Software Engineering and Applications 软件工程与应用, 2019, 8(2), 65-71
Published Online April 2019 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2019.82008

Hans汉斯

Research on Users Behavior Similarity Based on Bayesian Network

Jiamei Ye

Mathematics and Statistical Institute of Jiangxi University of Finance and Economics, Nanchang Jiangxi
Email: 502016755@qq.com

Received: Mar. 24th, 2019; accepted: Apr. 8th, 2019; published: Apr. 15th, 2019

Abstract

With the rapid development of mobile devices and mobile services, mobile social networks are integrated into people's daily lives, and people are also generating a large amount of data here. The research on this huge data source is very meaningful and necessary. User similarity in social networks is an important research field in social media data analysis. It also plays a very important role in the research of product recommendation and social network user relationship evolution. The similarity between users depends not only on the network topology, but also on the degree of dependence between users. In order to achieve the similarity measure between users in social network data, this paper proposes a basis based on topology and probabilistic reasoning. The user similarity measurement method of social network is adopted, and Bayesian network is used as the framework of this uncertain knowledge discovery. A user similarity discovery method based on Bayesian network is proposed.

Keywords

User Behavior Similarity, Bayesian Network, DBLP Dataset

基于贝叶斯网络的用户行为相似性研究

叶佳美

江西财经大学统计学院，江西 南昌
Email: 502016755@qq.com

收稿日期：2019年3月24日；录用日期：2019年4月8日；发布日期：2019年4月15日

摘 要

随着移动设备和移动服务的高速发展，移动社交网络融入了人们的日常生活。每时每刻人们都在这里生

文章引用: 叶佳美. 基于贝叶斯网络的用户行为相似性研究[J]. 软件工程与应用, 2019, 8(2): 65-71.
DOI: 10.12677/sea.2019.82008

成大量的数据，而对于这个巨大的社交媒体数据源的研究是非常有意义和必要的。但在对社交网络的数据挖掘中，发现存在大量的不确定性，以社交网络中的最广为人知的推荐算法为例，如何利用已知的用户信息为该用户更为精准地推荐其感兴趣的信息，这其中就蕴藏着大量的不确定性，如何清楚地展示和度量用户相似性这种不确定性知识，在商品推荐和社交网络用户关系演化等研究中一直是艰巨的挑战。因此本文提出采用贝叶斯网络这一结合拓扑结构和概率推理的重要的概率图模型作为发现这种不确定知识的框架，并基于此提出了一种用户相似性发现方法。

关键词

用户行为相似性，贝叶斯网络，DBLP数据集

1. 前言

1.1. 研究背景

当今世界，信息和网络技术的快速发展改变了人们的生活方式，根据中国互联网络信息中心(CNNIC)发布的第 42 次《中国互联网络发展状况统计报告》，截至 2018 年 6 月 30 日，我国网民规模达 8.02 亿，互联网普及率为 57.7%。手机网民规模达 7.88 亿，网民通过手机接入互联网的比例高达 98.3% [1]。大量的用户聚集到移动社交网络中，人们在社交网络上表达观点、交友互动，每天都产生数亿计的信息，使其成为一个新兴的具有高度研究价值的数据源。

所以近年来人工智能、机器学习、可视化技术、统计学等各学科联立起来以期利用这个数据源，在数据挖掘与知识发现领域获得更多发展。在社交网络建模中，综合考虑用户在网站点击或搜索行为、历史记录、以及用户之间行为的相关性等多维属性信息，帮助用户在移动社交网络中排除干扰因素、实现精准推荐，并预测社交网络中用户的社交属性、准确检测出移动社交网络的社区结构，是我们希望通过研究社交网络所能实现的目标。但在实际生活中，社交网络中由于规模巨大的用户群体的存在，庞大的数据中存在大量的不确定性信息，这些也会在为用户进行定位和划分过程中产生或直接或间接的影响，因此找到一种能够较好地表达这种不确定性的模型也是目前研究迫在眉睫的任务。

