

# 基于贝叶斯网络的信息提取技术研究

吴绍兵

(云南警官学院 信息网络安全学院, 云南 昆明 650223)

**摘 要:**随着互联网的飞速发展,公开获取可靠信息的不断增加,人们可从网络上获取各种各样的信息资源,这给人们的学习和利用信息带来了极大的方便。同时面对浩如烟海的海量信息,如何在短时间内获取人们感兴趣和有用的信息,成为目前关注的热点。同时信息提取活动是一个复杂的过程,基于此,文中提出了一种利用贝叶斯网络的方法来对信息进行有效提取的方法,得出了贝叶斯网络信息提取模型。通过 VC++6.0 编程,模拟实现了所提出的方法,实验结果表明该方法是可行的。

**关键词:**贝叶斯网络;信息提取;用户模型

**中图分类号:**TP309

**文献标识码:**A

**文章编号:**1673-629X(2012)11-0225-04

## Research of Information Extraction Technique Based on Bayesian Network

WU Shao-bing

(Institute of Information Security, Yunnan Police Officer Academy, Kunming 650223, China)

**Abstract:** With the rapid development of the internet, growing number of open access to reliable information, people can get all sorts of information from a network resource, which bring great convenience to the learning and use of information. Facing the vast mass of information, how to get the interested and useful information for people within a short time became the focus of attention. At the same time, information extraction activity is a complex process, for this gave a Bayesian network approach to the method of extracting information, come to a Bayesian network model of information extraction. Through the VC++6.0 program, simulated the approach proposed, simulan results indicate that the method is feasible.

**Key words:** Bayesian network; information extraction; user model

## 0 引言

随着互联网的飞速发展,公开获取可靠信息数量的不断增加,人们可从网络上获取各种各样的信息资源,这给人们的学习和利用信息带来了极大的方便。同时面对浩如烟海的海量信息,如何在短时间内获取人们感兴趣和有用的信息,成为目前关注的热点。

信息提取活动是一个复杂的过程,可以将此复杂的任务分解成几个子任务。这个分解带来如下好处:

- (1) 为每项任务可独立于其他任务选择最佳技术成为可能;
- (2) 作为一组独立的模块(每个任务一个),信息提取程序将会被开发出来,以便轻松执行本地调试;
- (3) 通过重新排序、选择、甚至任务的组合,很容易定义信息提取活动。

为此,通过对机器学习的相关技术进行研究,最终得出了采用贝叶斯网络来对所需信息进行提取的方法。

## 1 网络信息描述

人们在网络信息过滤方面进行了一定的研究,提出了一些信息描述模型<sup>[1]</sup>。

### (1) 布尔模型。

布尔(Boolean)模型是采用将集合论和布尔代数相关技术相结合的一种简单信息提取和检索模型。根据集合的简单和直观特性,Boolean模型提供了一个用户获取信息容易掌握的框架。对关键词之间的关系采用非、且、或等运算。结合下面的概率推理模型对信息进行筛选和过滤,以期快速、准确地得出用户所需要的信息。缺点是:布尔模型采用是和非两种状态,缺乏对文档的分级考虑。

### (2) 向量空间模型。

向量空间模型(Vector Space Model, VSM)是一种

收稿日期:2012-03-05;修回日期:2012-06-11

基金项目:国家社科基金课题(09XTQ004)

作者简介:吴绍兵(1976-),男,云南永善人,讲师,研究方向为算法、人工智能、信息安全和计算机取证等。

较著名的用于文档表示的统计模型,该模型以特征项作为文档信息表示的基本单位,特征项可以由字词或短语组成。当文档被表示为文档空间的向量,就可以通过计算向量之间的相似性来度量文档间的相似性。从而,可以用相似性度量方法来计算其余弦值。余弦为零表示待提取的词向量垂直于文件向量,即不符合信息提取要求,不能提取;余弦值为 1 表示待提取的词与文件向量重合,完全符合要求。张筱丹<sup>[2]</sup>等利用向量空间模型来对摘要进行自动冗余处理,取得了较好的效果。

