

文章编号:1001-9081(2006)12-3012-03

一种基于锚文本和改进 C4.5 决策树算法的主题爬行方法

刘金红, 陆余良

(解放军电子工程学院 网络系, 安徽 合肥 230037)

(wondergoldff@gmail.com)

摘 要:提出了一种基于锚文本和改进 C4.5 决策树算法的主题爬行方法:基于锚文本词项集训练决策树,然后基于决策树模型来计算网页的主题相关性和待爬行 URL 的优先级顺序。最后,应用该方法在四所大学网站网页数据集上针对“学术报告”主题进行了主题爬行实验,并与两种标准的网络爬虫进行了性能对比,实验结果验证了该方法的有效性。

关键词:主题网络爬虫;锚文本;决策树

中图分类号:TP391 **文献标识码:**A

Focused crawling method based on improved C4.5 exploiting anchor text

LIU Jin-hong, LU Yu-liang

(Department of Network, Institution of Electric Engineer, Hefei Anhui 230037, China)

Abstract: A new focused crawling method based on anchor text and improved C4.5 decision tree algorithm was proposed. It exploited the anchor text of URL to train the decision tree, and then applied the decision tree model to decide whether a downloaded page was on topic and how to choose the next URL to visit. Finally, a prototype system named DTFC based on this method was implemented, and experiments in four university websites were carried out in allusion to "academic report". The experimental results show that DTFC outperforms two standard crawlers for focused crawling.

Key words: focused crawler; anchor text; decision tree

0 引言

传统的 Web 信息采集的目标是尽可能多地采集信息页面,甚至是整个 Web 上的资源,而在这一过程中它并不太在意页面采集的顺序和被采集页面的相关主题。这样做的一个好处是能够集中精力在采集的速度和数量上,并且实现起来也相对简单。但是,这种传统的采集方法也存在着很多缺陷:随着 Web 信息的爆炸性增长,信息采集的速度越来越不能满足实际应用的需要;最近的试验表明,即使大型的信息采集系统,它对 Web 的覆盖率也只有 30%~40%^[1];待刷新页面数量的巨大也使得很多采集系统刷新一遍需要数周到一个月的时间^[2,3];并且传统信息采集需要消耗非常多的系统资源和网络资源,而对这些资源的消耗并没有换来采集到页面的较高利用率,事实上,它们中有相当大的一部分利用率很低。

一个好的解决方法就是采用主题采集,通过减小采集页面的数量,从而减小刷新一遍的时间和已采集页面的失效率。与传统搜索引擎不同,主题搜索引擎仅仅需要从 WWW 上爬行主题相关的网页,因此,研究者提出了高效抽取主题相关网页的主题网络爬虫^[4,5]。如何计算下载网页的主题相关性与待爬行 URL 的爬行顺序是主题网络爬虫需要解决的两个基本问题^[6],为此,本文提出了一种基于锚文本和改进 C4.5 决策树算法的主题爬行方法。该方法基于以下两个前提:1)大多数情况下,超链上的锚文本是目标网页最好的内容描述和摘要;2)传统的搜索引擎和主题爬虫都忽略那些低入度的网页,因此就忽略了许多低入度的主题相关网页。最后,以四所大学网站网页作为数据集,以“学术报告”为实验主题,通过对比基于本文方法的主题网络爬虫 DTFC 和其他两种标准网络爬虫的主题爬行性能,验证了本文方法的有效性。

1 基于锚文本和决策树的主题爬行方法

1.1 爬行空间假设

对本文的主题网络爬虫爬行的网络空间有两个假设:

1)假定网络爬虫在一个受限的 URL 领域内爬行,比如一所大学校园网或一个企业的内部网络;

2)认为存在一个到达主题相关网页 URL 的入口网页,比如一所大学校园网的主页。

因此,本文设计主题网络爬虫就是从入口网页开始,在一个受限领域的网页集合中进行主题爬行。本文使用 $G = (V, E, r)$ 来描述该受限领域的网页集合对应的 Web 图,其中 V 表示该领域包含的网页集合, E 表示 V 中网页之间的超链接集合, r 表示入口网页。

