

可在线增量自学习的聚焦爬行方法

傅向华, 冯博琴, 马兆丰, 何明

(西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 将 Web 爬行看作执行序列动作的过程, 结合改进的快速 Q 学习和半监督贝叶斯分类器, 提出一种新的具有在线增量自学习能力的聚焦爬行方法. 该方法从获取的页面中抽取特征文本, 根据特征文本评估页面的主题相关性, 预测链接的 Q 值, 然后基于 Q 值过滤无关链接. 当得到主题相关页面时产生回报, 将回报沿链接链路反馈, 更新链路上所有链接的 Q 值, 并选择相应的特征文本作为训练样本, 增量地改善主题评估器和 Q 值预测器. 实验结果表明, 该方法具有很快的自学习能力, 获取的页面数目和精度均优于离线聚焦爬行方法, 更符合 Web 资源发现的要求.

关键词: 资源发现; 聚焦爬行; 在线学习; 半监督学习

中图分类号: TP391 **文献标识码:** A **文章编号:** 0253 - 987X(2004)06 - 0599 - 04

Focused Crawling Method with Online-Incremental Adaptive Learning

Fu Xianghua, Feng Boqin, Ma Zhaofeng, He Ming

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Almost current focused crawling systems need volume of trained data samples and cannot learn persistently. Based on the principle of the reinforcement learning, the Web crawling is viewed as a process to perform sequential actions. Combining with the improved fast Q -learning and semi-supervised Bayesian classifier, a novel focused crawling method being able to make online-incremental adaptive learning is presented. Using the characteristic texts extracted from the retrieved pages, the topic-relevance of the new pages can be evaluated by topic evaluator, and the discounted cumulative reward (the value Q) of the links can be predicted by Q -predictor. The value Q is used to cut off the off-topic links, while the reward generated directly by the on-topic pages will be feedback along the link-chain to update all the value Q of the links. And then the characteristic text is selected as unlabeled trained data samples to incrementally improve the knowledge of the topic-evaluator and the Q -predictor. Experiment results show that this method is adaptive and rapid, and the number and accuracy of the retrieved pages can be higher than the off-line focused crawling method. So it is more suitable for discovering Web resources.

Key words: resource discovery; focused crawling; online learning; semi-supervised learning

Web 作为 Internet 环境下的开放信息源, 为广域异构环境下的信息交互提供了有效的手段. Internet 存储了海量信息资源, 如何有效地发现、过滤、处理和管理这些资源是亟待解决的问题. 解决海量 Web 信息收集的一种有效方法是聚焦爬行法, 典型的聚焦爬行系统有 Cora^[1]、IBM Focused

Crawler^[2]、Context Graphs Focused Crawler^[3]等. 聚焦爬行方法主要有两种: 基于页面之间的链接结构进行分析和通过机器学习方法并基于页面内容进行分析. 已有的聚焦爬行系统大多采用离线训练方式, 需要大量已标注的训练样本, 且不能使爬虫在爬行过程中增量学习新的知识, 因而很难符合

收稿日期: 2003 - 08 - 30. 作者简介: 傅向华(1977 ~), 男, 博士生; 冯博琴(联系人), 男, 教授, 博士生导师. 基金项目: 国家高技术研究发展计划资助项目(2003AA1Z2610).

Web 资源采集的需要. 在线学习新下载页面可加速聚焦爬行过程,改善下载页面精度^[4]. 文献[5]和文献[6]讨论用未标注训练样本来改善分类器性能,文献[7]讨论了一种快速在线增强学习方法. 受其启发,本文提出一种只需少量标注训练集、可在爬行过程中自动更新学习模型和不断提高爬行精度的新方法.