基于此，本文提出利用贝叶斯网络分析和研究社交网络用户相似性。贝叶斯网络学习是贝叶斯统计的重要前沿研究方向，它是不确定性知识表达和推理的框架，可有效地描述随机变量或对象属性间的相关性和相互依赖。能够在机器学习、计算机系统等不同领域得到广泛应用，其根本原因就是贝叶斯网络是图模型与概率论相结合的产物。贝叶斯网络是用一个有向无环图来表示变量之间的依赖关系，用条件概率变量表示对其父节点的依赖关系，贝叶斯网络的每个节点都有一个对应的条件概率分布[2] [3]。所以它能较好地帮助反映用户行为中所蕴含的用户直接相似性及利用模型的推理能力对用户间接相似性进行计算。

1.2. 国内外研究现状

移动社交网络是一种在信息网络上由社会个体集合及个体之间的连接关系构成的社会性结构，包含关系结构、网络群体与网络信息 3 个要素[4]，因此国内外对移动社交网络的研究多集中于此三块。

而利用贝叶斯网络解决社交网络问题，目前更多的人主要集中于利用 Hadoop 软件中的 MapReduce

编程模型，因其对于大规模数据的处理更加地有效，因其软件本身即可帮助完成数据分块，对后续分布计算更加高效便捷。徐娟[5]等结合社交用户贝叶斯网的拓扑结构和概率推理提出了一种基于 MapReduce 的用户相似性度量方法。最后，利用此用户相似性度量方法发现社交网络数据中的用户相似性。李青[6]等利用 MapReduce 框架对海量广告数据进行处理，接着基于贝叶斯网构造广告关键词之间的相似模型，在接下来对存储在 HBase 上的大规模贝叶斯网进行概率推理，进而得到待预测广告的点击率。郭俊[7]等根据模块度思想，结合图论、网络性质及近似优化理论，提出了新的社区发现模型——“多社区选择模型”，并设计了新的模块度增量更新方法，算法首先计算出所有节点间的模块度增量，然后选取网络中所有具有最大模块度增量的社区进行合并，并将本文所提的模型分别在仿真复杂网络和真实复杂网络数据上，同多个算法进行了对比，验证了本文所提算法的准确性和高效性。

但目前Python编程语言的普及性及其实用性，致使本文考虑利用Python构建用户相似性贝叶斯网络。基于 Python 的贝叶斯网络的话，方志鹏[8]等提出基于贝叶斯网的新广告点击率预测方法，通过构建关键词贝叶斯网的图结构，获得与新广告关键词存在直接相似关系的关键词，根据关键词贝叶斯网近似推理算法，发现与新广告关键词存在间接相似关系的广告关键词，进而发现广告之间的相似关系，为新广告预测点击率。Yan 等人[9]发现了点击同样广告的用户在网络上具有类似的行为，对于行为定向，使用短期用户行为来代表用户的行为会比长期用户行为来代表更有效。

2. 社交网络结构基本特征

2.1. 社交网络的理论基础

社交网络是由图表示的一种异构多关系数据集。这种图通常非常大，节点对应对象，边对应表示对象间联系或相互作用的链接，节点和链接都有属性。对象可以具有类标号，链接可以是单向的并且不必是二元的。一般认为图中的节点是社交参与者，在社交网络中任何一个社会单元或实体都可以看作节点，而边是各个参与者之间的社会关系或者交互行为，这种关系既可以是朋友关系、亲戚关系等强社会关系，也可以是因分享信息、资源等而产生的互动关系。显然，从概念可以看出，社交网络并不仅局限于像这样的在线社交网络，节点及其关系可以是传统的人与人之间面对面的交流或信件交流，也可以是现代的电话或者邮件交互，这些节点间的交互也是社会学研究的一部分。数据挖掘领域对社交网络的大多数研究在观察节点的度，即与每个节点相关联的边数，节点对之间的距离，通常是最短路径长度度量。

2.2. 社交网络的组成元素

社交网络中活动的用户是网络的基本节点，而社交网络正是由一个个用户之间相互联系而编织起来的巨型网络。网络中又可以依据兴趣爱好和关系类别的不同分为不同的组、群或圈子等社团结构。通常，把社交网络的主要组成元素归结为行动者、群体、关系和内容四类。

2.3. 复杂网络节点相关度(Node-Relevance)

