维普资讯 http://www.cqvip.com

一种面向 Web 日志挖掘的免疫网络聚类算法 *)

吕 佳
(重庆师范大学数学与计算机科学学院 重庆 400047)

摘 要 Web 日志隐含了用户访问网站的行为和特点，对其进行聚类分析可以获取用户的浏览模式，发现用户访问网站的偏好和兴趣，从而优化站点结构，实现个性化的服务。针对 Web 日志数据特点，本文提出免疫网络聚类算法。该算法将 Web 服务器看成生物机体，用户访问 Web 的请求序列看成需要检测的入侵抗原，模拟抗体学习抗原的生物机理，自动生成代表用户访问模式的记忆抗体，实现动态聚类。

关键词 Web 日志挖掘，克隆选择，独特型免疫网络，免疫网络聚类算法

An Immune Network Clustering Algorithm for Web Log Mining

LU Jia
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)

Abstract Web logs implicitly embody the users' actions and traits when they access Web. Therefore, through clustering analysis, user access patterns and user preferences and interests are automatically obtained, which in turn contributes to the optimization of Web station structure and the realization of personality services. Essentially there is high dimension feature in Web logs. Against such a background, this paper presents immune network clustering algorithm, in which algorithm Web is viewed as the biology body and the incoming requests as invading antigens that need to be detected. Simulating the biology mechanism of antibodies learning antigens, memory antibodies representing user access patterns can be automatically obtained and dynamic clustering can be achieved.

Keywords Web log mining, Clonal selection, Idiotype immune network, Immune network clustering algorithm

1 引言

随着 Web 站点自身信息日益丰富和拓扑结构日益复杂，导致用户寻找有效信息越来越困难。从用户角度来看，用户都希望能够在浩瀚的信息海洋中迅速快捷地定位到自己所需要的资源，得到个性化的服务；从网站经营者的角度来看，只有迅速满足不同用户的访问需求，提高信息服务质量，才能提高网站的声誉，创造更大的经济效益。而直接或间接地解决这些问题的主要途径就是 Web 日志挖掘。目前常用的 Web 日志挖掘算法包括神经网络[1]、关联规则[2]、模糊聚类[3,4]和 K-Means 聚类[5]等。聚类分析作为一种用户模式分析的重要手段，可用来发现相似客户群体，挖掘潜在客户。但是 Web 站点的内容动态变化，包含的网页数目成百上千，这类高维数据一般分布稀疏、含有较多的噪声，以至于应用传统的聚类算法往往效果不甚理想，需要做各种改进。

从生物免疫系统启发而来的人工免疫系统，由于其自组织、自学习和记忆的能力，为 Web 日志挖掘算法研究开辟了一条新的道路。一些学者相继将免疫算法应用于聚类分析中，其中以 De Castro 提出的人工免疫网络算法应用最为广泛[6]。该算法借助人工免疫系统的基本原理，将数据集视为抗原集合，让抗体不断地学习抗原特性来实现数据聚类。受其启发，本文提出面向 Web 日志挖掘的免疫网络聚类算法，以标记类别的抗体来学习抗原集的结构和分布特征，最终可得到能反映不同类特征的记忆抗体集即为聚类结果。该算法挖掘 Web 日志时，能压缩数据，与模糊 C-Means 聚类算法(Fuzzy C-Means Clustering Algorithm, FCM)的信息熵比较结果说明该算法能更好地将具有共同特征的数据聚为一类，适合于 Web 信息量大且动态增长的特点。

2 免疫网络聚类算法

2.1 基本原理[7]

克隆选择原理是免疫系统用来说明对抗原刺激产生特异性抗体的算法。其基本思想就是只有那些能够识别抗原的细胞才能克隆扩增。克隆选择原理描述了获得性免疫的基本特征，并且声明只有成功识别抗原的免疫细胞才得以增殖。经历变异后的免疫细胞分化为效应细胞(抗体)和记忆细胞两种。克隆选择对应着一个亲和力成熟的过程，即亲和力逐步提高而"成熟"的过程。而抗体的产生是免疫系统的学习过程，也是一个优化过程。

