基于非内容信息的网络关键资源有效定位

刘奕群,张 敏,马少平
(清华大学 智能技术与系统国家重点实验室,北京 100084)

摘 要:网络信息的爆炸式增长,使得当前任何搜索引擎都只可能索引到 Web 上一小部分数据,而其中又充斥着大量的低质量信息.如何在用户查询无关的条件下找到 Web 上高质量的关键资源,是 Web 信息检索面临的挑战.基于大规模网页统计的方法发现,多种网页非内容特征可以用于关键资源页面的定位,利用决策树学习方法对这些特征进行综合,即可以实现用户查询无关的关键资源页面定位.在文本信息检索会议(TREC)标准评测平台上进行的超过 19G 文本数据规模的实验表明,这种定位方法能够利用 20 %左右的页面覆盖超过 70 %的 Web 关键信息;在仅为全部页面 24 %的关键资源集合上的检索结果,比在整个页面集合上的检索有超过 60 %的性能提高.这说明使用较少的索引量获取较高的检索性能是完全可能的.

关键词:网络信息检索;关键资源页面;主题过滤;机器学习
中图分类号:TP181,TP391.3 **文献标识码**:A **文章编号**:1673-4785(2007)01-0045-08

Web key resource page selection based on non-content information

LIU Yi-qun, ZHANG Min, MA Shao-ping
(State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: Information growth makes it impossible for search engines to crawl and index all pages on the Web. Meanwhile indexed page set is filled with low quality information and spam. It is quite a challenge to select high quality Web pages (key resource pages) query-independently. With analysis in non-content features of key resources, a pre-selection method was introduced in topic distillation research. A decision tree was constructed to locate key resource pages using query-independent non-content features including in-degree, document length, URL-type and two novel proposed features involving site's self-link structure analysis. Although the result page set contained only about 20 % pages of the whole collection, it covered more than 70 % of key resources. Furthermore, information retrieval on this page set made more than 60 % improvement with respect to that on all pages. It shows an effective way to get better performance in topic distillation with a smaller data set.

Keywords: web information retrieval; key resource page; topic distillation; link structure analysis

当前网络信息检索技术面临的最大挑战来自网络信息环境本身.由于网络数据的大量膨胀,目前已经没有任何搜索引擎能够索引到 Web 上的所有网页.根据 Sullivan 等的评测[1],2003 年 2 月时,Google 是世界上索引量最大的搜索引擎(索引到 33 亿 Web 页面).但即使是按照加州大学伯克利分校 How Much Info 计划[2]的保守估计,当时的 Web 仅静态页面数目也将超过 200 亿.容易想象,在未来很长的时间里,随着网络技术在更多国家中得到发展,Web 数据将继续其增长趋势,这对数据收集、索引的建立更新及检索都带来几乎无法逾越的障碍.

然而,Web 信息的一大特点是其内容质量参差不齐.Google 公司的 Henzinger 等人指出[3]:大量冗余信息、低质量信息与 Web 中的有用信息鱼龙混杂,给现代搜索引擎的发展带来了极大的问题.一方面,由于网络数据的极大丰富,无法索引到所有的有用页面;另一方面,占用宝贵收集时间和存储空间的数据却有许多是低质量的.因此 Henzinger 等人将

收稿日期:2006-04-23.
基金项目:国家重点基础研究(973)资助项目(2004CB318108);国家自然科学基金资助项目(60223004, 60321002, 60303005, 60503064);教育部科学技术研究重点资助项目(104236).

寻找主题无关的判断页面质量的方法作为搜索引擎技术发展的最大挑战之一.

已有的网络信息检索研究,对用户笼统概念上的"高质量页面"给出了较科学明确的定义.能够为某个主题提供高质量信息或者链接的少部分页面被称为这个主题的关键资源页面,而寻找关键资源页面的任务称为主题过滤(topic distillation).反映国际信息检索研究最高水平的 SIGIR(International ACM SIGIR Conference on Research and Development in Information Retrieval)会议上,主题过滤技术无论从论文数目还是质量来看,一直都是近年讨论的热点.作为信息检索领域权威评测会议的 TREC,也从 2002 年开始,在其 Web 信息检索部分,使用主题过滤任务代替了传统的相关资源查找任务[4-6].查找关键资源,是当前网络信息检索的发展热点,也已经获得了一些卓有成效的理论研究和实验结果[4-8].但总的来说,主题过滤研究的发展还停留在一个比较低的水平上,作为评价标准的前 10 位结果检索精度(Precision at 10 documents, P@10)一直在 20%左右徘徊[4-5],而表现网络数据不同于普通数据的许多非文本内容特征也没有得到充分的考察.

