# 基于模型融合的分布式贝叶斯网络学习算法 *

缑葵香[a,b], 宫秀军[b], 冉 勇[b]
(天津大学 a. 理学院 数学系; b. 计算机科学与技术学院, 天津 300072)

**摘 要**: 提出了一个从同构数据集中学习贝叶斯网络结构的分布式算法。该算法首先使用搜索评分的方法学习每个局部贝叶斯网络结构,然后取节点对互信息变量和条件互信息变量的数学期望作为全局学习的评价标准,融合所有局部结构得到全局结构。由于只使用了数据集中变量间的互信息和条件互信息,没有直接获取局部个体数据信息,从而可以实现有效的隐私保护。该算法在 Alarm 数据集上进行测试,边的误差率小于 6%,运行时间比集中学习的算法的运行时间短,验证了算法的有效性。

**关键词**: 分布式数据挖掘; 隐私保护; 模型融合; 贝叶斯网络; 互信息
**中图分类号**: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0060-04
doi:10.3969/j.issn.1001-3695.2010.01.017

# Distributed Bayesian network learning algorithm based on model fusion

GOU Kui-xiang[a,b], GONG Xiu-jun[b], RAN Yong[b]
(*a. Dept. of Mathematics, School of Science, b. School of Computer Science & Technology, Tianjin University, Tianjin* 300072, *China*)

**Abstract**: For learning Bayesian network structure from homogeneous datasets, this paper proposed a distributed algorithm. It firstly learned each local structure using score method, then with the expectations of mutual information and conditional mutual information as evaluated criterion, it fused these local structures to obtain global structure. For using only mutual information and conditional mutual information of variation, without obtaining directly sample data, it might effectively protect privacy. Simulating the algorithm on Alarm dataset, the ratio of error is less than 6%, the running time of the algorithm is shorter than the running time of collective algorithms, the algotithm is valid.

**Key words**: distributed data mining; privacy-protecting; model fusion; Bayesian network; mutual information

## 0 引言

网络化时代数据呈现出爆炸式的增长态势,如何从这些结构上异质的、地理上分布的、内容上敏感的数据资源获取有用的知识,成为当前数据挖掘面临的重要挑战之一。挑战主要来自两个方面:a)数据挖掘算法需要快速,甚至实时地处理这些堆积如山的数据;b)数据挖掘过程需要对内容上敏感的隐私、安全方面的数据提供安全可靠的保护。基于上述两个方面的不同关注,出现了数据挖掘的两个重要分支,即分布式数据挖掘(distributed data mining, DDM)和保护隐私的数据挖掘(privacy-preserving data mining, PPDM)。DDM 强调如何利用分布的数据和计算资源及网络带宽等实现快速而高效的挖掘[1];PPDM 强调如何通过加密、特征重构等手段来保护用户的隐私[2,3]。

模型融合的方法是从分布的资源中获取全局知识的主要途径之一。事实上,正如不存在对所有问题都适用的技术一样,也不存在对所有模式识别问题都适用的单一模型。只有通过多个模型的融合才能获得对事物的全局认识。模型融合的方法已经在诸多领域中获得应用,如信息检索[4]、传感器网络感知[5]和生物信息学[6]等,其理论和方法已成为智能信息处理及控制的一个重要研究方向。贝叶斯网络是不确定性知识建模的重要工具,从数据中学习贝叶斯网络结构和参数是数据挖掘研究的热点问题之一。

## 1 相关研究工作

贝叶斯网络是表示变量间概率关系的有向无环图 $B=(G,\Theta)$。$G=(V,E)$,$V=\{X_1,\cdots,X_n\}$ 是一组节点的集合,每个节点表示一个领域变量;$E$ 是一组有向边的集合,每条边表示变量间的概率关系,$E=\{\langle X_i,X_k\rangle \mid X_i,X_k\in V, i\neq k\}$。其中$\langle X_i, X_k\rangle$表示有向边 $X_i\rightarrow X_k$,称 $X_i$ 是 $X_k$ 的父节点,本文用 $Pa(X_k)$ 表示 $X_k$ 的父节点的集合。$\Theta$ 为 $V$ 中节点对应的(conditional probability table, CPT)的集合,每个节点对应的 CPT 表明了该变量与其父节点之间概率依赖的数量关系。