1.2 主题爬行模型

与整个网页中的文字内容相比,锚文本中的词项(即特征)很少,并且噪音数据也比较少。该方法基于以下两个前提:1)大多数情况下,超链上的锚文本是目标网页最好的内容描述和摘要;2)传统的搜索引擎和主题爬虫都忽略那些“低入度”的网页,因此就忽略了许多低入度的主题相关网页。为了有效利用锚文本中所包含的主题相关信息,基于改进的 C4.5 决策树模型^[8]来预测目标网页的主题相关性和待爬行 URL 的优先级排序。改进的 C4.5 决策树模型算法(Improved C4.5 algorithm, I-C4.5)的基本思想如下:

1)对每一个属性特征(这里仅仅是指离散值的属性特征),属性特征的每个已知的值对应一个样本子集,计算样本子集的熵;

2)计算样本子集熵的平均值,并将样本子集熵的值不小于平均值的样本子集进行合并,形成一个临时的复合样本子

收稿日期:2006-06-05;修订日期:2006-08-14

作者简介:刘金红(1978-),男,山西永济人,博士研究生,主要研究方向:自然语言处理、Web 信息挖掘; 陆余良(1964-),男,江苏宜兴人,教授,硕士,主要研究方向:Web 信息挖掘、网络安全。

集,计算该复合样本子集的熵值;

3) 利用上述复合样本子集的熵值和未合并样本子集的熵值计算该节点的修正信息增益 (Information Gain, IG), 选择具有最高修正信息增益的属性特征作为当前节点的测试属性特征,其分枝对应于未合并样本子集和复合样本子集;

4) 树构造的其余部分与 C4.5 决策树模型算法相同。

I-C4.5 决策树模型对具有较高熵值的分枝进行合并,根据熵的定义,熵值越大,子集划分的纯度越小,将这些对划分无贡献的分枝合并,有效减少了无意义的分枝。在 C4.5 决策树模型中,由于需要对这些无贡献的分枝进行进一步的划分,会加速导致碎片的产生,从而在整个决策树模型中最终产生大量的空枝,以及具有极少节点的叶子节点,这些也是决策树学习导致过度拟合的主要根源之一,而改进后的 I-C4.5 决策树模型均有效避免了上述问题。

1.3 训练样本和特征选择

对于一个受限领域网页集合对应的 Web 图,本文通过以下三步来选择训练样本和特征集合:

步骤 1: 爬行 V 中的所有网页,并基于事先训练好的支持向量机 (Support Vector Machines, SVM) 分类器标识出所有主题相关网页, SVM 分类器是根据用户事先收集好的一些主题相关网页和主题不相关网页集合来训练得到的。为了方便起见,以下使用函数 C 表示 SVM 分类器,则对于网页 V 可以得:

$$C(V) = \begin{cases} \text{true} & \text{如果 } V \text{ 被 SVM 分类器分类为主题相关网页} \\ \text{false} & \text{否则} \end{cases}$$

步骤 2: 对于集合 V 中的任意网页 $t \in \{V | C(v) = \text{true}, v \in V\}$, 通过 Dijkstra 算法^[7] 计算从入口网页到 t 的最短路径,以下用集合 S 来表示所有具有最短路径的网页集合。

步骤 3: 设 $L = (b, e) \in E$ 表示一个起始链接,这里 b 和 e 分别表示源网页 (又称为父亲网页) 和目标网页 (即主题相关网页)。函数 $f(l)$ 返回超链接 l 相关联的锚文本。使用集合 $P = \{f(l) | l = (b, e) \in E \wedge b \in S \wedge e \in S\}$ 和集合 $N = \{f(l) | l = (b, e) \in E \wedge b \in S \wedge e \notin S\}$ 分别作为正样本集合和负样本集合,以 P 和 N 作为训练样本来训练决策树。这里,忽略那些锚本文为空的超链接。图 1 描述了本文使用的训练样本,其中灰色实心圆表示主题相关网页,虚线圈表示在最短路径上的不相关网页。实线边上的锚文本表示正样本,而虚线边上的锚文本表示负样本。这里用 $F = F_p \cap F_n$ 表示样本空间,其中 F_p 表示在 P 中至少出现一次的词项集合, F_n 表示在 N 中至少出现一次的词项集合。

