文章编号: 1673-1549(2011)01-0058-04

基于贝叶斯网络的网站信息抽取模型

谭龙江[1,2]

(1.华侨大学经济金融学院, 福建 泉州 362021; 2.西南财经大学经济信息工程学院, 成都 610074)

摘 要: 历史信息、即时信息以及流言往往冲淡网站中的主题思想, 导致信息隐藏等问题。为解决上述问题, 提出了网站信息抽取系统的结构模型、数据结构和处理流程; 该模型采用信息抽取技术, 从相关网页中抽取出带有主观倾向的主题信息; 采用贝叶斯网络对客户需求进行决策与预测分析。仿真测试与客户应用证明, 该模型能较准确的抽取网站中的客户倾向、有较好的及时性。

关键词: 网站; 贝叶斯网络; 信息抽取; 模型

中图分类号: TP393　　文献标识码: A

引 言

网站信息抽取属于网络内容挖掘(Web content mining)研究的一部分, 其主要内容包括: 结构化数据抽取(Structured Data Extraction)、信息集成(Information integration)和观点挖掘(Opinion mining)等。结构化数据抽取(Structured Data Extraction)的目标是从Web页面中抽取结构化数据[1-3]。这些结构化数据往往存储在后台数据库中, 由网页按一定格式承载着展示给用户。例如论坛列表页面、Blog页面和搜索引擎结果页面等。信息集成(Information integration)是针对结构化数据而言的。其目标是将从不同网站中抽取出的数据统一化后集成入库。其关键问题是如何从不同网站的数据表中识别出意义相同的数据并统一存储。观点挖掘(Opinion mining)是针对网页中的纯文本而言的。其目标是从网页中抽取出带有主观倾向的信息。大多数文献中提到的网站信息抽取往往专指结构化数据抽取。

网站信息抽取的流程通常如下[4-8]: 首先, 确立采集目标, 即由用户选择目标网站, 然后, 提取特征信息, 即根据目标网站的网页格式, 提取出采集目标数据的通性, 最后, 网络信息获取, 即利用工具自动的把页面数据把存到数据库[9]。根据工作原理的不同, 网站信息抽取工具可分为以下几大类: (1)开发包装器的专用语言(Languages for Wrapper Development): 用户可用这些专用语言方便地编写包装器。例如Minerva, TSIMMIS, Web-OQL, FLORID, Jedi等。(2)以HTML为中间件的工具(HTML-aware Tools): 这些工具在抽取时主要依赖HTML文档的内在结构特征。在抽取过程之前, 这些工具先把文档转换成标签树; 再根据标签树自动或半自动地抽取数据。代表工具有Knowlesys, MDR。(3)基于NLP(Natural language processing)的工具(NLP-based Tools): 这些工具通常利用filtering, part-of-speech tagging和lexical semantic tagging等NLP技术建立短语和句子元素之间的关系, 推导出抽取规则。这些工具比较适合于抽取那些包含符合文法的页面。代表工具有RAPIER, SRV, WHISK。(4)基于模型的工具(Modeling-based Tools): 这些工具让用户通过图形界面, 建立文档中感兴趣的对象的结构模型, "教"工具学会如何识别文档中的对象, 从而抽取出对象。代表工具有: NoDoSE, DEByE。(5)基于本体的工具(Ontology-based Tools): 这些工具首先需要专家参与, 人工建立某领域的知识库, 然后工具基于知识库去做抽取操作。如果知识库具有足够的表达能力, 那么抽取操作可以做到完全自动。而且由这些工具生成的包装器具有比较好的灵活性和

收稿日期: 2010-10-15

基金项目: 福建省社会科学基金项目(2010B064); 华侨大学科研基金项目(07HSK02)

作者简介: 谭龙江(1973-), 男, 新疆伊宁人, 讲师, 博士生, 主要从事电子商务方面的研究。

适应性。代表工具有: BYU, X-tract。

1 系统结构与运行流程

网页特征抽取的首要工作是网页特征向量化、归一化和标准化[8]。需要抽取的网页中的信息特征主要分为 4类:

