第17卷 第11期
2007年11月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.17 No.11
Nov. 2007

基于模式—区别方法聚类结构化的Deep Web源

陈 娟,王 贤,黄青松
(昆明理工大学 信息工程与自动化学院,云南 昆明 650051)

摘 要:近几年,网络被在线数据库迅速深化。在深网中,大量的资料提供了丰富的数据模式。这些模式详细说明了它们的目标领域和查询性能。因此对大规模数据的整合是当前面临的挑战。在数据挖掘中聚类分析是一个重要方法,为了发现通过这种统计分布管理的聚类,提出了一个新的目标函数:模型-区别(model-differentiation)。实验显示对于聚类Web查询模式,凝聚的层次聚类能正确地组织资料,区别模型函数胜过现有的凝聚的层次聚类。

关键词:数据整合;深网;分层凝聚聚类
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)11-0107-03

Clustering Structured Deep Web Sources: a Schema-Based Model-Differentiation Approach

CHEN Juan, WANG Xian, HUANG Qing-song
(Sch. of Info. Eng. and Automation, Kunming Univ. of Sci. and Techn., Kunming 650051, China)

Abstract: In the recent years, the Web has been rapidly deepened with the databases online. On this deep Web, numerous sources are structured, providing schema-rich data. Their schemas define the object domain and its query capabilities. The structured deep Web thus presents challenges for large-scale information integration. Clustering is one of the important approaches in data mining. To find clusters governed by such statistical distributions, propose a novel objective function: model-differentiation. Our evaluation shows that, on clustering the Web query schemas, the model-differentiation function outperforms existing ones with the hierarchical agglomerative clustering algorithm.

Key words: data integration; deep Web; hierarchical agglomerative clustering

0 引 言

在"深网"中,大量在线数据库通过它们的查询接口提供基于动态查询的数据[1]。信息数量是无限的,用户如何从大量的数据库中找到需要的资料并查询它们,因此对大规模的数据的整合是当前面临的挑战。

通过对深网的研究发现两个特征[1,2]:第一,查询模式(i.e,查询接口属性)对结构资料有很强的判别力。把查询模式作为一种明确的数据,可以把组织资料的问题抽象为对明确数据的聚类。此外,一些属性只在一个领域被观察到,这些主要特点(属性)使它们的领域更好识别。第二,一个领域资料的增长的同时,聚合的模式词汇的增长率却显著下降,趋于集中在一个小范围内。第一点建议在组织资料上把查询模式作为资料的代表来源,本质上是一个聚类问题。目标是聚类Web源到它们的层次领域。把查询模式看作一类特别的明确数据,组织资料的问题就抽象为对明确数据的聚类。第二点可使猜测每个领域都有隐藏的模型存在,可以产生查询模式。

文中的目的是:(1)通过与现有的用文本联接来比较评价MD目标的性能;(2)评价基于模式的聚类对组织结构资料到层次领域的效率。

1 MD-Based聚类

为了认识关于查询模式基于模型的聚类,设计了一个多项分布模型,并且把基于假设统计测试的模型-区别作为新的聚类目标函数。按照这个目标函数,采用普通的χ^2测试。把它应用在基于生成模型的明确数据上。既然采用一个层次聚类方法,使用HAC(凝聚的层次聚类)算法,它需要在两个聚类间测量相似度。于是从MD目标函数得到一个新的相似度量。

1.1 假设建模

首先,定义模式聚类工作的模型,需要描述什么是

收稿日期:2007-01-26
基金项目:云南省自然科学基金资助项目(Z2005-1-53004)
作者简介:陈 娟(1982-),女,山东临沂人,硕士研究生,研究方向为智能信息系统;黄青松,教授,研究方向为智能信息系统。

模式。认为一个查询模式是一个查询接口的一组属性，采用的是可归还选样(一个模式中属性可以重复选择)。通过这种方法，从一些聚类 C 中产生一个模式 Q。在 C 后的模型 M 是一个有着参数 $p_1,\cdots,p_N$ 的多项模型。多项模型 M 由一组概率为 $p_1,\cdots,p_N$ 的 N 个相互独立的事件(实际上是属性) $A_1,\cdots,A_N$(包含从 C 中发现的所有属性) 组成，$\sum_{j=1}^{N} p_j = 1$。把 M 表示为 $M = \{A_1:p_1,\cdots,A_N:p_N\}$，每个 M 从 N 个事件中的一个产生。从 M 中产生属性 A 的概率为：

$$P_r(A \mid M) = \begin{cases} p_i & \exists\, i: A = A_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

下一步讨论从聚类 C 中产生一个模式。在这个多项模型内，用观察到的属性及其出现次数来表示一个模式 Q，即为 $Q = \{A_1:y_1,\cdots,A_N:y_k\}$，$\sum_{i=1}^{N} y_i = n$，$y_i$ 为 Q 中属性 A_i 的出现次数，且 y_i 不是 0 就是 1。通过定义标准多项式，从 M 中产生的 Q(长度为 n) 的概率为：

$$P_r(Q \mid M,n) = n!\prod_{i=1}^{N} \frac{P_r(A_i \mid M)^{y_i}}{y_i!} \tag{2}$$

