一种提高搜索引擎检索质量的网页解析法*

钟楚玲 朱 丹 曹二堂
(河北经贸大学，河北 石家庄 050061)

摘 要：通过实验对网页结构和特点进行综合分析，给出对网页分块的原则和方法，在分块的基础上根据网页中噪音的出现规则提出了一种消除网页噪音的方法，使搜索引擎对网页的预处理阶段有效消除网页中的无关项和间接项的超连接，从而大大提高了搜索引擎的检索质量。

关键词：检索质量；分块模型；搜索引擎
中图分类号：TP306 **文献标识码**：A

A Web page parser to improve search engine retrieval quality

ZHONG Chu Ling, ZHU Dan, CAO Er Tang
(Hebei University of Economics and Trade, Shijiazhuang 050061, China)

Abstract: Through the comprehensive analysis of the Web page structure and features, the principles and methods for web segmentation are provided. Based on the patterns of web noise in web segmentation, a noise elimination method is also given, which can effectively eliminate the hyperlinks to irrelevant and indirect items, thus greatly enhancing the retrieval quality of search engines.

Key words: retrieval quality; segmentation model; search engine

随着 Internet 的快速发展，大量的信息呈现在用户面前，据统计，国内 Web 网页数量达 3 亿以上[1]，上网用户总人数达 8 700 万，将获取信息作为上网最主要目的网民所占比例最多，达到 42.3%[2]。数据表明，Internet 已成为人们获取信息的重要资源，而 Google、Yahoo、百度、新浪、天网等中英文搜索引擎是人们徜徉信息海洋、获取信息的工具。然而，人们面对如此丰富的 Web 资源，使用搜索引擎发现自己真正需要的信息却并非容易。一方面，各搜索引擎不断改进检索技术来提高返回结果的精度，在一定程度上解决了人们获取信息的问题；另一方面，由于搜索引擎自身的问题，返回的结果与用户的要求仍有一定的距离，用户对搜索引擎的满意度不太高。主要表现为查询结果中普遍存在大量的无关项和不含具体内容的间接项，造成搜索结果数量大、结果不精确、有用的结果淹没在无用的结果之中的局面。用户不得不花费大量的时间在查询结果中寻找相关项，使得用搜索引擎来查找信息的目的难以达到。这种结果的原因之一是目前的搜索引擎没有对网页进行处理或只做了简单的处理。

目前的搜索引擎采用以关键字检索为基础的检索技术[3-4]，即搜索引擎按关键字对整个网页进行索引和检索。在这种处理方法中，所有出现在网页中的字词都被用作索引项，但实际的网页中常常包含大量的与网页主题无关的文字。例如，图 1 和图 2 是以“河北人民出版社”为关键字的检索结果。图 1 所示网页的主要内容是关于 2004 十大印象图书介绍，其中包括上海人民出版社出版的《达芬奇密码》，在网页中注明的出处是新华网河北频道。在这个网页中包含了“河北”和“人民出版社”，搜索引擎误把它当做“河北人民出版社”的相关项。图 2 所示网页的主要内容是一些图书的介绍，在左边的导航栏中出现了河北人民出版社的连接，真正提供具体信息的应该是它指向的那个页面，而那个页面也应该能被检索到，因此，图 2 所示网页是多余的间接项。

* 基金项目：河北省教育厅科学研究计划项目(项目编号：2008202)

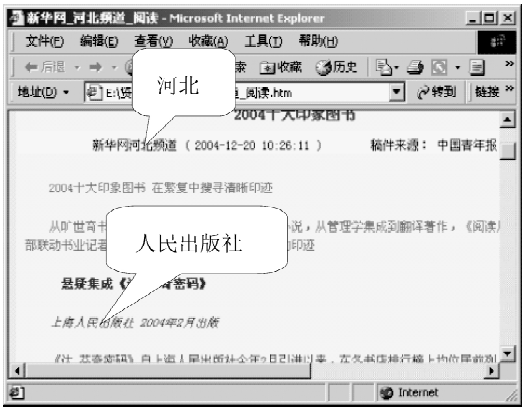

图1 无关项示例

图2 间接项示例

如果搜索引擎在对网页标引时，把整个网页上不同主题、不同作用的文字混合在一起进行处理，那么，在检索过程中根本无法排除如图1所示的无关项。使用站点聚类技术，把出现在同一个站点上的结果项进行合并，虽然可以排除大部分如图2所示的间接项，但是耗费了查询时间。本文提出一种在标引前对网页进行预处理的方法，能够排除上述的无关项和间接项。