1 聚焦爬行在线增量自学习原理

1.1 学习模型的建立

增强学习是在动态环境中通过回报与惩罚来学习最优策略的一种机器学习方法^[7,8]. 基于马尔可夫决策过程定义的增强学习模型包括环境状态集合 S 、动作集合 A 、状态转化函数 $T: S \times A \rightarrow S$ 和回报函数 $r: S \times A \rightarrow R$. 学习器的任务是学习一个策略 $\pi: S \rightarrow A$, 它基于当前观察到的状态 s_t 选择动作 $a_t = \pi(s_t)$, 使该策略具有最大的累积回报. 把爬虫在 Web 上根据链接爬行的过程看作是一个执行序列动作的过程, 每一个页面对应一个状态 s , 则从页面选择一个链接对应执行一个动作 a . 若所有 Web 页面的集合为 P , Web 页面上所有链接的集合为 U , 则可建立 Web 爬行的学习模型 (S, A, T, r) , 其中 $S = \{s_t | s_t \in P, t \in N\}$, $A = \{a_t | a_t \in U, t \in N\}$. 若链接页面与主题相关, 回报函数 $r(s_t, a_t) = 1$; 否则 $r(s_t, a_t) = 0$. 爬行从初始页面 p_t 开始搜索 Web, 通过选择链接序列 u_t, \dots, u_{t+n-1} 找到与主题相关的页面 p_{t+n} , 获得回报 r , 则序列 p_t, \dots, p_{t+n} 形成一个策略 π . 从 u_{t+n-1} 回溯, 形成一条链接链路, 按折算因子 $\gamma \in [0, 1]$ 计算, 链路上所有链接可获得折算回报, 链接获得的折算回报总和即为折算累积回报, 一般称为 Q 值. 将不同 Q 值的链接用于在线增量来训练多个 Q 值预测器, 用新搜索到的主题相关页面作为评估器的训练数据样本, 再用在线增量来改善评估器.

1.2 用快速 $Q(\cdot)$ 算法计算链接的最大累积回报

在 t 时刻, 页面 p_t 对应的状态为 s_t , 执行动作 a_t 获得直接回报 r_t , 那么根据策略 π 从初始状态 s_t 获得的折算累积回报可定义为

$$V(s_t) = \sum_{i=0}^{\infty} \gamma^i r_{t+i} \quad (1)$$

最优策略为 $\pi^* = \arg \max_{\pi} V(s_t), (\forall s_t)$. 为求得最优策略, 可令 $V^*(s) = \max_a Q(s, a)$, 这样当已知各动作的 Q 值后, 可直接选择具有最大 Q 值的动作来获得最优策略.

在链接的预测中, 除搜索直接链接的相关页面外, 还希望找到间接链接的页面, 则定义

$$Q(s_t, a_t) = (1 - \gamma) (Q_t^{(1)} + \gamma Q_t^{(2)} + \gamma^2 Q_t^{(3)} + \dots) \quad (2)$$

式中: $Q_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n V_{t+n}(s_{t+n})$ 为考虑前瞻 n 步的折算累积回报. 这种 $Q(\cdot)$ 函数使用常量 $(0 < \gamma < 1)$ 来合并从不同前瞻距离中获得的回报.

令 $e_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$, $e_t = r_t + \gamma V_t(s_{t+1}) - Q_t(s_t, a_t)$, 第 k 次更新 (s_t, a_t) 的学习率为 $\alpha_k(a_t, s_t)$, 当 α_k 较小时, $Q(\cdot)$ 的时间差分误差为 $e_t = e_t + \alpha_k(s_t, a_t) e_{t+1}$. 给定 s_t, a_t, r_t, s_{t+1} , $Q(\cdot)$ 算法可按

$$\hat{Q}(s_t, a_t) = \hat{Q}(s_t, a_t) + \alpha_k(s_t, a_t) e_t \quad (3)$$

更新规则迭代逼近 $Q(\cdot)$ 函数.

为了使式(3)顺利进行, 需引入适当跟踪 l 来增量计算时间差分误差. 令每个 (s, a) 对的适当跟踪 $l_t(s, a) = \sum_{i=1}^t \alpha(s, a)^i I_i(s, a)$, 其中 $I_i(s, a)$ 为指示函数. 如果 (s, a) 在时间 t 出现, 其值为 1; 否则为 0. 沿链接链路寻找相关主题页面, 并不必时时更新每个链接的 Q 值, 而是在找到相关主题页面时, 一次性更新链路上所有链接的 Q 值. 因而, 引入全局变量 $\tau_t = \sum_{i=1}^t \alpha(s, a)^i$, 用于记录累积时间差分误差; 并引入局部跟踪 $l_t = \frac{\tau_t I_t(s, a)}{\tau_t}$, 从而得到快速 $Q(\cdot)$ 的更新规则

$$\hat{Q}(s_t, a_t) = \hat{Q}(s_t, a_t) + \alpha_k(s_t, a_t) (l_t(s, a) (r_{t+1} - \tau_t)) \quad (4)$$

以快速计算链路上每个链接的 Q 值.