第一类: 信息位置特征: 信息片断所在的位置具有相对的确定性, DOM 树中具有层次结构的路径可以作为信息片断的"坐标"。

第二类: 信息上下文特征: 是指网页中信息片断的前后信息 (之前叫做前引导词, 之后叫做后引导词)往往与信息片断有一定的关系, 通常具有提示作用。

第三类: 信息一般特征: 是指不同信息片断之间的本质区别, 比如时间用数字表示而姓名用字母表示等。

第四类: 信息的可视特征: HTML标记不仅可以用来组织内容, 还可以用来表示网页的外观, 如字体的大小和颜色、段落的长短等。

本优化模型包括了以下模块, 如图 1所示。

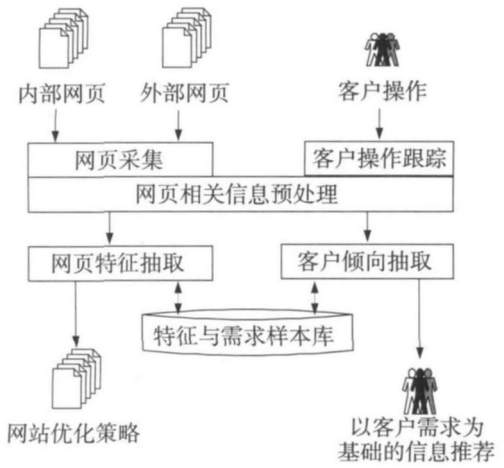

图 1 网站信息抽取系统模型

网页采集模块 (反向爬虫模块): 该模块定期对常用搜索引擎进行跟踪调查, 将表征本网站特征的主题词作为搜索依据, 在各搜索引擎查询其排名情况 (其基本过程为: 提取特征关键词 —输入关键词 (主题词) —下载搜索结果页面 —检索排名 —汇总处理); 并将所生成的优先关键词等信息注入多维度匹配与检索空间, 以便 SEO操作使用。

客户操作跟踪模块: 该模块主要收集两方面的信息, 并进行分析处理: (1)登录本网站的客户的内部需求倾向与检索目标; (2)反向爬虫提供的各搜索引擎的常用外部检索词。最终, 该模块将分析获取的与本网站内容相关的高频关键字注入多维度的"客户需求 VS网站特征"主题词匹配与检索空间。

网站相关信息预处理模块: 该模块主要用于关键词 (主题)的频次 (度)调整; 该指标对搜索引擎的优化起到重要作用。通常, 关键词密度一般在 1% - 7%较为合适, 超过这一标准就有过高或过低之嫌[7]。该模块从多维度匹配与检索空间中提取高频关键词样本, 进而通过学习该空间中的相关样本来确定本网站所依托的主题, 并将主题内容提交给网站宣传优化模块, 以此为依据, 对相关网页的内容作对应调整, 实现主题词密度的冲淡或浓缩。此外, 该模块还向反向爬虫模块提供相应的主题词 (关键词)作为其在搜索引擎中查找排名的依据。

特征与需求样本库模块: 该模块存储并计算外部搜索引擎中和网站内部的流行关键词 (以及链接)与本网站内部对应元素之间的相关度; 以便在实际的网页优化过程中判断这些网页 /链接与本网站主题相关的相关程度, 并最终为网站的宣传优化提供依据。

2 关键技术

网站系统中的网页结构繁复, 层次不清, 网页内的文件类型复杂; 因此, 网页优化系统在处理这类数据时, 很难兼顾实时性和处理效果。此外, 异构网站系统的信息还存在来源不同、排版格式不通、访问频率不同等差异, 导致网站优化具有相当难度。本系统为解决上述问题, 采用了贝叶斯网络进行相关信息的融合, 并就网站信息的整体客户接纳态势进行评估与测度。本模型采用的贝叶斯网络是一种因果推理网。通常, 贝叶斯网络的拓扑结构是包括节点和有向 (加权)边组成的有向无环图 (或树); 其中, 每个节点是用以表征一个或一类信息的随机变量, 其概率分布说明该变量处于该变量状态集合中每个状态的概率值, 而每条有向加权边代表两相邻节点之间的联合、推论或依赖关系, 这些边通常由条件概率矩阵存储和表述。

