

基于群聚个性化搜索引擎

石军,王儒敬,王志红

(中国科学院 智能机械研究所,安徽 合肥 230031)

摘要:在这十多年的发展过程中,搜索引擎技术日渐成熟。我们根据对搜索引擎历史和个性化搜索引擎的历史和现状研究,对当前的方法进行评价并提出一种基于群聚的个性化搜索引擎实现的方法。

关键词:搜索引擎;个性化;Web挖掘;用户模式

中图分类号: TP391,TP393

文献标识码: B

文章编号:1672-6251(2006)01-0045-03

Cluster-based personal search engine

SHI jun,WANG Ru-jing,WANG Zhi-hong

(Institute of Intelligent Machines of Chinese Academy of Science,Hefei 230031,China)

Abstract:In the last ten years,search engine had a rapidly development. It's technology becomes more mature .In this paper, we focus on history of search engine and it's personalization, offer a new cluster based approach in implement search engine personalization.

key word:Search engine;Personalization;Web mining;User pattern

1 引言

所谓的个性化,就是在商务中根据客户信息和客户爱好为客户提供独特的产品或者服务。人们对互联网中应用个性化技术有一定的研究,也取得了一些成果,传统的 Web 个性化实现一般有三种方法。

1.1 协同过滤

它的基本假设是经常访问相似资源的用户兴趣相似,相似兴趣的用户又会访问相似的资源。因此,通过对相似兴趣用户的判定,来确定某个用户对某一未知资源是否感兴趣^{[1][2]}。

1.2 基于内容的过滤

内容过滤主要是对用户兴趣的不断学习和反馈,以保证在任一时刻过滤的文本和当前用户兴趣相吻合^[3]。

1.3 互决策

网站管理员根据用户信息和统计资料为其指定规则,该规则确定对用户的服务内容^[4]。

在本文参考文献[5]中 Mobasher 等提出使用数据挖掘技术对 Web 日志以及用户会话的分析采用在线和离线两种模块结合使用的方法实现 Web 个性化。在本文

参考文献 [6] 中,作者使用 agent 技术建立 WBI(Web Browser Intelligence)来实现个性化的目标。微软的 Lumiere 项目则是建立贝叶斯模型来推断用户信息、需求、背景资料等,从而达到实现个性化的目的^[7]。本文参考文献[2]中使用一种称之为协调过滤的方法以及依靠用户输入以及用户资料来完成。

什么是搜索引擎?搜索引擎指自动从 Internet 中搜集信息,经过一定整理以后,提供给用户进行查询的系统。Internet 上的信息浩瀚万千,而且毫无秩序,所有的信息像汪洋上的一个个小岛,网页链接是这些小岛之间纵横交错的桥梁,而搜索引擎,则为你绘制一幅一目了然的信息地图,供你随时查阅。国内对网络个性化服务研究大多集中在基于 agent 的技术应用,如南京大学潘金贵教授使用个性化的 agent 信息收集系统 DOLTR12Agent (distance and open learning training resource information agent)^[8],清华大学李衍达教授的 Open Bookmark^[9]。

将搜索引擎的个性化可以提高搜索引擎的检索效率并且能提高搜索引擎相关参数以及实现结果的人性化。在这篇论文中,我们在第二章研究当前搜索引擎个性化实现的技术方法;第三章介绍一种基于集群的个

收稿日期:2005-08-17;修回日期:2005-09-13

基金项目:国家 863 高科技发展研究计划资助项目(编号:2001AA118070)。

作者简介:石军(1978-),男,硕士研究生,研究方向:Web 数据挖掘和信息抽取。

王儒敬(1964-),男,研究员,研究方向:智能决策支持系统。

性化实现方法;最后一章我们对个性化搜索引擎提出我们研究的技术成果并展望不久的将来个性化搜索引擎的发展趋势。

2 个性化搜索引擎方法

虽然搜索引擎技术发展较快,但是经过几年的研究试验,当前搜索引擎的个性化都还是处于个人化的阶段。搜索引擎个性化的完全实现需要经历个人化的阶段,下面是根据当前搜索引擎个性化实现总结出的一些实现方法:

