Review of Artificial Immune System in Web Personalization

Hamid Rastegari
Faculty of Computer Science and Information Systems
University Technology Malaysia
Skudai, Malaysia
e-mail: hamidrst@gmail.com

Siti Mariyam Shamsuddin
Faculty of Computer Science and Information Systems
University Technology Malaysia
Skudai, Malaysia
e-mail: mariyam@utm.my

Abstract— **This paper is a review on knowledge discovery in the field of web mining for the benefit of research on the personalization of web-based information services. The essence of personalization is the adaptability of information systems to the needs of their users. This issue is becoming increasingly important on the Web, as non-expert users are overcame by the quantity of information available online. This article investigates the application of artificial immune systems (AIS) to knowledge discovery as a web personalization tool. AIS are thought to confer the adaptability and learning required for this task.**

Keywords: Web Personalization, Artificial Immune System, Web Usage Mining , Pattern Discovery, Clonal Selection

I. INTRODUCTION

The World Wide Web is the largest distributed information environment that has grown to encompass diverse information resources. Since the web is a large collection of semi-structured and structured information sources, as the number of web pages increases dramatically, the problem of information overload becomes more severe when browsing and searching on the web. The full economic potential of the web will not be realized unless enabling technologies are provided to facilitate access to relevant web resources.

Web mining is a computational technique to recognize the useful patterns in web distributed information. Web personalization is one of the applications of web mining that affects on the web content. With personalization, web access or the contents of web pages are modified to better fit the desires of each user. This may involve actually creating new web pages that are unique per user or using the desires of a user to determine what web documents to retrieve. Personalization can be viewed as a type of clustering, classification, or even prediction [1]. With classification, the desires of a user are determined based on those for the class. Through clustering, the desires are determined based on those users to which he or she is determined to be similar. Finally, prediction is used to predict what the user really wants to see.

On the other hand, the biological immune system defenses against infectious organisms and other invaders into the human's body. Through a series of steps called the immune response, the immune system attacks organisms and substances that invade our systems and cause disease. The immune system is made up of a network of cells, tissues, and organs that work together to protect the body.

Also the artificial immune systems are considered for complex computational systems with a structure inspired biological immune system. It is a sub-field of computational intelligence, biologically-inspired computing, and natural computation, with interests in machine learning, data clustering, pattern discovery, and web personalization to the broader of Artificial Intelligence.

This paper, firstly, review the web personalization process and Summarize techniques used to support personalization. Secondly, survey artificial immune system in order to a model for web personalization, and then, we study some related works in this area. The end of this paper is mentioned challenges on web personalization in existing systems.

II. PERSONALIZATION PROCESS

In the personalization systems a variety of functions can be applied. These functions impose a number of requirements on the design of a personalization system, which aim at the development of a robust and flexible system (see figure 1). The following is a list of such generic requirements:

A. Data Collection

The first step in the web personalization process is gathering of the relevant data through the web, which will be analyzed to provide useful information about the users' behavior. There are two main sources of data for web usage mining, corresponding to the two software systems interacting during a web session: data on the web server side and data on the client side.

In the web server side, data are collected and stored in web log files. They consist primarily of various types of logs generated by the web server. These logs record the web pages accessed by the visitors of the site. Most of the web servers support as a default option the Common Log File Format, which includes information about the IP address of the client making the request, the hostname and user name, if

978-0-7695-3879-2/09 $26.00 © 2009 IEEE
DOI 10.1109/SoCPaR.2009.140

available, the time stamp of the request, the file name that is requested, and the file size[2].

Web mining tools use web server log files as the main data source for discovering usage patterns. However, log files cannot always be considered a reliable source of information about the usage of a site. The problem of data reliability becomes particularly serious for web personalization, where it is important to identify individual users, in order to discover their interests.