文中对关键资源页面的若干非内容特征进行了考察,这些特征包括已有一定研究[9-11],但尚未对关键资源页面进行专门考察的 URL 分级特征、网页文本长度特征、入链接个数特征,也包括了根据关键资源定义提出的新特征:站点自身出链接特征.关键资源页面在这些特征上的分布与普通网络页面有着明显的差异,从而使查询主题无关的判定关键资源页面成为可能.综合利用非内容特征进行决策树学习,使得人们能够从大规模测试页面全集中提取出一个关键资源页面集合.在关键资源页面集合上进行的检索证明:可以使用较少的页面索引量获得较高的检索质量.

1 相关研究工作概述

1.1 主题过滤与关键资源页面

主题过滤(topic distillation)技术以查找关键资源页面为目的,与查找相关页面的传统检索技术不同,它强调检索结果页面应当是用户获取信息的重要途径而非唯一来源.用户可以通过访问结果页面以及与其相链接的其他页面来获得完整的信息,而不是从有限的页面内资源中获取知识.主题过滤技术的提出,是为了用更有限的页面提供给用户更多的关键信息.

根据 TREC 网络信息检索部分的较权威的定义[5],关键资源页面应当是某个关键站点的入口页面,此站点需要提供关于某个主题的可靠信息.

需要特别指出的是,这里所提到的入口页面不一定是通常意义上的"主页",它可能是大规模站点的接入页面,也有可能是某个子站点或者某一类页面集合的接入页面.如 http://www.nida.nih.gov/drugpages/marijuana.html 是美国药物滥用治疗研究所(NIDA)关于大麻方面信息的关键资源页面,但它并不是一个通常意义上的"主页",这个页面只是为用户提供一个 NIDA 站点内部相关信息页面的访问索引.

从定义中可以看出,关键资源页面之所以"关键",是因为它提供给用户一个关于某个主题的可靠信息的入口.用户通过关键资源页面,可以比较快捷的查找到所需要的信息.同时,某个主题的关键资源页面数要比其相关页面数少得多(相关页面动辄成百上千,而关键资源页面往往只有几个到十几个),这也方便用户将注意力集中到少数一些与自己的查询主题最贴切的页面上.

关于所有可能的主题的关键资源页面的总和,就构成了网络环境中的关键资源页面集合.可见这个页面集合只占网络页面全集的小部分,其余大部分页面则是信息量较少的页面,以及包含错误和冗余信息的页面.这说明,定位关键资源页面集合,可以起到在网络信息环境中去粗取精的作用.这也就是文中关注关键资源页面判定,希望能够通过非内容特征进行用户查询主题无关的定位关键资源页面的原因.

1.2 Web 页面非内容特征研究的已有成果

文中考察 Web 页面非内容特征的目的,是试图利用这些特征进行主题无关的关键资源定位.尽管对于站点主页非内容特征的考察,已经有一定的基础,但是站点主页与关键资源页面的巨大差异,使文中的工作很少能够借鉴一些现有的研究结论.对于关键资源页面非内容特征的考察,只能是从头起步.

当前关于网页非内容特征的研究成果,大多是在主页查找工作的推动下得到的.主页查找即站点查找或导航类查找,按照 Broder 的统计[12],此类查找在网络搜索引擎的日常查询中占有约 20%的份额.由于主页有明显的非内容特征如 URL 特征、链接特征等,在主页查找中使用这些特征也成为了必然.

链接特征一直是人们考察非内容特征的重点,经典的链接分析方法,如 PageRank 和 HITS 算法,分别作为 Google 和 IBM CLEVER 的核心算法,在促进网

络信息检索工具的发展方面取得了极大的成功.

除链接特征之外,非内容特征还有多种形式:Westerveld 等人在 2000 年提出了 URL 分级特征[10],这种特征比较之前通常使用的 URL 长度特征更为合理,其核心思想是将页面的 URL 分为 Root, Subroot, Path 和 File4 级.由于统计结果发现,绝大部分的站点主页不属于 File 类,因此对非 File 类页面进行加权就可以提升主页查找的性能.Westerveld 与其同事 Kraaij 等在次年总结了包括 URL 分类特征在内的多种非内容特征,并在 SIGIR 2001 上发表[11];而 Craswell 等人则于 2003 年对各类主题无关的非内容特征包括页面长度特征、入度(入链接个数)特征、URL 分级特征和 PageRank 数值特征在主页查找中的应用进行了系统总结[9],他指出:入度特征和 URL 分级特征对于主页查找是有效的,而 PageRank 数值和文档长度特征的功能则不太明显.

