基于链接描述文本及其上下文的Web信息检索

张 敏[1] 高剑峰[2] 马少平[1]
[1](清华大学智能技术与系统国家重点实验室 北京 100084)
[2](微软亚洲研究院 北京 100080)
(zhangmin@s1000e.cs.tsinghua.edu.cn)

摘 要 文档之间的超链接结构是Web信息检索和传统信息检索的最大区别之一,由此产生了基于超链接结构的检索技术.描述了链接描述文档的概念,并在此基础上研究链接文本(anchor text)及其上下文信息在检索中的作用.通过使用超过169万篇网页的大规模真实数据集以及TREC2001提供的相关文档及评价方法进行测试,得到如下结论:首先,链接描述文档对网页主题的概括有高度的精确性,但是对网页内容的描述有极大的不完全性;其次,与传统检索方法相比,使用链接文本在已知网页定位的任务上能够使系统性能提高96%,但是链接文本及其上下文信息无法在未知信息查询任务上改善检索性能;最后,把基于链接描述文本的方法与传统方法相结合,能够在检索性能上提高近16%.

关键词 链接文本;链接描述文档;Web信息检索

中图法分类号 TP391

Anchor Text and Its Context Based Web Information Retrieval

ZHANG Min[1], GAO Jian-Feng[2], and MA Shao-Ping[1]
[1](*State Key Laboratory of Intelligent Technology & System*, *Tsinghua University*, *Beijing* 100084)
[2](*Microsoft Research Asia*, *Beijing* 100080)

Abstract One of the most important differences between traditional information retrieval (IR) and web IR lies in the hyperlink structure in web pages. This motivates the so-called link-based retrieval techniques for web IR. The concept of anchor description document is introduced, and then several methods of using anchor text and its context for web IR are proposed. The methods are evaluated using TREC2001 collection which contains over 1.69 million web pages. Several conclusions are drawn: Firstly, anchor text can represent precisely the topic of web page, but is insufficient in describing the web page content. Secondly, comparing with traditional content based IR technique, using anchor text on homepage finding task can get more than 96% improvement in terms of 11-point average precision, while it is not helpful on ad hoc task even with context information. Finally, combining anchor text-based and traditional content-based techniques, more than 16% improvement of performance can be obtained.

Key words anchor text; anchor description document; web information retrieval

1 引 言

Web信息检索和传统文本检索的最大区别在于文档集合的不同,即网页内有超链接结构而普通文本没有.网页内的超链接使得不同文档之间具有一定的连接关系.这种关系主要体现在两个方面:链接结构(link structure)和链接文本(anchor text).由此引出了两类不同的检索技术.

基于链接结构的检索方法目前已经发展得比较成熟了.根据链接结构,我们可以把整个文档集合看做一个有向的拓扑图,而每个网页就是图中的一

收稿日期:2002-09-25;修回日期:2003-06-27
基金项目:国家"九七三"重点基础研究发展规划项目(G1998030509);国家自然科学基金项目(60223004);国家"八六三"高技术研究发展计划基金项目(2001AA114082)

个结点,网页之间的链接就构成了结点间的有向边. 基于这个概念的一种最简单的方法就是根据每个结点的出度和入度来评价网页的作用[1,2]. 出度和入度越大,网页在整个数据集合中就越重要. 此外,Kleinberg 在 IBM 工作期间提出了网页的权威性(authority)和继承性(hub)的概念,并在此基础上提出了 HITS 算法[3]. 这一算法已经应用于 IBM 的 Clever 检索系统中,并在此后经过人们更深入的研究得到了改进[4]. 另外由斯坦福大学的研究机构提出的 PageRank 的算法[5,6]也是基于链接结构的经典方法之一,并在 Google 搜索引擎中成功地加以应用. 这些方法都不考虑链接本身的内容含义,而是根据文档拓扑图的边的方向结构,对文档在整个集合中的重要性加以度量和评价.

