Chapter VII

Towards a Danger Theory Inspired Artificial Immune System for Web Mining

Andrew Secker
University of Kent, UK

Alex A. Freitas
University of Kent, UK

Jon Timmis
University of Kent, UK

ABSTRACT

The natural immune system exhibits many properties that are of interest to the area of Web mining. Of particular interest is the dynamic nature of the immune system when compared with the dynamic nature of mining information from the Web. As part of a larger project to construct a large-scale dynamic Web-mining system, this chapter reports initial work on constructing an e-mail classifier system. The Artificial Immune System for e-mail Classification (AISEC) is described in detail and compared with a traditional approach of naïve Bayesian classification. Results reported compare favorably with the Bayesian approach and this chapter highlights how the Danger Theory from immunology can be used to further improve the performance of such an artificial immune system.
INTRODUCTION

Web-mining is an umbrella term used to describe three quite different types of data mining, namely content mining, usage mining and structure mining (Chakrabarti, 2003). Of these, we are concerned with Web content mining, which Linoff and Perry (2001) define as “the process of extracting useful information from the text, images and other forms of content that make up the pages” (p. 22).

The work in this chapter is concerned with performing text mining on the Web for the purposes of classification, but this is a hard task to achieve well. Firstly, the data contained on Web pages may be low quality, noisy and inconsistent in format, and secondly, the problem space may be vast. As of August 2003 the Internet’s largest search engine, Google, indexes $3.3 \times 10^9$ Web pages (Google, 2003). Finally, the ease with which pages are published, moved or removed gives rise to an extremely dynamic medium.

It is our ultimate goal to construct a system to mine from the Web pages that the user will find interesting. That is, the user may consider them novel, surprising or unexpected. This is a slightly different problem from the classic classification task, as the class assigned to the page will depend not only on its content, but some current context. Work in Liu, Ma and Yu (2001) describes a system to mine surprising pages from competitors’ Websites. At a high level it is possible that we may take inspiration from this, such as using a piece of user-specified information to infer the subject on which the user requires information and the interestingness of the retrieved result and thus lead to future work. Statistical techniques such as a naïve Bayesian algorithm (Mitchell, 1997) have proved successful when used for the classic classification task, but we propose the use of a system we believe may be more adaptable than a Bayesian algorithm: an Artificial Immune System.

Over the last few years, Artificial Immune Systems (AIS) have become an increasingly popular machine-learning paradigm. Inspired by the mammalian immune system, AIS seek to use observed immune components and processes as metaphors to produce algorithms. These algorithms encapsulate a number of desirable properties of the natural immune system and are turned towards solving problems in a vast collection of domains (deCastro & Timmis, 2002). There are a number of motivations for using the immune system as inspiration for both data mining and Web mining algorithms, which include recognition, diversity, memory, self-regulation, and learning (Dasgupta, 1999). Being based on an AIS algorithm, by its very nature the system will preserve generalization and forget little used information; thus giving a system such as this the ability to dynamically calculate interestingness based on context and adapt to changing user preferences. Being an adaptive learning system, it will not require expert set-up; instead it will learn, for example, a particular intranet structure and tailor itself to users’ tastes.

In itself, an AIS based Web mining system would be a significant advance in the field of immune inspired algorithms. However it is our ultimate goal to go further than the areas of both Web mining and artificial immune systems have to date, by taking inspiration from an immunological theory called “Danger theory” (Matzinger, 2002a). We believe that algorithms inspired by this theory are suited to continuous learning tasks on large and dynamically changing data sets. In this theory, an immune response is launched based on a notion of perceived danger based on a current context, thus inspiring the context-dependent measure of interestingness required in the final system.
Code Generators
www.igi-global.com/chapter/code-generators/8177?camid=4v1a