关键资源页面与站点主页有一定的关联,但二者发挥的功能具有很大的差异.一部分关键资源页面是站点主页,但关键资源页面又远不是所在站点的主页.这就决定了对关键资源页面非内容特征的考察,一方面要对站点主页查找中常用的非内容特征针对关键资源页面进行分析和筛选,另一方面则需要发现关键资源页面特有的非内容特征.

2　关键资源页面的非内容特征分析

文中采用基于大规模网页数据统计的方法考察关键资源页面的非内容特征,以便实现主题无关的关键资源页面判定,从而将关键资源页面从普通页面中区别开来.将要讨论的非内容特征除了上文提到的一些常用特征外,还包括根据关键资源的特性,提出的一种新特征——站点自身出链接特征.

作为基于统计方法的基础,本节使用的大规模语料库是.GOV 网页数据库*,而用于统计关键资源页面特征的"关键资源训练集合"则是指在 TREC 2002 主题过滤任务的标准答案基础上,用手工标出关键资源的方式加以筛选,而得到的符合关键资源定义的页面.

2.1　常用网页非内容特征在关键资源页面集合上的考察

常用于网络信息检索的网页非内容特征主要包括:页面长度特征、入度特征和 URL 分级特征.

1) 页面长度是指经过过滤无用字符等预处理之后的页面包含的单词数.

2) 入度特征是指某页面被多少个外部页面链接引用的度量.

3) URL 分级特征是表示页面 URL 种类的一个非内容特征,由 Kraaij 等人在文献[11]中提出,页面的 URL 划分为 ROOT, SUBROOT, PATH, FILE4 类,其中 ROOT 类的 URL 只包括域名,而从 ROOT 类到 FILE 类其 URL 中的"/"逐渐增多.

文中着重考察的是这些特征在网络页面全集与关键资源集合上的差异,经过对.GOV 和关键资源训练集合的统计,这种差异是确实存在的,如表 1 和表 2 所示.

表 1　.GOV 与关键资源训练集合在非内容特征值的分布差异

Table 1　Differences between .GOV corpus and key resource set in several non-content features　%

特征	.GOV	关键资源训练集
页面长度 ≤ 1 000	16.08	1.17
入度 ≥10	10.78	51.03
URL 分级 ≠ FILE	12.61	57.27

表 2　.GOV 与关键资源训练集合的非内容特征平均值差异

Table 2　Differences between .GOV corpus and key resource set non-content feature in mean values

特征	.GOV	关键资源训练集
页面长度	7 037.43	9 008.02
入度	9.94	153.12
URL 分级	/	/

由统计数据可以看出:

1) 关键资源训练集的平均页面长度与普通页面比较接近.但是表 1 的实验结果说明,2 个页面集合上的页面长度分布还是有差异,主要表现在关键资源页面是长度较短(≤1 000 词)页面的可能性很小,这个特征有助于去除冗余页面.

2) 表 2 显示,关键资源页面的平均入度比普通页面要大得多.这可以使用反映网页链接关系的内容推荐假设[13](recommendation assumption)解释:一个页面被其他页面所引用,反映这个页面的内容被其他页面的作者所推荐.拥有较多入度的页面,是内容质量比较高的页面,它成为关键资源页面的可能性也就比较大.

3) 根据表 1 的统计结果,FILE 类型的网页在.GOV 语料库中占了大多数;但在关键资源训练集中,FILE 与非 FILE 类型的页面数基本相等,非 FILE 类型的页面反而更多一些.这说明关键资源页面的 URL 更倾向属于非 FILE 类型.

*.GOV 是由超过 120 万个网页、近 20G 数据组成的实验性网络语料库,它是由 CSIRO(澳大利亚联邦工业研究组织)在 2002 年初基于真实网络环境抓取到的.可以认为它是一个真实网络环境的采样,详细信息参见 http://es.csiro.au/TRECWeb/govinfo.html

总之,这些常用的非内容特征在关键资源页面上有不同于. GOV 页面的分布,这种分布差异可以被用于关键资源页面的主题无关定位.

2.2 关键资源页面集合的站点自身出链接特征

关键资源页面是为用户提供高质量信息接入点的页面,这决定了它最重要的不是自身提供信息,而是作为关键站点内容的代表提供链向站点内其他高质量页面的链接. 例如美国药物滥用治疗研究所(NIDA)关于大麻滥用方面信息的关键资源页面 http://www.nida.nih.gov/drugpages/marijuana.html,它基本没有除去链接之外的文字内容,但这个页面提供了链向子站点内部其他与大麻信息相关的页面链接,这些链接对浏览者获取大麻滥用方面的信息是大有帮助的.

