基于改进决策树算法的网络关键资源页面判定*

刘奕群+, 张 敏, 马少平

(智能技术与系统国家重点实验室(清华大学),北京 100084)

Web Key Resource Page Judgment Based on Improved Decision Tree Algorithm

LIU Yi-Qun+, ZHANG Min, MA Shao-Ping

(State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62777699, E-mail: liuyiqun03@mails.tsinghua.edu.cn, http://www.tsinghua.edu.cn

Received 2004-07-26; Accepted 2005-06-02

Liu YQ, Zhang M, Ma SP. Web key resource page judgment based on improved decision tree algorithm. *Journal of Software*, 2005,16(11):1958−1966. DOI: 10.1360/jos161958

Abstract: Key resource page is one of the most important search target pages for Web search users. Decision tree learning is one of the most widely-used and practical methods for inductive inference in machine learning. Because of the difficulty in uniform sampling of Web pages, there are not enough negative instances for training a key resource decision tree. To solve the problem, the original algorithm is partly modified to learn from global instead of individual instance information. With the same evaluation method as TREC (Text Retrieval Conference) 2003, large scale retrieval experiments based on improved decision tree algorithm achieves more than 40% improvement than the ones based on the original algorithm. It not only offers an effective way for selecting Web key resource pages, but also shows a possible way to improve decision tree learning performances.

Key words: Web information retrieval; key resource page; machine learning; decision tree

摘 要: 关键资源页面是网络信息环境中一种重要的高质量页面,是用户进行网络信息检索的主要目标.决策树算法是机器学习中应用最广的归纳推理算法之一,适用于关键资源页面的判定.然而由于 Web 数据均一采样的困难性,算法缺乏有足够代表性的反例进行训练.为了解决这个问题,提出一种利用训练样例的统计信息而非个体信息进行学习的改进决策树算法,并利用这种算法实现了独立用户查询的关键资源页面判定.在 2003 年文本信息检索会议(Text Retrieval Conference,简称 TREC)标准的评测条件下,基于此种改进决策树算法的大规模网络信息检索实验获得了超过基本算法 40%的性能提高.这不仅提供了一种查找 Web 关键资源页面的有效方式,也给出了提高决策树算法性能的一个可行途径.

关键词: 网络信息检索;关键资源页面;机器学习;决策树

* Supported by the National Natural Science Foundation of China under Grant Nos.60223004, 60321002, 60303005 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318108 (国家重点基础研究发展规划(973)); the Key Project of Chinese Ministry of Education under Grant No.104236 (国家教育部科学技术研究重大项目资助)

作者简介: 刘奕群(1981−),男,山东济南人,博士生,主要研究领域为信息检索,机器学习;张敏(1977−),女,博士,讲师,主要研究领域为机器学习,信息检索;马少平(1961−),男,博士,教授,博士生导师,主要研究领域为知识工程,信息检索,汉字识别与后处理,中文古籍数字化.

中图法分类号: TP393 文献标识码: A

在网络信息极大丰富的情况下,网络信息检索工具越来越成为人们访问互联网资源的主要媒介.2004 年 7 月发布的第 14 次中国互联网络发展状况统计报告指出,搜索引擎是 64.4%的网络用户经常使用的功能,也是 71.9%的用户获取信息最常用的方法.然而当前网络信息检索工具存在的主要问题是,返回给用户的查询结果动辄成千上万,但这些结果又与用户的真实查询需求大相径庭.为此,从上个世纪末开始,关键资源页面查找越来越成为网络信息检索研究关注的重点[1−3].关键资源页面本身提供的信息有限,但它通过提供给用户一个关于某个主题的可靠信息的入口,协助用户快捷地查找到所需要的信息.根据 Broader 在文献[4]中的工作,当前网络信息检索中超过 80%的检索需求都可以用关键资源查找技术加以实现.在这一大部分检索需求中,只有关键资源页面是用户需要的结果,因此,事先判定关键资源页面对于网络信息检索工具提高其信息收集的有效性十分重要.

因此,关键资源判定成为网络信息检索研究中重点考察的问题之一.在研究中发现,网页的一些主题无关属性可以用于关键资源判定,如网页的链接特性、文档结构特性等等,而决策树学习的方式适用于综合这些特征进行判定关键资源的任务.传统的决策树学习算法需要一定数目的正例和反例样本进行学习.在判定关键资源的工作中,正例即关键资源页面样本的获得,可以通过手工收集的方式完成,相对容易,但由于页面不能成为关键资源页面的原因多种多样,而 Web 页面的均一采样本身就是世界性的难题(Google 公司的 Henzinger 等人在文献[5]中称其为网络搜索引擎面临的一个重大挑战),因此获得有充分代表性的反例样本集合十分困难.为了解决这个问题,必须从统计特征而不是个体特征的角度进行学习,对传统的决策树学习算法进行改进.