1.4 决策树训练

给定一个超链对应的锚文本 a , 设函数 $g(a)$ 返回 a 中出现的所有特征 (词项) 集合, 则有 $g(a) \subseteq F$, 这样, 本文构建的决策树可以用一个布尔函数 $B(g(a))$ 来表示, $B(g(a))$ 通过在 2.3 节中的正、负训练样本中应用改进的 C4.5 算法^[8] 训练得到。如果存在一个词项集 s 不能将正样本和负样本很好地区分开, 并且没有其他的词项可用来区分这些样本, 则基于概率来重新定义 $B(s)$, 具体定义如下:

设 P_s 和 N_s 分别表示包含词项集 s 中所有词项的正负样本集, 如果 $\frac{|P_s|}{|P|} > \frac{|N_s|}{|N|}$, 即正样本发生的可能性比负样本更大, 此时 $B(s) = \text{true}$, 否则 $B(s) = \text{false}$ 。

本文采用中科院计算所的 ICTCLAS 分词系统来实现其中的 $g(a)$ 函数功能, 以此来获取锚文本的 (特征) 集合。

本文实验使用基于以上训练方法得到的 $B(g(a))$ 函数来决定待爬行 URL 的优先级顺序和网页的主题相关性, 并将该主题网络爬虫实验系统称为决策树爬虫 (Decision Tree Focused Crawler, DTFC)。

2 系统实现及实验

2.1 系统实现与实验设计

基于上述主题爬行模型和 WebSPHINX^[9] 个性化爬虫框架, 在 Windows 系统下基于 Java 语言实现了一个基于锚文本和改进 C4.5 决策树算法的主题网络爬虫原型系统 (DTFC), 其中的主题评估器采用 SVM 分类器, 待爬行 URL 的优先级基于改进的 C4.5 决策树算法进行计算, 主题评估器的计算和训练采用页面特征文本, 而计算待爬行 URL 的优先级则采用链接文本特征。页面特征文本包括当前页面的标题、元数据和叙述正文, 链接特征即锚文本的词项集合。在计算 $g(a)$ 函数即获取锚文本的词项集合时, 先基于 ICTCLAS 系统进行分词, 然后取掉一些停用词以减少噪音数据和计算开销。

为了检验本文所提方法的有效性, 将本文主体爬虫系统 DTFC 的某些功能去掉, 分别形成宽度优先网络爬虫 (Breadth-First Crawler)^[6] 和标准的主题网络爬虫 (Standard-Focused Crawler, 即实现文献[4]提出的主题网络爬虫, 它对于主题相关网页包含的 URL 给予较高的优先级), 然后在相同数据集上比较三种模型的主题爬行性能。实验选择的评测指标为爬行页面中主题相关网页数目与爬行页面总数的比率, 实验平台为 Windows XP, CPU 为 2.0 GHz, 内存为 512M。

由于文献[4]中主题爬行实验使用的 Yahoo! (www.yahoo.com) 目录式网页数据已经不存在, 并且面向的是英文主题爬行, 而本文面向的是中文主题信息爬行, 所以实验选择四所熟悉的中文大学网站的网页作为实验数据集, 分别是: 清华大学 (TSH)、北京大学 (PKU)、浙江大学 (ZJU) 和哈尔滨工业大学 (HIT), 并选择“学术报告”信息作为目标主题进行主题爬行实验。主题爬行前预先用网站下载工具从浙江大学网站的“校园动态”类别 (<http://www.zju.edu.cn/xydt.htm>) 目录下下载了 1200 个典型页面作为离线训练数据来训练 SVM 分类器。实验结果如图 2 和图 3 所示。

2.2 实验结果分析

1) 相同数据集上的性能分析

使用 ZJU 大学的网页同时作为训练集和测试集, 图 2 给出了 DTFC 主题爬虫的性能与其他两种爬虫的性能比较曲线图。曲线图上每一点对应的 X 轴表示所有已经爬行的网页数目, 而 Y 轴表示主题相关网页的数目。图中还给出了仅仅爬行最短路径上网页的理想爬虫性能曲线图。由实验对比曲线可以看出, DTFC 主题爬虫的性能高于其他两种爬虫。为了爬行 50% (召回率) 的主题相关网页, DTFC 只需要爬行整个网页集合的 15%, 而且对实验数据统计发现大多数入口网页到主题相关网页的最短路径长度在 5 到 8 之间, 这说明基于改进的 C4.5 决策树算法导引的主题爬虫可以有效地爬行深层次的主题相关网页。

