基于聚类的智能网页推荐系统研究

Research of Intelligent Web Page Recommender System Based on Clustering Technique

王有为/WANG You-wei

复旦大学管理学院，上海 200433

School of Management, Fudan University, Shanghai 200433, China

[摘要] 设计了一种智能网页推荐系统的架构，其中包括数据预处理、聚类分析和网页推荐3个子系统，可以根据网站的访问日志来对用户进行自动分类，进而对网站的新用户在线提供网页推荐。提出了路径间距离的计算方法，进而研究了聚类子系统的结构，并通过对微软网站中用户访问日志的仿真实验，说明了所述方法的有效性。

[关键词] 网站访问日志；聚类分析；网页推荐；推荐系统

[中图分类号] TP391.9 [文献标识码] A [文章编号] 1000-7857(2006)10-0033-04

Abstract: With the size of WWW growing bigger, its structure is more and more complex. Web visitors with personalized information demands find it very difficult to locate information with ease. In this paper, we designed an intelligent web page recommender system, which include three sub-systems of data preprocessing, clustering, and web page recommendation. The proposed system can divide visitors into subgroups automatically, thus providing new users with online web page recommendation. The formula for computing the distance between visitor paths is proposed, followed by the study of the structure of clustering system. The simulation results based on web log of Microsoft website proved the validity of the proposed method.

Key Words: web log; clustering analysis; web page recommendation; recommender system

CLC Number: TP391.9 Document Code: A Article ID: 1000-7857(2006)10-0033-04

1 引言

随着万维网(WWW)越来越庞大和复杂，用户迅速获取所需的信息越来越困难。大型网站通常包含成千上万的网页，而具体用户感兴趣的信息可能仅仅包含在少数几个网页之中。对每个用户而言，所需要信息也各不相同，但目前绝大多数网站只能针对所有用户提供完全一样的服务[1]。因此，互联网庞大的信息量和用户个性化的信息需求之间存在着矛盾，这促生了各种各样的网页搜索技术。

搜索引擎是目前应用最广的网页搜索技术，用户可以通过输入与所需内容相关的关键词从而快速获取相关的网页。当用户提供的关键词比较准确，而且不存在同义词问题(同一个词有多种不同的含义)时，搜索的结果一般是比较满意的。但是，通用搜索引擎的主要问题是常常返回大量的网页集合，而其中可能仅有少数几个链接与当前用户的兴趣相关，这增加了用户搜索的成本[2]。因此，有的网站也提供站内搜索引擎，可以用于发现当前网站中与用户兴趣相关的网页集合。

解决个性化信息需求问题的另外一种方案是智能网页推荐技术。它通常集成在网站服务器中，通过观察用户已经访问过的网页，从而推测其未来可能感兴趣的网页集合。协同过滤(Collaborative Filtering)技术可以服务于这一目的，它通过参考与当前用户类似的用户的兴趣和偏好，从而为当前用户提供访问的建议。但是，协同过滤常常需要用户在网站中注册，并且要求用户提供偏好信息，并以此作为对其他用户推荐的基础[3]。不过，用户访问某个网站常常是一种临时行为，他们不愿意花费更多时间提供个人偏好或者反馈意见，传统的协同过滤推荐因此很难为非注册用户提供高质量的推荐，这种方法的应用范围受到一定的限制[3]。

还有一种智能推荐技术是基于网站服务器记录下的用户访问日志，它不需要用户主动提供访问偏好信息，而是通过观察用户访问的内容和次序来推测用户的行为，因此具有应用范围较广的特点[4-5]。举例而言，有一个用户甲正在浏览微软网站，他按顺序访问了MSDN、Windows Server 2003、.Net等内容，这种访问行为体现了他对网站内容的偏好。另外有一个用户乙，他访问了MSDN、Windows XP、.Net等页面。虽然这两个人的访问路径不完全相同，但直观上看两个人的兴趣比较相似。因此，如有一个新用户丙，通过分析发现他与甲和乙的行为非常类似，于是我们就可以根据甲和乙的访问路径来预测丙即将访问的网页集合。因此，智能推荐技术的关键就在于如何判断哪些用户具有相似的兴趣，以及将哪些网页推荐给当前访问的用户。

