基于聚类的个性化元搜索引擎设计

严莉莉,王倩倩,孟 杰,张燕平

(安徽大学 智能计算与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:Internet上信息资源的飞速膨胀造成用户在进行信息检索时的不便,传统的搜索引擎不能很好地解决这个问题。因此提出了一种基于聚类的个性化元搜索引擎模型,系统通过对用户建立个人模型,对此模型进行聚类形成不同用户群,并对检索到的结果进行聚类处理,同用户模型聚类相结合返回给用户个性化的搜索结果。分析了个性化元搜索引擎的系统构成,详细介绍了每个模块的功能,最后展望了它的发展前景。

关键词:元搜索引擎;聚类;个性化模型

中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2007)04-0186-03

Design of Personalized Meta-Search Engine Based on Clustering

YAN Li-li, WANG Qian-qian, MENG Jie, ZHANG Yan-ping

(Ministry of Education Key Lab. of Intelligence Computing and Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: It's difficult for users to search information because of the rapid expanding of information resource in Internet. Traditional search engines can't deal with that very well. So presents a personalized meta-search engine model based on clustering. This system constructs a personalized model for every user in order to form different custom crowd, and together with the clustering analysis of the searching results. The model can make search engine return more personalized searching results for users. Analyses system structure of personalized meta-search engines and introduces functions of each module in detail. Finally, look forwards its prospect.

Key words: meta-search engine; clustering; personalized model

0 引 言

搜索引擎是人们从Internet上获取信息的主要工具,但当前的搜索引擎存在的问题是:当人们查询某个关键字时,所有包含此关键字的页面都将作为搜索结果被下载到索引库中,再根据一定的算法排序后返回给用户。在这种情况下,没有考虑到关键词所包含的含义可能是不同的,不同的用户输入同一个关键字得到的结果是相同的,无法根据不同用户的要求给出符合不同用户要求的个性化搜索结果。而且单个引擎的覆盖面是有限的,不同的搜索引擎搜索的效果不一样,一般通过多个搜索引擎的搜索才能得到比较全面的搜索结果。

为解决上述情况,文中提出了一个基于聚类的个性化元搜索引擎框架,用户提出搜索要求,由元搜索引擎进行加工,转换成多个独立的搜索引擎一起搜索,并将搜索结果处理后返回给用户。整个系统的关键是在用户个人模型和搜索结果上运用聚类分析,形成不同的用户群及针对不同用户返回个性化搜索结果。

1 相关技术介绍

1.1 搜索引擎

传统搜索引擎系统由搜索器、索引器、检索器和用户接口四个部分组成[1]。搜索程序按照一定规律和方式对网上WWW站点进行搜索,将搜索到的Web页面信息存入到搜索引擎的临时数据库。搜索程序通常是指一种被称为"蜘蛛"的"机器人"程序,它可以高速不间断地执行某项任务的程序。"机器人"程序可以像蜘蛛一样自动沿着任意网页中的链接爬到其他网页,并自动收集因特网上千万到几十亿个网页信息,增加新的网页信息,去除死链接,并根据网页内容和链接关系的变化重新排序。

由分析索引系统程序对收集回来的数据进行分析,提取相关信息,包括网页所在URL、编码类型、页面内容包含的关键词、关键词位置、生成时间、大小、与

收稿日期:2006-07-04

基金项目:安徽省自然科学基金项目(0504200208);安徽省教育厅自然科学研究项目(2005kj053)

作者简介:严莉莉(1982-),女,安徽合肥人,硕士研究生,研究方向为智能计算;张燕平,教授,硕导,研究领域为人工神经网络、机器学习、人工智能在金融工程中的应用。

其他网页的链接关系等,根据一定的相关度算法进行大量复杂计算,得到每一个网页针对页面内容中及超链中每一个关键词的相关度,然后用这些相关信息建立网页索引数据库。

根据用户的查询信息在索引数据库中快速检索,并根据关键词进行排序,当用户查找某个关键词的时候,所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。在经过复杂的算法进行排序后,这些结果将按照与搜索关键词的相关度高低依次排列,然后由页面生成系统显示查询结果,将搜索结果组织起来反馈给用户。

