一个面向 Internet 的个性化信息检索系统模型

韩立新,陈贵海,谢 立
(南京大学计算机软件新技术国家重点实验室,210093 江苏南京)

摘 要: 随着 Internet 上的信息量急剧增加,如何使用户获得有用的信息已成为信息检索系统急需解决的问题. 文中提出了一个个性化信息检索系统模型(Parch). 该模型结合用户访问模式和类层次结构来检索用户需要的信息. 文中还提出了多个算法,这些算法综合运用数据挖掘、情报检索和机器学习等技术,较好地解决了在生成用户访问模式时人工干预较多、自适应性较差、准确性较差以及在构造类层次结构时出现计算量较大所造成聚集速度较慢的问题.

关键词: 个性化信息检索; 用户访问模式; 类层次结构; 数据挖掘; Agent
中图分类号: TP391; TP393 **文献标识码**: A **文章编号**: 0372-2112 (2002) 02-0240-05

A Model of Personalized Information Retrieval Systems for Internet Applications

HAN Li-xin, CHEN Gui-hai, XIE Li
(*State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210093, *China*)

Abstract: Nowadays the amount of information on the Internet is increasing dramatically. The ability of facilitating users to achieve useful information is more and more important for information retrieval systems. Parch, a model for personalized Internet information retrieval systems is proposed. This model can help users to retrieve what they need by combining class hierarchy with user profile. Several algorithms that use key technologies are proposed, such as data mining, information retrieval and machine learning. These algorithms can make a user profile less manually interfered, more self-adaptive and more accurate, and meanwhile achieve a faster congregating speed in building the class hierarchy by reducing the computing workload.

Key words: personalized information retrieval; user profile; class hierarchy; data mining; Agent

1 引言

随着 Internet 应用的日益普及,如何从如此众多的信息中发现有用的信息已经成为困扰网络用户的一大难题,而建立个性化的信息检索系统正是解决这种问题的一个重要途径.

Stanford 大学的 Smart[1], Kansas 大学的 Obiwan[2]提出将用户访问模式(user profile)构造成层次结构,但是它们对用户可能有多方面的兴趣考虑不太多. Stanford 大学的 SIFT[3]是新闻过滤系统,它需要用户通过提交反馈信息进行训练,因此自动化程度不太高. WBI[4]和 Personal WebWatcher[5]都是根据浏览行为进行数据获取,但是对 Web 页面上的内容考虑不太多. TREVI[6]和 Syskill &Webert[7]都是提供个性化推荐服务,然而服务质量不是很高.

上述系统主要存在以下不足:

(1) 综合考虑个性化检索和浏览式检索各自的优点不够;

(2) 自动化程度不太高,它们主要采用的是要求用户输入自己的信息需求或使用机器学习技术要求用户提供大量的反馈信息来进行训练的方法;

(3) 不能准确把握用户的信息需求并且不能很好适应用户需求的变化.

为解决上述问题,本文提出一个 Parch 系统模型. 该模型是一个主要面向 Internet 的个性化的中英文信息检索系统,它尤为适用于对英文信息进行检索.

Parch 系统模型的主要特点如下: (1) 为了更好地满足用户的复杂需求,本文利用用户访问模式,类层次结构和关键词等多种方式来构造个性化信息检索系统. 对于用户经常感兴趣的内容,使用用户访问模式,利用 push 技术主动向用户发送感兴趣的信息,而对于用户的随机查询,用户通过类层次结构和关键词相结合的方式获得需要的结果. (2) 在生成用户访问模式方面,它把关联规则挖掘法、情报检索技术、机器学习技术,以及对页面上的信息进行语法分析等方法结合起来. 这些技术不是简单地堆积在一起,而是有机地结合在一起,这样做可避免单纯使用某一技术造成的缺陷. 它具有不需人工干预、准确率较高、自适应性较强的特点. (3) 在生成类层次结构方面,本文将多个具有相似特征的文档聚集成类,同时利用分

收稿日期:2001-01-16;修回日期:2001-06-06
基金项目:国家自然科学基金(No. 60073029);国家 863 计划项目(No. 863-306-ZT02-0301)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

层聚类方法按照类的相似程度形成类层次结构. (4)我们对多 Agent 系统的管理采用的是完全分布式的管理模式. 每台加入到该多 Agent 系统中的机器上都有一个管理程序,该程序负责 Agent 的加入、退出、信息注册等. 每当某个 Agent 加入或退出时,都向本机的管理程序发出消息,而管理程序在更改本地信息的同时将该 Agent 的动作通知给其它管理器.