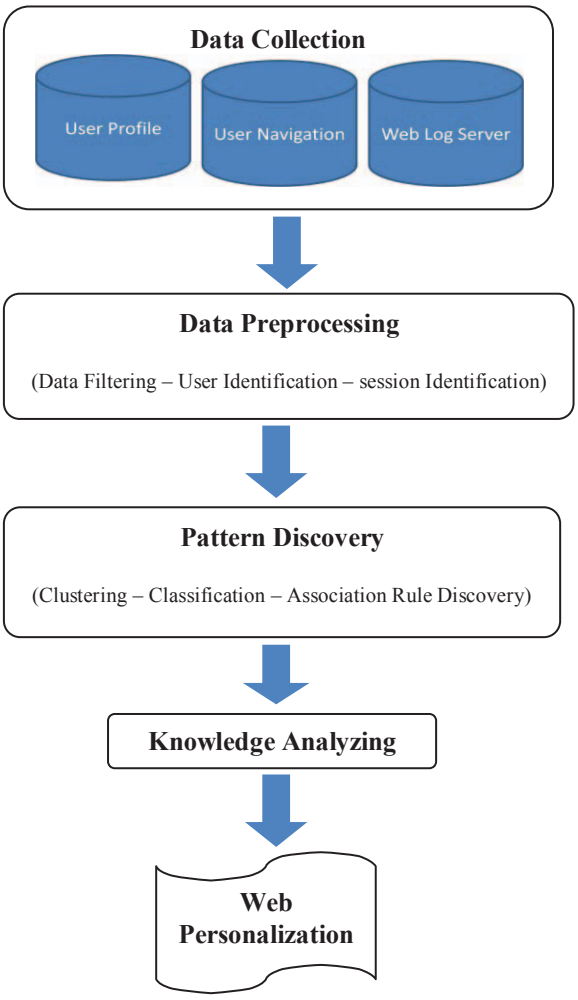

Figure 1. Web Personalization Process

Client side data are collected from the host that is accessing the web site. One of the most common techniques for acquiring client side data is to dispatch a remote agent, implemented in Java or JavaScript [4]. These agents are embedded in web pages, for example as Java applets, and are used to collect information directly from the client, such as the time that the user is accessing and leaving the web site, a list of sites visited before and after the current site, and the user's navigation history. Client side data are more reliable than server side data [3].

B. *Data Preprocessing*

Data preprocessing is a complex process in the data mining. Data preprocessing involves removing and filtering redundant and irrelevant data, predicting and filling in missing values, removing noise, transforming and encoding data, as well as resolving any inconsistencies. The task of data transformation and encoding is particularly important for the success of data mining. In web personalization, this stage includes the data filtering, identification of users and user sessions, which are to be used as the basic building blocks for pattern discovery.

C. *Pattern Discovery*

In this process, knowledge is discovered by applying machine learning and statistical techniques, such as clustering, classification, association discovery, and sequential pattern discovery to the data. The patterns required for web personalization, correspond to the behavior and interests of users.

Unlike the data preprocessing tools, the methods used for pattern discovery are domain-independent, meaning that they can be applied to many different domains, without concern about the content of the web site.

D. *Knowledge Analyzing*

In last stage, the extracted knowledge will be evaluated and presented to end user in a form that is like using reports, visualization techniques, suggested keywords or hyperlinks. For web personalization the extracted knowledge is incorporated in a Personalization module in order to facilitate the personalization functions.

III. ARTIFICIAL IMMUNE SYSTEM

The term artificial immune system refers to a group of computational intelligence techniques that are inspired by and attempt to emulate the information processing capabilities of the biological immune system. Dasgupta provides one of the earliest definitions for the term: "An Artificial Immune System is an intelligent methodology, inspired by the natural immune system, for real-world problem solving" [9]. DeCastro and Timmis provide more detail: "Artificial immune systems can be defined as abstract or metaphorical computational systems developed using ideas, theories, and components, extracted from the immune system [10]. Most AIS aim at solving complex computational or engineering problems, such as pattern recognition, elimination, and optimization." This distinguishes AIS from computational models used in biology to simulate and better understand the natural immune system itself (for more information see [10]).

This section, briefly mentioned some ability of AIS corresponding with web personalization process.

A. *Pattern Recognition Capability*

The natural immune system displays sophisticated pattern recognition capabilities. Self and non-self discrimination and the other means the immune system uses to identify and respond to threats are exercises in pattern recognition. The immune receptors can identify complex

molecular patterns, with the affinity giving a measure of the exactness of the match.

B. Learning Capability

If we consider a learning system to be one that can perform a task without having to be explicitly programmed for the specifics of that task, then the natural immune system qualifies as a learning system. The innate immune system appears to be largely 'preprogrammed' by the organism's genetic code. However, the adaptive immune system is not explicitly 'programmed' for its task. The huge number of possible pathogens, combined with the rapidity with which microorganisms and viruses change, make an enumeration in the genetic code of all possible threats infeasible. Any enumeration would inevitably have 'blind spots' consisting of unrecognizable threats. An immune system that could not learn and adapt to new threats would be defeated by some pathogen that found and exploited its blind spots.