目前的搜索引擎对网页的预处理较简单，几乎保留了HTML网页上所有的文字，这样固然可以保证查全率，但从目前的网络资源巨大丰富的角度来看，提高查准率对用户更具有实际意义。在研究领域里，有人提出了基于HTML标记结构的规律对特定网站进行信息抽取[5]，但不满足搜索引擎对多种多样的网站进行处理的要求；有人提出“语义块”的概念对网页内容分层，但没有具体的实现方案[6]；对于超连接的研究主要集中在对它所指向的页面在检索中的作用[7]，但很少有人研究超连接对网页的负面影响。

1 HTML网页的块结构模型和解析方法

1.1 HTML网页的块结构模型

通过对大量的网页进行分析，发现人们在设计网页时通常是把网页设计成几个区域，把不同主题、不同作用的文字安排在不同的区域。结合HTML标记的特点，认为网页是由块组成的，块中可以再嵌套块。因此，HTML网页的块结构模型是：{<块起始标记><块内容><块结束标记>[,<块起始标记><块内容><块结束标记>，…]}。其中，块内容中可以再包含块。实际的网页大多是由多层的块嵌套构成的。

1.2 分块原则及算法

HTML块标记有<hr>、 、<table>、<tr>、<td>、<p>等。在实际应用中，块的划分要合理。块划分得过多，会把相关的内容划分到不同的块区，这样将导致网页与查询关键字的相关度降低；块划分得过少，会把不相关的内容划分到同一个块区，这样将导致查准率的降低。例如，一篇文章由标题、作者、出处和多个段落组成，显然这些文字应划分在同一个块区。经过对大量网页的统计分析，不外乎两种情况。一种是网页中不包含<Table>标记，只有一篇文章，显然，这类网页只有一个块区；另一种是网页中包含多个<Table>标记，而一篇文章的标题、作者、出处和多个段落一般安排在某一个表格的一个或多个单元格中。因此，将网页中的表格(<Table>标记)做为块区比较合理。

分块原则如下：

(1)如果网页中包含水平线标记<hr>，首先按水平线分块；

(2)在上述分块的基础上，如果包含 、<table>标记，按 、<table>分块；

(3)如果在 、<table>中包含水平线标记<hr>，再按水平线分块。

分块算法如下：

查找水平线标记，插入块标记；

While(文件没有结束)

{查找块起始标记和结束标记，位置存入tableLoc()；

同时，在tableSym中简记为b和e； }

将tableLoc中的位置数据排序，同时调整tableSym中的b、e标记；

While(tableSym中的标记数不等于0)

{ 查找“be”；

提取块；

tableSym中的标记数减2；}

1.3 消除噪声的规则

人们在制作网页时，总是准备了一定的素材，这些素材是网页设计者希望通过网页传达给访问者的信息。但同时也会在网页中增加一些连接到其他网页的超连接，而这些超连接文字的作用仅仅起着向导作用，与页面主题无关，它们的加入会影响到页面的原貌，把这样的超连接文字定义为网页的“噪声”，把网页中原本要表达的内容定义为网页的“主题内容”。

通过对大量网页的统计分析，噪声主要来源于超

连接文字，但并非所有的超连接文字都是噪声，因此要准确地消除网页中的噪声也并非容易。

网页中的超连接文字可分为3类：

(1)超连接文字在网页中仅仅起着向导作用，其目的是提供一个访问目录。超连接文字在它所指向的网页中还会出现，这些页面能够被搜索引擎搜索到。因此，这类超连接文字是本网页的噪声。一般说来，这类超连接文字的前后还是超连接文字，所以噪声通常聚集成块。

需要说明的是索引网页中的超连接文字虽然是网页的主题，但是超连接文字在它所指向的网页中还会出现，这些页面通常能够被搜索引擎搜索到，所以，本网页不必出现在搜索结果中。

(2)超连接文字在网页中具有向导和陈述的双重功能，超连接文字引向另一个网页或本网页的其他位置的同时，本身也是网页主题内容的一部分，这样的超连接文字也是网页的主题内容，而不是噪声。一般说来，这类超连接文字的前后的文字不是超连接。

