网页中数据异常识别的非线性研究①

刘战东 戴玉刚 (西北民族大学 中国民族信息技术研究院 甘肃 兰州 730030)

摘 要： 由于Web上网页的急剧增加，信息搜索与挖掘越来越引起人们的重视。然而网页中除了主题信息外，还有噪声信息，从而网页净化技术受到越来越多的研究人员的关注，并提出了各种算法。借鉴人工免疫系统在计算机网络入侵检测中的应用，提出了一种基于AIS的网页去噪算法。同时对网页中的数据进行了异常识别的非线性研究。

关键词： 人工免疫系统; 异常下限; 非线性

Nonlinear Study on Web Data Anomaly Recognition

LIU Zhan-Dong, DAI Yu-Gang (China Minorities Information Technology Institute,Northwest University for Nationalities, Lanzhou 730030, China)

Abstract: Because the number of web pages is sharply increasing, information searching and excavation have drawn people's increasing attention. However, besides some thematic information, there is also inappropriate noise information on the Web pages, so Web purification technology has become a concern for more and more researchers, and various algorithms have been proposed. This paper refers to the application of artificial immune system in the computer network intrusion detection and proposes a Web page de-noising algorithm based on the AIS. At the same time, this paper does non-linear study on the data anomaly recognition of the Web.

Keywords: artificial immune system; abnormal lower; nonlinear

随着互联网的快速发展，Web上网页的数量急剧增加。面对如此巨大的信息资源，Web信息搜索与挖掘越来越引起人们的重视，在Web网页中，除了网页的主题内容之外,还常常包含一些噪声信息。所谓Web网页中的噪声信息,是指Web页面中与主题内容无关的导航条、广告信息、版权信息等[1-2]。因此网页净化技术成为网络信息检索特有的一个研究领域，受到越来越多的研究人员的关注，并提出了各种算法。较为典型的有基于一组启发式规则或页面布局信息的去噪算法、基于树形结构的去噪算法、基于模板的去噪算法和基于内容块思想的去噪算法等[1-3]。

人工免疫系统[4](AIS)是一种由理论生物学启发而来的计算范式，它借鉴了一些免疫系统的功能、原理和模型并用于解决复杂问题。近年来AIS在优化问题求解、欺骗检测、异常检测、计算机与网络安全、机器人控制、模式识别、数据挖掘及故障检测与耐受等诸多领域取得了成功的应用[5]。但尚未见AIS在网页数据去噪处理中的应用。本文主要应用现代数学非线性理论与方法，确定异常数据量；同时借鉴AIS在计算机网络入侵检测中的应用，构造新的免疫算法，它是将网页数据中的非主题数据看作入侵病毒加以清除，来达到网页数据去噪的目的。

1 人工免疫系统去噪

人工免疫是对生物免疫的模拟，免疫系统通过从不同种类的抗体中构造自己-非己非线性自适应网络，在处理动态变化环境中起作用。基于人工免疫系统提供了噪音忍耐、无教师学习、自组织、能明晰地表达

① 收稿时间:2009-12-30;收到修改稿时间:2010-01-27

学习的知识[6]。结合了分类器、神经网络和机器推理等学习系统的一些优点。本文使用的人工免疫算法思想是克隆选择算法[7]，其主要特征是免疫细胞在抗原刺激下产生克隆增殖，随后通过遗传变异分化为多样性效应细胞和记忆细胞。克隆选择对应着一个亲和力成熟的过程，即对抗原亲和力较低的个体在克隆选择机制的作用下，经历增殖复制和变异操作后，其亲和力逐步提高而成熟的过程。将网络中要处理的网页数据视为免疫系统的抗原，人们需要的网页主题数据视为抗体，依据免疫算法，找出记忆集合中的抗体，即不含噪声的主题数据。