然后，观察一个模式聚类，认为一个模式聚类 $C = \{Q_1,\cdots,Q_m\}$，每个模式 Q_j(长度为 n_j) 产生于同一个模型 $M = \{A_1:p_1,\cdots,A_N:p_N\}$。$Q_j = \{A_1:y_{j1},\cdots,A_N:y_{jN}\}$，$y_{ji}$ 表示 A_i 的出现次数。对于整个 C，有 $z_i = \sum_{j=1}^{m} y_{ji}$，$z_i$ 为一个属性出现次数的总和。因此 C 表示为 $C = \{A_1:z_1,\cdots,A_N:z_N\}$。

1.2 模型 - 区别：一个新的目标函数

聚类必须有一些详细说明理想聚类性质的目标函数作指导。基本的聚类思想是聚集相似的数据和分离不同的数据。对于基于模型的聚类，相似的数据可能产生于相同潜在的模型，不同的数据来自不同的模型。因而，当潜在的模型更易区别时，会得到较好的聚类结果[3]。因此，定义聚类的目标函数，聚类的目标函数 H 为在划分 P 下的异质模型的特性，表示为 $H(X;P)$。聚类的目的是划分 P 取函数 H 的最大值，i.e, $\arg\max_P H(X;P)$。在统计学中，如果有一个划分函数 P 划分 X 为 $C_k(1 \leqslant k \leqslant G)$ 个聚类，可以用标准测试方法检验假设“$C_k(1 \leqslant k \leqslant G)$ 从同一分类中取样”。测试的结果用一个盖然论的变量 λ 来表明认可这个假设的信心。从而模型的异质是 1—λ。基于 MD 聚类是从关于聚类 G 的划分 P 的假设测试结果 $\lambda(C_1,\cdots,C_G)$ 中发现。

$$\begin{aligned} \arg\max_P H(X;P) &= \arg\max_P H(C_1,\cdots,C_G) \\ &= \arg\max_P \{1 - \lambda(C_1,\cdots,C_G)\} \\ &= \arg\max_P \lambda(C_1,\cdots,C_G) \end{aligned} \tag{3}$$

给定一个关于观测数据 X 的划分 P，应用 χ^2 假设测试来计算 $\lambda(C_1,\cdots,C_G)$。假设有 m 个聚类 $C_1,\cdots,C_m$，每一个聚类产生于它们自己的多项分类。有 n 个不同的属性，$A_1,\cdots,A_n$。图 1 为表示这组数据的可能性表。O_{ij} 表示在聚类 C_i 中属性 A_j 的出现次数。X_i 为在第 i 行所有的 O_{ij} 的总和，Y_j 为在第 j 列所有的 O_{ij} 的总和。即 $X_i = \sum_{j=1}^{n} O_{ij}$ 和 $Y_j = \sum_{i=1}^{m} O_{ij}$。$S$ 为表中所有 O_{ij} 的总和。因而 $S = \sum_{i=1}^{m} X_i = \sum_{j=1}^{n} Y_j$。

	A_1	A_2	A_3	…	A_n	sum
C_1	O_{11}	O_{12}	O_{13}	…	O_{1n}	X_1
C_2	O_{21}	O_{22}	O_{23}	…	O_{2n}	X_2
…	…	…	…	…	…	…
C_m	O_{m1}	O_{m2}	O_{m3}	…	O_{mn}	X_m
sum	Y_1	Y_2	Y_3	…	Y_n	S

图 1 实验的可能性表

需要测试这个假设：$\forall j, 1 \leq j \leq n, p_{j1} = p_{j2} = \cdots = p_{jm} = \frac{Y_j}{S}$，$p_{ji}$ 为聚类 C_i 中属性 A_j 的出现概率。这个假设通过任意变量来测试：

$$D^2(C_1,\cdots,C_m) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij} - X_i \times \frac{Y_j}{S})^2}{X_i \times \frac{Y_j}{S}}\right] \tag{4}$$

D^2 渐进于一个有着 $(n-1)(m-1)$ 自由度的 χ^2 分布，记为 df。

用 D^2 和 df 来判定这 m 个聚类的相似程度。在统计学上，可以用 P - value 来特定计算 D^2 和 df，表示为 $PV(D^2,\text{df})$。P - value 是方程(3) 中 λ 值的概率，表示接受 m 个聚类从相同的分类中产生的假设的信心。目标函数 H 为

$$H(C_1,\cdots,C_G) = 1 - PV(D^2,\text{df}) \tag{5}$$

1.3 普通的 HAC 算法和 MD - Based 相似度量

对于构造层次领域，采用在数据聚类中广泛使用的普通凝聚层次聚类方法[4]。图 2 阐明了普通凝聚层次聚类算法框架。在凝聚层次聚类中，需要度量聚类的相似度[5]。即对给定的一组聚类 $C_1,\cdots,C_v$，成对地计算所有的 $s(k,l)$ 值，s 是来自聚类目标函数的一个相似度函数。反复合并有最小的 $s(k,l)$ 的两个聚类，合并聚类 C_k 和 C_l 用 $C_{<k,l>}$ 表示。当只有 G 个聚类剩余时算法结束。普通的 HAC 算法的过程见图 2。