本文的方法也属于基于日志的推荐技术，将综合运用人工智能和数据挖掘技术，识别相似的用户集合，并设计一个基于日志

收稿日期：2006-08-11

基金项目：国家自然科学基金资助项目(70401010)，国家留学基金资助项目

作者简介：王有为，男，上海市国顺路670号李达三楼复旦大学管理学院，讲师，研究方向为复杂系统建模与优化、决策支持系统、移动商务等；E-mail: ywwang@fudan.edu.cn

的推荐系统，从而提高网站的信息服务质量。本研究属于一种自适应网站技术[6]，使用这种技术，网站可以根据自身如何被使用而改变自身的服务方式。适应性可以通过多重形式来体现，包括对文本、链接或者页面格式的临时修改，也包括对链接的添加和删除[7-8]。

2 智能网页推荐系统架构

本文提出的网页推荐系统架构如图 1 所示，它包括以下 3 个子系统：① 数据预处理子系统，原始的网站访问日志数据虽然是格式化的，但是仍不能满足数据分析的要求，需要将日志数据预处理；② 数据聚类子系统，根据预处理过的数据，通过聚类技术发现典型的用户集合；③ 网页推荐子系统，推测一个用户可能属于什么样的用户集合，然后根据其可能存在的访问偏好对其进行网页推荐。其中只有推荐子系统是在线实时工作的，预处理子系统和聚类子系统都是离线工作，这种设计可以最大限度地减少服务器的压力，并提高系统的执行效率[9]。

图 1 所示的推荐系统实际上是一种 C/S 结构，智能推荐系统作为提供推荐服务的服务器端，网站服务器作为请求推荐的客户端，与推荐系统是相对独立的。2 个系统之间可以通过某种应用

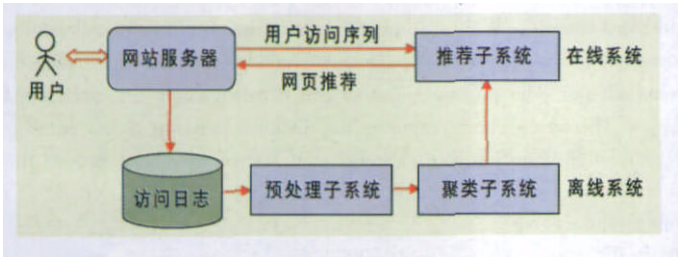

图 1　网页推荐系统架构图

Fig. 1　Architecture of web page recommendation system

接口（例如 TCP/IP，HTTP 等）交互。此构架的特点是推荐系统的运行环境不需要和应用系统相同，对应用环境的适应性好。网站服务器可以采用任何成熟的产品，比如 Apache HTTP Server。

当一个新用户访问网站的时候，网站服务器跟踪用户会话，记录用户访问路径，然后存储在访问日志中。日志中的数据是按照所有用户访问的时间顺序存放的，不适合于本文期望的以用户为单位的顺序。而且，日志数据也可能由于偶然的原因造成部分缺失，因此必须将日志数据进行处理后才能使用。数据预处理子系统定期从访问日志中抽取数据，然后把预处理过的数据提交给聚类子系统，聚类子系统根据最新的访问数据定期重新执行聚类算法，并将聚类结果保存下来。在线推荐子系统根据当前用户的访问序列，并结合聚类结果情况，判断此用户可能即将访问的网页集合，并反馈给网站服务器。然后，网站服务器再根据用户实际点击的链接，在目标网页上增加可能即将访问的网页链接集合，再反馈给用户供其浏览。以上就是整个推荐系统的执行流程。

3 智能网页推荐系统关键技术

3.1 数据预处理

现实世界中的数据经常包含噪声数据、空缺数据和不一致性数据。本文所设计的系统中，网站的访问日志是唯一的数据来源，它由网站服务器自动产生。一般情况下，网站服务器产生的数据是完整可靠的，但也有可能由于一些特殊的原因产生错误的日志，比如网络传输环境的不稳定、服务器意外关闭、用户端突发故障等等。另外，网站日志数据与本文需要的格式有一定的差距，因此在使用之前通常需要进行一定的预处理，才能为聚类子系统所用。本文具体使用的预处理步骤如下。