贝叶斯网络学习分为结构学习和参数学习,结构学习就是从数据集中学习有向无环图 $G$,参数学习是基于网络结构从数据中学习 $\Theta$。本文主要研究贝叶斯网络的结构学习。目前贝叶斯网络结构学习算法大体分为搜索评分的方法和相关性分析的方法两类。搜索评分的方法是定义一个打分函数,选定一个初始的网络模型,然后使用启发式搜索方法为网络加边,用打分函数对网络模型每一个可能的变化进行评估,取分数最优的网络模型为最终模型。常用的搜索评分方法有 K2 算法[7]、基于最小描述长度(minimal description length, MDL)的算法[8]

**收稿日期**: 2009-05-03; **修回日期**: 2009-06-28 **基金项目**: 天津市应用基础及前沿技术研究计划重点项目(07JCZDJC06700)
**作者简介**:缑葵香(1979-),女,河北衡水人,博士研究生,主要研究方向为生物信息学、分布式数据挖掘(gkxiang@tju.edu.cn);宫秀军(1972-),男,内蒙古赤峰人,副教授,博士后,主要研究方向为生物信息学、数据挖掘等.

等 。基于相关性分析的方法是通过使用条件独立性(conditional independence,CI)测试找到网络的依赖结构,CI 测试的方法一般有 $\chi^2$ 测试、互信息测试[9]等。

由于贝叶斯网络学习时间上的复杂性以及某些具体领域隐私保护等实际需求,使得分布式贝叶斯网络的学习受到越来越多的关注,相关研究有:Sterritt 等人[10]使用并行遗传算法来学习贝叶斯网络结构。文中针对复杂的大规模电信数据,提出一种并行因果遗传算法 P-CAEGA 来学习贝叶斯网络,并在大型并行虚拟机的局域网上实现了一个原型系统,该方法充分利用了遗传算法内在的并行性。Lam 等人[11]提出了利用异步分布搜索 nagging 方法将基于 MDL 评分的算法并行化来学习贝叶斯网络。Chen 等人[12]提出了一种基于收集(collective)的方法从分布的异构数据中学习贝叶斯网络。该方法首先在每个局部计算机节点构造一个局部贝叶斯网络,并标志出重要性的观测实例,传输到中心计算机节点;利用传输到中心节点的观测实例再构造一个中心的贝叶斯网络,将局部网络和中心网络相结合构成最终的贝叶斯网络。

以上算法或者从结构学习过程并行化的角度或者从传输部分数据的角度来实现网络结构的分布式学习,但缺乏隐私保护方面的考虑。MA Jian-jie[2] 和 YANG Zhi-qiang[3] 等人分别给出了从分布式异构数据中安全学习贝叶斯网络结构的方法,它们都是通过加密算法对局部数据进行加密后来集中学习贝叶斯网络,没有利用学习过程中的分布特性。

本文从模型融合的角度提出了一种从同构(不同数据集中含有完全相同的属性)数据集中学习贝叶斯网络结构的分布式算法(model fusion homogeneous Bayesian network structure,MFHomoBNS)。主要贡献包括:a)提出了一种基于相关性分析的模型融合方法,使得在信息损失最小的情况下有效保护数据的隐私;b)算法将学习贝叶斯网络的搜索评分的方法和相关性分析方法相结合,充分利用了分布数据集包含的信息。

## 2　基于模型融合的分布式贝叶斯网络结构学习算法

Cooper 等人[7]提出的 K2 算法是一个经典的基于搜索评分的算法。该算法在给定节点顺序这一先验信息的情况下,利用贝叶斯概率作为评分标准来评价模型与数据的符合程度,通过不断向网络中增加能提高评价指针的边的贪婪搜索方法来找出最佳网络结构。对于样本较小的数据集,K2 算法可以学习到较好的网络结构。基于相关性分析方法构建贝叶斯网络的算法以 Cheng Jie[9] 提出的基于互信息的 CI 测试的算法最具代表性。这是一种基于定量互信息检验的网络结构算法,算法假设将两节点 $A$、$B$ 间的非碰撞节点 $C$ 放入条件集中会减少 $A$、$B$ 节点间的互信息量,而将碰撞节点 $D$ 放入则会增加 $A$、$B$ 间的互信息量。通过这样的启发式评价函数来找到相应节点间的割集,进而确定节点间是否存在边。对于稀疏网络和具有较大样本数据集的系统来说,这种方法是非常有效的。