对基于 MD 的聚类，从 $H(X;P)$ 得到 $s(k,l)$ 如下所示：在每一个 HAC 的反复中，合并有最小 H 值的聚

类并定义为$s(k,l)$:

$$s(k,l) = H(C_k, C_l) \quad (6)$$

要求:模式集 X,目标函数 F,聚类数目 G
1:/* 初始的 V 个聚类 */
2:$C_k = X_k, (1 \leq k \leq V)$
3:/* 获得相似度 */
4:s = a similarity measure derived form F
5:/* 凝聚层次聚类算法的主要框架 */
6:for $K = V, V-1, \cdots, G$ do
7: /* 成对的计算相似度 */
8: $k^*, l^* = \arg\min_{k,s}(C_k, C_l), (1 \leq k \leq l \leq K)$
9: /* 合并最相似的两个聚类 */
10: $C_{<k^*,l^*>} = \text{Merge}(C_k^*, C_l^*)$
11:end for

图 2　普通凝聚层次聚类算法(HAC)

2 实　验

通过搜索引擎(Google)收集{汽车,书籍,手机,电影,游戏}这 5 个领域的资料,用选取的 312 个资料(如图1)一共包含 235 个属性来评价这种采用 MD-Based 相似度函数改进的凝聚层次聚类算法。对于每个资料,通过人工的萃取名词短语从它们的查询接口中抽取属性并判断其相应的领域。例如对于书籍,ISBN 是它的锚属性。采用有条件的平均信息量(conditional entropy)来测量聚类的结果。对于一个给定数量的聚类 G,有条件的平均信息量的值的范围在 0 到 log$G$ 之间,0 表示 100 的正确聚类,G 表示随机聚类的结果。因此,越靠值 0,结果越好。比较 MD-Based 方法和采用 HAC 算法的基于文本联接(ROCK)方法。

	汽车	书籍	手机	电影	游戏
C_1	96	0	4	0	0
C_2	0	34	0	3	2
C_3	0	0	101	0	0
C_4	0	0	0	1	31
C_5	1	2	0	37	0

(a)Conditional entropy of MD_{HAC}:0.35

	汽车	书籍	手机	电影	游戏
C_1	27	30		0	0
C_2	68	0	0	3	5
C_3	0	0	74	11	0
C_4	0	6	0	0	28
C_5	2	0	30	27	0

(b)Conditional entropyof CL_{hac}:0.62

图 3　HAC 不同方法的比较

首先,从 5 个领域比较 MD_{HAC}和基于文本联接(ROCK)方法(CL_{hac})聚类。图 3 表示比较的结果。结果显示:1)聚类结构资源看作是聚类查询模式是切实可行的;2)MD_{HAC}在对聚类 Web 模式上表现出比较好的性能。

第二,展示 MD-Based 相似度函数的凝聚层次聚类算法构造的层次领域。聚类 5 个领域后,再用同样的凝聚层次聚类方法来构造层次领域。结果如图 4 所示:汽车和手机合并在一个子树中,书籍、游戏和电影在另一个子树中。这个结果与观察到的真实情况一致:书籍、游戏和电影都是媒体产品并一般在一起销售。汽车和手机在一起是因为它们共享许多特定区域信息,如城市、省。

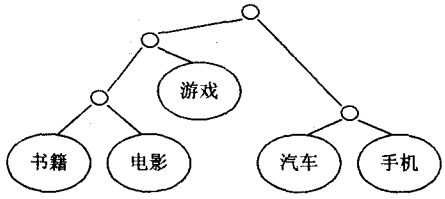

图 4　通过 MD_{HAC}构造的层次领域

3 结　论

研究对深网上的资料聚类的问题,进行大规模的整合这种资源是一项重要任务。通过对深网的观察,建议通过它们的查询接口来组织资料,进一步概要这个问题为对明确数据的聚类。对聚类应用了一个新的模型—区别目标函数。通过 MD 目标指导,得到一个适合于 HAC 算法的新的相似度量。通过与一些相关的已存在的技术比较实验显示模型—区别函数提取的效率。

参考文献:

[1] He B, Tao T, Chang K C C. Clustering structured web sources: A schema-based, model-differentiation approach[R]. Technical Report UIUCDCS-R-2003-2322, Dept. of Computer Science, UIUC, 2003.

[2] He B, Chang K C C. Statistical schema matching across web query interfaces[C]//Proceedings of the 22nd International Conference on Management of Data (SIGMOD). New York: ACM Press, 2003.

[3] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2005.

[4] 朱　明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2002.

[5] 邵峰晶,于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社,2003.