### (3) 概率推理模型。

概率推理模型是根据人们过去的经验和对相关信息的分析,结合专家知识和先验信息,由已知的变量信息来推导未知变量的信息的过程。

廖祥文<sup>[3]</sup>等将概率推理模型应用于博客倾向性检索中,提出了一个基于概率推理模型的博客倾向性检索算法。所提算法能够有效地识别博客空间中与给定查询相关的观点。

## 2 网络信息提取方法

### 2.1 网络信息提取和过滤方法

根据黄晓斌等<sup>[4]</sup>作者的讨论,有多种信息提取的方法:分级法、分割法、分类法、地址列表法、动态文本分析法、基于图像识别技术的信息过滤方法、动态跟踪技术法等。

#### 2.1.1 分级法

分级法首先是对网络信息进行分类过滤的方法,它是根据信息所呈现的关键词、特征和基本内容,运用一定的分级体系(国内正在建设中)分门别类地把要过滤的信息表示成一些分类目录或过滤模版,并做好分级标记,使用时与过滤模板进行比较以决定对所出示的信息是否过滤进行过滤处理。根据实际情况可以制定出符合自身实际的分级体系。

#### 2.1.2 分割

分割任务把文本划分为原子元素,称为段或标记。尽管这项任务是由于简单空白分离单词而使得西方语言变得大大简化,但仅仅有一些空白分段是不够的。通常,使用分割来确定如何处理每个个案的规则执行这些个案的分割。为了完成分割使用句法或词法分析的任务,可以使用词汇、语法。汉语中的分割方法的另一种方法使用的是基于统计的技术。

#### 2.1.3 分类

分类任务确定每个段分割任务中获得的类型。换句话说,它确定输出数据结构的字段与输入的段适合的地方。分类任务的结果是为类提供相关的域元素的实体。分类任务中使用的基于规则的技术通常基于语

言的资源,如词汇和语法等。一个最受欢迎的进行分类的方法是机器学习。在分类任务中使用的机器学习技术通常负责监督和附加说明。最常见的监督的学习技术是隐藏马尔科夫模型(HMM),条件随机域(CRF),支持向量机和决策树等。

#### 2.1.4 地址列表法

URL 地址列表法是采用预先编制好的 URL 地址列表来决定是允许用户访问还是禁止用户访问信息的一种方法,这种方法是信息过滤中最为直接也最为简单的方法。在日常的工作和生活中主要是通过管理者或第三方提供的黑白名单,或者是通过用户在日常工作中自行收集、总结并编制的黑名单。有了黑白名单以后,可以在网页的工具菜单上选择 internet 选项,打开“安全”内容,对“受信任的站点”和“受限制的站点”进行设置即可。

除以上讨论的四种以外,很多文献中还讨论了动态文本分析法、基于图像识别技术的信息过滤方法、动态跟踪技术等。

### 2.2 信息提取方法

当前信息提取技术在企业应用和网络安全领域有着很大的作用。赵金仿<sup>[5]</sup>等为了剔除网页中的无用信息,提出一种基于 HTML 自身结构特点的网页正文信息抽取方法并进行了实现,为我们提供了借鉴和参考。宫义山<sup>[6]</sup>等提出了一种融合于故障树和传统贝叶斯网络的新方法——诊断贝叶斯网络,并阐述了故障树和贝叶斯网络的故障诊断策略优化方法的基本思想和具体算法。马福晶<sup>[7]</sup>等通过论述信息检索的工作原理和其在网络环境下的作用,对比分析了基于网络的信息检索几种类型的特点,对高速而有效的信息检索系统的核心技术搜索引擎技术进行了分析,指出随之带来的亟待解决的快速有效获取信息的问题和搜索引擎技术符合时代要求的发展方向。张成伟<sup>[8]</sup>等提出了一种基于向量空间模型改进的文本信息检索方法。把本体技术引入到传统的文本信息检索系统中,利用领域本体中概念相似度计算对向量空间模型进行改进,从而实现一个高效的文本检索系统,并系统地对模型进行了阐述。

## 3 基于贝叶斯网络的信息过滤方法

### 3.1 贝叶斯网络的基本概念

贝叶斯网络的一个主要优势是它们允许一个概率分布分解成一系列小的分布。与网络拓扑相关的独立语义表明怎样组合这些局部分布以获取完全联合概率分布,在由网络节点所表示的所有随机变量。网络拓扑和独立性之间的关系通过一个分解属性来加以捕捉。