在一个网络中，两个节点之间的内容相关度可以看作是其所包含内容的相似程度。如果向量 i 和 j 分别为基于词出现的页面节点 i 和 j 的内容向量，则两个节点之间的相关度可以用简单的向量之间的夹角余弦来计算[10]：

$$R_{ij} = R_{ji} = \frac{X_i \cdot X_j}{|X_i| \cdot |X_j|}$$

显然，$0 \le R \le 1$，当两个节点的内容越相似，相关度越接近 1，反之趋向于 0。

2.4. 贝叶斯网络介绍

由于贝叶斯网络[11]只关联由某种因果依赖关系在概率上相关的节点，因此可以节省大量的计算。不需要存储所有可能的状态配置，需要存储和处理的是相关父节点集和子节点集(节点的族)之间所有可能的状态组合。这大大节省了储存空间和计算量。当然，对于今天的贝叶斯网络算法来说，有些模型还是太大了。但新的算法正在开发中，突破是有希望的。这是现代计算机科学研究的热点领域。贝叶斯网络被证明如此有用的第二个原因是它们的适应性如此强。你可以从很小的地方开始，对一个领域的知识有限，并且随着你获得新的知识而成长。此外，当我们去应用它们时，我们不需要完全了解您正在应用它的世界实例。我们可以尽可能多地利用现有的知识，而贝叶斯网络也会尽可能地利用现有的知识。

定义 2.1 贝叶斯网络定义

1) 存在一个变量集 $V=\{X_i\}, i=1,2,\cdots,n$，以及变量对应结点之间有向边的集合 E。

2) 每一个变量都取有限个离散值。

3) 由变量对应的结点和结点之间的有向边构成一个有向无环图 $G_B=(V,E)$。

4) 对于每个结点 X_i 和它的父结点集 Π_i，都对应一个条件概率分布表 $p(x_i|\pi_i,G)$，而且满足

$$p(x_1,\cdots,x_n)=\prod_{i=1}^{n}p(x_i|\pi_i,G)$$

其中 π_i 是 Π_i 的配置。满足以上四个条件的有向无环图称作贝叶斯网络。

2.5. 贝叶斯网络推理

在给定的贝叶斯网情形下，进行贝叶斯网推理，就是对要计算的联合概率分布进行分解，通过查询贝叶斯网的条件概率表，来计算回答查询变量的概率分布。贝叶斯网推理一类典型的问题就是计算后验概率问题，即已知某些节点变量的取值，来计算其他节点变量的后验概率分布的问题。从推理的精确程度来划分，贝叶斯网的推理可以分为精确推理和近似推理两类[12] [13]。

理论上精确推理方法可以解决任何贝叶斯网的推理问题，由于在实际问题中，多数的贝叶斯网络结构多属于多联通图网络，运用贝叶斯网络精确推理的复杂程度随着网络节点数的增加呈指数倍增长，为提高多变量下的复杂的贝叶斯网络的推理效率，相继出现许多近似推理方法。其中包括环信念传递算法、模型简化算法、基于搜索的算法等。随机抽样算法(random sampling algorithms)便是其中最重要的一种方法之一，它的基本思想是从某个概率分布随机抽样，生成一组样本，然后从样本出发近似估计要计算的量。随机抽样算法因算法所生成的样本性质不同分为两大类，一类是产生相互独立样本的重要性抽样(importance sampling)，另一类是产生相互关联样本的马尔可夫链蒙特卡洛(Markov chain Monte Carlo)算法，简称 MCMC 算法。本文基于构建好的贝叶斯网络的推理制，用以获取间接相似用户，故将精确推理和近似推理的一般过程说明如下。而 Gibbs sampling 是 MCMC 算法其中最基本的一种方法，所以下面就来介绍 Gibbs sampling 的后验概率计算。

3. 基于贝叶斯网络学习的用户相似度建模

3.1. 用户相似度贝叶斯网络模型的定义

$V=\{V_1,V_2,\cdots,V_m\}$、$K=\{K_1,K_2,\cdots,K_n\}$ 和 $T=\{T_1,T_2,\cdots,T_m\}$ 分别为用户集合(即该社交网络中的所有作者集合)、实物集合(即该社交网络中的论文标题集合)和与实物有交互关系的用户集合，其中 $T_i=\{T_{i1},T_{i2},\cdots,T_{il_i}\}$，$T_{ij}\in K, 1\le i\le n, 1\le j\le l_i$，$T_{ij}$ 表示属于用户 V_i 所著的某篇论文，用户 V_i 的所有文章以一个二元组 $A_i=\langle V_i,Q_i\rangle$，$Q_i=\{T_{ij}|i=1,2,\cdots,l_i\}$ 表示用户与论文之间的关系。