1.2　元搜索引擎

元搜索引擎是指将现有的多个搜索引擎集成为一个门户[2],为用户提供一个统一的搜索界面,通过链接多个独立搜索引擎,接收、分析与处理用户的信息需求与搜索提问。用户的查询请求被转换成多个成员搜索引擎所能识别的格式,然后按照搜索引擎调度管理,把规范的查询分送到成员搜索引擎,由这些搜索引擎完成实际的信息检索操作,最后元搜索引擎再以一定的格式返回给用户。

元搜索引擎与独立搜索引擎最大的区别在于[3,4]:后者拥有独立的网络资源采集、标引机制和一定的数据库,能直接提供用户需求的网络信息和数据;元搜索引擎没有属于自己独立的数据库,而是利用网络服务器和路由器之间的功能切换机理,来形成由分布于不同服务器与物理层上、具有独立搜索体系的多个独立搜索引擎构成的逻辑结合体。也就是说元搜索引擎是以各独立搜索引擎的分布式数据库为基础,将进入自己检索界面的用户搜索提问转移给其所链接的这些处于底层的独立搜索引擎去分析和处理。

1.3　聚类分析

所谓聚类,就是将一个数据单位的集合分割成几个称为簇或类别的子集,每个类中的数据都有相似性,它的划分依据就是“物以类聚”。聚类分析依据的原则是使同一聚簇中的对象具有尽可能大的相似性,而不同聚簇中的对象具有尽可能大的相异性,聚类分析主要解决的问题就是如何在没有先验知识的前提下,实现满足这种要求的聚簇的聚合。聚类分析称为无监督学习(Unsupervised Study)。无监督学习不依靠事先确定的数据类别以及标有数据类别的学习训练样本集合,需要由聚类学习算法自动计算。