1.3 基于简单贝叶斯分类器学习未标注页面

假定在爬行的特定主题中有类别集合 C , 词典集合 V , 用 $|C|$ 表示集合中元素的数目. 文档 d_i 由词序列 $w_{d_1}, w_{d_2}, \dots, w_{d_{|d_i|}}$ 组成, $|d_i|$ 为文档的长度. 概率模型认为, 任意文档 d_i 服从参数为 θ 的多项式混合概率分布, 那么文档 d_i 属于类别 c_j 的概率为 $p(d_i | c_j, \theta)$. 根据简单贝叶斯方法的假设, 则有

$$p(d_i | c_j, \theta) = p(I | d_i) \prod_{k=1}^{|d_i|} p(w_{d_k} | c_j, \theta) \quad (5)$$

由训练集 D 可计算任意词 w_t 在 c_j 中的产生概率

$$p(w_t | c_j, \theta) = \frac{1 + \sum_{i=1}^{|D|} I_{ij} n(w_t, d_i)}{|V| + \sum_{s=1}^{|S|} \sum_{i=1}^{|D|} I_{ij} n(w_s, d_i)} \quad (6)$$



和每个类的概率

$$p(c_j | \hat{c}) = \frac{1 + \sum_{i=1}^{|D|} l_{ij}}{|C| + |D|} \quad (7)$$

式中: l_{ij} 为类标注, 如果文档 $d_i \in c_j$, 则 $l_{ij} = 1$, 否则 $l_{ij} = 0$; $n(w_t, d_i)$ 为 w_t 在 d_i 中出现的次数.

用最大后验 (MAP) 似然标准可求得分类器的参数 \hat{c} , 然后按

$$p(c_j | d_i, \hat{c}) = \frac{p(c_j | \hat{c}) \prod_{k=1}^{|d_i|} p(w_{d_i k} | c_j, \hat{c})}{\prod_{r=1}^{|C|} p(c_r | \hat{c}) \prod_{k=1}^{|d_i|} p(w_{d_i k} | c_r, \hat{c})}$$

对新文档进行分类. 计算 $c = \arg \max p(c_j | d_i, \hat{c})$ 即得 d_i 的类别.

刚下载的页面为未标注样本, 若将其添加到训练数据集中, 训练集 $D = D_L \cup D_U$, D_L 和 D_U 分别为标注训练样本集和未标注训练样本集. 如果 D_L 中的任意文档 $d_i \in c_j$, 则其类标注 $l_{ij} = c_j$, L 表示所有 l_{ij} 的集合. 假设 D 中各样本相互独立, 即

$$P(D | \hat{c}) = \prod_{d_i \in D_U} p(c_j | \hat{c}) p(d_i | c_j, \hat{c}) \cdot \prod_{d_i \in D_L} p(l_{ij} = c_j | \hat{c}) p(d_i | l_{ij} = c_j, \hat{c}) \quad (8)$$

为 D_U 中的每一个未标注训练样本文档引入类别标注 u_{ij} , U 为所有 u_{ij} 的集合, 则随机变量 U 的概率分布依赖于未知参数 \hat{c} 和已知数据 L . 令 $Z = L \cup U$, Z 为 U 定义的一个随机变量, 则有

$$\ln p(\hat{c} | D, Z) = \ln(p(\hat{c})) + \sum_{d_i \in D} \sum_{j=1}^{|C|} z_{ij} \ln(p(c_j | \hat{c}) p(d_i | c_j, \hat{c})) \quad (9)$$

可采用期望最大化 (EM) 算法求解式 (9). 由于在将新页面添加到训练样本集之前, 已经由评估器计算了页面的相关度, 因而保证了 EM 算法具有较好的初始值, 使搜索过程可以较快收敛.