与克隆选择原理不同，独特型免疫网络理论将免疫系统视为相互影响、相互制约的网络。独特型网络的一个重要特征是：即使没有抗原的存在，抗体间的作用网依然存在。如果某种类型的抗体过多，会受到其他抗体的抑制，使整个免疫系统趋于平衡，同时又能抵御抗原侵入。独特型免疫网络理论主要包含两种抗体作用机制：克隆抑制和网络抑制，分别在同类和不同类之间进行。同类中相似或相同的抗体被确定性地抑制，促使产生具有相似度较低及细胞种类较多的免疫细胞群，起到维持群体多样性的作用，也有利于减轻网络抑制中抗体的选择压力。网络抑制机制主要体现了不同类抗体之间相互抑制和相互促进机理，若两个抗体相似，说明通过抗体对抗原的不断学习，都学习到了抗原的共性，造成不同类抗体之间边界交叉，为了更好地反映抗原集的类分布特征，网络抑制机制将其合并为同类。此外，免疫系统随时产生自我抗体插入到存活的抗体群，以便增强抗体群多样性并调节免疫系统中抗体数目。

2.2 算法思想

*)本文得到重庆市教委科学技术研究项目(KJ050802)和重庆师范大学研究项目(05XLY003)资助。吕 佳 讲师，主要研究方向：数据挖掘。

维普资讯 http://www.cqvip.com

首先将 Web 服务器看成生物机体，用户访问 Web 的请求序列即用户会话(可抽象为一个 N 维向量)看成需要检测的入侵抗原。随机产生一定数量的不同类别的初始抗体，利用人工免疫系统的抗体与抗原之间的关系，模拟克隆选择原理、独特型免疫网络理论，以使抗体充分学习并提取出抗原的结构和分布特征，从而自动生成用户会话的聚类模式。

抗体之间通过竞争获得生存的权利，竞争由亲和力来度量。亲和力分为抗原与抗体之间的亲和力以及抗体与抗体之间的亲和力两类。

前者的度量方法采用式(1)：

$$affinity(Ab_j, Ag_j) = sim(Ab_j, Ag_i) / |Ag_i| \quad (1)$$

其中，相似度 $sim(Ab_j, Ag_i) = \sum_{k=1}^{N} Ab_j^{(k)} \times Ag_i^{(k)}$，$Ab_j$ 为第 j 个抗体，Ag_i 为第 i 个抗原，$|Ag_i|$ 为抗原的长度。

后者的度量方法采用式(2)：

$$affinityofAbs = (1/sim(Ab_i, Ab_j)) \quad (2)$$

信息理论中的信息熵能够评估聚类质量[8]。同一聚类中的数据越相似越好，则其熵越低，说明这个类的数据相似性好，聚类效果就越好。因此，信息熵是度量聚类效果好坏的一种理论标准，信息熵计算方法见式(3)：

$$E = \sum_{j=1}^{N} \frac{N_j}{N} \times E_j \quad E_j = -\frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{S} P_{ij}^{k} \log P_{ij}^{k} \quad (3)$$

其中，N 为总的聚类数；N_j 为第 j 类的抗体数；E_j 为第 j 类的信息熵；M 为抗体的基因数，每个基因位上采用的字符集大小为 S(采用二进制编码，其字符集就为 $\{0,1\}$，$S=2$)；P_{ij}^{k} 为 N_j 个同类抗体的第 i 位为字符 k 的概率。

2.3 算法描述

免疫网络聚类算法具体描述如下：

输入：抗原集 Ag，初始抗体数目 $iniAbNum$，亲和力阈值 Qd，克隆抑制阈值 Qs，网络抑制阈值 Qn。

输出：记忆抗体集 Ab。

基本步骤：

step1：从 Ag 中随机选择一定数量的抗原作为初始记忆抗体集 Ab。每一个抗体单独作为一类，打上类别标记；

step2：将 Ag 中的每一个抗原提呈给网络；

step3：计算 Ab 中每一个抗体对抗原的亲和力；

step4：若抗体的亲和力都不大于 Qd，则将抗原作为新的一类抗体加入到 Ab 中，返回到 step2；否则，选择亲和力大于 Qd 的前 n 个抗体，执行下一步；