1.1 算法步骤

本算法分为五个步骤[8]：

① 输入原始数据(抗原 A_g)，随机初始化 p 个抗体。

② 对每一个抗原数据进行以下运算：a.计算 A_b 中所有个体与 A_{gi} 的亲和力 a_{ij}。b.选择其中亲和力 a_{ij} 大的 N 个抗体，对每个选取的抗体 a_{ij} 依据大小克隆 N_C 个。越大也越大。c.按下式对克隆的抗体进行变异操作，以产生具有更高亲和力的抗体:

$$A_{b_j} = A_{b_j} - a\,(A_{b_j} - A_{b_i}),\ (j = 1,2\,\mathbf{L}\ N_c)$$

其中 a 为突变率，其大小由随机函数及亲和力大小确定。d.重新计算各个 $A_b^{'}$ 与 A_{g_i} 的亲和力 a_{ij}，选择其中具有最高亲和力的抗体作为记忆集。e.选取下一代抗原数据，直到每一个抗原数据都进行上述克隆、变异及抑制操作，完成一代网络学习。

③ 将产生的全部记忆集合 $m^{'}$ 合并为记忆数据集 M。

④ 随机产生 r 个抗体替换抗体集合中亲和力较低的个体，以实现免疫系统的自组织功能。

⑤ 返回步骤②进行下一代网络学习过程，直到达到要求的学习次数或者满足设定的目标要求。

结束后，得到的记忆数据即不含噪声的输出结果。

1.2 算法描述

输入：原始数据(抗原 A_g)，输出：不含噪声的记忆数据，上述算法描述如下：

①Procedure 克隆选择算法

②Begin

②输入原始数据并随机初始化 p 个抗体 A_b；

④While(未达到要求的学习次数 or 未达到设定的目标要求)

⑤ Begin

⑥计算 A_b 中所有个体与 A_{g_i} 的亲和力 a_{ij}；

⑦选择亲和力 a_{ij} 大的 N 个抗体;

⑧对每个选取的抗体依据 a_{ij} 大小克隆 N_c 个;

⑨repeat

⑩ 按 $A_{b_j} = A_{b_j} - a\,(A_{b_j} - A_{b_i}),\ (j = 1,2\,\mathbf{L}\ N_c)$ 对抗体变异操作;

⑪选择具有最高亲和力的抗体作为记忆集;

⑫选取下一代抗原数据;

⑬Until

⑭每一个抗原数据都进行克隆、变异及抑制操作;

⑮End

⑯产生的全部记忆集合 $m^{'}$ 合并为记忆数据集 M；

⑰随机产生 r 个抗体替换抗体集合中亲和力较低的个体;

⑱输出不含噪声的记忆数据;

⑲End

2 确定异常下限

把网页中不同性质的数据[8-9]记为 (x_i, y_i, z_i)，其中 x_i 和 y_i 分别表示数据在网页中的位置，z_i 代表数据的含量（数据的大小）。我们把 (x_i, y_i, z_i) 构成的曲面称为含量曲面。先将含量曲面的投影平面用矩形网络分割为边长为 $\Delta x \times \Delta y$ 的矩形。第 i 个矩形记为 $abcd$ 这 4 个点的投影高度分别为 $h_{ai}, h_{bi}, h_{ci}, h_{di}$（即 4 个点处数据的含量），选取含量尺度 r，当 $h_{ai}, h_{bi}, h_{ci}, h_{di}$ 均大于或等于 r 时，计算该投影网格对应的小曲面面积（否则不计算其面积，即认为其面积为 0），图 1 为计算小曲面面积的示意图，近似面积计算公式为：

$$S_i(r) = \frac{1}{2}\left[\begin{array}{l}\sqrt{(\Delta x)^2 + (h_{ai} - h_{di})^2} \times \sqrt{(\Delta y)^2 + (h_{di} - h_{ci})^2} \\ +\sqrt{(\Delta x)^2 + (h_{bi} - h_{ci})^2} \times \sqrt{(\Delta y)^2 + (h_{ai} - h_{bi})^2}\end{array}\right] \quad (1)$$