贝叶斯推理,也称为贝叶斯定理或贝叶斯规则,是一种数学方法的派生概率,是一种基于不确定的证据。主要通过先验概率(即人们的经验或已有知识)、样本条件概率(通过样本数据算出的概率),来计算后验概率。贝叶斯定理求后验概率的方法是,它提供了一种通过使用  $P(A|B)$  计算  $P(B|A)$  的方法。

### 3.2 贝叶斯定理

贝叶斯网络作为因果关系图、因果关系网、简洁网络和概率网络而著名。贝叶斯网络是有限无环图:它们是有向的,以至于节点之间的联系是单向的,并且它们是无环的,因为它们不能包含圈。网络图中节点表示随机变量,弧表示变量之间的概率依赖关系。每个节点都有一个包含父节点状态的每个组合及其所给定的节点的每个可能状态的条件的概率表(CPT)。根节点表包含无条件的先验概率<sup>[9]</sup>。

在数学的领域, $E$  给定的情况下, $H$  的条件概率表示如下:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})} \quad (1)$$

$E$  表示一个证据,而  $H$  表示一种假设。这个方程其实代表的预测:如果观察到证据  $E$ ,那么假说  $H$  为真的可能性有多少?

### 3.3 贝叶斯方法的优点和主要应用

1. 以最佳的方式提供优雅简单和理性的方法,对于给定状态的信息,科学回答相关的问题。这与常规统计分析的方法形成鲜明对比:

(a) 明确说明您的问题与以前的信息或者称为先验信息。

(b) 适用之和与积的规则。起始点总是贝叶斯公式或贝叶斯定理。

2. 通过概率的假设和计算公式为贝叶斯的推理和统计推断提供理论基础。

3. 贝叶斯网络提供了更强大的方式,通过自动化奥卡姆剃刀评估竞争理论作为最前沿的科学。奥卡姆剃刀归因于中世纪哲学家奥卡姆的威廉原则。该原则指出一个不应比所需要的最多的假定。它包含所有科学建模和理论假设。它就会提醒我们选择从一组给定的现象出发,选择一个最简单的等效模型。在任何给定的模型中,奥卡姆剃刀有助于我们“剔除”那些不真的需要解释的变量现象。这曾经被认为是只有一个定性的原则。

贝叶斯网络的应用领域<sup>[10]</sup>包括:知识表示、因果推理、诊断推理、支持推理、情景分析、压力测试等。

邵继业<sup>[11]</sup>等把贝叶斯网络引入到模型诊断框架中,依据观测量,研究了一种建立系统贝叶斯网络观

测模型的方法。利用网络观测模型,可计算系统诊断解的后验概率,从而找出系统最可能的故障组件。刘家鹏等<sup>[12]</sup>将贝叶斯网络应用于银行操作风险管理中,取得了很好的效果。

### 3.4 利用贝叶斯网络来对信息进行过滤和识别的方法和步骤

根据贝叶斯网络推理的基本原理,可将信息过滤问题映射为一个多因素的推理分析问题,即建立用于信息过滤的贝叶斯网络。首先将信息过滤分级体系中的各级子目录都转化为贝叶斯网络中的节点,然后建立起所有节点之间的因果关系,这样信息提取问题就转化为一个贝叶斯网络信念更新问题。

#### (1) 准备阶段。

在准备阶段,通过软件代理自动获取某段时间内的网络信息,以关键词的形式反馈结果。可以通过分割、分类和分级的方法提取出关键词,按照一定的标准或格式进行筛选、添加和调整,然后输入到热点词库中,热点词库中的记录可以根据 keyword 字段进行分级处理,通过分级后对其进行分类存储。

#### (2) 处理阶段。

为便于处理,提高处理速度,在进入正式的信息提取和过滤处理阶段前,要对信息进行规范化预处理,统一格式。由于网络信息数据量非常大,所以提取出来的特征向量维数将会非常高,根据贝叶斯统计分析,计算出每个数据量的相关系统,利用相似特征来对有关数据进行有效的降维处理,根据出现的频率赋予一定的权值,保留权值较高的词条作为文档的特征项。将处理后的特征项存入文档特征库。