采用贝叶斯网络的网站信息抽取模型如图 2所示。该网络中的节点分为两类: 信息假想节点 (Hypothesis: H 节点)和事件节点 (Event: E 节点)。这些信息假想节点表示客户对某种信息的评估与测度的量化取值; 而事件节点表示网站管理方在一定的微观信息分类范畴内网页的相关事件。通常, 网页的相关事件可分为两类, 一类是可直接观测的 (例如: 页面内容更新、网页文件的删除与添加和站内主题词的更替等), 称为事件线索或事件征兆 (Event Cue), 另一类是不可直接观测的 (例

如: 客户对特定主题的偏好与浏览倾向)。模型中, 节点之间的有向加权边表示假想之间、假想和事件之间以及相关事件之间的彼此关系 (因果、推论与依赖); 进而采用了条件概率矩阵 (存储在文件中, 可以修改和更新) 描述两者之间的关联紧密程度。模型中, 设定了一个有向联结 $X^{\rightarrow} Y$, 其条件概率矩阵定义为:

$$M \triangleq p\left(\frac{y}{x}\right) \triangleq p\left(\frac{Y=y}{X=x}\right)$$

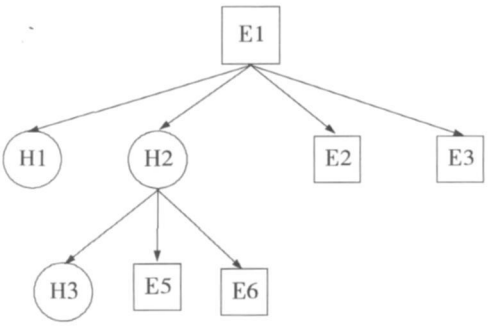

图 2 贝叶斯网络推理原理

该推理模型采用贝叶斯网络描述事件和假想元素之间的相互关系, 采用条件概率矩阵描述各个节点之间的关联程度。在网站信息抽取过程中, 根据该模型, 可以从观测到的客户或者网站事件出发, 逐层推理, 最终得到假想或相关事件, 并根据既定方案进行信息抽取。当模型中的分布式网站信息抽取终端发现高相关度网站事件时, 将发起信息抽取过程; 模型开始执行贝叶斯推理: 即所有事件节点的状态迁移概率都运用贝叶斯方法处理, 一方面将各类网页事件通过数据格式归一化处理, 使得所有数据进入统一的数据空间中待用, 另一方面通过数据库中保存的先验概率和条件概率, 对既得数据进行处理, 使得该模型能够保存节点每次更新的结果, 也能产生对应的、综合的信息抽取结果。由于采用贝叶斯网络获得的假想状态决策结果, 不但与现在最新获得的网页事件信息相关, 而且与以前一个时间段内网页事件管理累积的经验相关; 因此, 本模型具有信息抽取的时间经验记忆能力, 而这种记忆能力是传统的采用硬性产生式对照规则的模型无法实现的。

3 仿真实验结果

本模型通过某论坛的仿真在线监测系统进行了模拟。仿真环境通过虚拟聊天和发帖方式论证该系统的实际效果。仿真实验验证在 10个测试周期后进行; 管理员向测试客户 (实名和匿名)以论坛内外部邮件形式, 对新旧系统的客户满意度作在线对比调查。本次实验共发送在线调查问卷 300份, 收到回复 237份, 其中有效回复 217份。表 1给出了客户满意度调查结果的相关统计; 表 1主要从 6个主要调查项揭示出: 较之老系统, 新系统从网络中抽取到了客户的需求倾向信息, 大大提高了客户满意程度。

表 1 客户满意度调查对比

满意度调查项	新模型 (%)	旧模型 (%)
站内信息的接受程度	79	53
主题吸引程度	69	33
信息发布及时性	75	29
相关链接应时性	79	55
信息相关性	68	41
检索信息高效性	82	47
总体满意度	78	43