而基于链接文本的方法则从另外一个角度入手辅助检索. 链接文本是指当一个网页中具有指向另外一个网页的链接时,与此链接相对应的描述文字. 例如:在网页中有"〈a href ="http://www.moe.edu.cn"〉中华人民共和国教育部〈/a〉"这样的一条链接,则'中华人民共和国教育部'就是地址 http://www.moe.edu.cn 的链接文本. 在基于链接文本的检索方法中,有一个基本假设:链接文本是用来描述它所指向的文档的,而不是用来描述它所在的当前文档[5,7]. 根据这一假设,我们可以在链接文字和其目标文档之间建立联系以辅助检索,这既不同于基于文档自身内容的传统检索方法,又考虑到了链接关系的内容含义. 目前对单独使用链接文本的研究已经开始进行,并在 2001 年的 TREC 会议中有一些报告[8]. 进一步地,有人提出链接文字周围的上下文信息也和其目标文档有更紧密的相似性关系[9]. 而另外一些研究中,则指出这种上下文信息对检索没有什么帮助[10]. 两种观点都没有提供直接和明确的实验结论予以证明.

本文对链接文本在 Web 信息检索中的作用进行深入的研究,通过实验分析,考察使用链接文本及其上下文信息进行检索对系统性能的影响,并在此基础上将链接文本和传统检索技术相结合,从而有效改善系统的检索精确度.

2 方法描述与实验设计

2.1 检索任务

在 Web 信息检索中,通常有两大类不同的检索任务:已知网页定位和未知信息查询. 所谓'已知网页定位'是指用户已经知道(或者猜测)有某一个网页(通常是某个机构,某个门户网站等),但是不知道其具体的 URL 地址是什么,因此用户关心的是找到该网页的入口地址. 例如'中国国家图书馆'这样的查询. 而'未知信息查询'任务则是用户希望查找关于某一主题的信息,但是对这一信息究竟是否存在并不确定,用户关心的是网页的内容是否与自己想要查询的信息相关. 例如'研究生招生'这样的查询就属于这一类任务. 基于这两种任务的不同,某种检索方法在检索中的作用也可能有所不同.

本节首先介绍'链接描述文档'的概念,以及如何使用链接文本进行 Web 信息检索,然后说明实验环境和评价方法,最后给出两种不同任务的检索结果并进行分析.

2.2 链接描述文档

为了使用链接文本,我们首先在整个文档数据集内生成'链接描述文档集". 根据 Craswell 等人 2001 年提出的'链接描述文档'的概念[7],可以给出如下定义:

定义 1. 网页的链接描述文档. 由整个文档集合中指向该网页的所有超链的链接文本的集合构成. 它是在整个数据集中,其他网页在引用到当前网页时所使用的描述文字的集合.

例如,在整个文档集中,假设 Google(http://www.google.com)这个网页共被 4 个网页所链接,其中两个引做'Google 搜索引擎",一个引做'Google 的首页",还有一个对应的链接文本是'点击这里进入 Google",则它的链接描述文档就是"Google 搜索引擎,Google 搜索引擎,Google 的首页,点击这里进入 Google."

一个网页的链接描述文档描述的并不是该网页的作者自己对网页内容的说明,而是其他网页的作者对该网页描述的主要内容的观点概括,这是网页的链接描述文档和其原文档之间本质的区别. 链接描述文档在一定程度上反映了该网页在整个数据集中的作用. 被引用的次数越多,其链接描述文档也就有可能更长. 网页的内容主题或者功能越集中,则描述文档的内容就越突出(即关键主题被多次重复).

用每个网页的链接描述文档来代替原该文档内容,就得到新的链接描述文档集. 在这个新的数据集合上,我们可以建立索引,并用各种检索模型进行信息的检索.

2.3 实验环境及评价方法

文本信息检索会议(text retrieval conference,

TREC)是目前文本信息检索领域中影响最大的实验评测会议. 它对参加者提供统一的文档集和测试查询,使用 pooling 技术[11],用结果集中的文档构成评测集合,并对评测集合中文档的相似性进行人工标注,最后对参加者提交的结果进行统一评价. 因为其检索数据的规模很大,评价方法公正,结果可信,因此,我们选择 TREC2001 中网络检索任务(web track)提供的 WT10g 数据集和测试集[11]对本文提到的所有方法进行实验测试和验证.