依据在定义 2.1 中我们给出一般性的贝叶斯网络的定义，此处给出用户相似度贝叶斯网络模型的定义的内容。

定义 3.1 USBN 用一个二元组 $G=(G_B,W)$ 表示，其中：

1) $G_B=(V,E)$ 为 USBN 的 DAG 结构，每个用户对应网络结构的一个节点，用户 $V=\{V_1,V_2,\cdots,V_m\}$ 为 G_B 的节点集，V_i 的取值为 1 或 0，表示用户 V_i 是否著有文章。E 为网络结构中的有向边集合，表示用户间的相似性，若用户节点之间存在相似性，则用有向边 $V_i \to V_i$ 表示，并称 V_j 是 V_i 的一个父节点，V_i 是 V_j 的一个子节点，V_i 的所有父节点集合为 $Pa(V_i)$。

2) $W=\left\{p\left(V_i\middle|Pa(V_i)\right)\middle|V_i\in V\right\}$ 为网络结构中条件概率分布的集合，由各结点相对应的 CPT 中的值构成，$p\left(V_i\middle|Pa(V_i)\right)$ 表示结点 V_i 在其父结点发生情况下的条件概率，用来表示 $Pa(V_i)$ 的值对 V_i 的值的影响。

3.2. 用户相似性贝叶斯网络模型的构建

从给定的用户集合构建用户相似性贝叶斯网络，就是要构建用户相似性贝叶斯网络的 DAG 结构和计算每个节点的条件概率参数表 CPT。由 2.3 节可知，USBN 构建的关键是构建其图结构，即有向无环图的构建，这也是 USBN 构建的最困难的地方。一旦图结构构建成功后，便可以基于最大似然估计方法，利用 USBN 的有向无环图结构相对容易地计算各个关键词节点的条件概率参数表。所以我们接下来讨论图结构的构建，这包括如下两方面问题：

1) 首先需要确定判断用户是否相似，则能够明确两个用户节点间是否存在边；

2) 当明确用户节点之间存在相似关系时，接下来就需要确定用户节点之间有向边的指向。

针对问题 1，判断两个用户间是否相似，对于任何两个用户，我们首先计算出与两个用户共同著作的文章数占他们各自所著全部文章数的比例，该比例越高，则表明这两个用户在社交网络中的行为相似性就越高。依据复杂网络中对于节点相似性的定义，当我们设定好阈值时，若该比值高于预先设定的阈值，则表示这两个用户是相似的，即在这两个关键词之间存在一条无向边。下面给出基于对过去所著作论文情况已知的用户相似度计算方法，用户 V_i 和 V_j 间的相似度为：

$$\mathrm{sim}(V_i,V_j)=N(Q_i\cap Q_j)/N(Q_i\cup Q_j) \tag{3.1}$$

其中 $Q_i\cap Q_j$ 表示用户 V_i 和 V_j 共同著作的文章，$N(Q_i\cap Q_j)$ 表示共同著作的文章数，同理 $N(Q_i\cup Q_j)$ 表示用户 V_i 和 V_j 著作的所有文章数，$N(Q_i\cap Q_j)$ 和 $N(Q_i\cup Q_j)$ 可基于 T_i 计算得出。假设相似性阈值为 δ，当 $\mathrm{sim}(V_i,V_j)>\delta$ 时，可认为用户节点 V_i 和 V_j 之间存在一条无向边。

针对问题 2，考虑任意两个存在无向边的用户节点，比较两个用户所著文章数分别占他们各自所著全部文章数的比例，我们可以通过比较这两个比例的大小来确定指向关系。

$L(V_i|V_j)$ 表示用户 V_j 对用户 V_i 的影响程度，$L(V_j|V_i)$ 表示用户 V_i 对用户 V_j 的影响程度，则：

$$L(V_i|V_j)=N(Q_i\cap Q_j)/N(Q_j),\quad L(V_j|V_i)=N(Q_i\cap Q_j)/N(Q_i) \tag{3.2}$$

其中 $N(Q_j)$ 表示用户 V_j 所著全部文章数，$N(Q_i)$ 表示用户 V_i 所著全部文章数。

若 $L(V_i|V_j)>L(V_j|V_i)$，则表示用户 V_j 对用户 V_i 的影响程度低于用户 V_i 对用户 V_j 的影响程度，则节点之间的边的指向为由 V_j 指向 V_i。