本文第 1 节讨论相关研究工作,阐明在关键资源页面判定的现有研究成果和决策树学习算法的主要思路和实现方式.第 2 节简要介绍关键资源页面的特性和判定所使用的查询主题无关属性.第 3 节对改进的决策树学习算法进行推导,并说明利用这个算法进行关键资源页面判定的具体操作.第 4 节给出改进算法与传统算法的比较实验和检索性能评价结果.最后总结并列出主要结论.

1 相关研究工作概述

1.1 关键资源页面判定方面的已有成果

关键资源和主题过滤任务的概念来源于 Kleinberg 在文献[6]中提出的 HITS 算法以及内容权威度(authority value)和链接权威度(hub value)的概念.主题过滤是查找特定主题的关键资源的检索任务的特定称谓.而 Bhara[3],Chakrabarti[7,8]和 Amento[1]等人随后的努力则把关键资源的概念大为扩展,并给出了最初的主题过滤算法和评价算法性能的一整套评价方法.主题过滤实际上是对搜索引擎用户普遍行为的一个抽象描述.它的提出,实际上制定了网络搜索引擎提高性能的一个新方法和新目标.而当前几乎所有成功的网络搜索引擎也确实都已经在很大程度上依赖主题过滤的方法而不是传统的相关度检索方法来进行信息查找.一个有效的主题过滤算法应当尽可能地在检索过程之前实现查询主题无关的关键资源定位(类似于 PageRank 算法主题无关的计算页面的重要性),同时这个算法应当充分利用链接关系之外的页面主题无关特征来提高性能,这两方面就是本文所提出的利用决策树和主题无关信息的关键资源定位算法的出发点.

1.2 决策树学习算法的基本思路和实现

决策树算法本身的特点使其适合进行特征数较少情况下的高质量分类,因而适用于仅仅利用主题无关特征进行学习的关键资源定位任务.根据 Mitchell 在文献[9]中的论述,现阶段的大多数决策树学习算法是一种核心算法的变体,即采用自顶向下的贪婪搜索遍历可能的决策树空间.这种方法是 Quinlan 分别在 1986 年提出的 ID3 算法和 1993 年提出的 C4.5 算法的基础.

决策树算法的核心问题是选取在树的每个结点要测试的属性,争取能够选择出最有助于分类实例的属性.为了解决这个问题,ID3 算法引入了信息增益的概念,并使用信息增益的多少来决定树的不同结点需要测试的

属性.但这种做法存在着对取值情况较多的属性有所偏袒的问题.针对这个问题,C4.5 算法用信息增益率代替信息增益作为评价属性分类能力的度量,从而避免使算法倾向于优先选择分支多的属性.C4.5 还可以通过自动离散化的方式处理取值连续的属性,并借助决策树修剪的方式减少过学习情况的出现.

除了 ID3 和 C4.5 之外,比较著名的决策树学习还包括 Friedman 提出的 CART 算法以及 Kononenko 等人提出的 ASSISTANT 算法,此外还包括 Mingers 进行的关于属性选择和决策树修剪策略的详细比较研究等[9].国内的相关工作则包括刘小虎等人提出的基于考虑两层节点带来的信息增益的改进算法[10]以及洪家荣等人在文献[11]中提出的基于概率的决策树学习算法 PID.这些算法可能对某些特定的学习问题进行了优化,但其主要架构都与 ID3 及 C4.5 算法基本保持一致,因此不再赘述.

2 关键资源判定使用的 Web 页面属性

决策树判定所需要的训练实例是由〈属性,值〉来表示的,具体到关键资源判定的问题,则是以若干固定的 Web 页面属性和对应的网页是否关键资源页面的判定值作为实例的描述.

已有研究[12]表明,Web 页面的许多查询主题无关属性可以用于页面分类的依据.在网页普遍意义的质量判定上采用查询主题无关属性与内容相关属性比较有明显的优势:内容相关的属性往往是与网络信息检索用户的查询主题息息相关的,因此按照内容对某一个查询质量高的页面,不一定对其他查询有同样好的效果.而查询主题无关属性则是与用户查询主题相互独立的,它研究的是页面被各种需求的用户都看好的可能性.因此,利用查询主题无关属性进行网页分类,尤其是定位站点主页,成为近年网络信息检索研究的热点[13,14].