2　Parch 系统模型

2.1　Parch 系统的体系结构

Parch 系统是一个多 Agent 系统,整个系统有较明显的层次关系,自顶向下依次分为二层:人机交互层、信息处理层. 对 Parch 系统模型中的二个层次进行进一步细化就得到如图 1 所示的系统体系结构.

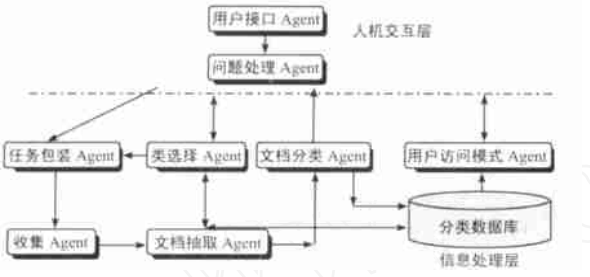

图 1　Parch 系统的体系结构

2.2　Parch 系统的系统功能

2.2.1　人机交互层

(1)用户接口 Agent:它向系统发出请求和接受系统的服务. 它主要是给用户提供一个友好的交互界面,该 Agent 按照用户的习惯以问题的方式提出查询请求,从而用户不需对复杂的搜索引擎的语法进行研究. 这样做避免了常用的搜索引擎所提供的关键字查询时所产生的用户对搜索引擎的查询语法并不熟悉,易导致用户无法对所要查询的问题抽取关键词或用户错误地表达自己的查询意图的缺陷. 返回查询结果主要采用二种方式. 方式 1:对于用户以问题的方式提出的随机查询,系统将查询结果以可扩展的类层次结构的形式返回给用户,以供用户进行浏览. 由于类层次结构具有一定的可扩展性,因此查询的结果较为直观. 方式 2:接受用户访问模式 Agent主动发送回来的用户经常感兴趣的内容.

(2)问题处理 Agent:由于此时的用户请求是以问题的形式提出,因此必须从用户提出问题的描述中抽取出关键词,按照这些关键词在问题中出现的次数赋于不同的权值,并从较为广泛使用的 WordNet[8]系统中找出与这些关键词相对应的同义词,并使用关联规则挖掘算法来获得这些关键词相关联的词,从而使检索的精度更高并使搜索信息的覆盖面更广. 最后使用 or 逻辑连接符将这些关键词和相应的同义词连接起来,使用 and 逻辑连接符将这些关键词和相关联的词连接起来,以便产生系统的查询请求. 限于篇幅,具体内容将在另一篇文章中作详细介绍.

2.2.2　信息处理层

信息处理层的各个 Agent 的具体功能如下:

(1)类选择 Agent:该 Agent 从问题处理 Agent 处接收查询请求,取出查询请求中的关键词,求出这些关键词和分类数据库中的每个类之间的相似度,找出相似度大于某一指定的阈值的一些类,使用类层次结构生成算法 Buildclassification 将这些类构造成类层次结构,返回给用户接口 Agent.

(2)任务包装 Agent:如果没有从类选择 Agent 中找出相对应的类,任务包装 Agent 就从问题处理 Agent 处接收已处理过的查询请求,由于此时的用户请求是系统内部格式,而各个搜索引擎的语法都不尽相同,因此任务包装 Agent 将其转换成符合各个搜索引擎语法要求的查询请求,并将包装好的查询请求转给收集 Agent 去进行具体的查询工作. 正是任务包装 Agent 使得用户不需要去了解各个具体的搜索引擎的语法就可以对多个搜索引擎进行查询.