3.1.1 将网页映射为标识符

预处理的第一步就是分析网站的结构。将每个网页用唯一的 URL 地址来标识，也可以将 URL 地址映射为唯一的数字编码，本文采用第二种方法。由于微软网页的数量过于庞大，可以按照内容把网页划分为一个个集合，每个集合代表一组相关的网页。例如，根据微软(www.microsoft.com)公开的 1998 年 2 月份一周的匿名访问日志[10]，网页集合用如下的形式表示：

A,1253,1,"MS Word Development","/worddev"
A,1109,1,"TechNet (World Wide Web Edition)","/technet"
A,1205,1,"Hardware Supprt","/hardwaresupport"
A,1076,1,"NT Workstation Support","/ntwkssupport"

其中，A 是每一行开始的标志符，接下来的序号表示集合的 ID，然后是集合的主题描述，以“/”开头的字符串表示各集合在网站上的路径。于是，用户的访问路径就不必以链接形式，而是以[1277,1253,1109]这样的集合 ID 序列来表示，这为聚类分析提供了很大的方便。

3.1.2 解析访问路径

从网站日志中直接发现知识比较困难，因为日志按照所有用户访问的先后次序排列，所有用户的访问过程彼此重叠在一起。但是，用户的访问路径适合以向量格式表示和分析，所以数据预处理的一个主要目标就是要把原始的日志数据转化为向量数据。微软网站访问日志中用户访问路径的原始数据格式如下[10]：

C,#10001, V,1038,1, V,1026,1, V,1034,1, C,#10002, V,1008,1, V,1056,1, V,1032,1

以上数据实际上已经经过一定的预处理。数据记录来自微软网站的 32 711 名匿名用户，每个用户是由一个序号作唯一的标识，比如 #10001，#10002。数据格式中，“C”开头的一行代表一个新用户的访问记录开始，“V”开头的一行代表这个用户访问的一个路径，其中的“1038”、“1026”等分别表示微软网站不同的虚拟目录，也就是上文所提及的网页集合。数据处理的目标是抽象出每个用户的访问路径，存储在一个向量中。比如对于 #10001 用户，我们可以用[1038,1026,1034]这样的向量表示它的访问路径，实际的数据处理通过编程的方式来实现。

3.1.3 删除访问路径中的重复结点

在解析出的路径中，经常有连续多次访问同一个网页的情况，这可能是由于用户端的 Cache 机制等所造成的，这对于本文的聚类算法没有意义，所以在解析出的路径里选择删除重复的结点。

3.1.4 删除太短或者太长的访问路径

太短的访问路径(比如长度为 1 或者 2)表示用户刚刚浏览了少数网页后就离开了，因此包含的信息量很少，对本文的意义不大；太长的访问路径(比如超过 50)通常是用户端使用代理服务器(Proxy)的结果，它通常是许多个用户的访问行为的重叠，对于本文的算法是一种噪声数据，因此也被删除。本文认为，只有用户访问的路径长于某个阈值的时候，此数据对系统而言才是有用的，这个阈值在数据预处理子系统中定义为 7。

原始数据一共有 32 711 个匿名用户访问记录，本文截取了 5 000 条记录作为测试数据用。根据筛选规则，一共得到 449 条用户向量，其格式如下：

[10003, 1064, 1065, 1020, 1007, 1038, 1026, 1052, 1041, 1028]
[10007, 1008, 1000, 1035, 1016, 1031, 1003, 1034, 1018]

[10037, 1008, 1016, 1152, 1052, 1040, 1034, 1014]

其中第一个分量是用户 ID，接下来是用户在微软网站的访问路径。

3.2 聚类分析

访问网站的用户数量通常比较庞大，如果参考所有用户的访问路径并对当前用户进行推荐，一方面服务器的负荷很大，另一方面推荐的准确性也不高。因此，为对每个访问用户提供有价值的推荐，需要缩小参考用户的范围，确定与当前用户访问兴趣最相似的用户集合。在基于访问日志的推荐系统中，用户的兴趣体现在用户访问过的网页集合，以及访问的先后次序上，即数据预处理获得的路径向量中[11]。而聚类分析算法可以按照访问行为的不同对先前的用户进行分类，从而达到缩小参考用户范围的目的。