由于关键资源页面是关键资源站点的代表和用户访问的接口,因此它应当具有能够直接链接到本站点/子站点内的大多数页面,也即必须具有 Craswell 等人在文献[13]中提出的"导航功能"(navigational function). 为了更清楚的表述关键资源页面的这个链接特征,把页面出链接划分为指向站点内部的出链接和指向站点外部的出链接2类. 由于关键资源页面首先是站点内部高质量页面的代表,因此指向站点内部的出链接是我们考察的重点,这种出链接数目的多少和质量的好坏,直接影响到关键资源页面本身的质量高低. 为区别一般出链接,在下文中称之为"站点自身出链接".

理想情况下,关键资源页面的站点自身出链接应当个数多,质量高. 具体到内容无关信息而言,无法考察这种出链接描述文本于某个主题是否相关,但链接描述文本的长度是可知的,通过长度标准,评判链接文本是否包含了足够的其他页面描述信息也是可行的. 因此,站点自身出链接个数的多少,以及站点自身出链接描述文字的长短是否可以作为考察关键资源页面质量的标准呢?下面将把站点自身出链接特征细化成此2个标准分别加以考查:

1)站点自身出链接个数(site self link number)

关键资源页面作为关键资源站点的入口页面和代表,应当拥有较多的站点自身出链接. 也就是说,关键资源页面的站点自身出链接个数从总体上讲应当比普通页面更多. 图1的统计数据中,横坐标表示站点自身出链接的个数,而纵坐标表示. GOV 语料库和关键资源训练集合上的不同分布百分比. 由统计数据可以看出,大部分(超过55%). GOV 语料库中的页面站点自身出链接个数小于10,而关键资源训练集中的这个比例在20%左右;而站点自身出链接数目较大(>20)时,关键资源训练集上的分布百分比明显高于. GOV 语料库.

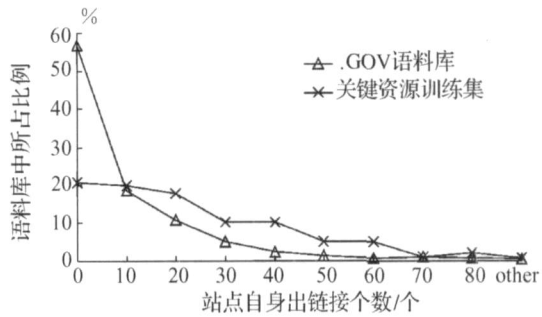


图1 站点自身出链接个数在关键资源训练集合和. GOV 语料库上的不同分布

Fig. 1 Site self link number of key resource training set and . GOV

在统计平均值上,关键资源训练集合的站点自身出链接个数是37,而. GOV 中的平均数目是18. 这也说明了站点自身出链接数目较小的页面更可能是普通页面,而这个数目较大的页面比较可能是关键资源页面.

2)站点自身出链接文本比率(site self link anchor rate)

关键资源页面的站点自身出链接文本可以认为是整个关键资源站点内容的概括. 这是由链接文本的特性所决定的,Craswell 在文献[13]中指出:对某个确定的页面 A 而言,指向它的链接文本可以看作是对此页面 A 的概括客观(通常来自其他作者)的描述. 因此对关键资源页面而言,它的站点自身出链接文本可以看作此页面作者对站点内其他页面的一个简要介绍,这些文本的集合,也就可以作为整个站点内页面的一个综述.

站点自身出链接文本比率的定义为

$$\text{site self link rate} = \frac{\text{WordCount(site self link anchor)}}{\text{Word Count(full text)}}. \tag{1}$$

对于关键资源页面而言,站点自身出链接文本比率与站点自身出链接个数是类似的非内容特征,2者的不同在于,站点自身出链接文本比率一定程度上反映了站点自身出链接质量的高低. 仅仅列出导航信息的站点自身出链接,与给出子页面简要介绍的站点自身出链接,其质量是显然不同的. 站点自身出链接文本比率较高的页面,可以认为其出链接质量较高.

关键资源页面一般都是关键资源站点的入口页面,其对应的站点自身出链接担负着引导用户访问的任务,因此链接质量较高,这类页面也相应的拥有较大的出链接文本比率.

对于关键资源页面而言,站点自身出链接文本

比率与站点自身出链接个数是类似的非内容特征,2者的不同在于,站点自身出链接文本比率一定程度上反映了站点自身出链接质量的高低. 仅仅列出导航信息的站点自身出链接,与给出子页面简要介绍的站点自身出链接,其质量是显然不同的. 站点自身出链接文本比率较高的页面,可以认为其出链接质量较高.

关键资源页面一般都是关键资源站点的入口页面,其对应的站点自身出链接担负着引导用户访问的任务,因此链接质量较高,这类页面也相应的拥有较大的出链接文本比率. 如图 2 所示.

