基于小波理论的 Web 挖掘模型研究

A Web Mining Model of Wavelet Theory – Based

邵良杉　刘飞飞　（辽宁工程技术大学系统工程研究所 辽宁阜新 123100）

摘　要： 为了对 internet 上的半结构化数据进行分析，发现其内在的关联模式，论文提出了基于小波理论的 web 挖掘模型，该模型支持 web 挖掘的全过程。Web 挖掘模型由数据采集器、预处理器、数据约简、挖掘综合器、挖掘方法库和系统维护六部分组成。该模型应用小波聚类分析方法，实现了对经过预处理的 Web 数据进行约简的功能。去除了一些冗余的无意义的数据，优化了系统的性能，提高了 web 挖掘质量。

关键字： 小波变换 小波聚类 小波变换去噪 Web 挖掘

1　引言

Internet 上包含了海量的数据。这些数据与传统的数据库中的数据不同，它们通常具有自述性和动态可变性，因此是半结构化的数据。要想对这些半结构化的数据进行分析，发现其内在的关联模式，必须先将数据统一格式，然后去除一些冗余数据和无意义数据，利用相应的分析方法对处理后的数据进行挖掘。小波理论作为一个新的数学分支，其数学理论和方法被应用于信号处理、图象处理、语言分析、模式识别等领域。而近年来，小波分析也越来越被广泛的应用于数据挖掘分析领域[1][2][4]。

本文在前人研究的基础上，利用 Mallat 塔式小波变换对数据进行多尺度分解，将不同尺度下或带宽下的噪声展现出来，通过聚类分析技术实现对系统噪声数据的去除，并结合 Web 挖掘技术，构造一个基于小波理论的 Web 挖掘模型。

2　小波聚类算法

2.1　Mallat 小波变换

Mallat 塔式算法的基本思想是，将一个分辨率为 1 的离散逼近 A_0F 分解为一个粗分辨率 2^{-J} 的逼近 A_jF 和逐次细节信号 $D_jf(0 < j \leq J)$，其分解可以不断地进行下去[3]。

2.2　小波聚类

小波聚类（Wave Cluster）方法是一种基于网格和密度的多分辨率聚类算法，能有效的处理大数据集合和多维度数据，发现任意形状的类，成功地处理孤立点[5]。

小波聚类方法是利用小波变换的多分辨率特性达到聚类的目的. 算法的主要思想为：一个具有 n 个属性的元组，它的 n 个属性值构成一个 n 维的特性矢量，该矢量可以看成是 n 维特性空间中的一个点. 数据集中所有元组的集合构成一个 n 维信号. 信号的高频部分对应于元组急剧变化的区域，即聚类簇的边界. 具有较大模值的低频成分对应于元组集中分布的区域，即聚类簇. 因此用小波变换将信号变换到频域后，通过寻找平滑概貌的密集区域来标识聚类. 在聚类算法中采取了网格技术，小波变换针对网格进行，其算法的步骤是：

1）数据空间的量化，将对象分配给单元；

2）在特性空间上应用小波变换去噪；

3）给单元分配标签；

4）建立查寻表，将对象映射到聚类；

2.2.1 数据空间的量化

量化特征空间：假设 n 维特征空间中每一维 i 都被等分为 m_i 个小区间，且所有维中 $m_i = m^*$，那么特征空间中将有 k = 个单元，然后数据集中的所有数据点将被映射到量化空间中去。数学描述如下：初始数据集中任一点在量化空间中的值为 $F_K = (F_1, F_2 \cdots, F_N)$，而 $v_i (1 \leq v_i \leq m, 1 \leq i \leq n)$ 是量化单元在特征

空间中 x_i 轴上的位置。假定是每个量化单元在轴上的大小，如果对所有 i 值，有 $(\upsilon_i - 1) \times s_i \leqslant f_i \leqslant \upsilon_i \times s_i \quad 1 \leqslant i \leqslant n$，那么数据点对应的量化值 $F_k = (f_1, f_2 \cdots, f_n)$ 就被定义到单元 $m_j = (\upsilon_1, \upsilon_2 \cdots, \upsilon_n)$ 上，其中，$1 \leqslant j \leqslant m_n$。