step5：按式(4)正比于亲和力进行克隆增殖，再按式(5)对每一个克隆体反比于亲和力进行高频变异；

$$Nc = round(KScale \times affinity) \quad (4)$$

其中，$round()$ 为四舍五入的函数，$KScale$ 为控制克隆的参数，用来调整抗体克隆的数目。

$$Mutation = round(O/affinity) \quad (5)$$

其中，O 为控制高频变异的参数。

step6：在 Ab 中具有同类标记的抗体之间进行克隆抑制，抑制掉抗体间亲和力小于 Qs 的抗体；在 Ab 中类别标记不同的抗体之间进行网络抑制，若抗体间亲和力小于 Qn，则将其中一个抗体的类别标记更新为另一抗体的类别；

step7：待所有抗原提呈完毕之后，对所有的记忆抗体再按照 step6 分别进行同类克隆抑制和不同类网络抑制；

step8：若满足预设迭代次数 T，计算信息熵，记忆抗体集 Ab 即为聚类结果，算法结束；否则再随机生成 m 个抗体作为新的 m 个类加入到 Ab，转到 step2 继续操作。

2.4 算法分析

从算法可知，在输入部分首先要确定初始抗体数目 $iniAbNum$ 以便生成初始抗体集。初始抗体集是从抗原中随机抽取 $iniAbNum$ 个形成的，每一个抗原独自作为一类抗体。抗体集在学习抗原的过程中，若抗体对抗原的亲和力都不满足给定阈值 Qd，则将此抗原作为新类加入到抗体集中。所有抗原全部提呈给网络之后，又会随机生成一些新类抗体补充到抗体集。新类的加入可以弥补在产生初始抗体集时可能造成的抗体选择不均而导致学习不全面的问题，增强抗体充分学习的机会，确保网络能够自动地搜索到新类，使得最终得到的抗体能表征抗原集的结构和类分布特征，反映数据集的全貌和特点。因此，初始抗体数目不会对算法结果产生任何实质性的影响，抗体的学习全部依赖于网络自身的动态调整。故亲和力阈值 Qd、克隆抑制阈值 Qs 和网络抑制阈值 Qn 这三个参数才是真正制约网络规模及类分布的重要因素。而从亲和力的定义得知，亲和力与相似度成一定的比例关系，式(1)的亲和力与相似度成正比，式(2)的亲和力与相似度成反比。相似度是抗体与抗体(或抗原)向量表示中对应位均为 1 的总和，故可事先根据实验目的估计抗体与抗体(或抗原)大致需要相似的程度来粗略确定几个相关参数，从而避免盲目的试算。

下面对上述算法做计算复杂度分析。

假定抗原数量为 n，抗体集最大规模为 m。将每一抗原提呈给网络的 $step3$ 到 $step5$ 总的计算量为 $O(m)$；所有抗原全部提呈给网络的计算量为 $O(n \times m)$。假定总的迭代次数为 T，则算法总的计算复杂度为 $O(T \times n \times m)$。

3 Web 日志挖掘：实验结果

3.1 Web 日志数据预处理

我们采用重庆大学 Web 服务器日志作为本文的实验数据(时间段为 2004 年 12 月 05 日 04：56：08 到 2004 年 12 月 06 日 20：06：52)，总共包含 1787006 条日志记录，共 4913 个不同网页。参考文[9]中的 Web 日志预处理方法对日志记录进行清洗，删除大量与挖掘无关的记录以净化数据。庞杂的数据经清洗后得到 12533 条记录，1673 个不同网页，有效日志记录占总数的 0.7%。

接下来再对数据进行用户识别和会话识别。将时间阈值超过 30 分钟的用户分割成不同的会话，将用户在网页的停留时间设置为至少 30 秒。为了更好地获得用户的浏览模式，我们在日志预处理阶段还做了如下处理：

①由于页面数目较大，经过数据清洗之后仍然有上千个，而在众多页面当中，用户的少量的几次点击并不能表明该页面对分析用户模式意义的重要性，因此我们将这类页面删除。在本实验中，有效网页即为点击次数不少于 5 次的网页。

