第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·博士论文· 文章编号：1000—3428(2008)13—0019—02 文献标识码：A 中图分类号：TP393

K-FTP 搜索引擎的核心技术

胡 亮[1]，傅泽田[1]，张小栓[1,2,3]，赵 明[2]，郭立力[1]，宫薇薇[1]
(1. 中国农业大学工学院，北京 100083；2. 中国农业大学信息电子工程学院，北京 100083；
3. 苏州大学计算机信息处理技术重点实验室，苏州 215006)

摘 要：传统 FTP 搜索引擎对检索结果优化程度不够，会降低检索质量。在 FTP 用户查询日志的统计分析基础上，采用双字节倒排索引、检索结果自动分类以及查询自动纠错等技术设计了一种高性能的智能化 FTP 搜索引擎。试验表明该方案能够有效地提高 FTP 文件检索效率与质量，平均检索响应时间低于 500 ms，检索准确率为 92.5%。

关键词：FTP 搜索引擎；倒排索引；自动分类；自动纠错

Kernel Technology of K-FTP Search Engine

HU Liang[1], FU Ze-tian[1], ZHANG Xiao-shuan[1,2,3], ZHAO Ming[2], GUO Li-li[1], GONG Wei-wei[1]
(1. Technical College, China Agricultural University, Beijing 100083; 2. College of Information and Electrical Engineering, China Agriculture University, Beijing 100083; 3. Key Laboratory of Computer Information Processing Technology, Suzhou University, Suzhou 215006)

【Abstract】The quality of query results in the traditional FTP search engines is low because it is not optimized. To solve the problem, this paper builds a high performance and intelligentized FTP search engine——KFSE, based on the analysis of FTP user query logs. The double bytes inverted index, automatic classification of query results, and automatic rectify mistake for users are adopted in the system. Validity of the scheme is proved in the real system and it can improve the query efficiency and quality for the FTP search engine. The average of response time is lower than 500 ms and the precision is 92.5%.

【Key words】FTP search engine; inverted index; automatic classification; automatic rectify mistake

FTP 搜索引擎是搜索引擎技术体系的一个研究方向，搜集匿名 FTP 服务器提供的目录列表以及向用户提供文件信息的查询服务。虽然 FTP 搜索引擎与 Web 搜索引擎都是对字符串进行匹配查找网络文档的链接，但它们所处理的数据对象在具体细节上有很多不同，例如：FTP 搜索不要求显示结果的内容摘要，对 FTP 站点各目录的数据刷新要求有不同的刷新速率，查询时需要文件信息、站点信息过滤等。

1 系统体系结构与设计

本文旨在设计一个高性能、大规模的 FTP 搜索引擎信息检索系统(KFSE)，系统体系结构如图 1 所示。

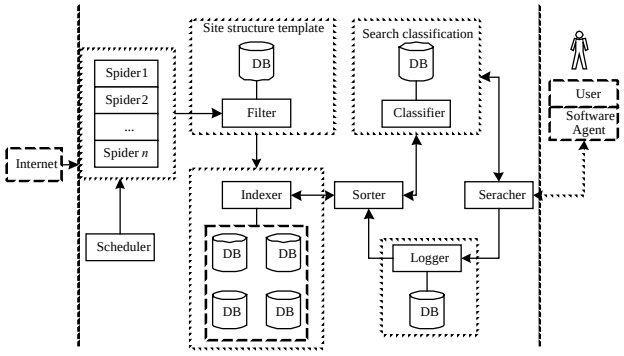

图 1 KFSE 体系结构

从网络用户对 FTP 搜索的实际需要出发，主要考虑解决传统 FTP 搜索引擎中存在的数据实时性(freshness)、搜索速度(response time)与智能化搜索(intelligentized search)等问题，因此，笔者从上述几个方面介绍 KFSE 应用中的核心技术，主要包括双字节倒排全文索引、检索结果自动分类以及查询自动纠错。

2 双字节倒排全文索引

Web搜索引擎采用的索引算法是先对网页进行中文分词，然后提取关键词项对网页数据文件建立倒排索引表，网页数据量大的特点限制了全文索引方式的应用。对FTP搜索引擎来说，采集的数据对象是FTP服务器上的文件名，与Web搜索引擎的数据对象网页相比数据量要少很多，FTP文件名字符串长度受文件系统的限制不超过 255 B，包括中文与英文字符。研究表明，网络上FTP文件名长度为 12 B的文件最多(约占 12.5%)，文件名的平均长度为 16.7 B，最长文件名长 200 B[1]，因此，在KFSE系统中索引方式采用全文索引结构。