2　系统框架和功能介绍

2.1　系统体系结构框架

系统体系结构框架如图 1 所示。

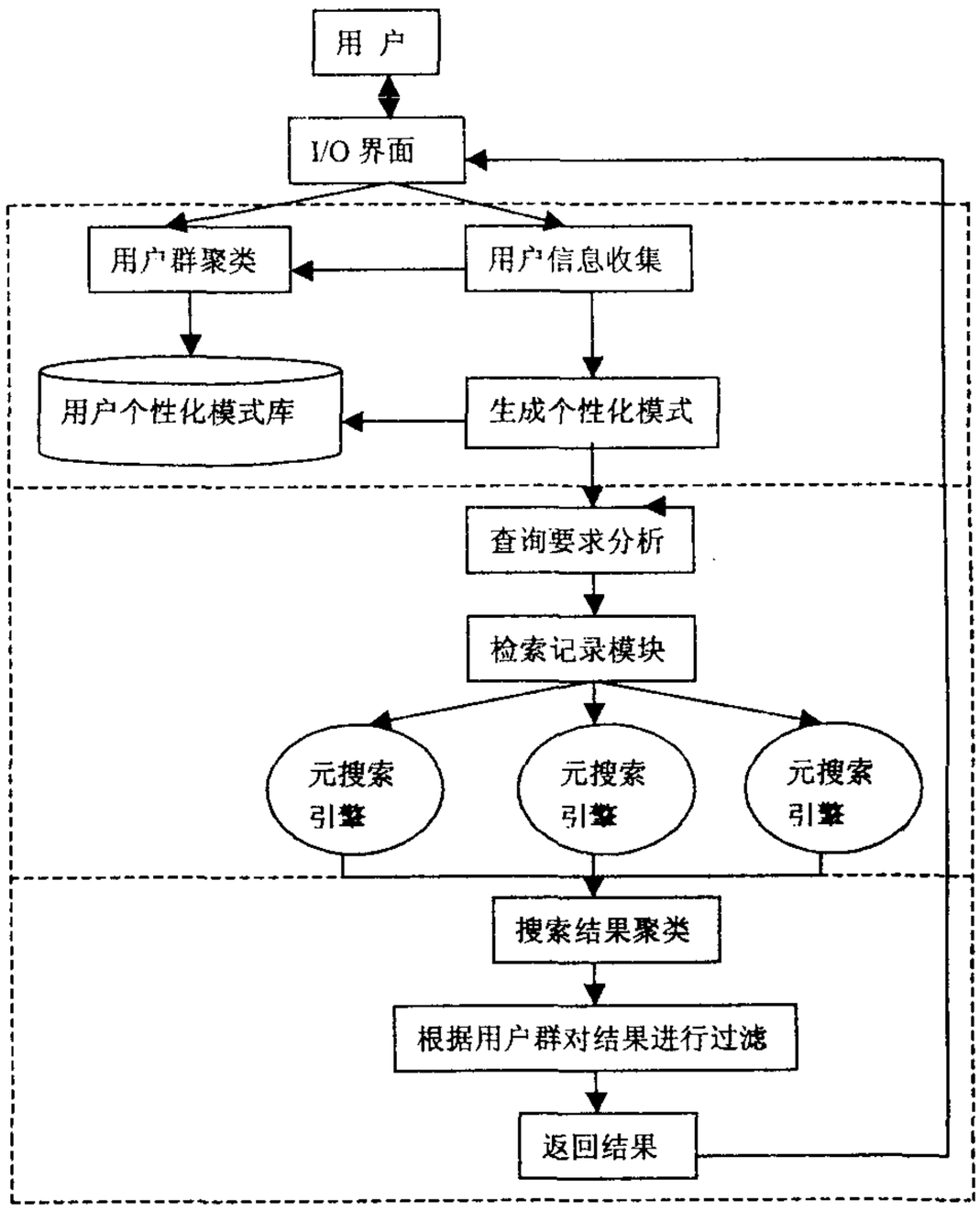

图 1　系统体系结构框架图

2.2　功　能

系统由用户模块、查询模块和结果处理模块三部分组成。其主要工作流程是:

第 1 步,用户通过 I/O 界面提交查询请求;

第 2 步,查询要求分析模块对查询请求进行分析,处理成规范化形成规范查询请求向量,检索记录模块根据查询请求向量结合一定的算法选择相应的搜索引擎对关键字进行搜索;

第 3 步,向各个元搜索引擎分发搜索请求;

第 4 步,结合用户特征对各元搜索引擎的返回结果进行过滤、排序等整合处理,包括去掉无链接的页面,去除冗余、重复、多余的数据,然后按照与关键字的匹配程度进行由高到低的排序,体现用户的个性,优化了搜索结果;

第 5 步,返回搜索结果并记录用户对结果的反馈信息,为用户群聚类提供信息,作为用户特征分析的依据。

2.3　对查询结果的聚类

对于模糊查询,传统的链接分析算法得到的查询结果,都不可能满足所有用户的需要。因为对于短语它包含的含义有很多,例如“狼”,或者是一种动物,或者是一种品牌,也可能是一首歌。不同的用户输入相同的短语希望得到的侧重点不一样。同样由于自然语言的丰富,同一概念可有不同的词语来表达,如“计算机”这个词,跟它同义的词语就包括“PC”、“电脑”等

等。这个时候不同用户输入不同的短语期望得到的结果应该是类似的。另一方面,由于使用不同的元搜索引擎,每个搜索引擎都有自己对搜索结果的排序算法,使得不同搜索引擎的搜索结果的排序无法比较,搜索结果与用户查询之间的相关程度也无法建立统一的度量标准,因此查询结果整合问题成了搜索引擎研究的核心问题。所以文中提出的模型利用聚类有关方法对网页集合按照语义信息进行初步归类,然后对每个类中的网页利用传统的链接分析算法计算网页权重。基本步骤是:

(1)对各搜索引擎的查询结果进行聚类分析,形成对查询结果的自动分类。

聚类是由计算机系统按照一定的要求将相似或者相同特征的对象聚合在一起的过程。自动聚类根据对象的不同特征划分成不同的类,使得同一个类中的对象之间的差别尽可能地小,而不同类中的对象之间的差别尽可能地大。聚类分析在机器学习、模式识别等领域已有不少应用,也已提出了不少聚类算法,如 C-均值聚类法、层次聚类法、密度聚类法、网格聚类法等,其中 K-均值聚类法以其计算的高效率而被广泛采用。