②对日志记录进行会话识别的目的，是为了得到由被访问的网页构成的页面序列，从中挖掘出用户浏览模式。而若用户会话中所包含的实际有效的网页数量过小，这将导致产生数目很多而成员很少的小类，或者没有共性的分类，将无助于用户浏览模式的分析，因此，我们将这类用户会话视为噪音删除。试验中，用户会话都是由至少 10 个不同网页形成的页面点击流组成。为用户会话中的每一个 URL 分配一个唯一的标识，$id \in (1,2,3,\cdots,u)$，其中 u 是所有有效 URL 的总和。这样用户会话就可用 u 位向量来表示，第 i 个用户会话向量的表示方法见式(6)：

$$S_{ij} = \begin{cases} 1 & \text{在第 } i \text{ 个会话中用户访问了第 } j \text{ 个 } URL \\ 0 & \text{在第 } i \text{ 个会话中用户没有访问第 } j \text{ 个 } URL \end{cases} \quad (6)$$

经数据预处理后，最终从原始日志文件中提取出了 56 个用户

维普资讯 http://www.cqvip.com

会话，165 个网页。

3.2 聚类结果

先在用户会话上应用 FCM 算法来考察信息熵的变化情况，从而与本文算法的信息熵做对照。权重指数取 2，迭代次数为 500，终止误差为 0.000001，表 1 为聚类数分别为 6～10 时得到的信息熵。

表 1 应用 FCM 算法不同类得到的信息熵

C	6	7	8	9	10
E	0.11204	0.12329	0.08994	0.10131	0.09409

再将本算法应用在用户会话上，表 2 为多次试算后理想初始参数取值。

表 2 初始参数设置

iniAbNum	N	Qd	Qs	Qn
30	10	0.036	0.1	0.2

免疫网络聚类算法区别于一般的基于划分的聚类算法，它不是将每一个对象归属到一个类别，而是经算法学习获得能代表类别特征的记忆抗体。这样，算法最终得到 11 个记忆抗体，按照同类别标记输出该类抗体，结果有 7 个类别，即聚类数为 7，信息熵为 0.0178197，远小于 FCM 算法聚类数为 6～10 的信息熵，证明本算法得到的同一聚类中的记忆抗体相似程度高，聚类效果好。

表 3 免疫网络算法聚类结果

类别	用户访问的 URL 类型
1	{main014 *. htm, guanlijigou/xiaoban/ *, yuanxishezhi/maofa/maoxing/ * }
2	{main013 *. htm, main02 *. htm, zhaoshengxinxi/ * }
3	{main05 *. htm, guanlijigou/kejiqiye/ *, yuanxishezhi/zcgl/index. asp, guanlijigou/kejiqiye/ *, guanlijigou/waishichu/ *, english/ * }
4	{main016 *. htm, xuexiaogaikuang/ *, guanlijigou/kejiqiye/zhidu/ * }
5	{main012. htm, guanlijigou/waishichu, guanlijigou/shiyuan/ * }
6	{main0 *. htm, tuanxuezuzhi/xstxs/ *, guanlijigou/chengshijianshe/ * }
7	{main0 *. htm, guanlijigou/chengshijianshe/ch/ *, yuanxishezhi/maofa/maoxing/new/ * }

表 3 为从用户会话中提取出的用户访问的模式，即用户访问的 URL 类型。表中只给出相同部分的链接，不同部分以 * 号代之，* 号也表示 0 到多个字符(实验中的 URL 则为每一个详细的网页地址)。

我们用 KNN 算法(K-Nearest Neighbor Algorithm)来验证聚类结果是否反映了用户会话的特征。从 56 个用户会话中随机选取 3 个会话作为检验抗体，在这 3 个会话向量中，根据其中为"1"的网页标识统计出它们分别代表了多次访问招生信息和偶尔访问学校概况、多次访问外事处和英语信息以及主要访问贸行学院和偶尔访问招生信息的 3 种用户浏览行为。分别计算 3 个检验抗体同 11 个记忆抗体的距离，距离最近的记忆抗体的类别作为检验抗体的类别。因为这 7 个类是由抗体不断学习原始用户会话的结构和分布特征抽象出的，相对于其它类，类别 2、类别 3 和类别 1 更能反映出这 3 个会话的特征，故将 3 个会话分别划归到这 3 个类中。