本模型能够对客户需求倾向与外部网络主流信息进行跟踪, 因此具有较高的自动化程度和良好的自适应能力; 较之传统网站信息管理系统采用的离线数据挖掘等方法, 贝叶斯网络使信息抽取过程速度更快, 时效性和准确性好。由于网站的客户无法评估信息抽取的实际速度和系统的反应速度, 在测试中, 通过服务访问计数等方式进行: 特定信息的网络抽取速度测试中, 新模型的处理速度平均缩短了 46%以上。

图 3显示了平均在 10个仿真实验监测周期内, 客户获得个性化网站信息抽取服务的平均时耗 (从进入网站开始到模型产生信息推荐为止)与平均浏览时间对比; 该实验证明, 本系统能够通过短暂的需求倾向分析, 采用贝叶斯网络产生决策结果, 能够吸引客户浏览本网站更多的信息, 延长客户的浏览时间, 从而提高本网站的流量业绩。

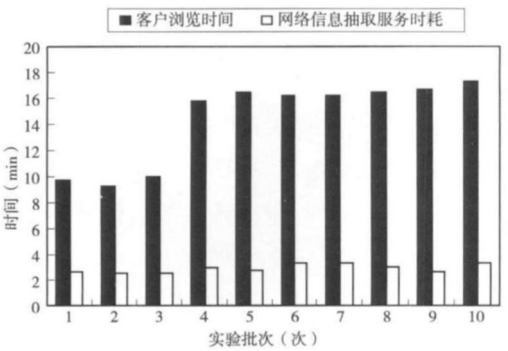

图 3 客户服务时耗实验结果

4 结束语

为解决信息隐藏等问题, 本文提出了网站优化系统的结构模型、数据结构和处理流程。该模型采用网站信

息抽取技术, 从相关网页中抽取出带有主观倾向的主题信息。采用贝叶斯网络对客户需求进行决策与预测分析。仿真测试与客户应用证明, 该模型能较准确的抽取网站中的客户倾向、有较好的及时性。进一步的工作将包括同类异构网站的模式识别、网络流言的识别与控制等。

参 考 文 献:

[1] 俞 琰. 基于隐马尔可夫模型的招聘网站信息抽取 [J]. 北京电子科技学院学报, 2008, 16(4): 93-98.
[2] 郭 岩, 王 宇. 网站信息抽取技术研究 [J]. 信息技术快报, 2008, 6(6): 15-24.
[3] 刘继红, 吴军华. 基于改进的网络蜘蛛算法抽取 Web站点结构的方法 [J]. 江南大学学报: 自然科学版, 2009, 8(5): 555-559.
[4] 杨 俊, 李志蜀. 基于 DOM 的 WEB主题信息抽取 [J]. 四川大学学报: 自然科学版, 2008, 45(5): 1077-1080.
[5] 刘 迁, 焦 慧. 信息抽取技术的发展现状及构建方法的研究 [J]. 计算机应用研究, 2007, 24(7): 6-9.
[6] 肖 明军, 张 巍. 一种多策略联合信息抽取方法 [J]. 小型微型计算机系统, 2005, 26(4): 614-617.
[7] 师雪霖, 程文涛. Web信息抽取与语义检索框架 [J]. 计算机应用, 2010, 42(3): 29-32.
[8] 肖建鹏, 张来顺, 任 星. 直推式支持向量机在 Web信息抽取中的应用研究 [J]. 计算机工程与应用, 2009, 45(2): 147-149.
[9] 何海东, 张文秋. 基于 Web的网络硬盘的设计与实现 [J]. 四川理工学院学报: 自然科学版, 2010, 23(2): 175-177.

Web Information Extraction Model Based on Bayes Network

TAN Long-jiang[1,2]

(1. College of Economics & Finance, Huaqiao University, Quanzhou 362021, China;
2. College of Information, Southwestern University of Finance and Economics, Chengdu 610074, China)

Abstract Historical information, real-time information and gossip can weaken the main ideas of Web sites and lead some information to conceal. In order to deal them, a Web site information extraction model is proposed based on Bayes Network. And its model structure, data structure and processing flows are given as following. The model utilizes information extraction technologies to get subjective key information and so on. Then Bayes Network is used to classify and assemble customers' requests. Simulation results and customers' applications show that the model can extract more accurate customers' tendency and have better real-time than the traditional does.

Key words Web site; Bayes Network; information extraction; model