

# 基于小型搜索引擎的个性化策略研究

王 萍, 刘 军, 姚笑秋

(浙江海洋学院 数理与信息学院 计算机系, 浙江 舟山 316004)

**摘 要:**随着全球信息化时代的到来和网络的普及,如何在互联网上获得有价值的信息已成为网民日益关注的问题。搜索引擎为人们收集信息提供了很大方便,如何对现存的搜索引擎进行优化成为新的课题。重点分析一个小型搜索引擎体现个性化服务的策略与方法,并主要介绍用户兴趣分析与判断算法及自然语言检索技术。

**关键词:**搜索引擎;个性化服务;自然语言检索

**中图分类号:**TP391

**文献标识码:**A

**文章编号:**1673-629X(2007)11-0036-03

## Research of Personalized Strategy Based on Small Scaled Search Engine

WANG Ping, LIU Jun, YAO Xiao-qiu

(Department of Computer, School of Mathematics, Physics & Information Science, Zhejiang Ocean University, Zhoushan 316004, China)

**Abstract:** Along with the global becoming an information based society time arrival and the network popularization, how to obtain the valuable information on the huge Internet to become the web cam daily matter of concern. The search engine development tendency is very fierce. And how to optimize them has been a new task. The search strategy and method to obtain personalized information service are analyzed. Furthermore, the user interest analysis and judgment algorithm and natural language retrieval technology are discussed.

**Key words:** search engine; personalized service; natural language retrieval

## 0 引 言

因特网的迅速发展和广泛普及导致网络信息爆炸性增长,搜索引擎作为因特网信息检索的主要手段,能够起到信息导航的目的。但目前的搜索引擎通常都是基于关键字匹配的方法,用户输入关键字,然后根据简单的匹配策略在索引库中进行查找,导致返回的结果过于庞大,而且用户很难简单准确地表达他真正需要检索的内容。因此人们迫切需要开发出一种更为个性化、智能化的网络信息检索工具,以便从因特网中快速准确地获取所需信息。在这里,针对小型搜索引擎,重点研究体现个性化服务的策略。

## 1 中文搜索引擎的现状分析

目前中文搜索引擎<sup>[1,2]</sup>已有较多,例如知名度较高的有:

1) 搜狐(<http://www.sohu.com.cn>)是由爱特信公司于1998年推出的,它是以提供分类目录为主的中文搜索引擎,分类质量较高,但更新速度慢,查全率较低。

2) 新浪(<http://www.sina.com.cn>)是最大的中文门户网站,收录了全球资讯逾万的中文网址,并分成15大类,其下更有多个小类,提供了中文关键词的搜索功能。

3) 百度(<http://www.baidu.com>)采用基于超链分析的方法进行相关度评价,为用户提供“网页快照”功能,在快照中用不同颜色在网页中标记用户的查询字符串,方便了用户的查询。

4) 雅虎中文版(<http://yahoo.com.cn>)是按照英文版的铺排方法,将1万多个中文网址以14个类别列出,提供Internet网上的中文站点目录信息检索服务,用户可以利用繁体或简体中文进行搜索<sup>[3]</sup>。

中文搜索引擎还有很多,但总体来说,它们都存在着各自应用时的局限性,分析如下:

(1) 对个性化的理解还没有明确的认识。

(2) 用户和检索系统的交互方式比较单调。

(3) 汉字的切分技术落后。

(4) 对自然语言没有理解能力。

收稿日期:2007-01-13

基金项目:浙江海洋学院2006年科研计划项目(X06LQ08)