3.3. 基于概率推理的用户相似性贝叶斯网络模型的构建

上节中，我们给出用户间相似度度量函数，可以得到相邻用户间的直接相关性，但是实际的情况是，对于网站的任一用户而言，与其相邻用户还存在一定程度的间接相似关系，所以本文中我们打算利用 USBN 的概率推理来获取节点间的间接相似关系，增加模型预测的准确性。由前文可知，贝叶斯网的精

确推理具有指数计算时间，不能够支持高效的发现大量存在于用户之间的相似关系。目前，马尔可夫链蒙特卡罗(MCMC)算法作为一种重要的近似推理方法，已经被广泛应用于贝叶斯网推理计算，Gibbs 采样是其中一种重要的马尔可夫链蒙特卡罗算法，它总是产生一条马尔科夫链，也是 MCMC 方法中最简单，使用最广泛的一种。基于 Gibbs 采样的贝叶斯网近似推理，能够高效的计算各个关键词节点的条件概率和后验概率为高效的近似推理提供了理论基础。因此我们给出一种基于采样的贝叶斯网近似推理的算法进行推断所有可能值的概率分布。在这里为了简化计算，对于每次的用户采样节点，我们只考虑采样用户节点的马尔科夫覆盖中的用户节点对它的影响，即考虑采样节点的父节点，采样节点的子节点及采样节点子节点的其他父节点对采样节点的影响。

3.4. USBN 模型构建效率测试

为了验证模型的最终效果，我们利用公式 3.1 和公式 3.2，对于 DBLP 数据集[14]进行了试验，分别从原始数据中选择100, 200, 300, ⋯, 1000 个用户，分别测试 USBN 的构建执行时间。对于每一个 USBN，我们记录了 20 次测试的平均执行时间。每次 DAG 构建执行时间，计算执行时间和总共执行时间都记录在图 1~2 中。

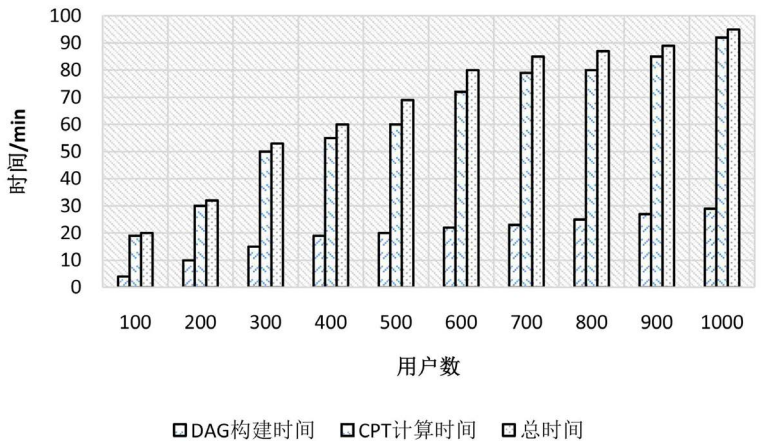

Figure 1. The construction efficiency diagram of USBN
图 1. USBN 构建效率示意图

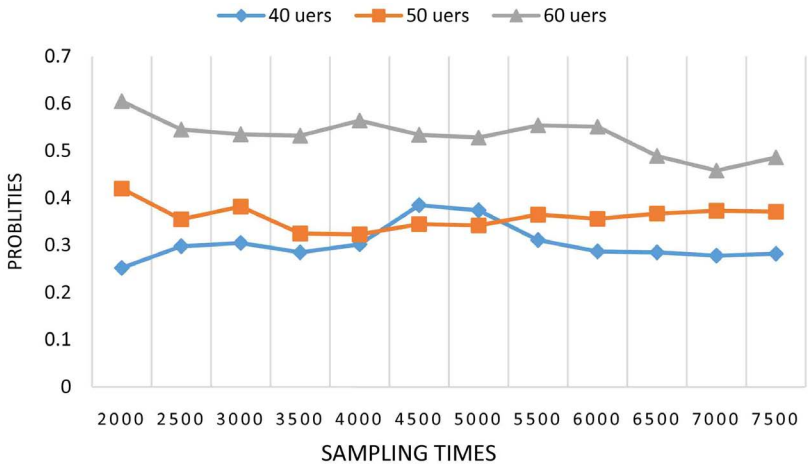

Figure 2. The convergence effect diagram of USBN
图 2. USBN 收敛效果示意图

见图 1，DAG 构建的时间远小于 CPT 计算的时间，同时 USBN 构建的总执行时间主要取决于 CPT 计算执行时间。这是与实际情况一致的，节点的 CPT 的计算执行的时间是随着父节点的增加而呈现指数增长。从实验结果可以分析出，对于小规模数据集，在当前计算机环境条件下(类似本文实验环境)，本文所提出的方法用于 DAG 的构建是可行的。基本上，USBN 的构建执行时间随着节点数量的增加而增加。