(3)收集 Agent:对于从任务包装 Agent 获得的查询请求,系统采用元搜索引擎技术向多个搜索工具如 Alta Vista 和 Infoseek 等提出查询请求,并将得到的 URL 列表进行合并,除去重复的内容,然后从各个不同的 WWW 服务器上获取页面内容,并将这些信息返回给文档抽取 Agent.

(4)文档抽取 Agent:对收集 Agent 返回的每个检索文档,利用单个文档中关键词的抽取算法 Singlesample 获得一组关键词,将这些关键词构成关键词集. 将所有检索文档和它们的关键词集一起返回给文档分类 Agent.

(5)文档分类 Agent:使用文档归类算法 MKM[9],将文档抽取 Agent 返回的检索文档归入一些类中,并将这些类归入包含所有类的分类数据库中,同时使用类层次结构生成算法 Buildclassification 将这些类生成类层次结构,将类层次结构返回给用户接口 Agent.

(6)用户访问模式 Agent:该 Agent 具有一定的学习性和反应性,它采用数据挖掘和机器学习等技术对日志文件、CGI 参数中的信息以及用户发出的查询请求和已浏览过的查询结果进行不断的分析和训练,使用构造用户访问模式的算法 Createprofile 归纳出用户的兴趣模型,使用户访问模式更好地适应用户的动态需求变化. 当分类数据库中收集到尚未发送给用户的信息时,用户访问模式 Agent 采用 push 技术主动将这些用户感兴趣的信息通过用户接口 Agent 发送给用户,从而更好地满足特定用户的要求.

3　Parch 系统中关键技术的实现

本节将介绍该系统中一系列关键技术的实现算法,并与相关算法进行相应的比较.

3.1　构造类层次结构

3.1.1　类层次结构生成算法 Buildclassification

Buildclassification 算法能根据各个类之间的相似程度形成类层次结构,从而能有效反映类之间的相似性. 因为考虑到在类层次结构中的类不是非常多,所以 Buildclassification 算法采用自底向上的分层聚类方法. Buildclassification 算法避免了一些系统必须按某种拓扑结构进行分层聚类的方法. 此外当相似度小于某一指定的阈值就不进行聚类分析,而不必象一些分层聚类方法那样直至合并到只剩下一个类为止的情况,这样做加快了聚集速度. 该算法的时间复杂度为 $O(n+\log_2 n)$.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Buildclassification 算法的处理步骤如下:

输入:n 个类

输出:类层次结构

{ classset = n 个类的集合;

height = H;

While 有相似度大于某一指定的阈值并且 height 大于 1

{使用余弦相似性公式,分别计算 classset 中的类之间的相似度;

按照相似度,把相似的类作为集合 set 中的一个元素;

classset = set;

pair = first(set);//first 函数是取出 set 中的第一个元素

While set < > nil

{if pair 中的类的相似度大于某一指定的阈值

then{把 pair 中的类作为子类,并将 pair 中的类进行合并,获得它的父类 fclass;

classset = classset - {pair};//从 classset 中将已合并的这些类删除

classset = classset + {fclass};

}

pair = next(set);//next 函数是取出 set 中的下一个子类对

}

height = height - 1;

}

}

3.2 构造用户访问模式

3.2.1 单个文档中关键词的抽取算法 Singlesample

Singlesample 算法是用来抽取单个文档的关键词集. 主要是从一些频率较高的单词开始,根据每个单词的前后词找出相互关联的词组或短语,最终构成一组关键词,这使得抽取关键词不需人工干预. 此外,为了提高自调节能力,本算法通过在每次循环中根据词组的聚集情况对最小支持度进行调整. Singlesample 算法和具有很大影响的关联规则算法 Apriori[10] 不同的是:Singlesample 算法在每次循环中根据词组的聚集情况对最小支持度进行调整,并结合抽取关键词的应用对 Apriori 关联规则法进行改进,从而具有较好地自调节能力. 该算法的时间复杂度为 $O(k \cdot n)$.