作者简介:王 萍(1979-),女,黑龙江齐齐哈尔人,讲师,研究方向为数据挖掘、控制与智能系统;刘 军,副教授,研究方向为计算机应用技术。

针对以上不足,在我们的“海网”搜索引擎中利用了自然语言检索技术和用户兴趣判断算法应用,从而实现了搜索引擎的个性化服务。

## 2 “海网”FTP 搜索引擎设计

### 2.1 搜索引擎的总体设计

FTP 搜索引擎是对用户提交的查询匹配串找到可以下载的 FTP 站点链接,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。FTP 搜索引擎不要求显示结果的内容摘要,查询时需要文件信息、站点信息过滤等,因而设计 FTP 搜索器时应从网络用户的实际需要出发,尽量重点实现数据的实时性、搜索的快速性和功能的强大性。

### 2.2 搜索引擎的系统结构

本 FTP 搜索引擎由数据采集、数据查询和站点维护等模块组成。首先要收集各个 FTP 站点上的文件信息,并把这些信息存储到数据库中;然后给用户提供一个查询界面,以收取用户要查询的信息,把这些查询信息转化为数据库语言,并进行数据库查询,把查询结果以友好的界面显示给用户;搜索引擎建立好以后,为了使数据库数据与 FTP 站点的数据保持一致,需要进行维护。其结构如图 1 所示。

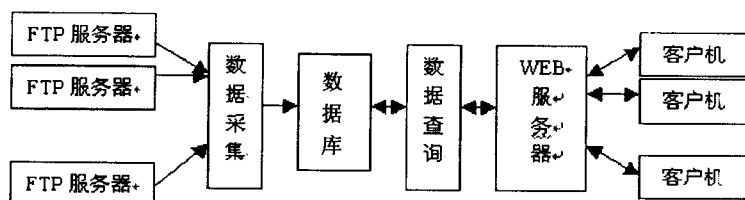


图 1 FTP 搜索引擎的系统结构

基于这种设计思想,在设计时,采用 Windows2000 操作系统,WWW 服务器采用 Apache v1.3.33,数据库采用 MySQL v4.0.24,编程语言采用 PHP v4.3.10,并在 PHP + Mysql 环境下给出了实现方法,从而建立起一个体现个性化服务的 FTP 搜索引擎。

在搜索引擎中输入待搜索的字符串,与之匹配的安装了 Server-U 的存在于引擎数据库的 FTP 站点将显示出来,通过链接找到相关并接近用户要求的 FTP 站点。

## 3 个性化策略处理

### 3.1 用户兴趣分析与判断方法实现

个性化学习方法有:

#### (1)直接学习。

提供用户可以操作的友好接口,允许用户主动地提出对用户模型的修改和维护意见,即通过用户直接

使用兴趣管理功能来改变,通过这些记录直接反应用户兴趣的变化。

#### (2)反馈学习。

通过用户对检索结果的历次反馈意见进行学习,对于结果输出的形式、结果相关度排列的顺序,以及结果记录的输出,用户都可以加以评价,给出是否满足自己需要的评定。

#### (3)历史学习。

通过用户历史查询记录的分析,经过一段时间的积累之后就可以发现用户需求的潜在规律,这些规律是用户没有意识到的。历史学习可以提醒或者是帮助用户发现自己的潜在需求,达到启发的效果。

#### (4)观察学习。

对于基于客户端的搜索工具和系统而言,利用与客户端环境的结合优势,可以从更多的方面观察并获取与用户相关的特征信息。例如用户浏览器中的历史记录和收藏夹记录了用户查看过的一些网址链接,尤其收藏夹内是用户喜好的网址链接,通过分析这些链接指向文档主题,可以得到用户的需求喜好,也有助于发现用户的新兴趣。