2 在线增量学习爬行系统模型

根据上述原理建立的在线增量自学习爬行 (OIAC) 系统模型如图 1 所示. 与通用 Web 爬行系统相比, 图 1 增加了两个组件: 主题评估器和 Q 值预测器. 主题评估器用于评估新下载页面的主题相关度, Q 值预测器用于计算链接提供相关主题页面的可能性.

该系统运行包括: 预计算链接的 Q 值、评估新下载页面的主题相关性和在线增量训练. 首先根据

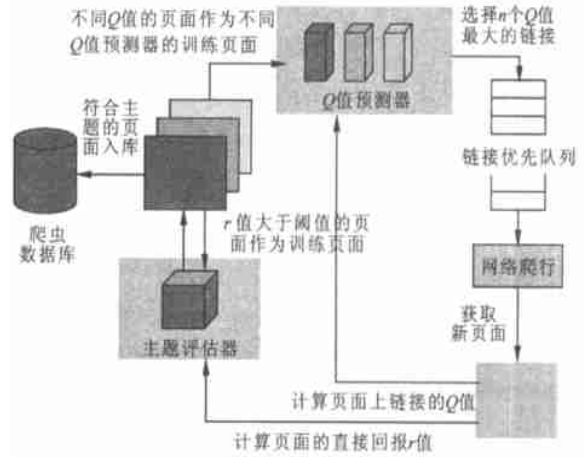


图 1 在线增量学习的聚焦爬行系统结构图

Q 值预测器预计算链接的 Q 值, 并根据预先设定的阈值进行过滤; 其次利用主题判别器评估新获取的页面的主题相关性, 如果找到主题相关页面, 则会产生回报, 再依据快速 $Q(\cdot)$ 算法将该回报值沿链路后向反馈; 接着选取已评估的页面作为新的训练样本来训练主题评估器, 选取链路上所有链接的特征文本训练 Q 值预测器. 为了避免主题评估器和 Q 值预测器的知识从零开始, 需预先用少量训练数据样本对主题评估器和 Q 值为 1 的预测器进行训练.

3 系统实现及实验

3.1 系统实现

根据上述的系统模型, 在 Windows 系统下用 Microsoft Visual Studio C++ 6.0 实现了一个在线增量自学习的网络爬行原型系统 WebBridge. 主题评估器采用简单多类贝叶斯分类器, Q 值预测器采用 5 个简单贝叶斯分类器, 对应的 Q 值分别为 0.2、0.4、0.6、0.8 和 1. 当爬虫从页面中提取出一个链接时, 根据该链接的特征文本计算 Q 值, 也就是分别计算链接的特征文本在 5 个 Q 值分类器中的概率, 然后以概率为权重求得加权平均值. 若链接反馈回多个 Q 值, 取其最大值. 主题评估器的计算和训练采用页面特征文本, Q 值预测器的计算和训练采用链接特征文本. 页面特征文本包括当前页面的标题、叙述正文, 链接特征文本包括锚文本、链接附近的文本, 根据 p 和 $1/p$ 进行区分. 在计算特征文本中的词频时, 先进行分词, 并去掉一些停用词. 为了减少开销, 对添加的训练样本采用批量处理. 系统实现参考了中国科学院计算所的汉语词法分析系统 ICTCLAS 和 bow 工具包.

3.2 实验

为了检验本文所提方法的有效性,将本文模型的某些功能去掉,分别形成标准聚焦爬行模型(STDC)和扩展标准聚焦爬行模型(STDC+),然后比较3种模型.实验中选择 $\alpha=0.5$, $\beta=0.2$,链接深度为5.选择的评测指标为搜索页面中主题相关页面的比率、访问链接数和下载页面的比率.实验平台为Windows2000,CPU为PIV 1.4 GHz,内存为256 MB,实验主题为“视觉艺术”.用网站下载工具从雅虎中国(<http://cn.yahoo.com>)分类目录中的“视觉艺术”类别目录(<http://cn.dir.yahoo.com/Arts-and-Humanities/>)下载了1347个典型页面,将其作为离线训练的数据.为了与STDC+进行对比,检测在线增量学习的有效性,OIAC随机选择50个初始页面作为主题评估器和Q值预测器的训练数据.每一种爬行模型均从雅虎中国的主页开始下载页面,在程序中记录访问的链接数和下载的页面数,以最终下载的最大相关页面数为1计算回调率.实验结果如图2和图3所示.