Singlesample 算法处理步骤如下:

输入:单个文档

输出:关键词集

{按单个词进行词频统计并排序;

剔除频率较高的介词,冠词等虚词;

set1 = nil;//set1 存放关键词集

set = {剩余的前 n 个频率较高的单词};//n 为给定的最大值

count = 1;

While count = k//k 为词组中包含词的最大个数

{frequent = nil;

words = first(set);//first 函数是取出 set 中的第一个元素

While set < > nil

{frequent = {words 和上文中前 count 个词所组成的词组} + frequent;

frequent = {words 和下文中后 count 个词所组成的词组} + frequent;

words = next(set);//next 函数是取出 set 中的下一个元素

}

n = n - m;//m 视词组的大小进行调整,本算法 m = 2

set = {计算 frequent 中频率较高的 n 个词组};//n 为给定的最大值

s = n

s = s - m;//s 为频率最高词组的数量

set2 = {计算 frequent 中频率最高的 s 个词组};

set1 = set1 ∪set2;

count = count + 1;

}

}

3.2.2 构造用户访问模式的算法 Createprofile

由于用户的兴趣点可能不止一个,因此可能涉及到不同的类,所以生成用户访问模式比较困难. 本文提出使用 Createprofile 算法来构造用户访问模式. 该用户访问模式由多个关键词集组成,每个关键词集表示用户的某一兴趣点,从而避免了一些系统中用户访问模式仅能反映用户单一兴趣的缺陷. 为了提高自动化程度,本算法主要采用数据挖掘和机器学习等多种技术对日志文件,CGI 参数中的信息以及用户发出的查询请求和用户已浏览过的查询结果进行不断的分析和训练,最终获得用户访问模式. 这样做避免了一些系统要求用户输入自己的信息需求或者采用机器学习技术要求用户提供反馈信息进行训练所造成的自动化程度不太高的缺陷. 为了减少用户访问模式算法的复杂度,本算法采用对页面上的信息进行语法分析等启发式规则来构造关键词集. 为了有较好的自适应性,本算法通过对用户访问模式中关键词集的添加和删除来修改用户访问模式的内容,从而使用户访问模式更好地适应用户的动态需求变化. 该算法的时间复杂度为 $O(n^3)$.

Createprofile 算法处理步骤如下:

输入:某一用户使用的文档集 set

输出:生成用户访问模式

{ sample = first(set);//first 函数是取出 set 中的第一个文档

While 文档集 set < > nil

{if 文档 sample 中已有的多个关键词

then{这些关键词构成关键词集};

else{if 找出文档 sample 中隐含的信息

then{对 HTML 页面上的信息进行语法分析;

if{出现标题字段中的内容(以 < h1 > ～ < h6 > 表示)或者出现下划线标记(以 < u > 表示)或者出现粗体标记(以 < b > 表示)或者出现强调标记(以 < strong > 表示)或者出现斜体标记(以 < i > 表示)或者锚元素(anchor element)中隐含的信息}

then{对上述标记后的内容进行分析后获得多个关键词,构成关键词集};

}

else{使用 Singlesample 算法来抽取文档 sample 的关键词集};

if 已存在用户访问模式

then{使用余弦相似性公式,求出该用户访问模式和从文档 sample 中获得关键词集之间的相似度;

if 相似度小于某一指定的阈值

then{使用 MKM 算法,找出文档 sample 和哪一类相似,获得该类的关键词集,将该关键词集并入用户访问模式};//对用户访问模式进行扩充

}

else{获得的关键词集构成用户访问模式};

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

sample = next(*set*) ;//next 函数是取出 *set* 中的下一个文档
}
if 该用户访问模式中的某些关键词集不相似于所有文档
then 删除这些关键词集;//删除用户访问模式中过时的关键词集
}

4 算法性能分析

使用 Singlesample 算法对下面各个站点的 web 文档进行测试,测试情况如表 1 所示.