根据对 TREC(Text Retrieval Conference)挑选出的关键资源页面集合的统计特性考察,以及对关键资源页面功能的分析工作[15],本文为判定关键资源页面所采用的查询主题无关属性,可以包括以下几种:

(1) 页面长度属性:是指经过过滤无用字符等预处理之后的页面所包含的单词数.

(2) 入链接个数属性:又称入度,是指某页面被多少个外部页面链接引用的度量.

(3) URL 长度:是表示页面 URL 种类的一个主题无关特征,由 Kraaij 等人在文献[16]中提出,页面的 URL 长度划分为 4 个级别,其中级别 1 的 URL 只包括域名,而从级别 1 到级别 4,URL 中的"/"逐渐增多.

(4) 站点自身出链接个数:是指页面指向其所在站点内部其他页面的链接数目的多少.

(5) 站点自身出链接文本比率:是指站点自身出链接的链接文本(anchor text)占页面内容文字的比例.

采用这些属性,是因为关键资源页面与普通页面在这几个属性上有明显的取值差别,见表 1.

Table 1 Differences in average values of non-content features between ordinary pages and key resource pages

表 1 普通页面与关键资源页面的主题无关特征平均值差异

Page attributes	Ordinary pages	Key resource pages
Page length	7 037.43	9 008.02
In-Link count	9.94	153.12
URL classification	3.851 6	3.073 4
In-Site out-link anchor text rate	0.061 8	0.124 0
In-Site out-link number	17.58	37.70

在表 1 中,普通页面集合特性的统计选用了页面量达 1.25M 的.GOV 语料库(http://es.csiro.au/TRECWeb/govinfo.html),而关键资源页面样例则选用了 TREC2002 主题过滤任务[13]的答案集,两者都是国际上通用的大规模网络信息检索研究语料库,有比较强的代表性和可信度.从表 1 可以看出,普通页面与关键资源页面在各个主题无关属性的取值上有明显的差别,这是采用决策树方法能够进行关键资源页面判定的基础.

3 决策树算法的改进和关键资源页面判定

3.1 传统决策树算法处理关键资源判定的优势与困境

决策树学习(decision tree learning)的方法可以适用于进行 Web 页面主题无关特征的综合,这是由决策树算法自身的一些特点所决定的.决策树学习适合解决目标函数具有离散输出值的问题,而且往往是特征数目较少

时解决此类问题的最简单、最有效的途径之一.此外,决策树学习中从根节点到下级节点选择特征的先后顺序,自然给予了这些特征排序关系,这也为评价特征定位能力的高低提供了重要的参考.

决策树算法与一般的机器学习算法一样,都需要利用训练数据进行学习.这就要求训练数据与真实需要进行判定的数据有较大的相似性,或者说,训练样例的组成要与真实样例尽可能地接近,以保证在将训练样例得到的决策树应用于真实样例处理时,真实错误率与训练错误率不会有太大的偏差.

在关键资源判定的问题上,训练样例的获取是主要的困难.由于关键资源页面有较为明确的定义,人们对这类页面也有较明确的感性认识,因此正例相对较易获得,特别是文本信息检索会议(TREC)从2002年开始将关键资源页面查找作为其网络信息检索任务的主要内容,可以简单地采用 TREC 的标准答案页面作为正例页面的可靠来源.相比而言,反例的获得就困难得多,Web页面多种多样,作为Web页面中绝大部分的非关键资源页面更是数目繁多,情况繁杂.内容不可信,内容太少,没有高质量的入链接或者出链接,有太多广告信息等等,都可能是页面无法成为关键资源页面的理由.Google公司的Henzinger等人近年来多次指出:Web页面的均一采样是当前在理论和应用上都无法克服的困难[5].因此,获得高质量的反映Web真实情况的反例绝非易事.当然,这种困难并非决策树学习的算法所特有,也是一切要处理关键资源判定的机器学习方法所必须面对的问题,因此它的解决也显得格外重要.

反例获取困难的问题既然短期不可能得到解决,则只能试图避开这个问题,这就要求对决策树学习的过程进行重新理解.在决策树建立的过程中,更多需要的不是训练样例的个体信息,而是像熵、信息增益(信息增益率)这样的统计信息.如果对于决策树的每个节点都可以给出正、反例的所有统计信息,如某个属性为某个特定值的样例的比例等,则完全没有必要使用个体信息.在关键资源判定中,由于正例页面较易获得,因此其统计信息也是不难得到的,而基于对大规模语料库的统计获得 Web 页面总的统计信息也是可能的.