2.2.2 在特征空间中应用小波变换去噪

对量化特征空间实施 Mallat 小波变换。单元 m_i 实施小波变换后能够得到一个体现信号特征的小波系数序列，在多分辨率分解中，信号在每一个分解尺度 2_j 上都有两个小波系数矩阵：一个是该尺度上的信号整体趋势部分的系数矩阵 $A_j = (a_{j1}, a_{j2}, \cdots a_{jn})$；另一个是该尺度上的信号细节信息的小波系数矩阵 $D_j = (d_{j1}, d_{j2}, \cdots d_{jn})$. 因此，对多分辨率下细节信息的小波系数分析可以帮助识别噪声数据. 对此给出多分辨分解噪声识别方法：在特征空间上应用小波变换，计算各尺度高频信号 d_j，根据 d_j 的最小值消除孤立点。然后需要选定阈值 t_n. 将 $d_{j1}, d_{j2}, \cdots d_{jn}$ 按 $|d_j|$ 的降幂排列成 $|d_{j1}'| \geqslant d_{j2}'| \geqslant \ldots \geqslant |d_{jn}'|$，设 $m = m = \dfrac{\sum_{t=1}^{k} \left| d_{jt}' \right|^2}{\sum_{t=1}^{N} \left| d_{jt}' \right|}$，选取满足条件 $m > w$ 的最小 m 值，则 $t_n = d_{jk} - 1$. 其中，对于数字数据信号而言，w 通常为 60%.，从而形成新的单元 T_k，从单元 T_k 所组成的集合中根据 $d_{jk} \leqslant t_n$ 来划分聚类。考察小波变换的每一个不同分辨率 r，将会有一组聚类 C_r。

2.2.3 给单元分配标签

对每个聚类 C，$C \in C$ 都有一个聚类序号 C_p。算法第三步中，给特征空间中的每个单元标上它所在聚类的序号，也即：对任意的 C 和 T_k，$T_k \in c \mid \rightarrow ft_k = C_p, c \in C_r$，此处 ft_k 即为单元 T_k 的标号。我们所找到的聚类都是在以小波系数为基础的特征空间中找到的，不能直接用来定义原始数据集中的聚类。

2.2.4 建立查寻表，将对象映射到聚类

小波聚类算法生成了一个查找表 LT 来给出量化空间中的单元和原始数据集中的数据之间的映射关系，这样从 LT 表可以容易地确定出原始数据集中的数据的聚类标识，这样聚类就确定下来。

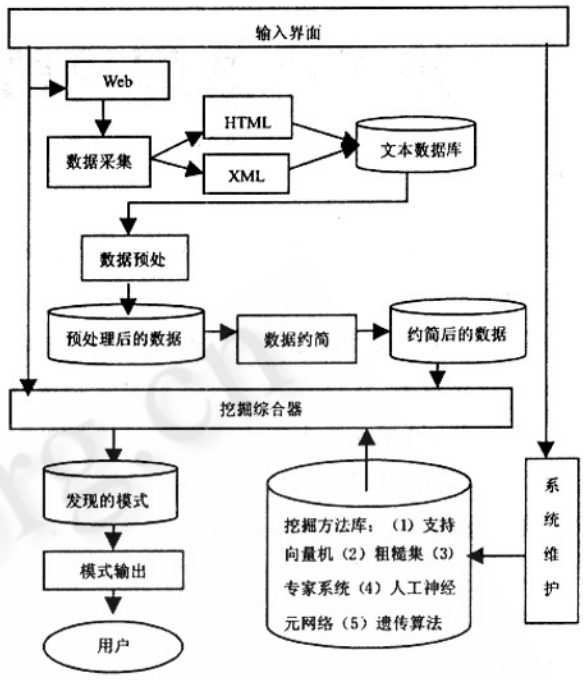

图 1　web 挖掘模型

3　Web 挖掘模型

Web 挖掘模型由数据采集器、预处理器、数据约简、挖掘综合器、挖掘方法库和系统维护六部分组成。

（1）数据采集器。按用户要求从网上采集数据，并将其存入文本数据库中。

（2）预处理器。将文本数据库中的 HTML 文档和 XML 文档转换成数据挖掘算法的目标格式，它包括几个方面，如数据标准化、处理缺省数据值和数值的离散化等。

（3）数据约简。去除一些冗余数据和无意义数据，本模型采用小波聚类分析进行数据约简与去噪。

（4）挖掘综合器。挖掘综合器是一个挖掘驱动引擎。它根据挖掘要求和挖掘方法来选择策略，从挖掘方法库中选择合适的挖掘算法，并使用该方法去执行挖掘任务。

（5）挖掘方法库和维护系统。挖掘方法库存放着各种挖掘方法，如支持向量机、粗糙集、专家系统、人工神经元网络、遗传算法、等算法和解决方案；并且维护系统还可以提供给用户一个增加新方法的接口。

