第 34 卷 第 19 期 计 算 机 工 程 2008 年 10 月
Vol.34 No.19 Computer Engineering October 2008
·软件技术与数据库· 文章编号：1000—3428(2008)19—0030—02 文献标识码：A 中图分类号：TP391

一种基于锚文本的并行检索策略

高 珊，何婷婷，胡文敏
(华中师范大学计算机科学系，武汉 430079)

摘 要：进行 Web 信息检索时，页面中的锚文本与正文存在较大相关性，多数检索系统忽视了锚文本对页面正文的贡献。该文提出一种提高检索精度的方法，为文档集建立一个基于页面正文的索引和一个基于锚文本的索引，对其采取并行检索策略。实验结果表明，该方法可以有效处理特定结构的网页集。

关键词：锚文本；并行检索；信息检索

Parallel Retrieval Strategy Based on Anchor Text

GAO Shan, HE Ting-ting, HU Wen-min
(Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】 Most retrieval systems ignore the anchor text, which is highly relevant to the page content in Web information retrieving. This paper proposes a method to improve the retrieval accuracy. It makes two indices, one for page content and the other for anchor text. A parallel retrieval strategy is utilized for the two indices. Experimental results show that this method is efficient for the special structure document collection.

【Key words】 anchor text; parallel retrieval; information retrieval

1 概述

人们经常利用互联网进行信息检索，用户关心的是如何快速、准确地检索到与其查询请求相关的内容。因此，信息检索系统的设计者在设计系统时应考虑检索效率和准确率。很多研究者致力于信息检索的研究，并提出了多种方法和技术手段以满足用户越来越高的查询需要。但在设计 Web 检索系统时，多数开发者只为提取的页面正文建立索引，忽视了锚文本对网页的重要性。一个网页中的锚文本和页面正文的相关性通常较大。因此，笔者在本文检索系统中为待查询网页集建立 2 个索引，并在检索过程中针对这 2 个索引采取一种并行检索策略。

2 系统描述

本文检索系统的检索过程如图 1 所示。

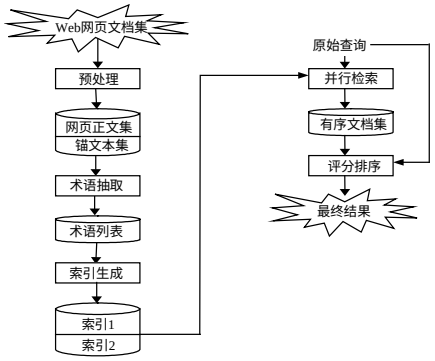

图 1 检索过程

检索步骤如下：

(1)对下载的网页进行预处理，得到 2 个文档集，即网页正文集和锚文本集。

(2)对 2 个文档集进行术语抽取，合并抽取到的术语，得到一个术语列表。

(3)根据第(2)步得到的术语列表对 2 个文档集分别建立索引。

(4)用原始查询在 2 个索引上进行并行查询，得到 2 个有序文档集。

(5)对 2 个有序文档集中的文档进行评分、合并和排序，得到最终的排序文档集合。

3 预处理

3.1 获取页面正文和锚文本

本文利用网络爬虫抓取大量网页，所选研究资源是从网页上摘取的文本内容。一个网页的内容可分为 2 类：供浏览器使用的标记信息和供用户阅读的信息，自然语言处理技术适用于后者。因此，必须去除网页中的标记信息。按内容划分，一个网页一般由导航信息、网页标题、网页正文、广告信息、相关链接等部分组成。

因为要为正文和锚文本分别建立索引，所以需要获取正文和锚文本。因为很多锚文本是链接到广告等与页面内容无关的信息，所以在获取锚文本时，要根据新闻网页结构的特殊性，抽取相关锚文本。例如，在正文下方通常有相关新闻的字样，笔者只抽取出现在相关新闻下方表格里的锚文本，因为在整个网页中，只有这部分的链接信息与正文内容相关度较大，所以可以有效剔除与正文内容不相关的锚文本，从而提高系统准确率。

在获取锚文本时，为该网页建立一个文本文件，用于存

基金项目：国家自然科学基金资助项目(60673040)；国家社会科学基金资助项目(06BYY029)；教育部科学技术研究基金资助重点项目(105117)

作者简介：高 珊(1982－)，女，硕士，主研方向：自然语言处理；何婷婷，教授、博士后；胡文敏，硕士

收稿日期：2007-12-22 **E-mail：**cecil@mails.ccnu.edu.cn

放与正文内容相关度较大的所有锚文本。待检索网页集经过预处理后，得到 2 个文档集合：页面正文集和锚文本集。后续处理将为这 2 种文档集分别建立索引。

3.2 术语抽取

由于中文与西方语言不同，词与词之间没有空格加以区分，因此如何将一篇中文文档切分成可检索的单位，是中文信息检索的难点。采取不同的检索单位对查询准确率和召回率具有一定影响。多数现有系统采用二元或词作为索引的基本单位，这可以保证检索的召回率，但在查询扩展时效果不理想。因此，本文采用从查询文档集合抽取的术语作为索引的基本单位以提高查询扩展有效性[1-2]。笔者采用一种基于聚类的方法[3-4]对文档进行术语抽取，得到一个术语列表。

3.3 数据建模

现有多种信息检索算法模型，如布尔模型、概率模型、向量空间模型、神经网络模型、遗传算法模型及模糊集合模型等，有其各自的适用范围和优缺点。其中，检索效果较好且较通用的是向量空间模型和概率模型。

