page which includes all the information you are interested in. For all these advantages, the only cost is the simple selection and insertion operation on web pages, but it's once for all.

3 Annotation Posting and Sharing Module (APSM)

This module is based on the web annotation technique, it provides a simple and convenient way for posting annotations on web pages and sharing them with other people. By annotations we mean comments, explanations, questions, advice or other types of external remarks that can be attached to any web document or a selected part of the document without actually needing to touch the document, it has the same definition with "Annotea Project".

Previous research has shown that making annotation text is an important companion activity for reading, users do it for manifold purpose. In an extensive field study of annotations in college textbooks, Marshall[3] found that annotations were used for purposes that included bookmarking important sections, making interpretive remarks, and fine-grain highlighting to aid memory. O'Hara and Sellen[4] found that people use annotations to help them understand a text and to make the text more useful for future tasks. Annotations are often helpful for other readers as well, even when they are not made with others in mind. Computer-based annotations can similarly be used for a variety of task. For example, Baecker[5] et al. and Neuwirh[6] state that annotations are an important component in collaborative writing system, where "collaborative writing" refers to fine-grained exchanges among co-authors creating a document. In the study reported here, it focuses on the later stage of the document generation process. When a relatively complete draft of the document is posted on the web, annotations are used to get coarser-grain feedbacks from a large group of people (beyond the original authors).

In our method, firstly, we should also analyze and generate the HTML DOM tree structure of the web page, divide the page into some blocks according to the DOM tree structure automatically. Afterwards, the users can select which part of the page they want to put annotations on. The annotations will be inserted into the DOM tree of the page as a "div" node in the same level of the selected block node. APSM will record the path of the insertion position and the URL of the page, it's same with what WICM does. When users load and browse a web page, APSM will search for the annotations attached to it from the server, if having relative annotations, it will load and display them. All the annotations will be displayed in the form of the notepaper, different type of annotation has different background color. We use an RDF (Resource Description Framework)

based annotation schema for describing annotation types, it's an extensible framework, we can define or delete types conveniently. The users can also control whether the annotations they post are public or private, once somebody loads the web page, he or she can only see all the public annotations on the page, but private ones.

Fig. 3. APSM

One of the most prominent characteristics is that it can still locate the position of the annotation when the layout of the web page changes. Let's take BBS website as an example, lots of people discuss about a same topic, when you annotate one of them, it will soon be moved to the hind page, because many people will post their opinions in a short time, the URL which contains the annotation you put will change too. In order to solve this problem, we adopt double-locating mechanism: first, besides the path of the annotation, we also consider the content which you select. Second, when you load an URL in browser, we'll generalize part of the URL to match it in annotation database of the server. By the method above, we can deal with most of the changes.

Compared with the common method for web pages sharing, such as social bookmark system. Our method, APSM, has several advantages, firstly, it enriches the content which the users share with other people, with APSM you can not only share the web pages(URLs) but also share your annotations on it. It can help you to share your unique comment or advice attached to the selected part of the page. Secondly, it provides a RDF metadata based extendible framework for rich communication about web pages while offering a simple annotation user interface. The annotation metadata

can be stored locally or in one or more annotation servers and presented to the user by a client capable of understanding this metadata and capable of interacting with an annotation server with the HTTP service protocol. Last, it runs much more easily than the social bookmark service, and can reduce much more operation costs. Please refer to the Fig.4.