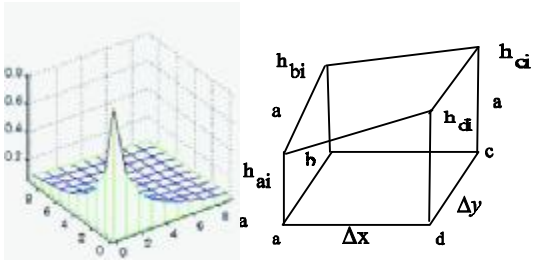

图 1　第 i 个小曲面面积

整个投影网络对应在曲面上的覆盖面积可粗略地表示为：

$$S(r)=\sum_{i}^{N}S_i(r) \tag{2}$$

其中为小矩形数目，模型为：

$$S(r)=Cr^{D},\ (r>0) \tag{3}$$

其中 r 表示含量尺度，C 为比例常数；D 为维数；s(r)是含量尺度下的投影面积，采用(2)和(3)可求。将采集的网页数据 $(S(r_1),S(r_2),\mathbf{L}\ S(r_n))$和 $(r_1,r_2,\mathbf{L}\ r_n)$ 绘在双对数坐标图上。如果其点多数分布在一条直线上，那么维数 D 可以用直线的斜率表示。化为线性回归模型：

$$\lg S(r)=-D\lg r+\lg C \tag{4}$$

可以利用最小二乘法求出斜率 D 的维数。如果其点分布在二段直线上时，可以采用分段拟合。模型如下：

$$\begin{aligned}E&=E_1+E_2\\&=\sum_{i=1}^{i_0}\left[\lg S(r_i)+D_1\lg r_i-\lg C_1\right]^2\\&+\sum_{i=i_0-1}^{N}\left[\lg S(r_i)+D_2\lg r_i-\lg C_2\right]^2\end{aligned} \tag{5}$$

D_1 和 D_2 为相应区间的斜率分维数。用所求得的维数确定出异常下限，把小于异常下限的网页数据删除，从而可以得到新的不含异常点的网页数据。

3 去噪效果实验

为检验本文方法的去噪声效果,我们使用该算法对各种含有噪声的网页进行处理。我们选取 http://news.xinhuanet.com/fortune/2009-09/05/content_12001982.html 这个网页来实验。在该网页中,从上到下依次有导航条、正文、广告信息、评论信息以及版权信息,而其中除去正文以外,其他的信息与正文本身没有任何联系。由于我们求得的网格对应的小曲面面积是一个近似解，小于异常下限的解舍去了，造成我们使用本文的算法进行去噪后,除了“热点专题”这个带链接的文字未能去除以外,其他的噪声信息均得到有效去除,而正文信息得到了有效保留。

4 结束语

本文应用现代数学非线性理论与方法对网页中的数据进行了异常识别的非线性研究，通过计算异常下限来近似计算小矩形的面积，进而确定异常数据的含量，在此基础上提出了一种基于人工免疫系统的网页去噪算法。通过去噪效果实验表明了此算法在网页去噪方面具有一定的优势。

参考文献

1 时达明,林鸿飞,杨志豪.基于网页框架和规则的网页噪音去除方法.第三届学生计算语言学研讨会.沈阳,2006.

2 Gupta S, Kaiser G, Neistadt D, et al. DOM-based content extraction of HTML documents. Proc. of the 12th International Conference on World Wide Web. New York: ACM Press, 2003:207－214.

3 Cheng QM, Agterberg FP, Bonham-Carter GF. A spatial analysis method for geo-chemical anomaly separation. Journal of Geo-chemical Exploration,1996,56:183－195.

4 Hunt JE, Cooke DE. Learning using an artificial immune system. Journal of Network and Computer Applications, 1996(19):189－212.

5 Ma L, Jiao LC, Bai L, Chen CG. Polyclonal clustering algorithm and its convergence. The Journal of China Universities of Posts and Telecommunications, 2008,15(3):110－117.

6 Grilo A, Caetano A, Rosa A. Agent-based artificial immune system. Proc. of the Genetic and Evolutionary Computation Conference 2001 (GECCO'01), San Francisco,USA, 2001,145－151.

7 李涛.计算机免疫学.北京:电子工业出版社,2004.

8 郭科,陈聆,唐菊兴.复杂地质地貌区地球化学异常识别非线性研究.成都理工大学学报, 2007,34(6):599－604.

9 冯长根.非线性科学的理论方法和应用.北京:科学出版社,1997.