MFHomoBNS 算法是从同构数据集中学习贝叶斯网络结构,各个数据集首先并行地学习各自的网络结构,这个学习过程称为局部学习。在局部学习中,由于每个数据集中包含的样本数量较小,采用基于搜索评分的 K2 算法来学习;然后将所有的局部结构融合成全局的结构,这个过程称为全局学习。由于将所有的数据集合并起来得到的数据集中包含大量样本,在全局学习中,使用基于相关性分析的方法来融合局部结构。

假设有 $s$ 个独立的同构数据集 $D_i$($i=1,2,\cdots,s$)分布在 $s$ 个计算机节点上,MFHomoBNS 算法包括以下两个过程:

a)局部学习。利用 K2 算法在计算机节点 $i$ 上获得局部贝叶斯网络结构 $G_i=(V,E_i)$,$i=1,2,\cdots,s$。

b)全局学习。融合 $s$ 个局部结构得到全局结构 $G=(V,E)$。将 $E_a=E_1\cap E_2\cap\cdots\cap E_s$ 作为全局结构的初始边集,将 $E_b=E_1\cup\cdots\cup E_s-E_a$ 作全局结构的待定边集。在该过程中,基于相关性分析的方法讨论待定边集 $E_b$ 中的哪些边可以加入到全局结构的边集中。

### 2.1　局部学习算法

在局部学习中,利用 K2[9] 算法在 $s$ 个计算机节点上同步学习,得到 $s$ 个贝叶斯网络结构 $G_1=(V,E_1),\cdots,G_s=(V,E_s)$。

K2 算法需要指定节点的输入次序和父节点的上界 $\mu$,对于不同的数据集,输入相同的节点次序和相同的 $\mu$。给定数据集 $D$,它包含 $N$ 个变量,$N$ 个变量对应贝叶斯网络的 $N$ 个节点,用 K2 算法找到与数据集 $D$ 最佳匹配的贝叶斯网络结构 $B$。

设 $X_i$ 为 $D$ 的一个变量,它有 $R_i$ 个可能取值,$Pa(X_i)$ 有 $T_i$ 个可能取值,$\beta_{ijk}$ 为数据集 $D$ 中 $X_i=k$,$Pa(X_i)=j$ 的记录条数,定义 $N_{ij}=\sum_{k=1}^{R_i}\beta_{ijk}$ 为构建贝叶斯网络结构,使用的打分函数为

$$S(X_i,Pa(X_i))=\prod_{j=1}^{T_i}\frac{(R_i-1)!}{(N_{ij}+R_i-1)!}\prod_{k=1}^{R_i}\beta_{ijk}! \tag{1}$$

### 2.2　全局学习算法

全局学习中,基于互信息检验方法融合局部模型得到全局模型。互信息和条件互信息是两个表示变量间相关程度的量。

**定义 1**　两个变量 $X$ 和 $Y$ 的互信息为

$$I(X,Y)=\sum_{x,y}P(x,y)\ \log\frac{P(x,y)}{P(x)P(y)} \tag{2}$$

**定义 2**　给定变量集 $C$ 中,两个变量 $X$ 和 $Y$ 的条件互信息为

$$I(X,Y|C)=\sum_{x,y,c}P(x,y,c)\ \log\frac{P(x,y|c)}{P(x|c)P(y|c)} \tag{3}$$

对于给定阈值 $\varepsilon$,若 $I(X,Y)<\varepsilon$ 时,称 $X$ 独立于 $Y$;当 $I(X,Y)>\varepsilon$ 时,称 $X$ 与 $Y$ 相关;若 $I(X,Y|C)<\varepsilon$,称 $X$ 与 $Y$ 在条件 $C$ 下独立;若 $I(X,Y|C)>\varepsilon$,称 $X$ 与 $Y$ 在条件 $C$ 下相关。

