Personally Inherited Features of Web Pages

Zheng Chen[1] Liu Wenyin[2] Fan Lin[1] Peng Xiao[1] Yin Liu[1] Bin Lin[1] Wei-Ying Ma[1]

[1]Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R. China, {zhengc, wyma}@microsoft.com

[2]Dept. of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, csliuwy@cityu.edu.hk

ABSTRACT

Inherited features are found among the pages on the same navigation paths a user has accessed and used to model her preferences and interests. The constructed user models can be used to build personalized web agents, which help the user navigate the Web by providing related pages suggestion and highlighting interesting content in a web page. Our experimental results show that the proposed techniques improve the user's web browsing experiences significantly.

1. INTRODUCTION

More and more Web agents [1] have been developed to facilitate users to utilize the Web. Personalized Web agents [2] can further provide better experience of the Web usage. However, these agents only utilize the content of web pages accessed by the user for building user's preference model. Although the content is the explicit information easily accessible to the systems, they are difficult for the user to remember and use. It is mainly due to two reasons. One is the gap between the editor's intention and the user's perception, and the other is the difference of semantic meanings of the same web page grasped by different users. Therefore, the content of web pages alone is not enough to represent the user's interests and favorites. The information associated with how and when the user finds and interacts with the content could also be useful to assist the search process. However, this implicit information is often ignored by most software agents. Hence, it is part of our motivation to build user models based on both explicit features extracted from the user's visited web pages and implicit features obtained from the usage pattern of the individual user. These user models can be used to build personalized Web navigation agents for helping the Internet users efficiently surf on the Web.

2. INHERITED FEATURES AND THEIR EXTRACTION

The user's navigation paths sometimes reveal useful information about what the user is looking for and how the user thinks the web pages are. This information is often more illustrative than the content itself in describing the semantics of the web pages because it reflects the user's view of the web pages. The following example demonstrates why implicit features are sometimes more illustrative than explicit features: When the user needs to find some information, she may submit a query to a search engine. She then follows several hyperlinks, starting from one of the search results and reaching the final target web page, which contains desired information. Frequently we found that the query words and the hypertexts over the hyperlinks are good resource for representing the meaning of the final target web page, though the content of the target web page may not always contains the query words. In this case, the target page could not be found if using only explicit features. But the query words, which are inherited by the target page on the navigation path, are implicit but highly related to the semantics of the page. A semantics hierarchy can be built based on the navigation pattern. For example, a user may go to http://www.sony.com to find information about the Sony digital camera. He may follow a sequence of hyperlinks http://www.sony.com ("digital cameras"), http://www.sonystyle.com/digitalimaging/cybershot.htm ("cyber-shot"), and http://www.sonystyle.com/digitalimaging/cybershot.htm ("DSC-P50") to get the information about Sony DSC-P50 from the target page containing the desired information. In this case, not only "DSC-P50" can be used to index the final web page, but "digital camera" and "cyber-shot" are also relevant text features to index the page. These two keywords can be inherited by their subsequent pages on the navigation path. This is the reason why this kind of implicit features is called "inherited features." Since different users may find the target page along different paths, the features are personally inherited. These kinds of features not only are semantic extension of the target web page, but also represent the semantic hierarchy of the web pages on the navigation path.

In order to extract the "inherited features" from the navigation path, one important task is to determine whether a web page can inherit the features from its precedent pages or not. For this purpose, we have developed two methods. One is based on the structure of the website; the other is based on the content similarity between the visited web pages. Analyzing the structure of a website is a simple

but efficient way to determine the inheritance property of a web page. First, we assign a level to each web page on the navigation path. The level of the root of a website is zero, the level of the hyperlinks within the root page is one, and the rest may be deduced by analogy. Suppose that the navigation path is $u_1, u_2, \ldots, u_n$ where u_i is the *i*th visited web page. $Level(u_i)$ is the level of the web page u_i. If $Level(u_i) > Level(u_{i-1})$, web-page u_i can inherit the features from its parent u_{i-1}, noted as $Inherit(u_i, u_{i-1})$. Furthermore, the web page can inherit features not only from its direct parent but also from its ancestors, as shown below.

$Inherit(u_i, u_{i-2}), if\ Inherit(u_i, u_{i-1})$ and $Inherit(u_{i-1}, u_{i-2})$

3. PERSONALIZED WEB NAVIGATION AGENTES USING INHERITED FEATURES

A user's interests and preferences can be learnt from the user's frequently visited information and described in user preference models. Once the user preferences models are obtained, better services can be provided to the Internet users, e.g., more appropriate suggestions or more preferred information collected automatically from all possible sources by an offline/online crawler, and information sharing with other users of similar user preferences models. The personally-inherited features actually present the personal view of the pages and therefore can be used to develop user preference models, which are further used to build the following two Web navigation agents.

(1) Related Pages Suggestion. This agent can help the user obtain a list of related pages while he is reading a web page. The underlying technique includes two levels. One is searching for the related pages that have been visited by the user before; the other is searching for the unseen but related pages. The agent first extracts the features for the current web page and calculates its similarity to each web page visited before. If the similarity exceeds a pre-defined threshold, the page is suggested to the user. To suggest unseen but related pages, the agent first selects top *K* dominant keywords from the current web page as the query and submit the query to a meta-search engine. The agent then re-ranks the returned web pages based on the user's preference model and suggests the top *N* web pages to the user. In our experiments, we asked five users to determine if the suggestion made by the agent is relevant. About 2000 related pages were suggested to the users based on 200+ pages they browsed. The accuracy of the top one suggestion is as high as 76%, and the average precision of suggested pages is over 69%.

(3) Interest Highlighting. This agent helps the user easily and quickly read the information of his interest in a web page. This is especially useful when the page is long and requires the user's substantial effort to read through it. Once highlighted, the user could quickly obtain those interesting information by scrolling to those highlights. This agent can provide both hyperlink highlighting and keyword highlighting according to the attribute of the page. If the visited page is a hub page, hyperlink highlighting is more suitable. On the other hand, if the visited page is an authority page, keyword highlighting could be more useful to the user. For hyperlink highlighting, the agent first extracts all the hyperlinks within the visited web page and their features and computes the similarity of each hyperlink to the user's preference model. If the similarity exceeds a pre-defined threshold and the link has not been visited before, the agent will highlight the link. For keyword highlighting, the agent first extracts the features of the current web page and then evaluates the current page using the preference model. If it belongs to one of the user's interesting categories, the agent will select the dominant keywords from this category and highlights them in the current web page. Otherwise, the agent selects the dominant keywords from all the user's interest categories and highlights them in the current web page. In our experiments, we asked the five users to count the highlighted keywords in which they were really interested. The 200+ pages mentioned above were also used to evaluate the performance of interest highlighting. The average accuracy of the highlighted keywords was approximately 71%.

4. CONCLUSION

The inherited features among web pages along the same navigation path have been found and utilized in user modeling and constructions of personalized web agents. From our experiments of related page suggestion and interest highlighting, the inherited features have been proved effective in helping the user navigating the Web.

5. REFERENCES

[1] Chen Z, Liu W, Yang R, Li M, Zhang H (2001) A Web Media Agent. In: *Poster Proc. WWW10*, pp 178-180, 2001, Hong Kong.

[2] Liu W, Chen Z, Li M, Zhang H (2001). A Media Agent for Automatically Building a Personalized Semantic Index of Web Media Objects. *Journal of the American Society for Information Science and Technology* 52(10):853-855.