该数据集是从 Internet 上下载的真实网页,共 169 多万个文档,所占磁盘空间大小约 10GB. 测试集有两个,分别用于前文提到的两种不同的检索任务. 其中用于未知信息查询任务的问题有 50 个,检索后对每个问题返回最多 1000 个结果文档,总相关文档数为 2590 个. 而已知网页定位任务则有 145 个问题,检索后对每个问题最多返回 100 个结果. 两种任务的相关文档集都由 TREC 提供.

结果评价方法使用 TREC 标准. 即对于未知信息查询任务,使用通用的 11 点平均精确度(11-point average precision) 对于已知网页定位任务,使用平均返回位置倒数、前 10 篇文档平均精确度和未找到比例共 3 项标准.

$$\text{平均返回位置倒数} = \frac{\sum_{q_i} \frac{1}{\text{正确文档在结果中排名}}}{n},$$

其中,q_i 表示第 i 个查询;n 为查询的总个数(这里为 145).

我们把传统的基于内容检索技术作为实验比较的基础. 这里基于内容检索技术就是以网页的原始文档内容作为数据集合,不考虑链接信息的影响.

3　单独使用链接文本

单独使用链接文件构成的链接描述文档集大小约 250MB,只是原文档大小的 1/40 . 在 TREC2001 中的实验结果表明,单独使用链接文本构成的链接描述文档建立索引进行检索,其效果对于两种不同任务的差别很大. 相关实验的数据分别如表 1 和表 2 所示.

表 1　已知网页定位任务检索结果

方法	平均返回位置倒数/%	前 10 篇平均精确度/%	未找到的比例/%
传统检索	22.46	44.1	25.52
基于链接文本	44.06	65.5	25.52

表 2　未知文档查询任务检索结果

序号	平均精确度/%	方法说明
1	20.08	传统基于内容的检索
2	3.12	基于链接文本的检索,短查询语句
3	3.42	基于链接文本的检索,中等查询语句
4	4.02	基于链接文本的检索,长查询语句
5	4.85	方法 2 + 扩展查询技术

对于已知网页定位任务,与传统的方法相比,只用了相当于原文档 1/40 大小的信息,却在平均返回位置倒数性能上的相对提高超过 96%,在前 10 篇文档的平均精确度上的相对提高超过了 48%. 而对于未知信息查询任务的效果则非常差,最好的结果也只有 4.85%,还不到传统方法得到的平均精确度的 1/4.

链接文本在已知网页定位方面非常有效的原因是很清楚的 通常一个网页的链接文本能够集中并准确地表达该网页的主题,甚至有不少的链接文本其本身就是网页的 URL 地址. 这就和网页定位任务的需求达到了一致. 而如果使用传统的网页原文档来进行查找,由于文档本身的长度相对比较大,会造成主题词在整个文档中的作用可能被大量的其他信息所减弱,因而在精确度上会有所降低.

分析基于链接文本进行未知信息查询检索效果很差的原因,主要可以归结为数据稀疏问题,也就是说链接文本无法充分覆盖网页本身所包含的所有主要内容. 统计结果表明(见图 1),在 TREC2001 提供的数据资源中,有 79%网页的链接描述文档长度小于 10 个词,其中占全部文档数 29%的网页是根本没有链接描述文档的(也就是说这些网页在整个数据集合中,从来都没有被其他的网页所链接引用过).这种链接描述文档太短的情况,使得文档内容和查询语句可能根本无法相互匹配,因此能够检索

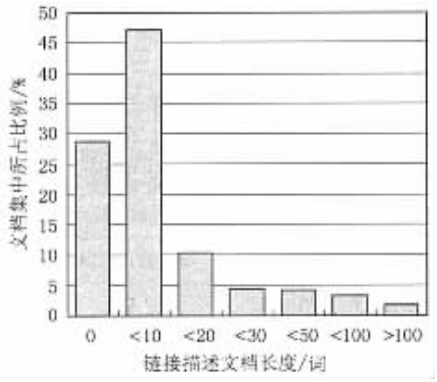

图 1　链接描述文档的长度分布图

到的相关文档数也非常有限. 即使进行了扩展查询,其检索性能的提高也是有限的.