(1)最早实现的个性化搜索引擎在于使用 Cookie 在客户端保存用户爱好,如保存用户爱好语种;保存返回页面搜索结果记录数目。

(2)根据用户对搜索结果反应获取查询语句与返回结果的相关度。用户可能会忽略某些返回结果页面;会点击浏览页面;会在某页面停留较长时间。

(3)根据用户资料来返回查询结果集。如用户资料中表示该用户对金融和电脑技术感兴趣,那么在用户在搜索关键字“图片”时显示的结果集应该只是在金融和电脑技术主题的图片,而不是返回给用户社会新闻等类别的结果。

当前(1)的功能大多数搜索引擎都已经实现。其中 Yahoo 的 My web 实现了(2)的功能,他将搜索引擎与用户的个人帐户结合起来,用户实现个人化搜索之前登陆 My web,随后可以实现如下功能:

①阻塞一些用户认为无效的网页(包括但不限于 spammer);

- ①存储搜索结果;
- ③重复搜索上次存储结果;
- ④和好友分享存储结果;
- ⑤将结果分类。

Google 的 personalized search 实现了(3)的功能,他主要是根据在客户端设置用户兴趣领域,随后根据此实现结果的精确化。在根据用户查询关键词返回一定数目的结果集后,用户还可以根据个性化工具条实现进一步的个性化,进行工具条的个性化时,只是改变返回结果集的排列顺序而不改变结果集的数量。在最新的 Google 应用中,他加入了个性化的 web 内容,主要是使用用户定制的方法来生成内容。

3 基于群聚的个性化技术

当前个性化搜索引擎有很多不足,其中包括:

- (1)需要用户给予明显的配合,如登陆或在本地存储 cookie 数据;
- (2)使用的数据都是静态的而不是动态交互的;

(3)搜索结果的个性化表现不明显;

(4)搜索结果的个性化结果不准确。

我们从网络粘度的实际出发,使用群聚的方法结合现有的个性化技术扩展实现个性化。所谓群聚,在这里我们将 IP 相关或者兴趣度相关的用户归为一个群,那么针对我们认定可以划分的群提供个性化服务。在 directhit 搜索引擎就用到使用搜索用户行为影响关键字排序结果的方法,每个人上网都会依附于一个 IP,我们假设每个 IP 对应的用户是一个独立的个体,那么我们可以根据记录 IP 及其浏览模式使用 Web 挖掘算法提供相应的个性化服务;对于相同兴趣度的用户我们假设对同一讨论组感兴趣的用户可以规划为一组,那么可以共用收集的相关信息。

Web 日志序列如

$\log\{w[1],w[2],w[3],\dots\}$,

在这里

$w\{ip, session [i], \dots\}$ 为日志记录内容,

label[i]为 $\text{label}\{\text{content}[1],\text{content}[2],\text{content}[3], \dots\}$,也就是页面内容的分类聚合;

首先我们定义 content[1]与 content[2]之间的相关度 $s\{\text{label}[1],\text{label}[2]\}$,再根据用户访问 IP 以及记录的 session 记录日志,对用户访问页面进行页面相关度的计算:

$\text{user}\{\text{lable}[1],\text{lable}[2],\text{lable}[3], \dots\}$,

其中 label[x]为用户访问页面所属类别。

定义 1:用户对 label[i]主兴趣度就是用户对 label[i]类别下内容访问的次数,公式为:

$$I[i]=\sum_{j=1}^m \text{label}[j] \times L[i] \quad (\text{在这里 } L[i] \text{ 为类别 label}[i] \text{ 用户点击数})。$$

定义 2:用户对 label[i]附加兴趣度就是用户访问所有类别的主兴趣度与 label[i]的相关度的乘积乘以一个衰减因子,公式为:

$$J[i]=c \times \sum_{j=1}^m I[j] \times s[i,j] \quad (c \text{ 为衰减因子})。$$

由于用户访问 lable [i] 的数目越多,那么用户对 lable[i]类别的兴趣度就越大;用户 lable[i]对类别的数目越多,lable[i]和 lable[j]两个类别的相关度越大,那么相应地用户对 lable[j]的兴趣度也就越多,反之越小。