#### (3) 分析阶段。

在前所述相关文献的基础上,对数据量进行分析,以  $H$ : 当前热点信息;  $R$ : 表示空格、逗号、句号、分号和阿拉伯数字当作分隔符;  $G$ : 非垃圾信息或恶意信息集;  $F$ : 表示用户感兴趣的信息集;  $E$ : 用户需要提取的信息。建立网络信息提取模型(见图 1)。

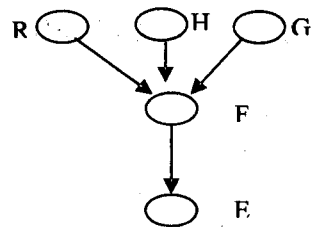


图 1 贝叶斯网络信息提取模型图

除了图 1 中表示的概率依赖外,以下关于网络的条件概率矩阵的假设似乎更简单。

1、如果  $F$  是真的,那么肯定  $E$ :  $P(E|F) = 1$ 。

2、如果  $F$  是假的,那么由有关信息的随机匹配概率  $f$  确定的  $E$  的概率:  $P(E|\bar{F}) = f$ 。

3、如果  $G, H$  为真, 则肯定  $F$ , 即:  $P(F|G, H) = 1$ 。

4、如果肯定  $H$ , 而排除  $G$ , 则:  $P(F|-G, H) = 0$ 。

5、如果排除  $H$  并且肯定  $G$ , 那么肯定的是排除  $F$ 。

由假设 1 及 2, 得出给定  $H$  的情况下,  $E$  的可能性由下列公式计算:

$$P(E|H) = P(F|H) + P(-F|H)f \quad (2)$$

然后就可以计算出  $H$  给定的情况下  $F$  的概率, 可以得出:

$$P(F|H) = P(F|G, H)P(G|H) + P(F|-G, H)P(-G|H) \quad (3)$$

在假设 4 的情况下, 这个公式简化为:

$$P(F|H) = P(F|G, H)P(G|H) \quad (4)$$

由假设 3, 以及  $H$  和  $G$  是概率独立的, 公式(4)简化成

$$P(F|H) = P(G) \quad (5)$$

因此, 第一个公式(1)的简化为:  $P(E|H) = P(G) + P(-F|H)f$

同样根据假设 3 和假设 4, 得到了  $H$  给定的情况下  $F$  不发生的概率等于  $G$  发生的概率:

$$P(-F|H) = P(-F|G, H)P(G|H) + P(-F|-G, H)P(-G|H) = P(-G) \quad (6)$$

利用似然比 (likelihood ratio) (LR) 公式

$$LR = \frac{P(E|H)}{P(E|-H)}, \text{ 得出该 LR 的分子(2) 可写为}$$

$$P(E|H) = P(G) + P(-G)f \quad (7)$$

利用假设 5, LR 的分母可以写成

$$P(E|-H) = P(F|-H) + P(-F|-H)f \quad (8)$$

即此式与公式(2)是相同的, 只是用了  $-h$  作为替代假设。

根据  $H$  不发生的情况下  $F$  发生的概率以及  $H$  不发生的情况下  $F$  不发生的概率, 得出如下的计算公式:

$$P(F|-H) = P(F|-G, -H)P(-G|-H),$$

$$P(-F|-H) = P(G|-H) + P(-F|-G, -H)P(-G|-H)$$

因此可以考虑,  $G$  和  $H$  仅仅是条件依赖于  $F$ , LR 由下式给出:

$$\begin{aligned} LR &= \frac{P(E|H)}{P(E|-H)} \\ &= \frac{P(G) + P(-G)f}{P(F|-G, -H)P(-G) + [P(G) + P(-F|-G, -H)P(-G)]} \end{aligned} \quad (9)$$

假定  $P(G) = r$ , 并且  $P(F|-G, -H) = p$ ; 就会获得一个简单的公式:

$$LR = \frac{r + (1-r)f}{rf + (1-r)[p + (1-p)f]} \quad (10)$$

如果 LR 要达到它的最大值, 则 LR 将被缩减到它

的最简单的形式,  $1/f$ 。

设  $P(G) = r, P(R) = q, P(F|-H, R, -G) = p, P(F|H, R, G) = 1$ , 通过计算, 其似然比为:

$$\frac{P(E|H)}{P(E|-H)} = \frac{rq + (1-rq)f}{rf + (1-r)[pg + (1-pg)f]} \quad (11)$$