**定义 3**　使两个节点 $X$ 与 $Y$ 相连接的一组有向边的集合称为连接 $X$ 与 $Y$ 的一条路径。

**定义 4**　在有向无环图 $G=(V,E)$,$X\in V$,$Y\in V$,$C\subset V$。在 $G$ 中如果去掉 $C$ 中的所有节点以及与这些节点相连的边,使得连接 $X$ 和 $Y$ 之间的所有路径中断,则称 $C$ 为 $X$ 与 $Y$ 的割集。如果一个割集任意去掉一个节点就不再是割集,这样的割集称为 $X$ 与 $Y$ 的最小割集。

在全局学习中,本文融合 $s$ 个局部结构 $G_1=(V,E_1),\cdots,G_s=(V,E_s)$ 得到全局结构 $G=(V,E)$。构造初始网络结构 $G=(V,E)$,其中 $E=E_a=E_1\cap\cdots\cap E_s$,称 $E_b=E_1\cup\cdots\cup E_s-E_a$ 为待定边集。全局学习的任务是讨论 $E_b$ 中的哪些边可以添加到 $E$ 中。

对于 $E_b$ 中边$\langle X_i,X_j\rangle$对应的节点对$(X_i,X_j)$,本文把 $X_i$ 和 $X_j$ 的互信息 $I(X_i,X_j)$ 和条件互信息 $I(X_i,X_j|C)$ 分别看做随机变量 $U_{(X_i,X_j)}$ 和 $U_{(X_i,X_j|C)}$,这些随机变量的取值为每个局部网络

结构中节点对$(X_i,X_j)$的互信息值和条件互信息值。由于一个随机变量的数学期望反映了该变量总体大小特征,本文取节点对$(X_i,X_j)$的互信息变量$U_{(X_i,X_j)}$和条件互信息变量$U_{(X_i,X_j|C)}$的数学期望作为全局学习的评价标准来衡量两个变量的整体相关性。

在实际问题中,对于取值为实数的离散随机变量,其总体的数学期望就是其算数平均值:$E(U_{(X_i,X_j)})=\sum_{t=1}^{s} I_t(X_i,X_j)$,$E(U_{(X_i,X_j|C)})=\sum_{t=1}^{s} I_t(X_i,X_j|C)$,这两个数学期望需要 $s$ 个计算机节点联合计算。对于给定阈值 $\varepsilon$,如果 $E(U_{(X_i,X_j)})>\varepsilon$,本文称 $X_i$ 和 $X_j$ 是整体相关的;否则,称 $X_i$ 和 $X_j$ 是整体无关的。如果 $E(U_{(X_i,X_j|C)})>\varepsilon$,称 $X_i$ 和 $X_j$ 在条件 $C$ 下整体相关;否则,称 $X_i$ 和 $X_j$ 在条件 $C$ 下整体无关。

全局学习算法描述如下:

a)构造初始网络 $G=(V,E)$。其中:$V=\{X_1,\cdots,X_n\}$,$E=E_a=E_1\cap E_2\cap\cdots\cap E_s$。

b)$E_b=E_1\cup\cdots\cup E_s-E_a$。对于 $E_b$ 中每条边$\langle X_i,X_j\rangle$对应节点对 $X_i$ 和 $X_j$,各计算机节点联合计算 $X_i$ 和 $X_j$ 的互信息变量的数学期望 $E(U_{(X_i,X_j)})$。如果 $E(U_{(X_i,X_j)})>\varepsilon$,判断当前网络结构 $G$ 中是否存在连接 $X_i$ 和 $X_j$ 的路,如果不存在路,则添加边$\langle X_i,X_j\rangle$到 $E$ 中;如果存在路,则把 $X_i$ 和 $X_j$ 这对节点加入到一个空队列 $L$ 中。