聚类是一个将数据对象分组成为由类似的对象组成的多个类的过程，它的输出是一些簇，即数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异[2]。通过聚类，能够识别密集和稀疏的区域，发现数据的全局分布模式，以及数据属性之间的相互关系。聚类算法需要以一定的格式输入，即基础数据结构。

3.2.1 数据结构

以下是聚类算法 2 种有代表性的输入数据结构[2]。

3.2.1.1 数据矩阵(data matrix)

它用 p 个变量(也称为度量或者属性)来表现 n 个对象，例如用年龄、身高、体重、性别、种族等属性来表示对象“人”，这种数据结构是关系表的形式，是一个 n×p(n 个对象×p 个对象)的矩阵。

3.2.1.2 相异度矩阵(dissimilarity matrix)

存储 n 个对象两两之间的近似性，表现形式是一个 n×n 维的矩阵。在相异度矩阵中，d(i, j)是对象 i 和对象 j 之间的相异性的量化表示，它是一个非负的数值，当对象 i 和 j 越相似或“接近”，其值越接近 0；两个对象越不同，其值越大。显然有，d(i, j)=d(j, i)，而且 d(i, i)=0。

本文采用相异度矩阵作为聚类算法的基础。根据数据对象的类型不同，使用的相异度计算方法也不相同。由于本文所讨论的对象(访问路径)不属于通常的区间标度变量、二元变量、标称型变量、序数型变量、比例标度型变量，或者一般的混和类型变量，因此必须设计新的变量相异度计算方法。

3.2.2 路径的相异度计算方法

假设用户甲的访问路径为 v1=[1008, 1000, 1035, 1016, 1031, 1003, 1034, 1018]，用户乙的访问路径为 v2=[1008, 1009, 1058, 1017, 1018, 1042, 1021]，这 2 个访问路径都以向量形式表示，如何确定 2 个用户对于网站内容的偏好是否相似？最简单的方法是计算代表甲和乙访问路径的 2 个向量间的距离。

由于 2 个向量的长度不同，所以要先对比较的 2 个向量进行处理，然后才能进行其间距离的运算。对于甲和乙的访问路径 v1、v2，如果分量 a 既属于 v1 也属于 v2，那么 a 将属于新向量 v3，令 v3=[a1,a2,…,ai]；如果分量 b 属于 v2，同时也属于 v1，那么 b 将属于新向量 v4，令 v4=[b1, b2,…,bi]。显然，v3 和 v4 的维度相同，而且包含的分量也是相同的，只是分量的排列顺序可能不同。对于甲和乙而言，v3=v4=[1008,1018]。

对于长度相同的向量 v3 和 v4，它们之间的距离只与其分量的排列顺序有关，可以采用如下计算公式：

$$d(v3, v4)=|\{(i, j): i<j \text{ 并且 } v4.indexOf(v3(i))>v4.indexOf(v3(j))\}| \quad (1)$$

公式(1)中，v4.indexOf(v3(i))表示 v3(i)这个分量在 v4 中的位置；符号 1·1 表示集合中元素的个数。如果把所有的向量间距离都转换到共同的值域区间[0.0,1.0]上，得到如下距离计算公式[12]：

$$d(v3,v4)=\frac{d(v3, v4)}{N\times(N-1)}\times 2 \quad (2)$$

公式(2)中，N 表示对象 v3、v4 的长度。对于甲和乙而言，v3 和 v4 是完全相同的，则它们之间的距离为 0。如果 v3=[1008,1018]，v4=[1018,1008]，则他们之间的距离为 1。

显然，在上述计算方法中，没有考虑原始的路径向量的长度，因此单纯用 v3、v4 之间的距离来衡量用户的兴趣差异具有一定的偏差。与文献[12]中的顺序距离(Rank Distance)不同，本文考虑给 v3、v4 的相异度计算公式加一个调整量，这个调整量将反映 v1、v2 区别于 v3、v4 的特征。假设 v1 的长度为 m_1，v2 的长度为 m_2，考虑如下的调整量：

$$m=\frac{\sqrt{m_1\times m_2}}{N} \quad (3)$$

于是，加入调整量后对象 v1、v2 间的距离计算公式为：

$$d(v1,v2)=\frac{d(v3, v4)+m}{N\times(N-1)}\times 2 \quad (4)$$

考虑调整量以后，路径间的距离有可能大于 1.0，但这并不影响本文的聚类分析算法。特殊情况下，当 N 值为 0 或者 1 的时候，甲和乙访问路径没有一点相同或者只有一个集合相同，网站浏览者甲和乙的偏好很不相似，他们很少可能会访问相同的网站链接。