通过比较(10)和(9)式, 如果  $q = 1$ , 那么公式(11)变为(10)式。如果  $r = 1$ , 那么

$$LR = \frac{q + (1-q)f}{f} \quad (12)$$

在图 1 所示的贝叶斯网络中, 通过公式(10)和(12)来判断可能是有帮助的, 因为它提供了一个直观的模型: 它形象地说明了由专家作出的依赖性和独立性假设, 以及由公式所计算出的评估概率。

### 3.5 实验结果与分析

作者通过 VC++6.0 语言编程, 模拟实现了上述贝叶斯算法, 然后向其提供大量的素材文件供其学习, 从而建立起了一个小型的单词特征数据库。为了检测该模型的实际应用效果, 作者给出了一组测试样本文件, 规模在 500 份左右, 利用该算法逐一判断每份文件为不良信息文件还是相关信息文件。测试语料同建立学习数据库的语料不重合, 测试数据和结果如表 1 所示。

表 1 样本数据的识别结果

学习样本数 Ln(个)	测试样本数 Sn(个)	识别结果 ln(个)	识别率 D/%
2058	523	504	96.31

## 4 结束语

这篇文章在信息过滤和提取相关内容的基础上, 通过比较网络信息过滤和提取的相关方法。在此基础给出了利用贝叶斯网络来进行信息过滤和提取的方法。利用似然公式来解释变量之间的概率关系, 通过假设和计算得出简洁和直接的模型公式, 解释了概率公式背后的合理性, 使变量的意义更清楚, 以提取出满足一定概率需要的信息。并进行模拟实验, 效果好。

### 参考文献:

- [1] 邹萍, 纪沙. 网络信息过滤机制的研究[J]. 哈尔滨师范大学自然科学学报, 2008, 24(2): 75-80.
- [2] 张筱丹, 胡学钢. 基于向量空间模型的自动摘要冗余处理研究[J]. 合肥工业大学学报(自然科学版), 2010(9): 1355-1358.
- [3] 廖祥文, 曹冬林, 方滨兴, 等. 基于概率推理模型的博客倾向性检索研究[J]. 计算机研究与发展, 2009, 46(9): 1530-1536.
- [4] 黄晓斌, 邱明辉. 网络信息过滤方法的比较研究[J]. 大学图书馆学报, 2005(1): 42-48.

(下转第 234 页)

WSN 的安全数据融合的轻量级检测机制是以后研究的重要内容;

(5) 当前协议方案都是基于同构的 WSN, 如何开发出适应于异构 WSN 的安全数据融合技术值得进一步研究。

#### 4 结束语

物联网的兴起势必会带来无线传感器网络的又一次革新, 而安全数据融合作为一种有潜力的基本的安全服务应用, 也必将引起更大的重视。各种新的安全数据融合方案也将被提出和得到应用, 轻量级的安全数据融合方案是下一个阶段要进行的工作。