因此，从效率的角度来看，我们可以得知通过合并数据，减少节点个数，密集型计算聚合查询处理技术在未来进一步改善构建将是可行的，这也正是我们下一步的工作。

3.5. USBN 收敛性测试

为了测试算法的效率和收敛性，我们采用由部分数据集作为测试数据集生成的。随着采样次数的增多，我们记录了在不同用户个数情形下，算法每次执行时间和推理结果，并分别记录在图中。见图 2，在近 5000 次采样以后，近似推理结果收敛于一个相对稳定的值，所以说明算法近似推理在一定程度上是有效的并且是收敛的。

在本文中，针对社交网络数据中的用户相似性发现问题，提出了一种有效的社交网络用户相似性发现方法。为了发现社交网络用户之间的直接相似性，我们通过对社交网络数据进行分析提出了一种基于贝叶斯网络的社交用户贝叶斯网构建方法。为了发现社交网络用户之间的间接相似性，我们同时考虑了 USBN 结构以及 USBN 中包含的不确定语义对社交网络用户相似性的影响，最终结合结构相似性和并行推理提出了一种基于贝叶斯网络的社交网络用户相似性度量方法。为了支持基于贝叶斯网的用户相似性度量方法的高效性和高的可扩展性，使算法更适合于海量分布式数据，我们提出利用对贝叶斯网络进行加权的方法作为改进的方向。为了测试本文所提出社交网络用户相似性发现算法的执行性能，我们利用 Python 执行了相关的实验，通过对实验结果分析发现本文提出的基于的社交网络用户相似性发现方法具有良好的可扩展性和较高的精确度。

参考文献

[1] 中国互联网络信息中心(CNNIC)发布的第 42 次《中国互联网络发展状况统计报告》[R]. http://www.cnnic.net.cn/.

[2] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

[3] Russell, S.J. and Norvig, P. (2010) Artificial Intelligence: A Modern Approach. *Applied Mechanics & Materials*, **263**, 2829-2833.

[4] 方滨兴, 贾焰, 韩毅. 社交网络分析核心科学问题, 研究现状及未来展望[J]. 中国科学院院刊, 2015(2): 187-199.

[5] 徐娟. 基于贝叶斯网的社交网络用户相似性发现[D]: [硕士学位论文]. 昆明: 云南大学, 2015.

[6] 李青. 基于 MapReduce 的广告点击率预测系统设计与实现[D]: [硕士学位论文]. 昆明: 云南大学, 2016.

[7] 郭俊. 大规模复杂网络社区发现与社区进化分析技术研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2017.

[8] 方志鹏. 基于贝叶斯网的新广告点击率预测[D]: [硕士学位论文]. 昆明: 云南大学, 2015.

[9] Yan, J., Liu, N., Wang, G., *et al.* (2009) How Much Can Behavioral Targeting Help Online Advertising? *Proceedings of the* 18*th International Conference on World Wide Web*, Madrid, 20-24 April 2009, 261-270. https://doi.org/10.1145/1526709.1526745

[10] 程学旗. 信息网络拓扑结构与内容相关性研究[D]: [博士学位论文]. 北京: 中国科学院研究生院(计算技术研究所), 2006.

[11] 胡笑旋, 杨善林, 马溪骏. 面向复杂问题的贝叶斯网建模方法[J]. 系统仿真学报, 2006, 18(11): 3242-3246.

[12] 王双成. 贝叶斯网络学习、推理与应用[M]. 上海: 立信会计出版社, 2010.

[13] 贝叶斯网引论[M]. 北京: 科学出版社, 2006.

[14] DBLP 数据集[R]. http://dblp.uni-trier.de/xml/.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2286，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sea@hanspub.org