c)对于 $L$ 中的每一对节点 $X_i$ 和 $X_j$,在当前网络 $G$ 中寻找它们的所有最小割集,设它们为 $C_t,t=1,2,\cdots,m$。各计算机节点联合计算 $X_i$ 和 $X_j$ 在这些最小割集下的条件互信息变量的数学期望 $E(U_{(X_i,X_j|C_t)})$。$E(U_{(X_i,X_j|C)})=\min_{1\leqslant k\leqslant m}\{E(U_{(X_i,X_j|C_k)})\}$,如果 $E(U_{(X_i,X_j|C)})>\varepsilon$,则添加边$\langle X_i,X_j\rangle$到 $E$ 中。最后,本文得到全局贝叶斯网络 $G=(V,E)$。全局学习算法程序的伪代码如下:

```
input: datasets D1,D2, …,Ds threshold ε
output: graph structure G = (V,E)
begin
1   V = (X1,X2,…,Xn); E = ϕ;
2   for i = 1 to s do Ei = localBayesian(Di);
3   Ea = ∩Ei; Eb = ∪Ei - Ea;
4   E = Ea; L = ϕ; EI = 1;
5   for each ⟨Xi,Xj⟩ ∈ Eb do
6     for t = 1 to s do
7       MIi = calcuMI(Xi,Xj);
8       EI = calcuExpection(MIi,s);
9       if EI > ε and hasPath(Xi,Xj) in E
10          then L = L + (Xi,Xj);
11      if EI > ε and nothaspath(Xi,Xj)
12          then E = E + ⟨Xi,Xj⟩
13  C = ϕ;
14  for each (Xi,Xj) ∈ L do
15      C = find_MinimumCut_Set(Xi,Xj,E)
16      for i = 1 to s do
17          CMIi = calcuCMI(Xi,Xj|C);
18      CEI = maxC{calcuExpection(CMIi,s)};
19      if CEI > ε then E = E + ⟨Xi,Xj⟩;
20  return G = (V,E)
end
```

本文利用逐点置换法来求最小割集,用一个具体的例子来说明本文的算法。

例如要在图 1 中寻找节点对$(A,F)$的最小割集。$A$ 称为初始节点,$B$ 称为目标节点。

a)求出图 1 中节点的联络矩阵 $M$,矩阵中的行和列均对应图中的节点。若两节点间有有向边,值为 1;若无有向边,值为 0。

$$M=\begin{array}{c} \\ A\\B\\C\\D\\E\\F \end{array}\begin{array}{c}\begin{array}{cccccc} A & B & C & D & E & F \end{array}\\ \begin{bmatrix} 0&1&1&0&0&0\\0&0&0&1&1&0\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&0&0&0&0&1\\0&0&0&0&0&0 \end{bmatrix}\end{array}$$

b)本文从初始节点 $A$ 出发,与节点 $A$ 相连的节点构成最小割集$\{B,C\}$。

c)进行逐点置换,分别用与割集$\{B,C\}$中的节点相连的节点置换该节点,得到新的割集。如果出现目标节点则结束 置换。节点 $B$ 用 $D$、$E$ 置换,得到割集$\{D,E,C\}$;节点 $C$ 用 $E$ 置换,得到割集$\{B,E\}$;对于割集$\{D,E,C\}$,用 $E$ 置换 $C$ 得到割集$\{D,E\}$。以上共得到四个最小割集:$\{B,C\}$,$\{D,E,C\}$,$\{B,E\}$,$\{D,E\}$。

在 MFHomoBNS 算法的局部学习中,本文使用的是 K2 算法,它的时间复杂性为 $O(2^uN)$。其中 $N$ 为数据集中节点的数目,$\mu$为贝叶斯网络中父节点的上界。对于全局学习,通过实验,需要进行 CI 测试的节点数目小于 $0.2N$,全局学习的时间复杂性为 $O(0.2^2N^2)$,所以 MFHomoBNS 算法的时间复杂性为 $O(2^uN)+O(0.2^2N^2)$。

## 3　实验结果

以 Alarm 网络为例进行实验。Alarm 网是贝叶斯网络学习领域应用最广泛的基准测试集,Alarm 网有 37 个变量,每个变量有 2 ~4 个取值,它有 46 条有向边,产生实验数据的方法采用了文献[9]中的方法。利用 WEKA 软件中的 K2 算法进行贝叶斯网络学习,利用 Cheng Jie's Bayesian belief network software 实现 TPDA 算法的学习。