为了有效地优化设计 KFSE 的索引结构与算法，笔者给定了一个统计前提，假设使用 FTP 搜索引擎的用户查找的文件大多集中在某一范围内。为了更具体地描述这个命题，可以从在线问卷与用户查询日志这两个不同的角度分析。关于用户查询需求问卷方面，2004 年北京大学天网文件搜索引擎组曾组织一次有 1 050 人参加的投票，结果表明用户的检索集中在电影、软件、歌曲与数据几个主要的方面。

基金项目：国家“863”计划基金资助项目(2006AA10Z239)；欧盟亚洲信息技术与通信项目(CN/ASIA-IT&C/005 (89099))；国家科技支撑基金资助项目(2006BAH02A16)；江苏省高校省级重点实验室开放课题基金资助项目

作者简介：胡 亮(1980－)，男，博士，主研方向：大规模搜索引擎，用户个性化服务；傅泽田，教授；张小栓、赵 明，副教授；郭立力、宫薇薇，博士

收稿日期：2007-09-20 **E-mail：**holypku@163.com

在用户查询日志方面，分析结果也很好地支持了这个前提。从FTP搜索引擎用户查询日志中随机提取了一个星期的检索记录 36 337 条，其中包含不同的检索词 12 931 个。为了比较不同类型文件查询次数的关系，设检索词集$Q=\{q_1, q_2,\cdots,q_n\}$，其中，$q_j$表示不同的查询关键词；集合$S=\{S_1, S_2, \cdots, S_n\}$，其中$S_j$表示对应检索词集$Q$中元素$q_j$的查询次数。KFSE将所有文件划分为软件、歌曲、电影、程序、数据、图像、文档、游戏与其他共 9 种文件类型，即

$$T=\{T_{软件},T_{歌曲},T_{电影},T_{程序},T_{数据},T_{图像},T_{文档},T_{游戏},T_{其他}\}$$

其中，T_j表示对应文件类型的查询频数，则有计算公式为

$$T_j=\sum_{q\in T_j}^{n} S_k \Big/ \sum_{k=1}^{n} S_k$$

软件、歌曲、电影、程序、数据、图像、文档、游戏与其他等类型文件的查询频数比例分别为：21%, 15%, 34%, 5%, 10%, 5%, 5%, 3%, 2%。可以看出，电影、软件、歌曲与文档类是用户查询频率最高的。

一般来说，FTP 文件名可以分为 4 种类型：纯英文(一定有英文字母但没有中文字符，可以包含数字与符号)，纯中文(一定有中文字符但没有英文字母，可以包含数字和符号)，中英文混排(即有英文字母又有中文字符，可以包含数字和符号)，其他(即没有中文字符也没有英文字母，只由数字和符号组成)。

虽然网络上纯英文文件名比其他类型多，但带有纯中文以及中英文混排文件数量是纯英文命名文件数量的 2 倍，而电影与音乐类型中文件名主要是中文或中英文混排[2]，从这个角度考虑，应该在设计KFSE索引文件结构时对中文进行更多的优化。由于中文字符集的常用汉字数比英文字母多，因此两者的编码方式不同，英文是单字节，而中文采用 2 B来存储，而且双字节索引比单字节性能要高，因此，在考虑兼容性与性能的基础上，KFSE采用双字节倒排索引技术，基本原理是对文件名中每两个字节建立倒排索引表。

3 检索结果自动分类

传统的FTP搜索引擎一般根据用户提交的查询词在索引数据库中查找，然后对检索结果相关度进行排序。虽然相关度排序能够将比较重要的结果输出给用户，但是由于一词多义与多词同义问题的存在，使得检索结果中含有许多不同主题，这些数据即使经过相关排序，但很难满足不同用户的要求[3]。为了改进检索结果的质量，KFSE先根据FTP文件扩展名标识，然后利用K-近邻算法对检索结果的文件名按行业分类[4]，这样检索结果可以按门类划分层次结构，既方便了用户的查找又提高了效率。

KFSE 根据 FTP 文件扩展名，将其分为视频、音频、文档、程序、压缩、图像与其他，每种类型的扩展名见表 1。

表 1　文件扩展名分类

类型	文件扩展名
视频	rm, ram, rmvb, avi, wmv, mpg, mpeg, mov, asf, asx, swf, vob, …
音频	mp3, wma, mp4, swf, midi, ape, wav, vqf, …
文档	txt, doc, dot, rtf, wps, wtf, ppt, pot, pps, …
程序	exe, bat, vbs, js, css, asp, aspx, c, cpp, php, php3, jsp, …
压缩	rar, iso, zip, tar, gz, …
图像	bmp, jpg, jpeg, gif, png, …
其他	…