结束语 针对 Web 日志数据自身特点，提出免疫网络聚类算法。与一般的聚类算法不同，免疫网络聚类算法不需要具体确定用户会话所属的类别，而是利用免疫系统自身的自组织、自适应的特性，让带有类别标记的抗体学习抗原特性并提取其分布特征，使之尽可能地反映一个类的结构特点，用这些能代表一个类的较少的几个典型的抗体来概括出一个类的特征。通过在一个真实的网站日志数据集上进行实验，结果表明该算法是有效的。

参 考 文 献

1 董一鸿，庄越挺. 基于新型的竞争型神经网络的 Web 日志挖掘[J]. 计算机研究与发展，2003，40(5)：661～667

2 Krishnapuram R, Joshi A, Nasraoui O, et al. Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining. IEEE Trans. on Fuzzy Systems, 2001, 9(4): 595～607

3 宋擒豹，沈钧毅. Web 日志的高效多能挖掘算法[J]. 计算机研究与发展，2001，38(3)：328～333

4 刑东山，沈钧毅，宋擒豹. 从 Web 日志中挖掘用户浏览偏爱路径[J]. 计算机学报，2003，26(11)：1518～1523

5 王实，高文，李锦涛，等. 路径聚类：在 Web 站点中的知识发现[J]. 计算机研究与发展，2001，38(4)：482～486

6 De Castro L, Vou Zuben F J. AiNet: an artificial immune network for data analysis. In: Data Mining: A Heuristic Approach, H. A. Abbass, R. A. Sarker, C. S. Newton, eds. Idea Group Publishing, USA, Chapter XII, 2001. 231～259

7 李涛著. 计算机免疫学[M]. 北京：电子工业出版社，2004

8 Hang Xiaoshu, Dai Honghua. An immune network approach for Web document clustering [A]. In: Proc. of the IEEE/WIC/ACM International Conf. on Web Intelligence [C]. Beijing: China, 2004. 278～284

9 郭岩，白硕，于满泉. Web 使用信息挖掘综述[J]. 计算机科学，2005，32(1)：1～7

(上接第 153 页)

Kashyap 提出的方法[3]：该方法与 Stojanovic 等提出的方法类似，而且它要求用户进行语义标注工作，抽取语义信息时的自动化程度不高，该方法无法产生公理。

Dogan 和 Islamaj 提出的方法[4]：该方法提供了简单、自动的模式转换和数据映射过程，在该方法中一个关系映射成一个概念，其中的属性映射成概念的性质，一个关系的元组映射成本体的一个实例。该方法没有考虑继承关系，不能对关系数据库的结构进行优化，因此所建立的本体看起来很像关系。

结论 本文提出了一种新的从关系数据库中获取本体的方法。通过对主键、数据、属性的相关性的分析，允许用户从关系数据库中抽取更多的语义信息，包括继承关系，甚至可以对结构进行优化；减少用户干预，仅仅要求确认抽取的语义信息，为它们命名。以后可以将该方法应用于关系数据库向语义网的转换。

参 考 文 献

1 Studer R, Benjamins V R, Fensel D. Knowledge engineering, principles and methods. Data and Knowledge Engineering, 25(1-2): 161～197

2 Stojanovic L, Stojanovic N, Volz R. Migrating data-intensive Web Sites into the Semantic Web. In: Proc. of the 17th ACM symposium on applied computing. 2002

3 Kashyap V. Design and creation of ontologies for environmental information retrieval. In: Proc. of the 12th Workshop on Knowledge Acquisition, Modeling and Management. 1999

4 Dogan G, Islamaj R. Importing Relational Databases into the Semantic Web. http://www. mindswap. org/webai/2002/fall/Importing_20Relational_20Databases_20into_20the_20Semantic_20Web. html (2002)

5 F-Logic. http://www. informatik. uni-freiburg. de/~dbis/Publi-cations/95/flogic-jacm. html

6 Codd EF. A relational model of data for large shared data banks. CACM, 1970, 13(6)

7 Maedche A. Ontology learning for the Semantic Web. Boston: Kluwer Academic Publishers, 2002

8 Astrova I. Reverse Engineering of Relational Databases to Ontologies. In: Proceedings of the 19th ACM Symposium on Applied Computing (SAC), 2004