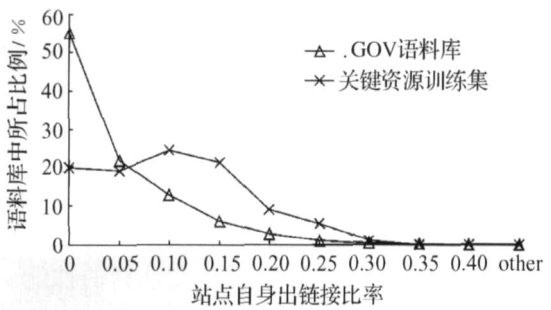


图 2　站点自身出链接比率在关键资源训练集合和. GOV 语料库上的不同分布

Fig. 2　Site self link anchor text rate of key resource training set and . GOV

图中横坐标表示站点自身出链接文本比率的大小,而纵坐标表示. GOV 语料库和关键资源训练集合上的不同分布百分比. 由数据可以看出,站点自身出链接文本比率较高的页面更可能是关键资源页面. 此特征的分布曲线与站点自身出链接个数特征非常类似,但在区分关键资源的能力上更强一些. 有超过 76 %的. GOV 语料库页面出链接文本比率不足 0.1,但在关键资源训练集中,这个比率只有 39 %.

3　非内容特征与关键资源页面判定

决策树学习(decision tree learning)的方法被选择用于进行 web 页面非内容特征的综合,这是由于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有效的途径之一. 文中使用的决策树学习算法,是由 Quinlan 在 1986 年提出的 ID3 算法[14]. 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结点需要测试的属性.

具体到关键资源页面判定的问题,我们把页面非内容属性的取值离散化为一个布尔变量,即对于某个非内容属性 A 而言,某页面 P 的取值只有"0"和"1"2 类. 离散化属性取值的目的,是为了适应 ID3 算法处理的要求,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. 离散化的具体方式是选取特征阈值,比阈值大的样例特征取值为 1,否则取值为 0. 阈值的选取总体遵循保证信息增益最大的原则,但也可以作调整以得到满足不同需要的决策树.

取值离散化后,根据 ID3 算法的要求,信息增益最大的非内容特征(页面入度特征)选作决策树的根节点. 训练样例集在根节点被分类后,每个子集重复计算信息增益的过程,并选取信息增益最大的非内容特征作为这个子集对应的决策树节点的分类特征. 当满足下列条件之一时算法结束:

1)所有样例都具有近似相同的分类结果;

2)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.

由上述步骤生成的决策树如图 3 所示,利用此决策树,判定任意 web 页面是否可以归入关键资源页面的范畴就成为可能. 如果对全体 web 页面施行决策树判定算法,就可以得到一个 web 页面全集的子集——关键资源页面集合.

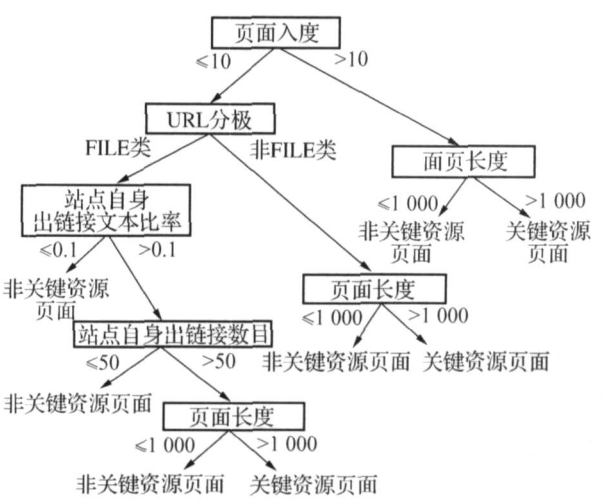


图 3　利用非内容特征进行关键资源页面判定的决策树

Fig. 3　Key resource decision tree constructed with ID3 and non-content features

4　实验与结果分析

本节将重点讨论根据. GOV 数据得到的关键资源页面集合在实验中的若干统计与检索特性. 在讨论这些特性之前,还将介绍文中实验所使用的训练与测试集合.

4.1　训练集与测试集

文中所采用的实验数据均来源于. GOV 语料库中的页面,为获取可信的关键资源页面训练集与测

试集,实验基本沿用了 TREC2002 与 TREC2003 主题提炼任务的查询主题及标准答案. 由于 TREC2002 主题提炼任务的目标是查找关键资源站点所包含的页面而不一定是关键资源页面[4,6],因此专门对这部分的答案集合进行了手工筛选,以找出标准答案页面对应的站点/子站点入口页面.