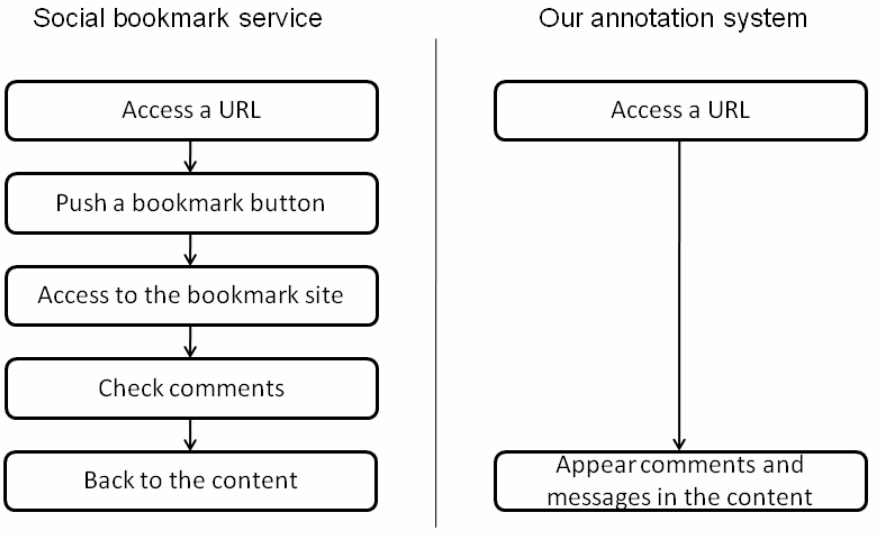

Fig. 4. Operation costs comparison

4 Features and Problems of Our Tool

Our tool has three salient features: low requirement for the system, providing an easy-to-use set of function, and has a user-friendly interface. First and foremost, the tool is based on the B/S architecture, it's just a plug-in for the explorer (IE or Firefox), so it requires less system resources, such as memory, CPU, or hard disk space and so on. It's similar with the common-used toolbar for IE, such as yahoo and Google toolbar. In the second place, although the tool is so small, it provides many practical functions. With it, you can feel some new experiences different with traditional internet surfing. You can put annotations on the selected part of the web page, and see other people's annotation on the page, if you are the administrator, you can also predefine some key words, when the users see them, it will pop up your unique annotation attached to them automatically. It also allows you to recombine many different parts of different web pages into one part of an existed page or a new page, then you can avoid frequent switches between different pages in order to browse the interested part, it will save a lot of time for you and it also makes your work much more efficiently. As a light-weighted application of web annotation and web scraping technique, it's simple but powerful. Lastly, the tool has a user-friendly interface, it's simple and easily demonstrated, it provides both push-button and right-click menu operation, similar with some common-used software, so the user can get used to it fast.

However, in spite of all the features above, it still has some problems, which are needed to be improved in the future. First of all, it can't deal with the part of the page which is generated by the script language in the running time automatically, such as the part which is generated by the javascript code, because we can't get the source code of that part when we analyze the DOM tree structure of the page. Secondly, for the WICM, when it remixes the pages which have different language encoding, it doesn't work normally. So we still have much work to do in order to improve the performance of the tool.

Furthermore, we have installed this plug-in for explorer (IE or Firefox) in our company and collect some feedbacks from the users. From the user's feedbacks, we know that they think the tool itself is useful, by it, they can gain many extra things related to their interested part of the web page, such as the other people's comment, suggestion, and recommendation, these information is very helpful for them. And they feel very convenient to recombine their interested part into one part. However, they also encountered the problems above, this influences their initiative to use it to a certain extent. They also want to annotate or remix some multimedia information, such as video, flash, which can't be dealt with by our tool now.

5 Conclusion

In this paper, we present a tool for web information customization and sharing, which is based on the web scraping and web annotation technique. It has a user-friendly interface, extendible framework and is easy to use. By it, you can customize all the information you are interested in on the web, and reorganize them according to your requirement. You can also share your unique comment or advice attached to the selected part of the page conveniently. Just as the tool's name: Web Orchestration, the tool itself is like a stage, all the web pages act as players, and the user is like a conductor, the conductor waves his baton, directs a perfect orchestration.

References

1. W3C Annotea Project, `http://www.w3.org/2001/Annotea/`
2. XML Path Language(XPath) Version 1.0, `http://www.w3.org/TR/xpath`
3. Marshall, C.: Annotation: From Paper Books to the Digital Library. In: Proceedings of the 1997 ACM International Conference on Digital Libraries (DL 1997) (1997)
4. O'Hara, K., Sellen, A.: A Comparison of Reading Paper and On Line Documents. In: Proceedings of the 1997 ACM Conference on Human Factors in Computing Systems (CHI 1997) (1997)
5. Baecker, R.M., Nastos, D., Posner, I.R., Mawby, K.L.: The user-centered iterative design of collaborative writing software. In: Proceedings of the INTERACT 1993 and CHI 1993 (1993)
6. Neuwirth, C.M., Kaufer, D.S., Chandhok, R., Morris, J.H.: Issues in the design of computer support for co-authorng and commenting ACM (1990)
7. Delicious, `http://del.icio.us/`

8. Luff., P., Heath, C., Greatbatch, D.: Tasksin-interaction: Paper and Screen Based Documentation in Collaborative Activity. In: Proceedings of the 1992 ACM Conference on Computer Supported Cooperative Work (CSCW 1992) (1992)
9. Cadiz, J.J., Gupta, A., Grudin, J.: Using Web Annotations for asynchronous collaboration around documents. In: Proceedings of the 2000 ACM Conference on Computer supported cooperative work (CSCW 2000) (2000)
10. Lee, K.J.: What goes around comes around: an analysis of del.icio.us as social space. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW 2006) (2006)
11. Hansen, F.A.: Ubiquitous annotation system: technologies and challenges. In: Proceedings of the seventeenth conference on Hypertext and hypermedia (Hypertext 2006) (2006)
12. Kume., T., Nishino, F.: Data Matching Technique Between a Blog Content and a XPath. In: Proceedings of the Institute of electronics, information and communication engineers, Web intelligence and Interaction (WI2 2007) (2007)