由于在Web页面集合中,每一个页面是否为关键资源是一个确定的论断,因此,如果对Web页面集合的全体进行考察,是可以计算出关键资源页面所占的数目以及比例的.关键资源页面在全体页面集合中的比例因而是一个确定的常数,这个常数的大小取决于页面集合的页面质量,页面质量高的集合对应的这个常数也就高.更一般地,如果定义“关键资源比率”为某个页面集合中,单位页面数所包含的关键资源页面数目,则这个指标可以衡量某个集合中关键资源的“密度”,也就是页面集合质量的高低.关键资源比率的形式定义为

$$Key\ Resource\ Rate = \frac{\#Key\ resource\ page}{\#Page} \tag{1}$$

在下文中,我们将整个.GOV 数据集合的关键资源比率设为 K,则根据并集的特性有下面的公式成立:

$$R^{Whole} = K \cdot R^{Key} + (1-K) \cdot R^{Non-Key} \tag{2}$$

其中,R^{Whole}是属性为某个特定值的Web页面的总数,而 R^{Key}和 $R^{Non\text{-}Key}$表示属性值也是这个特定值的正例和反例页面分别在正例页面集合和反例页面集合中所占的比例.则 $K{\cdot}R^{Key}$表示属性值为这个特定值的正例页面在全体 Web 页面中的比例.由式(1)易得到:

$$R^{Non-Key} = (R^{Whole} - K \cdot R^{Key})/(1-K) \tag{3}$$

其中各符号的含义与式(3)相同,这就意味着尽管反例页面的个体信息很难得到,但其统计信息却是可以通过正例页面和全体页面的统计信息计算得到的.这是在关键资源判定中对决策树算法进行改进的基础.

3.2 连续属性值的离散化

在推导改进的信息增益公式之前,需要先进行属性值离散化的工作,离散化属性取值的目的,是为了适应决策树算法处理的要求;而本文的改进算法将取值类别局限在布尔变量上,则是出于减少算法复杂度的需要.事实上,将连续特征取值离散化也是使用决策树进行学习的一般通用步骤.

离散化的方法是采用选取阈值的方法,阈值的选取要使 Web 页面全集在该属性进行离散化后可能的信息增益值最大.在本文的实验研究中,阈值选取按照以下方法实现:先按照属性值的分布特征选取若干可能成为阈值的取值点,而后逐一计算这些阈值待选点对应的信息增益,最后选择信息增益最大的一个待选点作为离散化的阈值点.比如,根据页面入度进行分布统计后,发现按照对数规律选择阈值待选点可能比较合适,于是选择

1,10,100,1000作为可能的阈值待选点,进行信息增益计算比较后,发现属性阈值选择为10引起的信息增益最大,因此,属性值就根据“页面入度是否大于10”划分成“1”和“0”两类.

进行离散化后,各个属性的阈值选取与统计数据见表2.统计数据所采用的语料库与表1相同,均使用.GOV语料库作为普通页面统计的标准,而把TREC2002的关键资源页面答案集合作为关键资源样例的代表.

Table 2 Statistical data from different attributes of ordinary and key resource pages after discretization

表2 离散化后各个属性取值对应的普通页面和关键资源页面的统计数据

Page attributes	Ordinary pages (%)	Key resource pages (%)
Page length (>1000)	16.08	1.17
In-Link count (>10)	10.78	51.03
URL classification (≠FILE)	12.61	57.27
In-Site out-link anchor text rate (>0.1)	45.14	80.23
In-Site out-link number (>10)	43.31	79.07

由表2可以看到,离散化后,在各个属性取值上,普通页面与关键资源页面有较大的差别,这一方面说明属性选取比较恰当,同时也说明了属性值的离散化是合理的,既保证了统计特征的分布差异,又避免了属性分支过多而造成的过学习现象.

3.3 信息增益计算方法的改进

进行离散化后,各个属性分支的数目相同,信息增益率的比较结果与信息增益的比较结果类似,因此可以只进行运算量较小的信息增益的比较.下面将对原有的信息增益公式进行推导,以寻找适合关键资源页面判定的算法结构.