例如 TREC2002 主题提炼任务的 599 号主题“scientific research misconduct”,原有的答案包括子站点 http://ori.dhhs.gov/html/misconduct/内的 24 个页面,手工筛选后,将这些页面用这个子站点的入口页面 http://ori.dhhs.gov/html/misconduct/casesummaries.asp 代替.

关键资源测试集直接采用了 TREC2003 主题提炼任务的查询主题与标准答案,此任务共提供了 50 个查询主题和对应主题的 516 个标准答案. 任务的目的是查找与主题相对应的关键资源页面,查询主题来源于真实网络搜索引擎的用户查询,包含的内容领域涉及社会政治、经济生活的方方面面,因此具有较高的权威性.

4.2 基于关键资源页面集合的统计结果

特征取值离散化时判定阈值的选取不同,得到的决策树形式也会有不同;对应的,关键资源页面集合的规模也有差异,实验中不同的实验结果集对应的非内容特征阈值如表 3 所示,而图 4 则给出了这些实验结果集合覆盖关键资源页面的相关实验数据.

表 3 对应不同实验结果集合的非内容特征阈值

Table 3 Corresponding non-content feature thresholds for different result sets

	结果集 1	结果集 2	结果集 3	结果集 4	结果集 5
站点自身出链接数目	50	50	30	10	10
站点自身出链接文本比率	0.1	0.05	0.05	0.1	0.05

图 4 中的纵轴标志着不同比例的数值,分别是:测试集在实验结果集中的比例 R_1,训练集在实验结果集中的比例 R_2 和实验结果集在. GOV 语料库中的比例 R_3. 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,也就是说要在 R_3 尽量小的情况下,保证 R_1 和 R_2 较大.

从图 4 的实验结果中可以得到如下 2 个重要结论:

1) 使用文中提出的关键资源定位算法,在多种特征判定阈值下,都可以用 20 %左右的网页数量,覆盖超过 70 %的关键资源页面. 这说明依靠非内容特征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. 这也说明仍有约 30 %的关键资源页面不被实验结果集合所包括,因此在实验结果集合上进行主题提炼查找的性能上限即为 70 %左右. 但是,当前主题提炼任务按平均精度计算的性能一般都在 20 %上下浮动[4-5],因此这个上限对主题提炼任务的性能影响甚微.

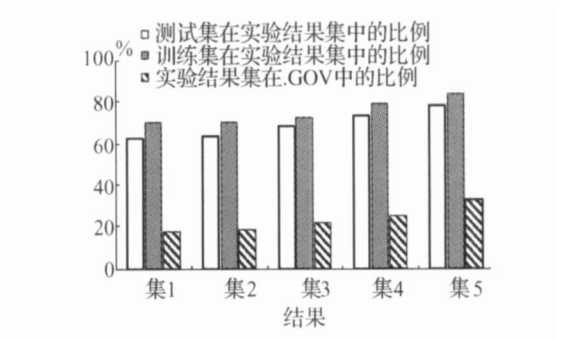


图 4 不同特征判定阈值下的实验结果集合数据

Fig. 4 Key resource coverage and result set size with different non-content feature threshold

2) 实验结果得到的高质量集合覆盖关键资源页面的比例是随着这个集合的规模而增加的. 与. GOV 集合规模相等的关键资源页面集合可以覆盖所有的关键资源,但这个集合显然不能称之为“高质量”,因此必须定义一个评价标准,从而能够在关键资源覆盖率与页面集合大小之间找到较好的平衡点,从而筛选出质量较高的关键资源页面.

4.3 关键资源页面集合的评价标准

关键资源页面判定结果的评价与一般的分类问题有类似之处,其问题可以归结到“用最小规模的关键资源页面集合覆盖最大数量的关键资源页面”. 对于此类问题,一般采用精度—召回率的评价标准. 召回率和精度的一般定义为

$$\text{recall} = \frac{\#(\text{相关页面集合} \cap \text{检索结果页面集合})}{\#\ \text{相关页面集合}}. \tag{2}$$

$$\text{precision} = \frac{\#(\text{相关页面集合} \cap \text{检索结果页面集合})}{\#\ \text{检索结果页面集合}}. \tag{3}$$

由于无法判断实验结果集合中所有的关键资源页面,因此在精度和召回率的计算中必须进行关键资源数目的估计. 为此引入如下假设:

1) 关键资源训练集合是. GOV 中所有关键资源页面的一个均一采样 k;

2) 关键资源页面占. GOV 页面总量的比例为 k.