网页聚类算法如下:

首先生成初始类,在网页集中任意选取一个网页作为第一个类,计算其余网页与此网页的相似度(链接相似度),选取相似度最小的一个网页作为第二个类,计算其余网页与此二网页的相似度,取与这两个网页相似度之和最小的网页作为第三个类。依此直到产生 K 个类(K 为事先给定的正整数),然后根据网页对类的隶属度反复迭代直到所有类不再发生变化为止。

(2)分析对各类和用户查询请求之间的相关度,并根据相关度确定类排序。

在确定了各个类之后需要进一步对类内各结果项进行排序,查询结果排序类似于类排序的策略,通过计算每一个结果项和查询请求之间的相似度并结合在元搜索引擎中的排序,最终确定各个结果项在类中的排序,最后根据和用户特征之间的相似度,确定类排序。

(3)将相关度最大的类中的结果返回用户。

2.4 用户个性化模型的建立

最终确定的类排序与用户特征有着密切的联系,系统根据不同的用户特征给出的输出结果是不一样的,从而体现个性化的输出结果,这一功能的实现依赖于用户个性化模型的建立。系统可以使用户在 I/O 界面中输入自己的账号和密码,用户信息收集模块用来记录用户的历史搜索记录,以及 Web 日志中包括像 IP、访问时间、请求页、访问页、会话号、浏览器版本等信息,同时用户浏览返回的结果,确认是否为感兴趣文档,若为感兴趣文档则做一标记,每个被确认的文档都有其对应的关键词,将这些关键词添加到用户所对应的数据库中,用于下次搜索时的用户行为分析。用户群聚类模块可以根据这些关键词的集合以及日志信息对不同的用户进行聚类,如果两个用户的关键词集合的相似度在给定的阈值 f 之上,则认为两个用户是同类的用户。所有这些信息被记录在用户个性化模式库中。

3 结 论

随着 Internet 的不断发展,搜索引擎已经成为人们获取网上信息最重要的途径,但存在着命中率低、成本高等不足[5]。文中介绍的个性化搜索引擎针对传统搜索引擎的不足进行了相应的改进,将元搜索引擎技术和 聚类技术结合到一起,元搜索引擎技术提高了传统搜索引擎的查准率和查全率,聚类技术充分考虑到用户的个性化,因此有很强的理论意义和实际意义,具有一定的智能性,能够应用到 Internet、信息检索、电子商务等诸多领域。

当然系统中也有一些问题:如何用更加优化的算法提高搜索的效率;如何使用多种显示方式对搜索结果进行显示,以满足不同用户的浏览习惯,增加可读性和生动性,这是今后工作需要进一步研究的问题。

参考文献:

[1] 曹二堂,刘玉林. 基于语义理解的智能搜索引擎的研究[J]. 情报杂志,2005(6):58-59.

[2] 胡 亮,许永诚,高 文,等. 个性化高效元搜索引擎的设计与实现[J]. 计算机工程与设计,2005,26(4):896-899.

[3] 刘 丽,孙燕唐. 智能型元搜索引擎的设计与实现[J]. 计算机工程,2003,29(6):118-120.

[4] 刘 炜,陈俊杰. 一种基于 Agent 的智能元搜索引擎框架[J]. 计算机工程与应用,2005(3):137-138.

[5] 左雄辉,糜 麒. 个性化搜索引擎研究[J]. 计算机工程与应用,2005(17):190-192.

(上接第 185 页)

[4] Saldhana J A, Shatz S M. UML diagrams to Object Petri Net Models: An approach for Modeling and Analysis[C]//Proceedings of Twelfth International Conference on Software Engineering and Knowledge Engineering (SEKE2000). Chicago, USA: Knowledge System Institute, 2000:103-110.

[5] Booch G, Rumbaugh J, Jacobson I. UML 用户指南[M]. 北京:机械工业出版社,2001.