2) 不同数据集上的性能分析

为了评估 DTFC 的泛化能力, 基于 ZJU 的网页作为训练样本来训练 DTFC 决策树模型, 而基于其他三所大学的网页作为测试数据集设计了 DTFC 主题爬行对比实验。通过实验表明, 基于 DTFC 的主题爬虫性能远高于宽度优先搜索主题网络爬虫, 通过与标准的主题爬虫进行比较, DTFC 主题网络爬虫在 HIT 和 TSH 两所大学网页集合中爬行性能整体要高于标准的主题爬虫, 而在 PKU 大学网页上爬行效果比较差。图 3 给出了在 HIT 大学网页数据集上 DTFC 与其他两种标准

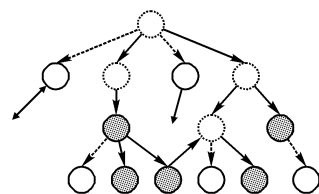


图 1 主题相关网页与最短路径

网络爬虫的爬行性能对比实验数据。

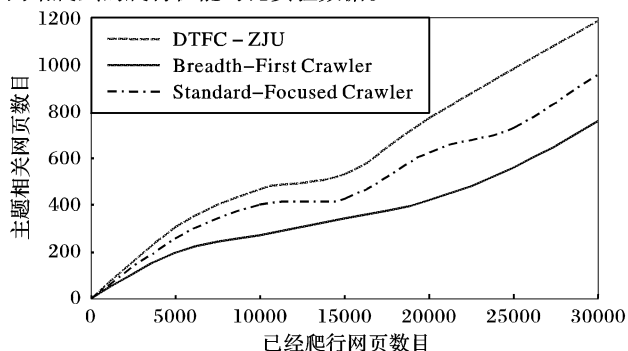


图2 相同数据集上的爬行性能对比

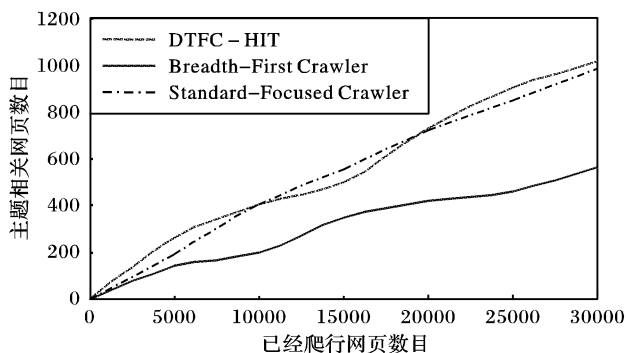


图3 不同数据集上的爬行性能对比

3 结语

研究了面向主题信息采集的主题爬行方法,设计实现了一个实验系统 DTFC。通过在四所大学网页数据集上与其他两种典型的网络爬虫进行对比实验,实验数据初步验证了基

于锚文本和改进 C4.5 决策树算法的主题爬行方法的有效性。但当前实验数据规模还比较小,并且决策树的泛化能力还比较弱,下一步需要对该决策树模型进一步改进,以适合处理大规模的、质量参差不齐的 Web 海量数据的要求。

参考文献:

- [1] 李盛韬,赵章界,余智华. 基于主题 Web 信息采集系统的设计与实现[J]. 计算机工程, 2003, 29(17): 102-104.
- [2] AGGARWAL C, AL-GARAWI F, YU P. Intelligent Crawling on the World Wide Web with Arbitrary Predicates[A]. Proceedings of the 10th International WWW Conference[C]. Hong Kong, 2001.
- [3] BRIN S, PAGE L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[A]. Proceedings of the Seventh International World Wide Web Conference[C]. Brisbane, Australia, 1998.
- [4] CHAKRABARTI S, VAN DEN BERG M, DOM B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery [A]. The Eighth International World Wide Web Conference[C]. Toronto, Canada, 1999.
- [5] DILIGENTI M, COETZEE FM, LAWRENCE S, et al. Focused Crawling Using Context Graph[A]. Proceedings of the 26th VLDB Conference[C]. 2000. 527-534.
- [6] ALTINGÖVDE IS, ULUSOY Ö. Exploiting Interclass Rules for Focused Crawling[J]. IEEE Intelligent Systems, 2004, 19(6): 66-73.
- [7] ZHAN FB. Three Fastest Shortest Path Algorithms on Real Road Networks[A]. Journal of Geographic Information and Decision Analysis, 1997, 1(1): 69-82.
- [8] 刘奕群,张敏,马少平. 基于改进决策树算法的网络关键资源页面判定[J]. 软件学报, 2005, 16(11).
- [9] MILLER RC, BHARAT K. SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers[A]. Proceedings of WWW7[C]. Brisbane Australia, 1998.

(上接第 2949 页)

是最优解,即此时积分参数采用 3.18,微分参数采用 0.89。

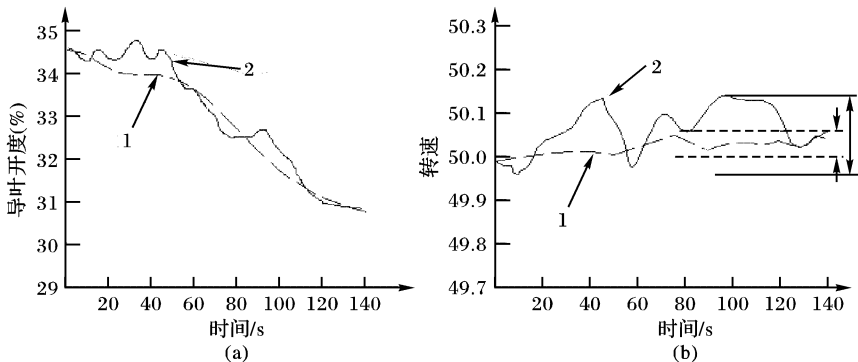


图4 两种模型性能对比

图4 为仅采用定值 PID 参数和采用融合策略的 PID 模型的水轮机系统在同样的机组初始条件下,在甩负荷过程中转速和导叶开度变化曲线。由对比可知,采用融合决策的调速器模型在收敛速度及动态误差等方面均优于传统的调速器模型,有效克服了原 PID 模型参数整定困难而导致的系统转速和频率振荡等问题。

4 结语

在多传感器信息融合过程中,引用寻求交互博弈解的方法来解决不同传感器或局部决策之间信息的冲突或矛盾。本

文基于博弈融合模型的融合过程是由局部决策构建交互策略,由交互策略经融合决策得出全局决策两个变换完成。并

且在融合算法上采用了博弈的算法。在丰满水电控制系统的调速器 PID 模型的参数调节中,通过应用上述算法明显改善了原模型收敛慢和动态误差大等问题。在后续研究中将对博弈过程中的期望效用函数进行研究,以实现大时滞、非线性复杂控制系统中 PID 参数的全局最优动态调节。

参考文献:

- [1] FEDOTOV GA. Information fusion for turbulence measurements in hydrophysical applications[A]. Proceedings of the 4th International Conference on Information Fusion [C]. Montreal: LM Canada, 2001. 3-9.
- [2] KOPISAARI P, SAARINEN J. Bayesian networks for target identification and attribute fusion with JPDA[A]. Proceedings of the Second International Conference on Information Fusion[C]. Sunnyvale: Omnipress, 1999. 763-770.
- [3] GOEBEL KF. Conflict resolution using strengthening and weakening operations in decision fusion[A]. Proceedings of the 4th International Conference on Information Fusion[C]. Montreal: L M Canada, 2001. 1-3.
- [4] MYERSON R. 博弈论[M]. 于震,费剑平,译. 北京: 中国经济出版社, 2001. 295-313.
- [5] 杜庆东,徐凌云,赵海. 基于分布式结构的判决反馈数据融合算法[J]. 东北大学学报(自然科学版), 2001, 22(4): 385-388.

表1 调速器 PID 参数

组别	1	2	3
积分参数	3.18	3.25	3.60
微分参数	0.89	0.85	0.77