表 1　使用 Singlesample 算法的实验结果

站点名	抽取的文件数	抽取的文件量(kB)	平均关键词数	获得关键词集准确的文档数	关键词集的正确率
Tsinghua.edu.cn	40	758	4.33	34	85 %
Ogi.edu	50	938	5	43	86 %
Marshall.edu	40	732	4.67	35	87.5 %
Engr.umd.edu	30	648	5	27	90 %

依次对表中的网站进行实验.从实验结果中可以看出:本算法获得的关键词集较准确,这主要是由于采用从一些频率较高的单词开始,根据每个单词的前后词找出相互关联的词组或短语,最终构成一组关键词的方法.此外,随着抽取文件数的不断增加,关键词集的正确率也在不断增加.因为在最初查找关键词时,由于没有预先将一些常用的词存放在库中,结果造成频率较高的介词、冠词等虚词被误认为是关键词,从而影响了关键词集的正确率.随着抽取文件数的不断增加,库中的常用词不断增加,从而正确率也随之上升.

使用 Createprofile 算法对下面各个站点的 web 文档进行测试,测试情况如表 2 所示.

表 2　使用 Createprofile 算法的实验结果

实验项目 / 实验结果	分别在一个网站上进行的实验	综合在多个网站上进行的实验
测试文档平均长度	11.18kB	10.28kB
测试文档数	30	20
得出结论的文档数	26	16
成功率	86.7 %	80 %
结果准确的文档数	24	14
准确率	92.3 %	87.5 %

注:这些网站是 Yahoo,金陵热线,中国教育和科研计算机网

从实验结果可以看出:成功率并不很高,但准确率还是比较高的.这主要是一些测试文档并不是用户感兴趣的信息,而是用户随机浏览的信息.但当我们剔除用户并不感兴趣的文档后,从用户感兴趣的文档中构造用户访问模式,所产生的准确率还是比较高的.准确率比较高的具体原因是由于采用数据挖掘和机器学习等多种技术对日志文件,CGI 参数中的信息以及用户发出的查询请求和用户已浏览过的查询结果进行不断的分析和训练,最终获得用户访问模式的方法.此外,实验在一个网站上进行比实验在多个网站上进行准确率高,这是因为从同一网站上的一些 WEB 文档中抽取关键词可能更为容易,因为有时这些 WEB 文档中的关键词可能用相同或相似的颜色、字体.

5 与国内外同类工作的比较

与一些个性化信息检索系统模型相比,本原型系统把用户访问模式,类层次结构和关键词等多种方式结合起来构造个性化信息检索系统,从而更好地满足用户复杂的需求.并且可以为用户提供交流合作的机制,实现人机优势互补.

在文献[1,2]中,因为用户可能有多方面的兴趣,所以无法很好地形成用户访问模式的层次结构.而本原型系统提出用户访问模式和类层次结构相结合的方法可以较好地解决这一问题.

由于在多 Agent 系统中主要存在一个通讯和协调的问题,现有的一些多 Agent 系统大多采用集中式的管理模式.即在系统中有一个固定的管理器,所有 Agent 的通讯或请求都要通过这个管理器来协调.一旦这个管理器崩溃,那么整个多代理系统也就瘫痪了,为了避免这个缺陷,我们对多 Agent 系统的管理采用的是完全分布式的管理模式.

6 结束语

本文提出个性化检索和浏览式检索相结合的一个系统模型 Parch.从已发表的一些国内外文献来看,虽然在此方面已作了一些工作,但是我们的系统是通过对一系列关键技术的改进,从而使得系统的性能得到明显的提高.目前该系统的设计思想已运用在国家 863 高科技项目"Java 开发环境的开发"的研制过程中,并且这一项目已通过了国家 863 专家组组织的鉴定和验收.

参考文献:

[1] T Kurki, S Jokela, R Sulonen, M Turpeinen. Agents in delivering personalized content based on semantic metadata [A]. In Proc. 1999 AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace [C], Stanford, USA, 1999:84 - 93.