C. Information Storage Capability

The immune system stores knowledge by means of the immune memory. If pathogens represent the problems for the immune system, then the immune memory represents a library to solutions to previously solved problems.

IV. AIS Model for Web Personalization

Biological immune system has two types of cells for detection and destruction of infection that named B-cells and T-cells. Considering first a B-Cell, whose task is to tag an antigen for destruction by another immune cell. This B-cell must bind to the antigen, and stay bound until the antigen can be destroyed. It is quite possible for no B-cells in the body to have a high affinity with this antigen. For this reason, an activated B-cell will begin a process of cloning and receptor mutation called clonal selection. Strong selective pressures during this proliferation process have the effect of maximizing affinity with the antigen, and so increasing the effectiveness of the immune response. In AIS, an activated immune cell may adapt to new web page in a similar way. Upon activation the artificial cell may undergo a process of cloning with a rate proportional to the antigenic affinity. Each new clone is mutated with a rate inversely proportional to the affinity with the antigen. Both of these processes have the goal of moving the cell closer to the antigen within the solution space. An adaptation process such as this is a common paradigm found in many evolutionary algorithms (EAs), but while mutation with rate dependent on some fitness measure is very common with AIS, other evolutionary paradigms use it rarely. After activation, a small number of clones with high affinities will survive to provide some memory of the event in the form of memory cells.

V. EXISTING SYSTEMS

The analysis of the systems presented here is based on the personalization solution they offer in terms of policy and functionality and the web usage mining methodology they adopt in order to support this personalization solution. The examined systems are shown in TABLE 1 briefly. For more detail see references.

VI. Challenges

In the artificial immune field, web personalization system is a teenager and it has the potential to increase the quality of system's output were identified [13]. The most important challenges in recent works are:

Recall that pages retrieved are thought to be interesting to a user if they are relevant to the user's search but significantly different from the original pages and therefore contain known information. This situation is conceptually close to one found in the natural immune system. That is, for T-cells to mature they must recognize the self-protein, but not bind to any other self-protein (causing an autoimmune reaction). The biological process that performs this task is called negative selection [10]. In the case of AISIID, negative selection could be used to remove mutated cells that contain a large number of interesting words that are also found on the original pages. It is acknowledged that negative selection based AIS has been criticized in the literature, especially regarding its application to the task of classification [14], for being a random search method when used in isolation without being integrated with other immune processes. Negative selection in this context would not be used in isolation rather, would be used as an enhancement to clonal selection, an enhancement that can ensure retrieved pages do contain information that is novel to the user.

The ability of AISIID to retrieve information only from hypertext (HTML) documents is an acknowledged limitation and to read the content of files such as Adobe's Portable Document Format or Postscript would be a great improvement as it would allow users to find more results.

Also, attribute weighting is a problem encountered throughout the data mining literature, especially when considering instance based learning [15]. Two weighting strategies could be advantageous to the running of AISIID: first, weighting words in mini-documents based on the proximity of the word to the hyperlink it describes. It could be hypothesized that words closer to a hyperlink are more likely to describe that hyperlink. Second, weight interesting words generated by WebIC based on the transformation used to generate them and/or their location in the hyponym/hyponym hierarchy with respect to the base word.

In language, some WebIC transformations may produce words that are fundamentally likely to be more interesting than others.

The second problem is improving preprocessing. Whilst AISIID has been designed to be robust against noise by using multiple start pages, it could be further improved if it were to disambiguate noise from content on a web page.

Furthermore, it should be noted that when viewing a webpage, two content sections rendered closely on the screen may not appear in close proximity to each other in the raw HTML. Thus the mini-document generation which relies on the proximity of content to hyperlinks may become confused. However, associating a hyperlink with text as it appear on the screen rather than as it appears in the raw HTML of the page would be a significant research topic in itself.