定义 ${R_1}^{Whole}$ 是某个样例页面集合 S 中主题无关特征 A 为一特定布尔值的页面的比例.对应地,${R_1}^{Key}$ 和 ${R_1}^{Non\text{-}Key}$ 分别是关键资源页面集合与非关键资源页面集合中 A 为同一个特定值的页面比例.再假设 K 表示全体页面中关键资源页面所占的比例,则有类似于式(2)的式(4)成立.

$$R_1^{Non-Key} = (R_1^{Whole} - K \cdot R_1^{Key})/(1-K) \tag{4}$$

类似地,可以定义内容特征 A 的值为另一个布尔值的页面统计数值 ${R_2}^{Whole}$,${R_2}^{Key}$,并计算 ${R_2}^{Non\text{-}Key}$ 的取值,则可以计算出主题无关特征 A 对应此样例页面集合 S 的信息增益.

$$Gain(S,A) = Entropy(S) - {R_1}^{Whole} Entropy(S_1) - {R_2}^{Whole} Entropy(S_2) \tag{5}$$

其中,S_1 和 S_2 分别对应特征 A 的取值为“0”和“1”的样例页面集合.又因为

$${R_1}^{Whole} Entropy(S_1) = {R_1}^{Whole}\left[\frac{{R_1}^{Whole} - K{R_1}^{Key}}{{R_1}^{Whole}}\log_2\frac{{R_1}^{Whole}}{{R_1}^{Whole} - K{R_1}^{Key}} + \frac{K{R_1}^{Key}}{{R_1}^{Whole}}\log_2\frac{{R_1}^{Whole}}{K{R_1}^{Key}}\right]$$

$$= {R_1}^{Whole}\log_2 {R_1}^{Whole} - ({R_1}^{Whole} - K{R_1}^{Key})\log_2({R_1}^{Whole} - K{R_1}^{Key}) - K{R_1}^{Key}\log 2(K{R_1}^{Key}),$$

利用条件 $\begin{cases}{R_1}^{Key} + {R_2}^{Key} = 1\\ {R_1}^{Whole} + {R_2}^{Whole} = 1\end{cases}$ 可以实现上式的化简.如果我们令

$$F_{entropy}(r) = r\log_2\frac{1}{r} + (1-r)\log_2\frac{1}{1-r},$$

则式(5)可以化为

$$Gain(S,A) = F_{entropy}({R_1}^{Whole}) - (1-K)\cdot F_{entropy}({R_1}^{Non-Key}) - K\cdot F_{entropy}({R_1}^{Key}) \tag{6}$$

对于式(6)可以从另一个角度考虑,如果把“是否是关键资源”作为一个特征,而把原有的主题无关特征 A 作为取值为布尔值的目标函数,则也可以由式(1)直接得到此式.

3.4 算法描述与决策树生成

按照上一节的计算方法,取 K=1/6(这是根据对.GOV 语料库质量进行估计得到的经验值),可以计算出各个主题无关特征对应的信息增益,如图1所示.根据信息增益数值的大小,可以对各主题无关特征判定关键资源的能力进行排序.由图1可以看出:首先,入度数值和URL分级特征的信息增益值最大,可以认为这两个特征是判

定页面是否关键资源页面的可靠依据;其次,站点内出链接文本比率的信息增益要大于站点内出链接数目,这反映了在判定关键资源页面方面,链接文本比率是更有效的站点内出链接特征.

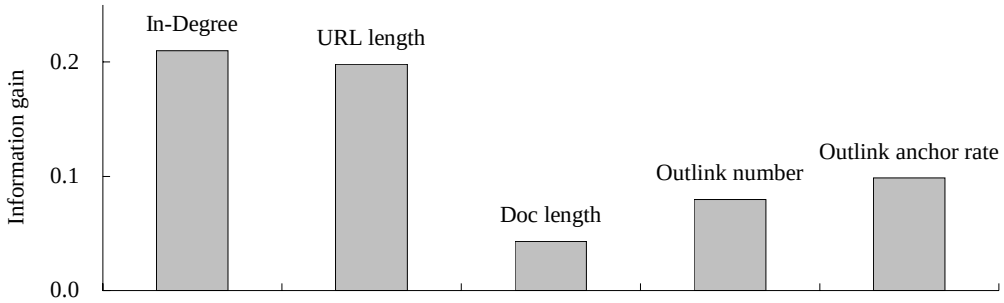

Fig.1 Information gains of different non-content features

图 1 不同主题无关特征对应的信息增益

总结上述步骤,可以得到改进算法的具体描述如下:

(a) 决策树根节点设定为当前节点.

(b) 利用式(6)计算各主题无关属性对应当前节点页面集合的信息增益值.

(c) 信息增益最大的主题无关属性选作决策树的当前节点.

(d) 训练样例集在当前节点根据主题无关属性的取值进行分类.

(e) 判断是否满足下列条件之一:

- 每个节点对应的样例集中所有样例都具有相同的分类结果.
- 所有属性都已在每条从根节点到叶子节点的路径上被测试.

如果满足条件,则算法结束,输出决策树;否则,转(b).