3.2.3 聚类算法

已经开发出许多种聚类算法，比如，划分的方法、层次的方法、基于密度的方法、基于网格的方法、基于模型的方法，等等[2]。每种算法在伸缩性(Scalability)、对数据类型的要求等方面具有不同的特性。算法的选择取决于数据的类型、聚类的目的和应用。上述算法中，许多只能发现球形的簇。在本文的推荐系统中，需要发现任意形状的簇，因此采用了基于密度的方法。其主要思想是：只要临近区域的密度(对象或数据点的数目)超过了某个阀值，就继续聚类。也就是说，对给定类中的每个数据点，在一个给定范围的区域中必须至少包含某个数目的点。这样的方法可以用来过滤孤立点数据，发现任意形状的簇。

DBSCAN (Density - Based Clustering of Applications with Noise)是一个有代表性的基于密度的聚类方法[13]。该算法将具有足够高密度的区域划分为簇，并可以在带有“噪声”的空间数据库中发现任意形状的聚类。它定义簇为密度相连的点的最大集合。

DBSCAN 算法不但在二维和三维的欧几里德空间，在高维空间工作得也很好。关键是对于聚簇中的每一个点，在给定的半径内至少需要包含一定数量的点，也就是说密度需要超过某个阀值。点相邻的形态取决于所采取的 2 个点 p 和 q 之间的距离计算函数，表示为 dist(p,q)。基于密度的聚类算法涉及一些新的定义[13]。

1) ε 邻域：给定对象半径 ε内的区域成为该对象的 ε 邻域。

2) 核心对象：如果一个对象的 ε 邻域至少包含最小数目 MinPts 个对象，则称该对象为核心对象。

3) 直接密度可达：给定一个对象集合 D，如果 p 是在 q 的 ε 邻域内，而 q 是一个核心对象，我们说对象 p 从对象 q 出发是直接密度可达的。

4) 密度相连：如果对象集合 D 中存在一个对象 o，使得对象 p 和 q 是从 o 关于 ε和 MinPts 密度可达的，那么对于对象 p 和 q

是关于 ε和 MinPts 密度相连的。

4 仿真实验

4.1 实验参数

DBSCAN 算法的参数必须手动选择，在聚类子系统中，确定参数值是一个需要非常谨慎的过程，这直接影响到聚类结果。对于微软网站的数据，本文选取了多组参数值，得到了多组参数组合的不同聚类结果，根据不同的聚类结果得到的核心对象数目等选择最优的参数组合。本文实验的参数组合如表 1 所示。

表 1 参数实验组合
Tbl. 1 Different parameter settings

参数组别	ε	MinPts	核心对象
1	0.1	5	17
2	0.1	10	5
3	0.1	15	0
4	0.2	10	85
5	0.2	15	46
6	0.2	20	26
7	0.3	15	141
8	0.3	20	109
9	0.3	25	80

模拟系统中处理的对象总数为 449。经过实验测试，发现核心对象数取 40~50 左右较为合适，所以选取了 ε为 0.2, MinPts 为 15 这样的参数组合。

4.2 实验结果

本文所采取的推荐算法分为以下 3 个步骤。

步骤 1: 当一个新用户浏览网站的时候，应用系统向推荐系统发送一个推荐请求，这个推荐请求中包含用户在此之前的访问路径。

步骤 2: 推荐系统从聚类子系统获取聚类结果，然后根据聚类结果对用户访问路径进行分析，看该用户可能属于什么样的簇。

步骤 3: 选取用户所属聚类中的一个距离最近的核心对象，根据该核心对象对网站的访问路径，对新用户进行下一步访问的链接推荐。

本文从微软网站所提供的数据中选取一些对象对推荐算法进行模拟，以验证算法的有效性。从预处理过的数据中任取一个向量用户访问路径 v1 =[1017, 1018, 1008, 1046, 1060, 1052, 1003, 1034, 1002]，发现向量 v=[1017, 1156, 1004, 1018, 1008, 1027, 1009, 1046, 1038, 1006, 1026, 1052, 1060, 1001, 1041, 1034]可以作为对向量 v1 进行推荐的核心对象。