#### 参考文献:

- [1] Sang Y, Shen H, Inoguchi Y, et al. Secure data aggregation in wireless sensor networks; a survey [C]//Proc of the Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies. [s. l.]: [s. n.], 2006: 315-320.
- [2] Vu H, Mittal N, Venkatesan S. THIS: THreshold Security for Information Aggregation in Sensor Networks [C]//Proc of Fourth International Conference on Information Technology. Washington: IEEE Computer Society Press, 2007: 89-95.
- [3] Przydatek B, Song D, Perrig A. SIA: Secure Information Aggregation in Sensor Networks [C]//Proc of 1st Conference on Embedded Networked Sensor Systems. Amsterdam: IOS Press, 2003: 255-265.
- [4] Mahimkar A, Rappaport T S. SecureDAV: A Secure Data Aggregation and Verification Protocol for Wireless Sensor Networks [C]//Proc of the 47th IEEE Global Telecommunications Conference (Globecom). Dallas, TX: [s. n.], 2004.
- [5] Wang Y Y, Zhu X S, Cao G. SDAP: a secure hop-by-hop data aggregation protocol for sensor networks [C]//Proc of the ACM MOBIHOC'06. [s. l.]: [s. n.], 2006.
- [6] Ichikawa H, Ozdemir S. Secure and reliable data aggregation for wireless sensor networks [J]. LNCS, 2007, 4836: 102-109.
- [7] Du W L, Deng J, Han Y S. A Witness-based Approach for Data Fusion Assurance Wireless Sensor Networks [C]//Proc of IEEE Global Telecommunication Conference. Washington: IEEE Computer Society Press, 2003: 1435-1439.
- [8] Cam H, Ozdemir S, Muthuavinashiappan D. ESPDA: Energy Efficient and Secure Pattern-based Data Aggregation for Wireless Sensor Networks [C]//Proc of IEEE Sensors. Washington: IEEE Computer Society Press, 2003: 732-736.
- [9] Sanli H O, Ozdemir S, Cam H. SRDA: secure reference-based data aggregation protocol for wireless sensor networks [C]//Proc of the IEEE VTC Fall Conference. Los Angeles, CA, 2004: 4650-4654.
- [10] He W B, Liu X, Nguyen H. PDA: Privacy-preserving Data Aggregation in Wireless Sensor Networks [C]//Proc of 26th IEEE International Conference on Computer Communications. Washington: IEEE Computer Society Press, 2007: 2045-2053.
- [11] Li H, Lin K, Li K. Energy-efficient and High-accuracy Secure Data Aggregation in Wireless Sensor Networks [J]. Computer Communication, 2011, 34(4): 591-597.
- [12] 杨庚, 王安琪, 陈正宇, 等. 一种低能耗的数据融合隐私保护算法 [J]. 计算机学报, 2011, 34(5): 792-800.
- [13] Girao J, Westhoff D, Schneider M. CDA: concealed data aggregation for reverse multicast traffic in wireless sensor networks [C]//Proc of IEEE International Conference on Communications. Washington: IEEE Computer Society Press, 2005: 3044-3049.
- [14] Castelluccia C, Mykletun E, Tsudik G. Efficient Aggregation of Encrypted Data in Wireless Sensor Networks [C]//Proc of Second Conference on Mobile and Ubiquitous Systems. Washington: IEEE Computer Society Press, 2005: 109-117.
- [15] Mykletun E, Girao J, Westhoff D. Public Key Based Cryptoschemes for Data Concealment in Wireless Sensor Networks [C]//Proc of IEEE International Conference on Communications. New York: IEEE Communications Society Press, 2006: 2288-2295.
- [16] Rodhe I, Rohner C. n-LDA: n-Layers Data Aggregation in Sensor Networks [C]//Proc of the 28th International Conference on Distributed Computing Systems Workshops. Beijing: IEEE Computer Society Press, 2008: 400-405.
- [17] Zhang W, Liu Y, Das S K, et al. Secure Data Aggregation in Wireless Sensor Networks: A Watermark Based Authentication Supportive Approach [J]. Elsevier Pervasive Mobile Computer, 2008(4): 658-680.

(上接第 228 页)

- [5] 赵金仿, 赵艳, 缪建明. 网页信息抽取及其自动文本分类的实现 [J]. 计算机技术与发展, 2008, 18(10): 37-39.
- [6] 宫义山, 高媛媛. 基于信息融合的推断贝叶斯网络研究 [J]. 计算机技术与发展, 2009, 19(6): 106-108.
- [7] 马福晶, 葛润霞. 基于网络信息检索的研究 [J]. 计算机技术与发展, 2008, 18(8): 111-114.
- [8] 张成伟, 郑诚. 基于改进 VSM 的文本信息检索研究 [J]. 计算机技术与发展, 2009, 19(1): 71-73.
- [9] 林士敏, 王双成, 陆玉昌. Bayesian 方法的计算学习机制和问题求解 [J]. 清华大学学报 (自然科学版), 2000, 40(9): 61-64.
- [10] 张连文, 郭海鹏. 贝叶斯网引论 [M]. 北京: 科学出版社, 2006: 18-36.
- [11] 邵继业, 王日新, 徐敏强. 贝叶斯网络在模型诊断中的应用 [J]. 吉林大学学报 (工学版), 2010(1): 234-237.
- [12] 刘家鹏, 詹原瑞, 刘睿. 基于贝叶斯网络的银行操作风险管理 [J]. 计算机工程, 2008, 34(18): 266-271.