基于这些理论知识,可编写出以下的用户兴趣判断算法。判断用户兴趣算法的功能是根据用户不断更新的用户兴趣描述文件来判断所接受到文件是否为其兴趣所在。

算法流程如图 2 所示,其中  $TEXT_i$  是用户输入的实义词组集(若某一词组集  $TT$  中除名词、动词及形容词外不含其它性质的词,则称  $TT$  为实义词组集), $E_i$  为用户感兴趣的实义词组出现的次数, $E_u$  为用户无兴趣的实义词组出现的次数, $T_{im}$  为用户第  $i$  次输入词组中的第  $m$  个词, $E_{ti}$  为词  $T_{ti}$  出现在用户感兴趣词组中的次数, $E_{tu}$  为词  $T_{ti}$  出现在用户无兴趣词组中的次数。 $P_g$  为用户对实义词组  $TEXT_i$  感兴趣的概率, $P_s$  为用户对实义词组  $TEXT_i$  的第  $j$  个实义词  $T_{ij}$  感兴趣的概率, $Dim$  为用户对实义词组  $TEXT_i$  的第  $m$  个实义词  $T_{im}$  的熟词度, $i_{im}$  为用户对实义词组  $TEXT_i$  的第  $m$  个实义词  $T_{im}$  的兴趣度,则有:

$$P_s = P\{TEXT_i, T_{ij}\}$$

$$P_g = P\{TEXT_i\}$$

$$D_{ij} = |P_s * \log(P_s/P_g) - (1 - P_s) \log((1 - P_s)/(1 - P_g))|$$

实义词组的兴趣度:

$$I_i = \sum_j D(P_s, P_g) (i=1, 2, 3, \dots, k; j=1, 2, 3, \dots, n)$$

判断算法的过程是从库中读取用户记录,如  $TEXT_i, E_i, E_u, E_{ti}, E_{tu}$ , 来计算  $P_g = E_i / (E_i + E_u)$ 。

对实义词组  $TEXT_i$  中的每一个实义词  $Tim$  判断是否已在实义词集  $TEXT_i$  中,若不在让  $Ps$  等于零,在则再判断  $Emi + Emu$  是否小于设定的阈值(如 35),否: $Ps = Emi / (Emi + Emu)$ ,是: $Ps = 0$ 。再计算  $Dim$  和  $lim$ ,得到兴趣度的大小值,这样在用户查询时按兴趣度的大小进行反馈。

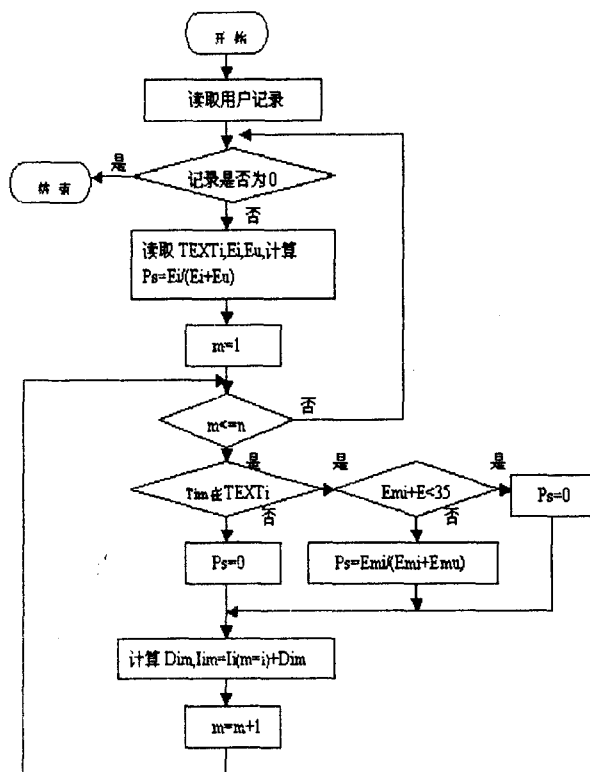


图2 用户兴趣判断算法

### 3.2 自然语言检索技术

优化引擎还可以在用户接口处使用自然语言检索<sup>[4]</sup>。自然语言就是最接近人类语言的一种语言,它基本上包括人类使用的所有语言、专有名词、方言等。自然语言理解<sup>[5]</sup>就是如何让计算机能够正确处理人类语言,并据此做出人们期待的各种正确响应。自然语言理解研究不但要运用语言学中的词汇、语法、句法和语义学知识,而且还要涉及到大量的客观世界的知识以及与其相关学科的知识。这里提出的自然语言理解步骤如图3所示。