因为在数据预处理系统中只有用户访问的路径长度的阀值定义为 7，所以取 v1 的前 7 个分量模拟用户的第一个可聚类访问路径 v2=[1017, 1018, 1008, 1046, 1060, 1052, 1003]。计算 v2 与核心对象 v 的距离，计算结果为 0.2。这表示，在 v 的 ε 邻域内，v2 与 v 有相同的偏好。可以推测 v2 下一步访问的页面可能为包含在 v 中，但是没有被 v2 所访问过的页面，这样的网页包括 [1156,1004,1027,1009,1038,1006,1026,1001,1041,1034]。显然，v2 所模拟的用户在下一步访问了 1034 页面，这个页面在当前推荐的网页集中，因此推荐是有效的。

5 结论

本文介绍了网页链接推荐系统的概念，设计了智能推荐系统的架构，并对 3 个子系统进行了系统和算法的设计。由于预处理子系统提供的数据中各向量的维度并不相同，本文设计了计算向量间距离的方法。在比较研究数据挖掘的各种聚类技术后，在聚类子系统中选择了基于密度的聚类方法，并根据相异度矩阵来进行聚类分析。最后，通过仿真实验说明了本文所述方法的有效性。

作为未来的研究方向，还可以在以下方面进行提高。① 在实验部分增加实验的数据量，将访问数据分为训练集(用于聚类分析)和测试集(用于测试实际推荐效果)，通过统计分析结果来说明推荐算法的精度。② 在路径间距离的计算方法部分，可以与其他方法(比如基于图论的距离计算公式[14])进行实验比较，从而寻求更好的计算方法，为实际应用打下基础。

参考文献(References)

[1] BERGHEL H. Dealing with information overload [J]. Communications of the ACM, 1997, 40(2): 19-24.

[2] HAN J W, KAMBER M. Data mining: concepts and techniques [M]. 范 明, 孟小峰等译. 北京: 机械工业出版社, 2002.

[3] BURKE R. Hybrid recommender systems: survey and experiments [J]. User Modeling and User Adapted Interaction, 2002, 12(4): 331-370.

[4] SRIKANT R, YANG Y H. Mining Web Logs to improve website organization. the 10th Proceedings of International WWW Conference [C]. Hong Kong: 2001.

[5] SCHAFER J B, KONSTAN J, RIEDL J. Recommender systems in e-Commerce, Proceedings of the First ACM Conference on Electronic Commerce [C]. Denver: 1999.

[6] PERKOWITZ M, ETZIONI O. Adaptive Web sites: automatically learning from user access patterns. Proceedings of the Sixth International World Wide Web Conference [C]. Santa Clara, CA: 1997.

[7] 王有为, 汪定伟. 电子超市中网页设计策略的研究 [J]. 系统工程学报, 2003, 18(2): 104-108.

[8] WANG Y W, WANG D W, IP W H. Optimal design of link structure for E-supermarket website [J]. IEEE Transactions: Systems, Man and Cybernetics - Part A, 2006, 36(2): 338-355.

[9] YAN T W, JACOBSEN M, GARCIA-MOLINA H, et al. From user access patterns to dynamic hypertext linking [J]. Computer Networkds and ISDN Systems, 1996, 28(7-11): 1007-1014.

[10] Microsoft Anonymous Web Data, The UCI KDD Archive, 1998 [2006-07-20]. http://kdd.ics.uci.edu/databases/msweb/msweb.html

[11] CHAN P. Constructing Web user profiles: a non-invasive learning approach. Web Usage Analysis and User Profiling [M]. Springer-Verlag:2000.

[12] TSAPARAS P. Application of non-linear dynamical systems to Web searching and ranking. WWW2003 [C]. Budapest: Hungary, 2003.

[13] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]/Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996.

[14] BUNKE H, MESSMER B. Recent advances in graph matching [J]. International Journal of Pattern Recognition and Artificial Intelligence, 1997, 11(1): 169-203.

(责任编辑 王 芷)