从这样的结果中我们可以得到这样的结论:链接文本对目标网页的主题概括具有高度的准确性,而对内容描述的信息覆盖具有严重的不完整性.

进一步地,我们考虑是否能够采用一种有效的方法提高链接文本的信息覆盖面,解决数据稀疏问题. 也就是我们下面所要描述的使用链接文本的上下文信息来辅助检索的根本想法.

4 使用链接文本的上下文信息辅助检索

4.1 方法描述

使用链接文本上下文信息方法的提出建立在一个基本假设之上:链接文本的上下文,通常与链接文本具有内容上的一致性. 例如我们常见的网页中有这样的内容:"…云南你好 提供云南旅游景点介绍,线路报价,酒店预订,自助旅游指南等综合服务.",其中"云南你好"就是关于网页 http://www.hiyunnan.com 的链接文本,而其下文"提供云南旅游景点介绍,线路报价,酒店预订,自助旅游指南等综合服务"则是对这个链接文本的内容有补充说明的作用. 另外,我们也经常遇到这样的例子:"了解更多健康方面的知识,点击这里". 其中"点击这里"作为目标网页惟一的链接文本,不具有任何信息. 但是通过其上下文则可以把网页的内容限定在"健康知识"方面,对检索的进行将有所帮助.

基于这样的考虑,如果不仅使用链接文本而是同时抽取一定长度的上下文来建立链接描述文档,也许能够在一定程度上解决数据稀疏问题,从而改进检索的效果. 这里关于上下文的定义,采用链接文本同一段内前后(不包括链接文本)大小为 n 个词的窗口. 在我们的实验中,$n=20$. 但是这种做法显然在增加信息的同时,也必然会引入噪声. 究竟有用的信息和噪声哪一个影响更大,则有待于实验的验证说明. 因此,我们在同样的数据集和测试集上,加入链接文本的上下文信息,生成了扩充的链接描述文档集合,并在未知信息查询任务上进行了实验比较分析.

4.2 实验结果

4.2.1 文档长度分布

使用链接文本的上下文(前后 20 个词),得到的链接描述文档的统计分析如表 3 和图 2 所示. 可见使用链接文本的上下文,使得链接描述文档的长度有了很大的提高. 小于 10 个词的文档数由原来的 79% 减少到了 70%,而长描述文档的个数也有很大增加.

表 3 链接描述文档的长度

链接描述文档的长度/词	文档个数	
	只用链接文本	使用链接文本的上下文
0	487866	487866
0～10	795892	705453
10～20	173882	181978
20～30	73388	99393
30～50	71695	95470
50～100	57235	71433
100～300	24460	37431
>300	7678	13072
总文档数	1692096	

注:考虑链接文本上下文信息 vs 只用链接文本

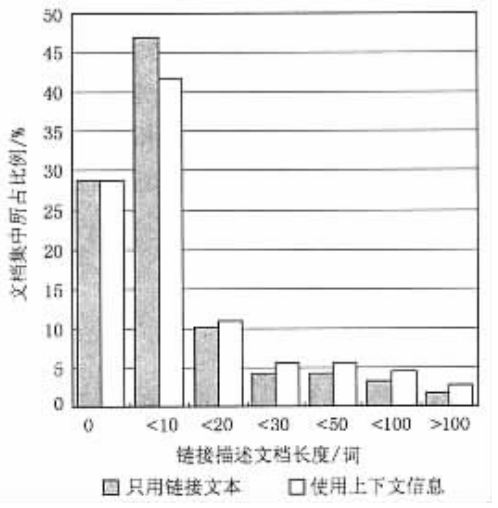

图 2 链接描述文档的长度分布比较图

4.2.2 对检索的效果

遗憾的是使用链接文本的上下文信息并不能有效地提高检索性能. 表 4 是在同样条件下,使用两种描述文档的检索效果(平均精确度和查全率)的比较. 其中方法 4～6 中的相关反馈(m,n)表示采用伪反馈(pseudo relevance feedback)技术,把第 1 次检索结果中最相似的前 m 篇文档中权值最高的 n 个词扩展为查询词.