向量空间模型是一种易于理解且广泛用于信息检索领域的检索模型。本文实验采用向量空间模型，在此模型中，将一个文档表示为一个由 N 个术语组成的术语串向量。向量中第 i 项的值表示 T_i(第 i 个term)在文档中的权值。用户的查询请求也表示为一个由 N 个术语组成的术语串向量。通过计算向量之间的距离计算文档和查询的相似度。相似度值越大，表明文档与该特定查询的相关性越高。

对数据建模过程算法的描述如下：

输入：经过分析处理的所有页面正文文件集 F_1 和锚文本文件集 F_2。

输出：用向量表示的页面正文文件集 D_1 和锚文本文件集 D_2。

根据式(2)计算术语 T_j(j＝1, 2,…, n)在 D_{ki}(文档集 D_k 中的第 i 个文档，k=1, 2)中的DF-IDF权值 $w_k(i, j)$，生成所有文件的文件向量。

$$idf_{kij}=\lg\left(\frac{n}{df_{kj}}\right) \tag{1}$$

$$w_k(i,j)=(1+\lg(tf_{kij}))\times idf_{kij} \tag{2}$$

其中，n 为检索空间的大小，即用来生成索引的术语个数；idf_{kij} 为倒排文档频率；tf_{kij} 表示 T_j 在文档 D_{ki} 中出现的次数；$w_k(i, j)$ 表示 T_j 相对于 D_{ki} 的倒排文档权重。

4 索引生成

对原始数据建立索引是为了快速定位查询词所在位置，为了达到此目的，索引结构很关键。目前的主流方法是以词为单位构造倒排文档表[5]。在本文系统中，对 2 个文档集以术语为单位构造 2 个倒排文档表，其结构类似，如图 2 所示。

$term_1$	invertedlist
$term_2$	invertedlist
⋮	⋮
$term_n$	invertedlist

→

$docid_1$	tf_{11}	pos1	pos2	…	$postf_{11}$
$docid_2$	tf_{12}	pos1	pos2	…	$postf_{12}$
⋮	⋮	⋮	⋮	⋮	⋮
$docid_n$	tf_{1n}	pos1	pos2	…	$postf_{1n}$

图 2 倒排文档表

5 检索文档

5.1 并行检索策略

为页面正文集和锚文本集分别建立索引，针对这 2 个索引，在查询中采取并行检索策略。当用户提出查询请求时，分别对 2 个索引进行检索。

检索步骤如下：

(1)对用户的查询请求 Q 进行术语抽取，抽取 m 个术语。

(2)利用步骤(1)得到的 m 个术语，分别检索 2 个倒排索引表，得到一个正文文件结果集 R_1 和锚文本文件结果集 R_2，其中，R_1, R_2 分别是 D_1, D_2 的子集。

(3)用式(3)分别计算 Q 与 R_1, R_2 中所有文件的相似度 Sim_1 和 Sim_2。

$$Sim_k(Q, R_{kj})=\frac{\sum_{i=1}^{n} w_k(i,j)\times w_q(i)}{\sqrt{\sum_{i=1}^{n}(w_q(i))^2}\sqrt{\sum_{i=1}^{n}(w_k(i,j))^2}} \tag{3}$$

其中，k=1, 2；$w_q(i)$ 表示术语 T_i 在查询 Q 中的权值。

(4)选择并输出 R_1 中 $Sim_1(Q, R_{1j})$ 值排在前 1 000 的文件集 S_1，选择并输出 R_2 中 $Sim_2(Q, R_{2j})$ 值排在前 1 000 的文件集 S_2。

5.2 合并结果集

查询结束时可得 2 个相关文档集，利用它们可得 2 个与之关联的网页集。因为这 2 个网页集中有大量重复网页，所以不能直接将其作为相关网页返回给用户，应对它们进行合并。出现在 2 个集合中的网页其重要性不同，本文给其中的文档以不同的权值，再进行评分、合并和排序。由于提取的锚文本和网页内容的相关度较高，因此对于出现在第 2 个集合中的网页应给予一个较高的权值。

用式(4)对 2 个集合中的文档进行评分、合并，并以此作为排序的依据。

$$Score(i)=\lambda Sim_1(Q,D_{1i})+(1-\lambda)Sim_2(Q,D_{2i}) \tag{4}$$

一个网页如果只出现在第 1 个集合中，其 *Score* 值为该网页对应的页面正文和查询的相似度值与 λ 的乘积；如果只出现在第 2 个集合中，其 *Score* 值为该网页对应的锚文本和查询的相似度值与(1-λ)的乘积；如果同时出现在 2 个集合当中，那么此网页的 *Score* 值较大，在进行排序时，其位置会相对靠前。本文 λ 的值设置为 0.35。

6 实验及评估

本文实验采用从新浪网下载的 2007 年 1 月~2007 年 3 月的新闻网页作为文档集。该文档集共有 95 536 篇文档，人工构造 50 条查询。

实验以 SMART 11.0 为基准系统，对每条查询分别用网页正文和锚文本进行检索。SMART 信息检索系统是基于向量空间检索模型的信息检索系统，其根本目的是为信息检索研究提供一个研究框架，包括建立索引、检索和评价等基本功能。比较本文方法和 SMART 11.0 基准系统的实验结果，如表 1 所示。实验结果表明，本文方法显著提高了检索的平均精度。

表 1 平均精度比较

基准系统	本文方法	本文方法相对基准系统的提高/(%)
0.251 2	0.322 6	28.4
0.284 4	0.350 1	23.1

7 结束语

在信息检索系统中，如何提高检索文档的准确率是开发者必须关注的问题。本文通过增加一个索引，即基于锚文本索引，提高检索文档和查询的相关度。采用一种并行查询策略提高查询效率。为锚文本集建立索引可极大提高检索的平均精度。

（下转第 34 页）