实验中分别产生了含有 10 000、20 000、30 000、50 000 条样本的四个数据集。在集中式学习下,分别用 K2 算法、TPDA 算法在每个数据集上进行测试。在分布式 MFHomoBNS 算法学习中,将含有 10 000 条记录的数据集随机平均分布在五个计算机节点上进行测试,对于其他三个数据集的测试方法相同。实验结果如表 1 所示。

表 1　Alarm 网上实验结果

| 样本数目 | 学习算法 | 集合状态 | 结果 M. A. | 结果 E. A. | 运行时间/s |
|---|---|---|---|---|---|
| 10 000 | K2 | 集中 | 1 | 3 | 25 |
| | TPDA | 集中 | 2 | 2 | 85 |
| | MFHomoBNS | 随机平均分布在 5 个节点 | 2 | 1 | 12 |
| 20 000 | K2 | 集中 | 3 | 2 | 53 |
| | TPDA | 集中 | 2 | 1 | 148 |
| | MFHomoBNS | 随机平均分布在 5 个节点 | 1 | 2 | 18 |
| 30 000 | K2 | 集中 | 3 | 3 | 115 |
| | TPDA | 集中 | 1 | 2 | 321 |
| | MFHomoBNS | 随机平均分布在 5 个节点 | 1 | 1 | 33 |
| 50 000 | K2 | 集中 | 4 | 3 | 278 |
| | TPDA | 集中 | 1 | 1 | 713 |
| | MFHomoBNS | 随机平均分布在 5 个节点 | 1 | 1 | 62 |

注:与正确的网络相比较,E. A. 表示添加边数,M. A. 表示丢失边数。

由表 1 可以看出,分布式 MFHomoBN 算法学习的结果优

于集中式学习结果,边误差率低于 6%,而且随着样本的增加,边的误差率越低,运行时间也比集中式学习算法短。

图 2 中下方的曲线是四个数据集分别在集中式学习下 K2 算法运行时间与分布式学习下 MFHomoBNS 算法运行时间比值的连线;上方的曲线是四个数据集分别在集中式学习下 TPDA 算法运行时间与分布式学习下 MFHomoBNS 算法运行时间比值的连线。从图 2 中可以看出,随着记录条数的增多,这两个比值呈近线性增长。

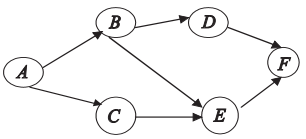


图1　一个有向无环图

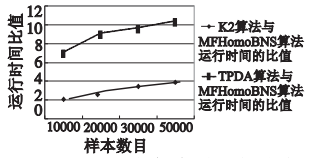


图2　Alarm网上集中学习与分布学习运行时间的比值

## 4　结束语

本文研究了如何从分布的同构数据中学习贝叶斯网络结构的问题。在 MFHomoBNS 算法中,将贝叶斯网络结构的搜索评分的方法和相关性分析的方法相结合来学习贝叶斯网络结构,充分利用了分布数据中所包含的信息。融合局部结构时,把贝叶斯网络中的节点对的互信息和条件互信息看做变量,它们的取值为各个局部数据集中节点对的互信息和条件互信息,取节点对的互信息变量和条件互信息变量的数学期望作为全局学习的评价标准,融合所有局部结构得到全局结构。这样只需要传递局部互信息和条件互信息到融合中心节点,而没有直接获取局部个体样本数据信息,从而有效地保护了个体的隐私。通过实验表明,MFHomoBNS 算法学习的结果优于集中式学习结果,边误差率低于 6%;而且随着样本的增加,边的误差率越低,运行时间也比集中式学习算法短,而且随着记录条数的增多,集中学习的运行时间与分布学习的运行时间比呈近线性增加。实验证明 MFHomoBNS 算法是有效的。

笔者下一步将研究计划如何运用模型融合的方法从异构分布的数据集中学习贝叶斯网络,如何从时序数据中基于动态贝叶斯网络重构基因调控网络,以及融合多个不同实验数据集中重构基因调控网络。

**参考文献:**


[1] ANTONIO C, DOMENICO T, PAOLO T. Distributed data mining services leveraging WSRF[J]. Future Generation Computer Systems,2007,23(1):34-41.