在上述假设下,召回率可以通过关键资源测试集在实验所得到的结果集合中的覆盖度来估计,即:

$$\text{recall} = \frac{\#(\text{实验结果集合} \cap \text{关键资源测试集})}{\#\ \text{关键资源测试集}}. \quad (4)$$

在精度的计算中,由于关键资源页面的总数可以用. GOV 页面总数和 K 来计算,而 #(. GOV 页面集合) $\times K \times$recall 则表示了关键资源页面在实验结果集中的数目,因此精度表达式为

$$\text{precision} = \frac{\#(.\ \text{GOV 页面集合}) \times K \times \text{recall}}{\#(\text{实验结果集合})}. \quad (5)$$

为了在关键资源覆盖率与页面集合大小之间找到较好的平衡点,利用通常使用的均衡评价精度与召回率的 F-measure 评价[15],它的定义为

$$F_\alpha(r,p) = \frac{(1+\alpha)\text{recall} \times \text{precision}}{\text{recall} + \alpha \times \text{precision}}. \quad (6)$$

式中:precision 的权重为 1,而 recall 的权重为 α. 在关键资源页面性能判定的评价中,由于试图用实验得到的关键资源页面集合代替原有页面集合进行检索,因此关键资源页面的召回率应当得到更多的重视,以保证原有页面集合信息尽量少丢失. 由此设定 $\alpha = 2$. 取 $K = 1/6$,则根据此实验数据得到的精度—召回率评价结果如下(与图 4 中的实验结果集合一一对应):

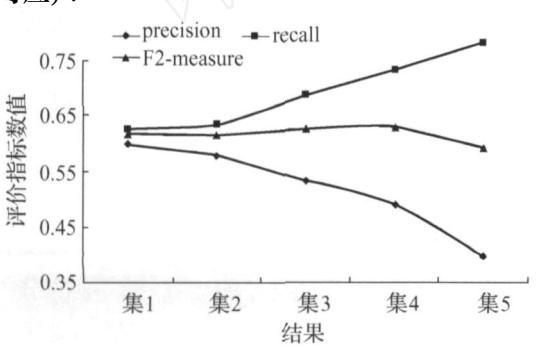


图 5　不同实验结果集合的精度、召回率和 F-measure 评价数值比较

Fig. 5　Recall, Precision and F-measure values of different result sets

从实验结果可以看出,随着召回率的上升,实验结果集合的精度是逐步下降的,而 F-measure 值则先增后减,结果集 4 的 F-measure 评价最高. 此结果集的页面数占页面总量的 24.89 %,但其包含的关键资源页面却占测试集的 73.12 %,满足用较少页面覆盖较多关键资源信息的要求,下面的检索实验中就是基于这个页面集合完成的.

4.4　基于关键资源页面集合的检索实验结果

关键资源页面定位的最终结果评价,还要落实到关键资源页面集合检索的效能提高上. 实验结果说明,基于关键资源页面集合的检索效果比全部页面检索的效果有明显的提高,如下面的结果列表所示.

表 4　不同页面集合上的检索效果比较

Table 4　Content retrieval results for different result sets

评价方式	全部页面集合	关键资源页面集合	TREC2003 最优结果
Precision @10	0.072 0	0.124 0	0.124 0
R-precision	0.114 5	0.167 0	0.163 6

实验比较了 TREC2003 主题提炼任务在 2 个页面集合上的性能,可以看出关键资源页面集合的检索效果明显好于页面全集. 为了方便比较,2 组实验都只采用了 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). 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:关键资源页面集合中用少量的页面集中了大量的关键资源,在这样的集合里进行主题提炼检索的难度要远小于在页面全集上进行检索. 从另一个角度,也可以认为关键资源页面定位的过程去除了 web 信息环境中的大量冗余信息,在一个信息有效性高的页面集合上进行检索的效果自然会好.

为了验证方法的有效性,还把这 2 组结果与 TREC2003 的最优结果[5]进行了比较,实验证明,关键资源页面集合上的检索效果与 TREC2003 主题提炼任务的最优结果性能相当,在 R-precision 评价上还优于这个结果. 这也充分说明了基于非内容信息进行关键资源定位对于主题过滤任务是行之有效的.

5　结论与未来工作

网络数据的爆炸性增长与低质量信息的泛滥给网络信息检索技术的发展带来了巨大的挑战,文中提出了一种综合利用 web 页面的非内容信息进行关键资源页面提取的方法,利用这种方法得到的关键资源页面集合,可以用 20 %左右的 web 页面数量,覆盖超过 70 %的关键信息. 基于关键资源页面集合的检索,也获得了远远超过在页面全集上检索的效果. 这说明利用 web 页面正文内容以外的信息,去除冗余页面,在保证检索效果的前提下,将搜索引擎索引的页面控制在少量高质量页面上是完全可能的. 这对于在索引量一定的条件下提高搜索引擎的信息覆盖率至关重要;同时也为在信息覆盖率一定的情况下减少搜索引擎维护索引的成本提供了一个解决途径.