由上述步骤生成的决策树如图 2 所示.如果对全体 Web 页面施行决策树判定算法,则可以得到一个 Web 页面全集的子集——关键资源页面集合.

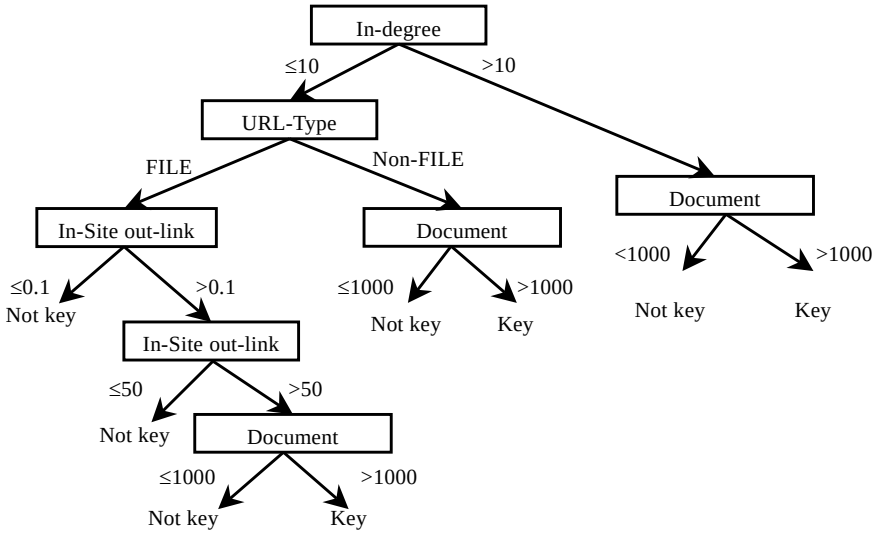

Fig.2 Decision tree for key resource page judgment with non-content features

图 2 主题无关特征进行关键资源页面判定的决策树

在判定过程中可以发现决策树算法的另一个优势,即判定算法本身具有较低的时间复杂度.尽管决策树的训练过程是一个复杂的过程,一旦训练结束,则利用诸如图 2 所示的决策树进行关键资源判定就非常简便了.决策树等价为析取范式的表达形式,因此,判定关键资源的时间复杂度实际上是线性 $O(N)$的.与使用迭代方式的

PageRank,HITS 算法相比,决策树算法的时间需求大为降低,这也更有助于将此类算法应用于实际检索系统.

4 实验与结果分析

4.1 实验环境和方法简述

本文所采用的实验数据均来源于.GOV 语料库中的页面.利用决策树算法生成的关键资源页面决策树对.GOV 语料库的所有页面进行筛选,就可以得到关键资源页面的实验结果集合,在这个集合的基础上可以进行检索实验.

检索实验使用了 TREC2003 网络信息检索任务的查询主题及标准答案,此任务一共提供了 50 个查询主题和对应主题的 516 个标准答案,任务的目的就是查找与主题相对应的关键资源页面.其查询主题来源于真实搜索引擎的用户查询,而标准答案的标定也经过了多位评测人员的反复验证,因此具有较高的权威性与可靠性.

实验采用了两种不同的决策树判定方法进行关键资源页面的判定,分别是传统的基于反例对象个体特征的方法和基于反例统计特性的方法.由于方法的不同,这两种方法分别对应的训练集合也不尽相同.两种方法采用的正例训练集合都是 TREC2002 网络信息检索任务的标准答案,由 344 个页面组成.在反例训练集合方面,传统的决策树学习方法采用随机抽取反例页面的方式构成训练集合,按照实验中获得的关键资源页面与非关键资源页面的大致比例(1:5),反例训练集合的规模在 1 700 个页面左右.改进的决策树学习方法基于正例样本数据和全体页面数据计算反例数据的统计信息,因此不涉及反例样例训练集合的问题.

4.2 基于不同决策树判定方法的关键资源页面覆盖率实验

利用实验得到的结果集合的大小及其覆盖关键训练集,测试集的比例可以衡量不同算法的优劣.一个理想的实验结果,应该用较小的页面数包括较多的关键资源页面.按照式(1)中关键资源比率的定义,比率较高的结果是实验效果较好的结果.基于传统决策树判定算法和改进算法的关键资源页面覆盖率实验结果如图 3 所示.