与只用链接文本相比,其检索性能在平均精确度和查全率两个评价标准上的效果是不同的. 一方面,除了使用长的查询以外(参见表 4 中的方法 3),大多数情况下使用上下文信息都使平均精确度降低了(参见表 4 中的方法 1,2,4,5,6). 另一方面,在所

有的测试方法中,使用链接文本的上下文信息检索到的相关文档数都比原来只用链接文本进行检索得到的相关文档数多,也就是说查询率有所提高.

但是无论用哪个标准来评价,即使考虑了链接文本的上下文信息,使用链接描述文档进行检索的效果都依然远差于传统的基于网页本身内容的检索方法. 于是我们可以知道,在上下文信息中引入的噪声产生的影响大于所加入的有用信息的作用.

表 4　检索的结果比较

方法	平均精确度		返回的相关文档数		方法描述
	包含上下文/%	只用链接文本/%	包含上下文	只用链接文本	
1	3.01	3.12	281	269	1 次检索,短查询语句
2	3.47	4.02	309	290	1 次检索,中查询语句
3	3.85	3.25	266	248	1 次检索,长查询语句
4	4.15	4.94	332	297	相关反馈(5,10)
5	3.29	4.48	323	292	相关反馈(5,20)
6	3.59	4.85	317	291	相关反馈(5,30)

注:使用链接文本上下文 vs 单独使用链接文本

4.2.3　分析和讨论

分析使用链接文本的上下文信息在大多数情况下无法改进系统性能的原因可能在以下 3 个方面:

(1) 正如我们所担心的,使用链接的上下文虽然引入了更多的有用信息,但是同时也带入了很多噪声,这些噪声在检索的过程中起到了相当大的负面影响. 例如‘友情链接 新浪网站’等这样的情况,其上信息‘友情链接’就是一个干扰因素. 尤其是当一个网页内有多个链接并列的时候,上下文信息会混在一起,因此,噪声的干扰可能会更严重.

(2) 使用上下文信息,只能够使已有的链接描述文档长度在一定范围内增加,并且增加的幅度有限(容易想到,如果选取太大的窗口定义上下文,必然引入更多的噪声). 而对于那些本身就没有链接描述文档的网页来说,这种方法无法带来任何帮助. 前面的分析中已经提出,这样的文档共有 487866 个,约占总数据集总文档数 1692096 的 29%,这就使得仅仅使用链接相关技术,相当于丢弃了接近 30%的文档,这些文档无论如何也不可能被检索到. 这种数据稀疏性问题仍然是造成检索效果不好的主要原因.

(3) 使用链接的上下文使得检索到的相关文档数增加了. 但是检索精确度反而下降. 说明有更多的相关文档已经被检索出来,但是根据文档的相似性排序,有很大一部分的排名都变得靠后了. 根据这样的情况,可以考虑在结果列表中按照不同的标准对结果重新排序,有可能得到更好的结果.

5　基于链接文本的技术与传统方法相结合

虽然使用链接描述文本及其上下文信息,都无法对未知信息查询任务提供帮助. 但是从前面已知网页定位检索任务的结果中,仍然可以清晰地看到链接描述文本本身对网页主题描述的准确性. 于是有理由认为,如果能够把链接描述文档和传统的检索技术相结合,则有望改进系统的检索性能.

为了验证这一想法,我们提出一种简单地将两种方法相结合方法:将链接描述文档和网页原有的文档进行内容上的简单合并,然后在扩充了的文档集合上建立索引,进行检索,最后对检索的性能进行测试. 进一步地,我们还根据单独使用链接文本得到的结果,对合并文档集得到的结果进行重新调权,生成新的文档结果集合,并进行性能测试. 如表 5 所示,测试的结果是令人满意的,两种方法都使系统性能得到了很大的改善. 和传统的基于原文档内容的方法(平均精确度 20.08%)相比,使用合并后的文档进行检索,系统性能相对提高了 10.7% 和 13%,再经过重新调权,系统性能能够提高 15.8%.