最后用户对网站访问的 i 类别内容的兴趣度 $T[i]$ 为:

$$T[i]=I[i]+J[i]$$

根据对用户兴趣度序列 T 的从大到小的排序,我们可以选取部分或者全部用户感兴趣的内容作为个性化用

户页面的证据。

在这里由于开始我们定义类别相关度 $s[\text{label}[1], \text{label}[2]]$ 可能会有误差,对此,我们在后续期间,利用数据挖掘的关联挖掘算法获得类别关联信息,再根据一定支持度和置信度所得到的数据挖掘结果来回归此参数。

上面说过可能一个 IP 下可能会存在多个访问用户,对于每个单独的用户个体访问的资料不可能完全一致,可能存在一些私有的访问信息。在对用户访问信息做分类和聚类的数据挖掘,选择用户共同特征作为分析内容。这样基于群聚的方法为了保障用户的隐私权。对于 IP 为单用户的或者对于无隐私问题的站点可以不采用过滤技术。

在使用搜索引擎时,不同兴趣爱好的用户对于相同页面内容可能观测的重点不一样,由于实现用户的特征兴趣分类,基于群聚的个性化技术在搜索引擎中使用会提高极大地搜索引擎的精确度。

我们在数据挖掘研究院网站^[10]使用这种方法进行基于 Web 数据挖掘技术的个性化网站建设取得了一定的效果,在用户调查中用户对其效果给予肯定的评价,这还是处于测试初期,在后续中加入用户回归数据会使得结果更为精准。

4 个性化搜索引擎展望

成功的个性化的搜索引擎实现首先需要的是人性化,人性化的表现应该多为用户考虑,让用户尽可能少的简单操作实现最大可能的结果精确化。前面已经说过个人化是个性化的一个步骤,当前其他相关学科的技术处于瓶颈时期,把个人化作为个性化的一个突破口不失为一个好的办法。当前一些搜索引擎都在加大力度的收集用户信息和上网模式,如对用户的邮件进行搜索或者提供个人门户空间,记录用户网页浏览日志,随后根据 Web 用户使用挖掘(Web Usage Mining)对用户进行聚类和分类以及模拟预测用户的可能行为,以此为依据提供个性化的搜索结果;在今后搜索引擎将会充分利用到这些资料以及会利用各种手段获取更多资料(如 Google 最近发布的网络加速器),随后结合用户其他信息,如根据用户 IP 和群组来进行相应的数据挖掘,由于可能存在多个用户通过同一网关访问互

联网,那么记录用户的 IP 将会是多个用户上网的结果,我们假设通过同一网关访问互联网的用户具有相同属性,如:可能为同一软件公司的员工或者同一系的学生等;至于同一群组的用户也可以这样认为具有相同属性,如爱好访问同一网站;爱好讨论同一主题等等。这样再根据现有的个人化搜索引擎技术作为辅助方法进一步处理。

个性化搜索引擎的发展始终是体现在搜索引擎的个性化上,而不是人的个性化。随着 Web 数据挖掘广泛应用和 Web 标准的实施,个性化搜索引擎技术必然得到进一步的发展。

参考文献

- [1] 王 斌,许洪波.大规模内容计算[J].信息技术快报,2004,(3).
- [2] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., GroupLens: applying collaborative filtering to use net news. Communications of the ACM (40)3, 1997.
- [3] Shardanand, U., Maes, P., Social information filtering: algorithms for automating "word of mouth." In Proceedings of the ACM CHI Conference, 1995.
- [4] Bamshad Mobasher¹, Honghua Dai, Tao Luo, Yuqing Sun, Jiang Zhu ,Combining Web Usage and Content Mining for More Effective Personalization ,In Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000).
- [5] Bamshad Mobasher ,Robert Cooley, Jaideep Srivastava ,Automatic Personalization Based on Web Usage Mining ,Communications of the ACM,2003.
- [6] Rob Barrett Paul P. Maglio Daniel C. Kellem ,How to Personalize the Web ,Proceedings of the Conference on Human Factors in Computing Systems CHI'97.
- [7] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, Koos Rommelse,The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users ,In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence ,1998.
- [8] 潘金贵,胡学联,李 俊,张灵玲.一个个性化的信息搜集 Agent 的设计与实现[J].软件学报,2001.
- [9] 冯 翱,刘 斌,卢增祥,路海明,王 普,李衍达.Open Bookmark ——基于 Agent 的信息过滤系统[J].清华大学学报,2001.
- [10] <http://www.dmresearch.net>.