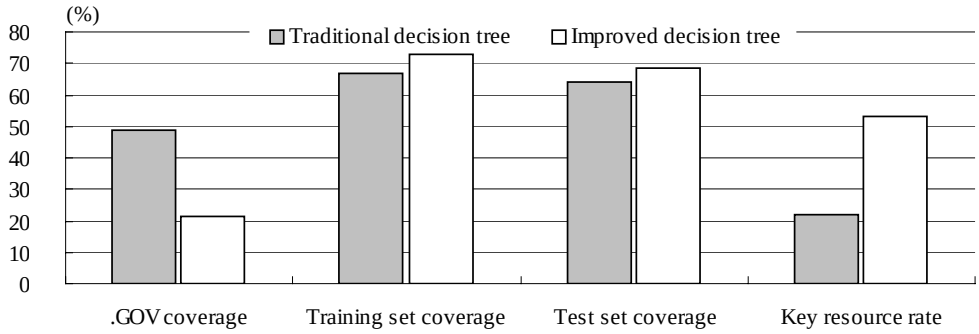

Fig.3 Key resource page coverage of result sets based on different decision tree learning algorithms

图 3 不同决策树学习算法结果集合对应的关键资源页面覆盖率

实验结果说明,改进算法判定得到的结果集合质量明显高于传统决策树学习算法.传统方法得到的结果集合是改进方法得到的结果集合的两倍多,但其关键资源页面集合的覆盖率无论是在测试集合还是在训练集合上却都低于改进方法得到的结果集合,改进决策树学习方法的关键资源覆盖率则是传统学习方法的 2.4 倍.这说明改进的决策树学习算法利用反例的统计性能能够比较理想地估计真实情况.而由于受到处理能力的限制,基于随机抽取少量反例的方法很难保证抽取到的反例样本的特性能够与真实的网络环境一致.

同时可以发现,利用改进的决策树学习算法,可以用 20%左右的页面数量覆盖超过 70%的关键资源页面,这说明依靠主题无关属性和决策树学习进行关键资源页面的定位是完全可能的.实验得到的"关键资源页面集合"确实能够覆盖大部分关键资源页面.

4.3 基于不同决策树判定方法的检索实验结果

由于关键资源页面决策树最终生成的页面集合要用于网络信息检索,因此页面判定的最终结果评价还要

落实到信息检索的效能提高上.实验结果说明,基于改进决策树学习算法的检索效果比传统检索的效果有明显的提高,见表 3.

Table 3 Retrieval performance comparison with different key resource judgment methods

表 **3** 使用不同关键资源判定方法的检索效果比较

Evaluation metrics	Without key resource judgment	Key resource judgment based on traditional decision tree algorithm	Key resource judgment based on improved decision tree algorithm	Best run in TREC 2003
Precision @ 10	0.072 0	0.086 0	0.124 0	0.124 0
R-Precision	0.114 5	0.119 1	0.167 0	0.163 6

实验比较了 TREC2003 网络信息检索任务在两个页面集合上的性能,可以看出,虚拟站点入口页面集合的检索效果明显好于页面全集.为了方便比较,各组实验都只采用了信息检索中常用的 BM2500 权重计算公式和此公式默认的实验参数.评价方式采用的是 TREC 网络信息检索任务通用的前 10 位结果平均精度(Precision@10)和 R-精度(R-precision).前 10 位结果平均精度用于反映检索用户的满意程度,而 R-精度则是综合考察检索系统精度与召回率的评价指标,这两个指标都是公认的评价主题过滤任务的权威评价指标[13,14].

在 Precision@10 评价上,关键资源页面检索比较全部页面集合检索有 72.22%的提高,而在 R-precision 评价上性能提高的比例是 45.85%.检索性能的差异可以作如下解释:关键资源集合用少量的页面集中了大量的高质量信息,在这样的集合里进行检索的难度要远小于在页面全集上进行检索.

为了验证方法的有效性,我们还把这两组结果与 TREC2003 的最优结果[14]进行了比较.实验证明,关键资源集合上的检索效果与 TREC2003 主题过滤任务的最优结果性能相当,在 R-precision 评价上还优于这个结果.这说明基于关键资源提取进行主题过滤的方法与传统的主题过滤算法相比是有优势的:检索对象集合从页面全集变为关键资源页面集合,从而大大提高了检索的效率,而检索的效果甚至还优于原有方法.所付出的代价,仅仅是在检索进行之前,利用时间复杂度为 $O(N)$的判定算法过滤了一下页面集合,可以说,这是一个进行高质量主题过滤的事半功倍的有效途径.