3.3 实验结果讨论

由图2可以看出,OIAC最初下载的主题精度比较低,但随着下载页面的增加,由于可以不断地将新下载页面作为评估器的训练数据样本,评估器性能得到改善.当下载到2400个页面时,OIAC的精度已接近STDC的;下载到4400个页面时,精度接近STDC+的,最后其精度超过了STDC+的.另外

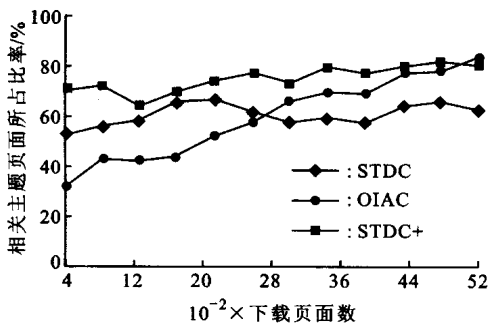


图2 相关主题页面占下载页面的比率

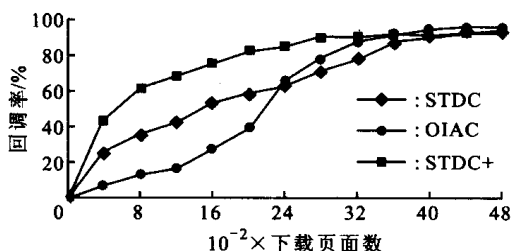


图3 相关主题页面的回调率

还可以看出,由于引入了增强学习方法,使爬虫可以搜索到具有间接相关性的页面,帮助爬虫更准确地确定页面主题.无论是OIAC还是STDC+,都具有比STDC更高的页面下载精度.从图3可以看出,由于OIAC最初的训练集较小,其所具有的主题相关知识较少,所以在开始的一段时间Q值预测器准确度较低,因而回调率比STDC、STDC+的都低,访问链接数目较多.随着Q值预测器准确度的提高,OIAC的回调率慢慢增加,随之搜索到更多的相关页面.

4 结论

本文将半监督简单贝叶斯分类器和在线快速Q()增强学习方法引入Web爬行,提出了一种新的具有在线增量自学习能力的聚焦爬行方法.采用该方法,无需大量标注训练样本,可在线增量学习新知识,积累爬行过程中的经验.实验结果表明,采用在线增量学习的聚焦爬行方法可以更好地提高下载页面的精度和回调率,相对于离线训练的聚焦爬行方法,它更符合Web信息采集的要求.

参考文献:

- [1] McCallum A, Nigam K, Rennie J, et al. Building domain-specific search engine with machine learning techniques [A]. AAAI Spring Symposium on Intelligent Agents in Cyberspace, Stanford University, USA, 1999.
- [2] Chakrabarti S M, van den Berg H, Dom B. Focused crawling: a new approach to topic-specific Web resource discovery [J]. Computer Networks, 1999, 31(11-16): 1623-1640.
- [3] Diligenti M, Coetzee F M, Lawrence S, et al. Focused crawling using context graphs [A]. 26th International Conference on Very Large Database, Cairo, Egypt, 2000.
- [4] Chakrabarti S, Kunal P, Mellela S. Accelerated focused crawling through online relevance feedback [A]. The Eleventh International Conference on World Wide Web, Hawaii, USA, 2002.
- [5] Nigam K. Using unlabeled data to improve text classification [D]. Pittsburgh, USA: School of Computer Science, Carnegie Mellon University, 2001.
- [6] 宫秀军, 史忠植. 基于Bayes潜在语义模型的半监督Web挖掘 [J]. 软件学报, 2002, 13(8): 1508-1514.
- [7] Wiering M, Schmidhuber J. Fast online Q() [J]. Machine Learning, 1998, 33(1): 105-115.
- [8] Jing Peng, Williams R. Incremental multi-step Q-learning [J]. Machine Learning, 1996, 22(1-3): 283-290.

(编辑 苗凌)