表 5　合并文档前后的检索结果比较(平均精确度)

方法描述	传统方法/%	合并文档方法/%	性能提高/%
1 次检索	20.08	22.23	10.7
查询扩展	20.21	22.84	13.0
结果调权	<20.08	23.28	15.9

表 5 中根据结果调权的传统方法中的结果,“< 20.08% ”是说明无论采用怎样的权值,其平均精确度都小于 20.08%. 相应的相对提高 15.9%就是相对于这个上限 20.08%的性能提高.

6 结 论

本文主要研究链接文本及其上下文在 Web 信息检索中的作用. 首先描述了链接描述文档的概念,生成不同于原文档的新的数据集合,考察链接文本对检索的作用. 进一步地,我们研究了引入链接文本上下文信息辅助检索的效果,最后给出了一种简单地将传统检索技术与基于链接文本技术相结合的方法. 所有方法均使用 TREC2001 中 Web Track 任务的数据及查询进行测试. 根据实验结果,得到以下结论:① 链接描述文本对目标网页的主题概括具有高度的准确性,而对目标网页的内容描述的具有严重的不完全性;② 单独使用链接文本,对于已知网页定位任务,其效果远远好于传统的基于文本内容的检索方法;③ 对于未知信息查找任务,由于连接描述文本的数据稀疏性问题,则不能有效提高系统的性能;④ 使用连接文本的上下文信息,和单独使用链接文本相比,虽然检索到的相关文档数有所增加,但是系统的平均精确度却降低了,因此也无法提高系统检索性能;⑤ 把基于链接描述文档的技术与传统的基于内容检索的技术相结合,能够得到最大接近 16%的相对精确度的提高,大幅度地改善了系统检索性能.

在未来的工作中,我们将进一步研究更多的将两种技术相结合的方法,以更大程度地发挥链接文本的特点.

参 考 文 献

1 R Botafogo, E Rivlin, B Shneiderman. Structural analysis of hypertext: Identifying hierarchies and useful metrics. ACM Trans on Information System, 1992, 10(2): 142~180

2 J Carriere, R Kazman. WebQuery: Searching and visualizing the Web through connectivity. The 6th Int'l WWW Conf (WWW6), Santa Clara, 1997

3 Jon M Kleinberg. Authoritative sources in a hyperlinked environment. The 9th Annual ACM-SIAM Symp on Discrete Algorithms, California, 1997

4 K Bharat, M R Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. The 21st Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 98), Melbourne, 1998

5 S Brin, L Page. The anatomy of a large-scale hypertextual web search engine. The 7th Int'l WWW Conf (WWW7), Brisbane, Australia, 1998

6 L Page, S Brin *et al*. The pagerank citation ranking: Bringing order to the web. 1998. http://dbpubs.stanford.edu:8090/pub/1999-66

7 N Craswell, D Hawking, S E Robertson. Effective site finding using link anchor information. The SIGIR 2001, Louisiana, 2001

8 Gao Jianfeng *et al*. TREC-10 Web track experiments at MSRA. The 10th Text Retrieval Conf, Gaithersburg, 2001

9 S Chakrabarti, B Dom, D Gibson *et al*. Automatic resource compilation by analyzing hyperlink structure and associated text. The 7th Int'l WWW Conf (WWW7), Brisbane, 1998

10 B D Davison. Topic locality in the web. The 23rd Int'l ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 2000), Athens, 2000

11 Ellen M Voorhees, Donna Harman. Overview of TREC2001. The 10th Text Retrieval Conf, Gaithersburg, 2001

张 敏 女,1977 年生,博士研究生,主要研究方向为信息检索.

高剑峰 男,1971 年生,研究员,主要研究领域为统计自然语言处理、信息检索.

马少平 男,1961 年生,教授,博士生导师,主要研究领域为模式识别、信息检索、网络数据挖掘等.