关键资源页面提取方法也带来了新的问题,从文中的实验结果中可以看出,提取出的关键资源页面集合在检索特征上与 web 页面全集有明显的不同,因此系统考察已有检索方法在关键资源页面集合上的表现,从而确立这个集合上可以应用的方法体系是很有必要的.需要考察的可能内容包括:各种已有检索模型的性能如何;各种链接分析算法是否有效;通常使用的链接文本检索方法是否能取得性能的提高等等.此外,尽管文中在评价关键资源页面集合本身的质量上完成了一些工作,但仍然缺乏从检索性能层次评价集合质量的尝试.这些可能都是未来研究工作的方向.

参考文献:

[1] SULLIVAN D. Search engine sizes [EB/OL]. From search engine watch web site http://searchenginewatch.com/reports/article.php/2156481, 2005-01-28/2005-06-18.

[2] LYMAN P, HAL R V. How much information 2003 [EB/OL]. On line at: http://www.sims.berkeley.edu/how-much-info-2003, 2003-10-30/2005-06-18.

[3] MONIKA R H, MOTWANI R, SILVERSTEIN C. Challenges in web search engines [A]. Georg Gottlob, Toby Walsh eds. IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence [C]. San Francisco: Morgan Kaufmann Press, 2003.

[4] HAWKING D, CRASWELL N. Overview of the TREC-2002 web track [A]. In Voorhees and Buckland [6] [C]. [s.l.],2002.

[5] HAWKING D, CRASWELL N. Overview of the TREC 2003 web track [EB/OL]. On line at: http://trec.nist.gov/pubs/trec12/papers/WEB.OVERVIEW.pdf, 2004-02/2005-01.

[6] VOORHEES E M, BUCKLAND P L. The eleventh text retrieval conference (TREC-2002), volume 11 [M]. National Institute of Standards and Technology, NIST, 2003.

[7] DAVISON B D. Topical locality in the web [A]. Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval [C]. [s.l.],2000.

[8] BHARAT K, HENZINGER M. Improved algorithms for topic distillation in a hyperlinked environment [A]. In 21st International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. [s.l.], 1998.

[9] CRASWELL N, HAWKING D. Query-independent evidence in home page finding [J]. In ACM Transactions on Information Systems (TOIS), 2003, 21(3): 286-313.

[10] WESTERVELD T, HIEMSTRA D, KRAAIJ W. Retrieving web pages using content, links, URLs and anchors [A]. In Voorhees and Harman [7] [C]. [s.l.], 2000.

[11] KRAAIJ W, WESTERVELD T, HIEMSTRA D. The importance of prior probabilities for entry page search [A]. In 25th annual international ACM SIGIR conference on research and development in information retrieval [C]. pages 27-34.

[12] BRODER A. A taxonomy of Web search [J]. SIGIR Forum, 2002, 36(2):1-8.

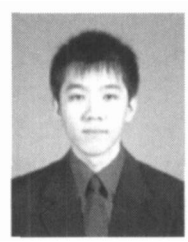
[13] CRASWELL N, HAWKING D. Stephen robertson. effective site finding using link anchor information [A]. In 24th ACM-SIGIR Conference on Research and Development in Information Retrieval [C]. pages 250-257.

[14] MITCHELL T M. Chapter 3: Decision Tree Learning, in Machine Learning [M]. McGraw-Hill International Editions, 1997.

[15] RIJSBERGEN C J. Information Retireval [M]. Butterworths, London, 1979.

[16] HAWKING D, CRASWELL N. Overview of the TREC-2001 web track [A]. In Voorhees and Harman [7] [C]. [s.l.],2001.

作者简介:



刘奕群,男,1981 年生,博士研究生.主要研究方向为信息检索、机器学习与网络用户行为分析.发表学术论文 10 余篇.

E-mail: liuyiqun03@mails.tsinghua.edu.cn.

张 敏,女,1977 年生,助理研究员.主要研究方向为信息检索、机器学习、自然语言处理、基于认知的信息处理,以及在网络环境下用户行为模式的抽取和分析,及其对相关网络信息获取技术.发表学术论文 40 余篇.

马少平,男,1961 年生,教授,博士生导师.主要研究方向为知识工程、信息检索、汉字识别与后处理以及中文古籍数字化.承担过多项国家自然科学基金、"863"高技术项目、"973"项目及国际合作项目.在脱机手写体汉字识别和后处理方面达到了国际先进水平."脱机手写体汉字与数字识别系统"1998 年 1 月获得国家教委科技进步二等奖.发表论文 60 余篇,出版教材 2 部.