[2] Alexander Pretschner, Susan Gauch. Ontology based personalized search [A]. Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence [C], 1998.

[3] T Yan, H Garcia-Molina. SIFT-a tool for wide-area information dissemination [A]. In Proc. 1995 USENIX Technical Conf [C]. 1995:177 - 186.

[4] R Barrett, P Maglio, D kellem. How to personalize the Web [A]. In Proc. ACM CHI'97 [C], Atlanta, USA, 1997.

[5] D Mladenic. Personal WebWatcher: design and implementation [R]. Technical Report IJS.DP.7472, J. Stefan Institute, Department for Intelligent Systems, Ljubljana, 1998.

[6] Ismael Sanz, Luigi Mazzucchelli. Distributed Objects in a Large Scale Text Processing System [A]. Proceedings of the International Symposium on Distributed Objects and Applications [C], 1998.

[7] M Pazzani, J Muramatsu, D Billsus. Syskill&Webert: identifying interesting web sites [A]. In Proc. 13th Natl. Conf. on Artificial Intelligence [C], 1996.

[8] G Miller. WordNet: A lexical database for English [J]. Communications of the ACM, 1995, 38(11):39 - 41.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[9] Chih-Ming Su, Shian-Shyong Tseng, Mon-Fong Jiang, Joe C S Chen. A Fast Clustering Process for Outliers and Remainder Clusters [A]. In Proc. of the 3th Pacific-Asia Conference on KDD [C], Beijing, Springer Press (1999): 360 - 364.

[10] R Agrawal, R Srikant. Fast algorithms for mining association rules [A]. In Proc. of the 20th Int. Conf. on Very Large Databases (VLDB'94) [C], Santiago, Chile, Sep. 1994: 478 - 499.

[11] Eduard Hoenkamp, Onno Stegeman, Lambert Schomaker. Supporting content retrieval from WWW via basic level categories [A]. Proceedings on the 22nd annual international ACM SIGIR conference on research and development in information retrieval [C], 1999: 311 - 312.

[12] Michael J Pazzani. Representation of electronic mail filtering profiles: a user study [A]. Proceedings of the 2000 international conference on intelligent user interfaces [C], 2000: 202 - 206.

[13] 胡舜耕, 刘晓宇, 钟义信. 基于多 Agent 技术的自动文摘系统的研究和设计 [J]. 电子学报, 2001, 29(2): 247 - 249.

作者简介:

韩立新 男, 1967 年生于江苏南京, 南京大学计算机系博士研究生. 主要研究方向: 数据挖掘和分布式计算.

陈贵海 男, 1963 年生于广西贵县. 南京大学计算机系教授、博士生导师. 主要研究方向: 分布与并行系统, 网络计算.

中医舌像分析仪

(北京工业大学信号与信息处理研究室)

北京工业大学信号与信息处理研究室与北京市中医院合作开发的中医舌像分析仪已在北京市中医院投入使用, 并在临床应用中得到了专家及用户的好评. 其中部分关键技术已申请国家专利.

中医舌像分析仪是一种无创、定量和客观的中医舌像智能分析仪器, 为中医的临床诊断、教学和研究服务. 它能够采集、察看、存储数字化彩色舌图像, 实现彩色舌图像的真实重现, 并具有自动分析常见舌像特征(舌色、苔色、苔厚、湿度、裂纹等)的功能. 中医舌像分析仪可以有效地提高舌诊的准确度、客观性和工作效率.

中医舌像分析仪在标准化的采集环境下, 采集舌图像并传送到计算机, 实时地进行彩色校正、舌体分割、舌像特征自动分析, 经过医生诊断, 最后将舌图像、分析结果、医生的诊断结果分类归档存储, 通过高分辨率显示器或彩色打印机输出. 可以快速检索查询病人信息, 也可通过计算机网络共享医疗信息.

中医舌像分析仪对提高舌诊理论水平和临床诊断能力, 进一步丰富中医舌诊的医学宝库, 开创中医舌诊的创新之路具有十分积极的意义.

中医舌像分析仪外观图

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net