(3)超连接文字所指向的目标文件中不会出现此超连接文字，目标文件是搜索引擎不能直接搜索到的文件。例如，超连接文字指向的目标是MP3格式文件、.exe格式文件或图片格式文件等，这些超连接文字不能视为网页的噪声。

从网页的结构上看，(1)类超连接文字聚集成块，超连接文字与块区内所有文字的比值R接近于1；(2)类超连接文字处在主题内容块区，超连接文字与块区内所有文字的比值R远小于1。通过实验确定两个阈值$R1$和$R2$。若$R>R1$，则确定为噪声；若$R<R2$，则确定为网页的主题内容。

根据上面的分析，在对网页分块的基础上确定消除网页噪声的规则：

(1)在块区中扫描超连接，如果超连接指向的目标是网页，则将此超连接文字标记为准噪声；如果超连接指向的目标不是网页，则在网页中保留此超连接文字。

(2)统计块区内超连接文字数量及文字的总数量并计算其比值R，若$R>R1$，保留准噪声标记；若$R<R2$，删除准噪声标记；若R介于$R1$与$R2$之间，转(3)进一步检测。

(3)检查超连接前后相邻的文字是否是超连接，如果相邻的超连接数S大于某一阈值，将此超连接文字的准噪声标记删除。

2 实验及结果分析

本文开发了一个HTML网页解析器实现了上述算法。实验中使用的网页都是根据著名搜索引擎的搜索结果下载的真实网页。实验中参数的取值分别是：$R1=0.9$；$R2=0.3$；$S=3$。由于文章篇幅的限制，在此略去实验结果的图片。

实验一是网页的分块实验，实验中对数十个网页进行了分块，正确率达100%；实验二使用100个网页进行了消除(1)类超连接文字噪声的实验，其中98个网页的无关项超连接和间接项超连接都被消除；实验三和实验四是保留(2)类超连接文字和(3)类超连接文字的实验，正确率达100%。

实验二的正确率与$R1$、$R2$、S的值有关。对于参数S而言，如果值过小，就会把一些有用的超连接文字消除，例如文章的标题、作者、出处都有超连接时，这些文字是网页的重要内容，不应消除；如果S的值过大，会将一些噪声保留。通过对大量网页的统计分析，认为S取值为3较合适，这样即使在网页中保留一些噪声，由于数量较小，对网页的影响也不大，同时对网页有用的超连接文字也不会被误认为是噪声而消除。

本文介绍的网页解析方法在搜索引擎和数据挖掘方面具有重要的意义和应用前景。通过消除网页的噪声，使网页的主题更加突出。在搜索引擎的返回结果中排除了无关项和间接项，提高了搜索引擎的查准率；在网络使用行为挖掘领域，分析用户感兴趣的网页方面，由于排除了噪声的干扰，使得分析结果更准确。

参考文献

[1] 中国互联网信息中心. 2003年中国互联网络信息资源数量调查报告.信息资源开发利用调查报告[DB/OL]. http://www.cnnic.net.cn/download/ manual/report20030330.doc: 60.

[2] 中国互联网信息中心. 第十四次中国互联网络发展状况调查统计报告(2004年7月) [DB/OL].http://www.cnnic.net.cn/download/2004/2004072002.pdf

[3] 杜阿宁，方滨兴，胡铭曾，等. 中文交互式网络搜索引擎及其自学习能力[J].计算机工程与应用，2003(10):148-150.

[4] 陈俊杰，薛云，宋翰涛，等. 基于Agent的元搜索引擎的研究与设计[J].计算机工程与应用，2003(10): 33-36.

[5] KUSH M N, WELD DS,DOOREMBOS. Wrapper Induction for Information Extraction,proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, 1997: 729-735.

[6] CARCHIOLO V, LONGHEU A, MALGERIM. Malgeri,M.,Structuring the Web,Database and Expert Systems Applications,2000. Proceedings.llth International Workshop on,1123-1127,2000.

[7] N.Cras well, D. Hawking,S. e. Robertson,Effective Site Finding Using Link Anchor Information,SIGIR 2001,2001.

(收稿日期：2008-11-30)