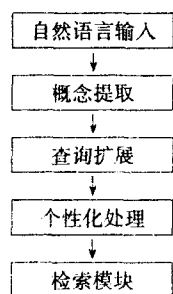


图3 自然语言理解步骤

首先用户输入自然语言,如“查找搜索引擎相关概念”,或者“搜索引擎”这样的关键词,系统对所输入的自然语言进行概念提取,获得能够表达用户意愿的一组概念,再对该概念进行查询扩展(主要是同义词、近义词扩展),然后从用户信息表中读取该用户的兴趣表示,对该概念再一次提炼,把提炼后的概念提交给检索模块进行检索。

自然语言理解的关键问题是语料知识库的建立,知识库是自然语言处理、用户兴趣学习与推理必需的理论依据。本系统中的知识库<sup>[6]</sup>主要由以下几部分组成:

- 1) 语义语用知识:对概念的定义与解释、概念之间复杂的语义关系;
- 2) 词法句法知识:构词规则,词根、词缀处理规则;常用句法结构,句子成分划分规则等;
- 3) 常识:常用的词间关联知识,如数码与电子、运算器与控制器等存在着常识上的关联;
- 4) 语料库:包括由大量的未经加工的文本文件组成的原始语料库和经过词性标注、词语切分等技术处理的多级加工语料库;
- 5) 词典数据库:同义词典、反义词典、多语种对应词典、词的层次关系词典等这部分知识含量高、存储格式规范,是知识库必不可少的组成部分;
- 6) 反向词频统计表:通过对大量文档进行分析统计,按预设算法计算词的权值,存入统计表中,文档中出现频率高的词语具有较高的权值。

## 4 结 论

搜索引擎的未来就是互联网的未来。从某种意义上讲,谁在搜索上占有先机,谁就有可能赢得未来的互联网。文中提出并设计了一个小型个性化搜索引擎,重点在于体现“个性化”服务,并对其中的关键部分做了详细的理论阐述、算法描述及数据库设计和基于PHP的编程。引擎的个性化主要体现在用户接口和信息获取两个方面:

#### ① 用户接口处的自然语言检索。

自然语言检索的使用不仅提高了查准率,而且简化了用户输入时的复杂性,用户不需了解检索系统使用规范就可使用系统。

#### ② 信息获取时用户兴趣的概念合成。

此处的个性化表现信息反馈时考虑了用户兴趣,给出了用户兴趣学习方法和用户兴趣判断算法。

对于本系统,还有一些工作要做。比如语义分析部分,由于汉语本身的复杂性和汉语词义的多变性,这

(下转第45页)

为时间段,纵坐标为网络流量。与实际流量数据进行比较,可以看出整个预测流量在变化趋势、变化快慢和分散程度上都较好地反映了网络实际流量。因此,小波分析对自相似网络流量的预测性能是比较好的。近年来也不断有学者结合小波技术对网络流量进行建模预测,都取得了不错的预测效果<sup>[8,9]</sup>。

#### 4 结束语

网络流量的自相似性决定了网络的行为特征,只有基于网络重要特征—自相似的建模才能准确描述网络实际情况。文中对主要的自相似网络模型预测方法进行了分析和总结,并应用多分形小波模型对网络流量进行了验证和预测,表明实际网络流量确实具有自相似性。

网络流量行为随着时间和地域的不同会呈现出很大的变化,因此想提供一种针对网络流量的通用化预测模型是很困难的。今后的流量预测会考虑实际网络流量存在的长、短相关性,根据不同网络分量的特点分别选用合适的方法进行预测,然后合成得到网络预测流量。