不同于网页自动分类，KFSE采用K-近邻算法对FTP文件名字符串进行分类。FTP文件名是字符数不超过 255 的字符串，而网页对象一般都远远超过这个值；FTP文件名一般都有扩展名，不同扩展名对应不同文件类型，所以，FTP搜索引擎的K-近邻算法运算量相对较小，性能要比Web搜索引擎高。K-近邻算法基本原理是：假定所有的FTP文件名字符串对应于N维空间R_n中的点，坐标点的最近邻是根据标准欧氏距离定义的。设坐标点x表示为特征向量$(\boldsymbol{a}_1(x), \boldsymbol{a}_2(x),\cdots,\boldsymbol{a}_n(x))$，其中，$\boldsymbol{a}_j(x)$表示$x$的第$j$个属性值，那么任意两个$x_i$与$x_j$间的距离定义为

$$D(x_i,x_j)=\sqrt{\sum_{r=1}^{n}(\boldsymbol{a}_r(x_i)-\boldsymbol{a}_r(x_j))^2}$$

目标函数为离散值的K-近邻算法的基本思想为：在训练时，对于每个训练样例$(x, f(x))$，将其加入列表training-examples中，在对具体的x_q进行分类时，从training-examples中选出最靠近x_q的K个元素，并用$x_1,x_2,\cdots,x_k$表示，设$f'(x_q)\leftarrow \arg\max\delta(v,f(x_i))$，其中，$v\in V$，如果$a=b$，那么$\delta(a, b)=1$，否则$\delta(a, b)=0$，即在训练集中，选取在欧氏空间中离$x_q$最近的$K$个实例，求出$f'(x_q)$为对目标函数$f(x_q)$的近似值，它就是最靠近$x_q$的$k$个训练样例中最普遍的$f(x)$值。基于上述原理的自动分类的体系结构见图 2。

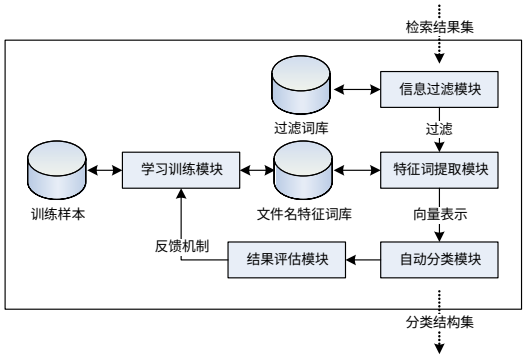

图 2　自动分类

4 查询自动纠错

在搜索引擎的使用过程中，用户常常难以给出准确的查询词，或者因为笔误而输入错误，主要表现在输入对应英文通常是字母错误，中文通常是同音异字的错误，尤其是英文软件名与中文名字：(1)英文查询词的拼写错误，如查询eclipse，却写为eklipse；想查询serv-u，却拼写为server-u。(2)中文查询词的拼写错误，如查询词“模板”与“模版”。一般来说，采用词典能够纠正常见的拼写错误，但不能自动添加新词，为此，笔者从FTP用户查询日志的角度考虑，假设$q_1,q_2,\cdots,q_n$为n个不同查询词序列，$S_1,S_2,\cdots,S_n$为对应q_k的查询次数，其中$q_1,q_2,\cdots,q_n$按对应的S_k值降序排列，拼写错误查询词统计如图 3 所示，查询次数多的高频词错误率极低，用户的查询词输入很少有拼写错误，大多数拼写错误查询词都集中在查询次数平均不超过 2 次的低频词区，这些错误查询词大多可以在日志里找到相对应的正确查询词。这表明查询次数多的高频检索词是正确的查询词，拼写错误的查询词通常能在高频词里找到与之对应的正确查询词，而且它们之间的字符串相似程度较高，可以考虑采用某种算法计算查询词之间的相似度来实现自动纠错。高频词可以作为纠错词典的数据项，但是如果高频词数量太多而且不具备相对稳定性，那么就影响词典的数据量与系统性能，因此有必要对高频词的规律进行统计，见图 4。其中，查询词频率为$S_k/\sum_{j=1}^{n}S_j$。

（下转第 23 页）