[2] MA Jian-jie, SIVAKUMAR K. Privacy-preserving Bayesian network learning from heterogeneous distributed data[C]//Proc of International Conference on Data Mining. 2006.

[3] YANG Zhi-qiang, WRIGHT R N. Privacy-preserving computation of Bayesian networks on vertically partitioned data[J]. IEEE Trans on Knowledge and Data Engineering, 2006,18(9):1-12.

[4] WANG Xuan-hui, SUN Jian-tao, CHEN Zheng,*et al*. Latent semantic analysis for multiple-type interrelated data objects[C]//Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006:236-243.

[5] NAKAMURA E F, LOUREIRO A A F, FRERY A C. Information fusion for wireless sensor networks: methods, models, and classification [J]. ACM Computing Surveys, 2007,39(3):1-9.

[6] GEVAERT O, VOOREN S van, De MOOR B. A framework for elucidating regulatory networks based on prior information and expression data[J]. Annals of the New York Academy of Sciences, 2007, 1115:240-248.

[7] COOPER G F,HERSKOVITS E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning,1992,9(4):309-347.

[8] SUZUKI J. Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique [C]//Proc of International Conference on Machine Learning. 1996.

[9] CHENG Jie, GREINER R, KEUY J,*et al*. Learning Bayesian networks from data: an information theory based approach[J]. Artificial Intelligence, 2002,137(1-2):43-90.

[10] STERRITT R, ADAMSON K, SHAPCOTT M, *et al*. Parallel data mining of Bayesian networks from telecommunications network data [M]//Proc of IPDPS Workshops on Parallel and Distributed Processing. London:Springer-Verlag, 2000:415-426.

[11] LAM W, SEGRE A M. Distributed learning algorithm for Bayesian inference networks[J]. IEEE Trans on Knowledge and Data Engineering, 2002,14(1):93-105.

[12] CHEN R, SIVAKUMAR K, KARGUPTA H. Collective mining of Bayesian networks from distributed heterogeneous data[J]. Knowledge and Information Systems, 2004,6(2):164-187.


(上接第 59 页)


[5] PONTE J, CROFT W B. A language modeling approach to information retrieval[C]//Proc of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval. 1998:275-281.

[6] SINGHAL A. Modern information retrieval: a brief overview[J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001,24(4):35-43.

[7] ROBERTSON S E, WALKER S, BEAULIEU M. Okapi at TREC-7: automatic Ad hoc, filtering, VLC and interactive track[C]//Proc of the 7th Text Retrieval Conference, NIST Special Publication 500-242. 1999:253-264.

[8] LAFFERTY J, ZHAI Cheng-xiang. Document language models, query models, and risk minimization for information retrieval[C]//Proc of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval. 2001:111-119.

[9] SANDERSON M. Retrieving with good sense[J]. Information Retrieval, 2000,2(1):49-69.

[10] SCHUTZE H, PEDERSEN J O. A cooccurrence-based thesaurus and two applications to information retrieval[J]. Information Processing and Management, 1997,33(3):307-318.

[11] VOORHEES E. Query expansion using lexical-semantic relations [C]//Proc of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval. 1994:61-69.

[12] CHIRITA P A, PAIU R, NEIJDL W. Using ODP metadata to personalize search[C]//Proc of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval. 2005:178-185.

[13] Google personalized search[EB/OL]. http://www.google.com/psearch.

[14] BAI Jing, NIE Jian-yun, BOUCHARD H, *et al*. Using query contexts in information retrieval[C]//Proc of the 30th CAM SIGIR. 2007:15-22.

[15] 黄连恩,张燕,李晓明. 互联网上信息报道的最早发布时间检测[J]. 计算机科学与探索,2009,3(1):5.

[16] ZHAI Cheng-xiang. Statistical language models for information retrieval a critical review[J]. Foundations and Trends in Information Retrieval, 2008,2(3):137-213.

[17] 彭波."网络信息体系结构"课程讲义[R/OL]. (2008). http://net.pku.edu.cn/~wbia/.