近几年,有研究人员也将自相似流量和混沌理论、模糊理论和神经网络理论结合起来研究网络的行为特性,这些新理论的引入都将对流量预测产生重要的影响。

(上接第 38 页)

部分仍有待提高。另外对搜索的结果可以进行优化、分类,双管齐下,能最大限度地提高引擎效率,体现个性化服务的特点。

#### 参考文献:

- [1] 陈笑辉,范晓红. 搜索引擎的分类体系及性能评价[J]. 中国信息导报,1999(7):18-20.
- [2] 朱俊卿. 搜索引擎 google 研究[J]. 现代图书情报技术,2001

(上接第 41 页)

的识别都有着很好的借鉴作用。

#### 参考文献:

- [1] Lee Y H, Kassam S A. Generalized median filtering and related nonlinear filtering techniques[J]. IEEE Transactions on Acoustics, Speech, Signal Processing, 1985, 33(3):672-683.
- [2] 胡小峰,赵 辉. Visual C++ + MATLAB 图像处理与识别实用案例精选[M]. 北京:人民邮电出版社,2004.
- [3] 求是科技. Visual C++ + 数字图像处理典型算法及实现

#### 参考文献:

- [1] Leland W E, Taqqu M S, Willinger W, et al. On the Self-similar Nature of Ethernet Traffic (Extended Version)[J]. IEEE/ACM Transactions on Networking, 1994, 2(1):1-15.
- [2] Paxson V, Floyd S. Wide-area traffic: the failure of Poisson modeling[J]. IEEE/ACM Transaction on Networking, 1995, 3(3):226-244.
- [3] Crovella M, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes[J]. IEEE/ACM Transactions on Networking, 1997, 5(6):160-169.
- [4] 赵佳宁,李忠诚. 基于模拟的网络流量自相似现象分析[J]. 计算机科学,2001,28(11):57-61.
- [5] Willinger W, Taqqu M S, Sherman R, et al. Self-similarity through high-variability: statistical analysis Ethernet LAN traffic at the source level[J]. IEEE/ACM Transactions on Networking, 1997, 5(1):71-86.
- [6] Norros I. On the Use of Fractional Brownian Motion in the Theory of Connectionless Traffic[J]. IEEE JSAC, 1995, 13(6):953-962.
- [7] Riedi R H, Crouse M S, Ribeiro V J. A multiscale wavelet model with application to network traffic[J]. IEEE Trans. on Info. Theory, 1999, 45(3):992-1018.
- [8] 李 捷,刘瑞新,刘先省,等. 一种基于混合模型的实时网络流量预测算法[J]. 计算机研究与发展,2006,43(5):806-812.
- [9] 洪 飞,吴志美. 基于小波的多尺度网络流量预测模型[J]. 计算机学报,2006,29(1):166-170.

(4):34-36.

- [3] 武助宇. 中文搜索引擎的发展现状、问题与对策[D]. 湘潭:湘潭大学,2002.
- [4] 王开铸. 自然语言理解[M]. 哈尔滨:哈尔滨工业大学出版社,1996.
- [5] 姚天顺. 自然语言理解,一种让机器懂得人类语言的研究[M]. 北京:清华大学出版社,2002.
- [6] 袁占亭,张爱民,张秋余. 基于概念的 Web 信息检索[J]. 计算机工程与应用,2003(36):173-175.

[M]. 北京:人民邮电出版社,2006.

- [4] Gonzalez R C, Woods R M. Digital Image Processing[M]. 2nd ed. 北京:电子工业出版社,2002.
- [5] Liu Ji-lin, Song Hong-tao, Ding Li-ya. Vehicle License Plate Recognition System with High Performance[J]. 自动化学报,2003,29(3):457-465.
- [6] 夏 勇,田 捷,邓 翔. 一种高效的自适应指纹图像压缩算法[J]. 计算机学报,1999,22(5):525-528.
- [7] 王树禾. 图论及其算法[M]. 合肥:中国科学技术大学出版社,1990.