（6）用户输出与评估界面。将挖掘结果以直观友

好的方式提交给用户，用户对挖掘结果进行评估，如果满意则挖掘过程结束，如果用户对挖掘结果不满意，则可以重提挖掘要求，再选择新的方法重新进行挖掘。

3.2 Web 挖掘模型的实现

作为一个系统，各个元素之间是相互关联协同工作的。用户首先通过用户输入界面输入自己的挖掘要求，包括挖掘哪类网页，希望进行哪些挖掘操作等。数据采集器根据用户要求搜集网页并存入文本数据库中，文本数据库存储着网页的内容和地址等信息。然后预处理器将文本数据库中的内容取出，按照挖掘需要将数据组织成数据挖掘算法的目标格式。

运用小波分析对数据进行约简，是从信号处理的角度来观察多维数据空间. 在多维数据空间中选取数据构成一个 n 维信号，其中最关键的是运用信号处理技术进行空间变换，在变换后的空间内寻找密集区域。

小波聚类方法是利用小波变换的多分辨率特性达到聚类的目的. 对于经过预处理的高维数据而言，由于维数较高而不能简单地使用小波分解的高频部分来进行挖掘. 在多维数据中，考虑到数据的稀疏性和量化空间中拥有大量的非空单元，我们采用如 2.2.1 节所述方法进行数据空间的量化，然后在特征空间中应用小波变换。. 这样在应用小波变换后，所有量化列表中的元素都将获得新值 $D_l - (d_{l1}, d_{l2} \cdots, d_{ln})$. 大多数情况下，大量的新非空单元都有小的计数值，这些值是由孤立点所引起，而不是实际的聚类所造成的. 由于卷积运算，实际簇的形状也会有所扭曲. 应用阈值 $m = \mathrm{m} = \dfrac{\sum_{t=1}^{k} \left| d'_{jt} \right|^2}{\sum_{t=1}^{N} \left| d'_{jt} \right|}$，选取满足条件 $m > w$ 的最小 m 值可有效地去除孤立点，并有助于保持簇的原形. 这样在应用小波变换所构造的单元 T_K 中，只有有意义的单元被保存，而噪声和孤立点被去除. 我们所找到的聚类都是在以小波系数为基础的特征空间中找到的，不能直接用来定义原始数据集中的聚类。小波聚类算法生成了一个查找表 LT 来给出量化空间中的单元和原始数据集中的数据之间的映射关系。

去除噪声数据后，挖掘综合器根据挖掘要求和挖掘方法来选择策略，从挖掘方法库中选择合适的挖掘算法，并使用该方法去执行挖掘任务。用户输出和评价界面将挖掘的结果包括关联规则、重要页面、趋势预测等呈现给用户。用户根据自己的满意程度，可以结束任务，或调整挖掘要求并进入新一轮挖掘。考虑到不断有新的挖掘方法出现，为了用户对系统升级方便，我们提供给用户一个维护接口，用户可以把新的方法加入到挖掘方法库中。

4 实例

该方法对某大学教育在线学生选课数据进行了分析处理. 在该学生选课关系中包含有学号、姓名、授课院系等 7 个属性，有 3 万多条记录. 经数据预处理后，将 7 个属性可转换成 5 个属性，组成 5 维空间. 在此系统中选用 *Mallat* 塔式小波变换. 在每个属性上采用聚类方法对其进行聚类分析，划分成不同的簇，便于挖掘有用信息。

1) 对原始数据空间进行预处理及量化，结果如表 1。其中 m_i 代表第 i 个学生的学号，例如在此例中 m_1 = 01240138，在这里有 30000 个学生所以 m = 30000. 这里有 5 个属性，$n = 5$，$\upsilon_i = (\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \upsilon_5)$. 其中 υ_1 代表学号，υ_2 代表性别，1，2 分别代表“男”、“女”；υ_3 代表专业，1，2，3，…，8 分别代表测量，土木，. . . ，环境；υ_4 代表课程，1，2，…，8 分别代表测量、土木、环境等专业所对应的课程；υ_5 代表教师职称，1，2，3 和 4 分别代表“教授”、“副教授”、“讲师”和“助教”见表 1。