5 结论与未来工作

网络数据的极大丰富给传统的信息检索任务带来了巨大的挑战.为了使计算机能够更加智能地帮助用户查找有用信息,激起学习的方法越来越多地被引入网络信息检索的研究.但由于面临着一个有史以来从未有过的庞大处理对象——网络信息,传统的机器学习方法必须进行一定程度的改进才能适合处理大规模且质量参差不齐的数据的要求.针对关键资源判定的问题,传统的决策树学习的方法需要进行改进,以应付反例样本缺乏的困境.在这方面,本文的主要贡献与结论是:

(1) 提出一种利用关键资源反例页面的统计特性进行判定的决策树学习算法,可以成功地进行关键资源页面的判定和关键资源集合的提取.

(2) 大规模 Web 数据上的关键资源集合覆盖率实验和检索性能实验说明,改进的决策树算法与传统的基于个例学习的决策树算法相比有明显的优势.

(3) 本文的算法推导和分析说明,基于主题无关属性和决策树判定算法可以提高网络信息检索中的一个重要任务,即关键资源页面查找任务的性能.

对决策树学习算法的改进也带给我们更多的思考:其他机器学习方法能否也利用类似的思路进行改进,以进行关键资源判定?如果有可能,何种机器学习方法进行关键资源判定的效果更好?关键资源页面判定的方法应该如何应用,以提高网络信息检索工具(如搜索引擎)的检索性能?这些都将是我们今后需要考察的问题.

References:

[1] Amento B, Terveen L, Hill W. Does authority mean quality? Predicting expert quality ratings of Web documents. In: Belkin NJ, Ingwersen P, Leong MK, eds. SIGIR 2000: Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval 2000. New York: ACM Press, 2000. 296−303.

[2] Davison BD. Topical locality in the Web. In: Belkin NJ, Ingwersen P, Leong MK, eds. SIGIR 2000: Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval 2000. New York: ACM Press, 2000. 272−279.

[3] Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In: Croft BW, Moffat A, van Rijsbergen CJ, Wilkinson R, Zobel J, eds. SIGIR'98: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1998. 104−111.

[4] Broder A. A taxonomy of Web search. SIGIR Forum, 2002,36(2):1−8.

[5] Henzinger MR, Motwani R, Silverstein C. Challenges in Web search engines. In: Gottlob G, Walsh T, eds. IJCAI 2003, Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2003. 1573−1579.

[6] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604−632.

[7] Chakrabarti S, Dom B, Kumar R, Raghavan P, Rajagopalan S, Tomkins A. Experiments in topic distillation. In: Brown E, Smeaton A, eds. Proc. of the ACM SIGIR Workshop on Hypertext Information Retrieval. New York: ACM Press, 1998. 13−21.

[8] Chakrabarti S, Joshi M, Tawde V, Bombay IIT. Enhanced topic distillation using text, markup, tags and hyperlinks. In: Croft BW, Harper DJ, Kraft DH, Zobel J, eds. SIGIR 2001: Proc. of the 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2001. 208−216.

[9] Mitchell TM. Machine Learning. New York: McGraw-Hill, 1997. 55−64.

[10] Liu XH, Li S. An optimized algorithm of decision tree. Journal of Software, 1998,9(10):797−800 (in Chinese with English abstract).

[11] Hong JR, Ding MF, Li XY, Wang LW. A new algorithm of decision tree induction. Chinese Journal of Computers, 1995,18(6): 470−474 (in Chinese with English abstract).

[12] Craswell N, Hawking D. Query-Independent evidence in home page finding. ACM Trans. on Information Systems (TOIS), 2003, 21(3):286−313.

[13] Hawking D, Craswell N. Overview of the TREC-2002 Web track. In: Voorhees EM, Buckland LP, eds. NIST Special Publication 500-251: The 11th Text REtrieval Conf. (TREC 2002). Washington: Department of Commerce, National Institute of Standards and Technology, 2002.

[14] Hawking D, Craswell N. Overview of the TREC 2003 Web track. In: Voorhees EM, Buckland LP, eds. NIST Special Publication 500-255: The 12th Text REtrieval Conf. (TREC 2003). Washington: Department of Commerce, National Institute of Standards and Technology, 2003. 78−92.

[15] Liu YQ, Zhang M, Ma SP. Effective topic distillation with key resource pre-selection. In: Myaeng SH, *et al*., eds. Proc. of the AIRS 2004. LNCS 3411, Berlin/Heidelberg: Springer-Verlag, 2005. 129−140.

[16] Kraaij W, Westerveld T, Hiemstra D. The importance of prior probabilities for entry page search. In: Ricardo BY, ed. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2002. 27−34.

附中文参考文献:

[10] 刘小虎,李生.决策树的优化算法.软件学报,1998,9(10):797−800.

[11] 洪家荣,丁明峰,李星原,王丽薇.一种新的决策树归纳学习算法.计算机学报,1995,18(6):470−474.