TABLE I. SUMMARY OF SOME WEB PERSONALIZATION AND RECOMMENDER SYSTEMS

systems	Data collection	User Identification	Session identification	Data mining tools	Output
WUM (Spiliopoulou 1999)	Log file	Done	Time based	Sequential pattern discovery	Customization of hyperlinks
web Personalizer (mobasher 2000)	Log file and cookies	Done	Context based and time based	Clustering and association rule	Recommendation of hyperlinks
(Schwarzkopf 2001)	Server Log File	Done	Time based	Association	Customization of hyperlinks
Yoda (shahabi et al. 2001)	web usage data from Client	-	-	Clustering	Customized recommendation
webIC (Zku et al. 2005)	web log file	-	-	association rule	Information Content words
AISIID (secker 2008)	web log file	-	-	Artificial immune system	User Interesting Information

It is necessary to substantially reduce the number of irrelevant "hits" returned in web searches. One of the most serious problems in delivering personal information today is the inability of the technology to allow people to easily and precisely specify the information they want, or to understand the semantics of the contents of multimedia documents.

The last lack is improving the "search keyword" to a "general object", including a free-form text, an HTML or XML document, and eventually a multimedia object such as an image, video clip, or a speech fragment. Web search engines will then retrieve "general objects" on the internet that exactly match a query object or "similar" to the query object.

VII. CONCLUSIONS

This article provides a survey of the artificial immune system, focusing on its application to web personalization. The survey aims to serve as a source of ideas for people working on the personalization of information systems, particularly those systems that are accessible over the web. Web personalization is seen as a fully automated process, powered by operational knowledge, in the form of user models that are generated by a web usage mining process. A number of systems following this approach have been developed, using methods and techniques from web usage mining, in order to realize a variety of web personalization functions. The combination of recommendation and customization functionality has been seen as the main solution to the information overload problem and the creation of relations between the web site and its visitors. Also, artificial immune system is suitable for web personalization problems.

ACKNOWLEDGMENT

We would like to thank reviewers for their value comments. This work is supported by the Research Management Centre (RMC), University Technology Malaysia (UTM).

REFERENCES

[1] Udi, M., P. Ash, and R. John, *Experience with personalization of Yahoo!* Commun. ACM, 2000. 43(8): p. 35-39.

[2] Rastegari, H. and M.N. M. Sap, *Data Mining and E-Commerce: Methods, Applications, and Challenges.* Journal of Information Technology, 2008. 20(2): p. 116-128.

[3] Pierrakos, D., et al., *Web usage mining as a tool for personalization: A survey.* User Modelling and User-Adapted Interaction, 2003. 13(4): p. 311-372.

[4] Shahabi, C., et al., *Yoda: An Accurate and Scalable Web-Based Recommendation System*, in *Cooperative Information Systems*. 2001. p. 418-432.

[5] Ansari, S., et al., *Integrating e-commerce and data mining: Architecture and challenges.* ICDM 2001 Proc. of the 2001 IEEE Intl. Conf. on Data Mining, 2001: p. 27-34.

[6] Cooley, R., B. Mobasher, and J. Srivastava, *Data Preparation for Mining World Wide Web Browsing Patterns.* KNOWLEDGE AND INFORMATION SYSTEMS, 1999. 1: p. 5-32.

[7] Schwarzkopf, E., *Personalized Interaction with Semantic Information Portals.* 2001.

[8] Spiliopoulou, M. and C. Pohle, *Data mining to measure and improve the success of web sites.* J. Data Mining and Knowledge Discovery, 2000.

[9] Dasgupta, D. *Artificial neural networks and artificial immune systems: similarities and differences*. in *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on*. 1997.

[10] Castro, L.R.d. and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Paradigm*. 2002: Springer-Verlag New York, Inc. 368.

[11] Mobasher, B., R. Cooley, and J. Srivastava, *Automatic personalization based on Web usage mining.* Commun. ACM, 2000. 43(8): p. 142-151.

[12] Zhu, T., et al., *Off-line Evaluation of Recommendation Functions*, in *User Modeling 2005*. 2005. p. 337-341.

[13] Secker, A., A.A. Freitas, and J. Timmis, *AISIID: An artificial immune system for interesting information discovery on the web.* Applied Soft Computing, 2008. 8(2): p. 885-905.

[14] Freitas, A. and J. Timmis, *Revisiting the Foundations of Artificial Immune Systems: A Problem-Oriented Perspective*, in *Artificial Immune Systems*. 2003. p. 229-241.

[15] Wang, H. and S. Wang, *A knowledge management approach to data mining process for business intelligence.* Industrial Management and Data Systems, 2008. 108(5): p. 622-634.