表 1　预处理及量化结果

V_1	V_2	V_3	V_4	V_5
01240138	1	1	1	3
99350419	1	2	2	2
02360534	2	8	3	3
…	…	…	…	…
00170111	2	5	3	3

2) 在特征空间上应用小波变换，计算各尺度高频

信号 d_j，根据 d_j 的最小值消除孤立点. 以属性 v_5 为例，d_j 的计算结果见表 2。

表 2　分类结果

V_1	d_{j1}	d_{j2}	d_j3	C_r
…	…	…	…	…
99350419	2.16500	1.307177	0.641837	4
02360534	-1.288520	-1.439760		2
00350533	1.101780	1.272668	0.452690	4
00320531	-1.812100			1
99350533	0.557768	-0.530990	-0.601430	3
…	…	…	…	…

3) 选定阈值 t_n. 将 $d_{j1}, d_{j2}, \cdots d_{jn}$ 按 $|d_j|$ 的降幂排列成 $|d_{j1}'| \geq d_{j2}'| \geq \ldots \geq |d_{jn}'|$，设 $m = \frac{\sum_{t=1}^{k}|d'_{jt}|^2}{\sum_{t=1}^{N}|d'_{jt}|}$，选取满足条件 $m > w$ 的最小 m 值，则 $t_n - d_{ji} - 1$. 其中，对于数字数据信号而言，w 通常为 60%. 属性的阈值分别为 -1.684 89，-1.254 33 和 0.

4) 簇的划分. 根据来划分，小于的属于一个簇. 结果属性 v_5 被分成 4 类，如表 2 所示. 同理类推，应用此方法可成功地将上述 v_2, v_3, v_4 三个属性分别划分成 2,8,8 个类. 结果如图 2 所示。

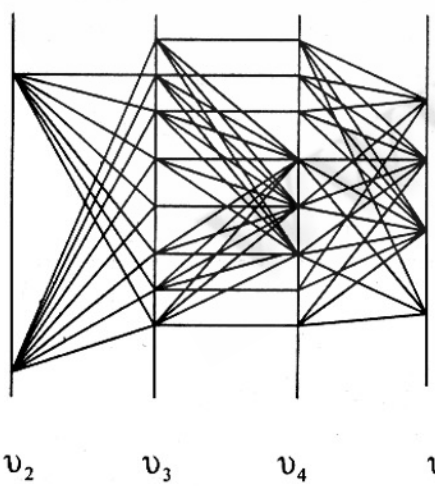

图 2　选课结果

由于聚类算法的时间复杂度为 O(N)，遍历数据集的次数为 1 次，因此在多维数据集中，因函数被频繁地使用而要耗费时间。此时，通过去除冗于和无意义的点后，整个算法的时间复杂度为 O(Ndlgn)。

聚类分析的结果并不是最终目的. 而是通过聚类分析，将数据划分为若干类，然后在每一类中寻找模式或各种潜在的有用信息. 这时，挖掘综合器根据挖掘要求和挖掘方法来选择策略，从挖掘方法库中选择合适的挖掘算法去执行挖掘任务。

5　结论

针对互联网上的海量数据提取出其中隐含的信息是一项非常有意义的工作。本文提出了利用小波聚类分析对经过预处理的 web 数据进行约简，并结合 web 挖掘技术，构造一个基于小波理论的 web 挖掘模型。实验证明，其中小波聚类分析方法具有较高的效率和较好的结果。基于小波理论的 web 挖掘技术还有很多值得深入研究和开发的领域。

参考文献

[1] 高雷. 任慧玉. 基于小波分析的上证综指预测. 统计与决策，2006，(7)：116—117.

[2] 侯守国，张世英. 基于小波分析的股市高频互相关研究. 中国管理科学，2006，14(3)：1－6.

[3] 周小勇，叶银忠. 基于 Mallat 塔式算法小波变换的多故障诊断方法. 控制与决策，2004，19(5)：592－594.

[4] 王亮. 基于小波分析的空间数据挖掘方法的研究. 合肥：硕士毕业论文，安徽大学，2006.

[5] 段利国，李爱萍，曹啸. 基于小波聚类的数据集简化算法研究. 太原理工大学学报，2006，37(5)：532—535.

[5] 邵良杉. 基于 Web 挖掘的虚拟企业合作伙伴选择决策支持系统研究. 计算机系统应用，2006，15(10)：2－5.