个性化搜索引擎研究

欧建斌
(暨南大学,广东 广州 510632)

摘 要：对搜索引擎个性化服务技术中的用户描述文档、资源描述文档、个性化推荐技术、个性化服务体系结构以及该领域的主要研究成果进行了综述。通过比较现有原型系统的实现方式，详细讨论了实现个性化服务的关键技术。

关键词：数据挖掘;搜索引擎;个性化;网络资源
中图分类号：TP393 **文献标识码**：A **文章编号**：1674-7720(2010)11-0006-04

Research of personalized search engine

OU Jian Bin
(Jinan University, Guangzhou 510632, China)

Abstract: In this paper, personalized service search engine technology, user profile, resource description files, personalized recommendation technology, personalized service architecture and the main research achievements in this field were reviewed. By comparing the current prototype implementations are discussed in detail the realization of personalized services to key technologies.

Key words: data mining; search engine; personalized; network resources

随着网络的迅猛发展,互联网已成为人们获得信息的重要手段[1]。海量信息在给人们带来方便的同时也带来了许多问题。人们经常要耗费大量宝贵的时间在大量组织松散的信息中寻找有用信息,因此搜索引擎网上检索信息的重要工具[2-3]越来越流行。如何准确高效地为用户提供需要的信息,是搜索引擎信息推荐研究的核心[4]。在搜索引擎中提供个性化服务技术就是针对这个问题而提出的,它为不同用户提供不同的服务,以满足不同的需求。个性化服务通过收集和分析用户信息来学习用户的兴趣和行为,达到主动推荐的目的。个性化服务技术能充分提高站点的服务质量和访问效率,从而吸引更多的访问者[5]。个性化服务系统根据其所采用的推荐技术可以分为2种:基于规则的系统和信息过滤系统。信息过滤系统又可分为基于内容过滤的系统和协同过滤系统[6]。

基于规则的系统,如IBM的WebSphere,其优点是简单、直接,缺点是规则质量很难保证,而且不能动态更新。此外,系统随着规则的数量增多,变得越来越难以管理。基于内容过滤的系统有:PersonalWebWatcher,Letizia等[7],其优点是简单、有效,缺点是难以区分资源内容的品质和风格,而且不能为用户发现新的感兴趣的资源,只能发现与用户已有兴趣相似的资源[8]。协同过滤系统如WebWatcher、GroupLens等,其优点是能为用户发现新的感兴趣的信息,缺点是存在2个很难解决的问题:一个是稀疏性,亦即在系统使用初期,由于系统资源还未获得足够多的评价,很难利用这些评价来发现类似的用户;另一个是可扩展性,随着用户和资源的增多,系统的性能会越来越低[9]。还有一些个性化服务系统如WebSIFT、FAB等,结合了基于内容过滤和协同过滤2种技术。为了克服协同过滤的稀疏性问题,可以利用用户浏览过的资源内容预期用户对其他资源的评价,这样可以增加资源评价的密度。利用这些评价再进行协同过滤,从而提高协同过滤的性能[10]。

1 国内外的研究现状

为了实现个性化服务,首先需要跟踪和学习用户的行为和兴趣,并设计一种合适的表达方式。必须组织好资源,选取资源特征,采用合适的推荐方式。此外还必须考虑系统的体系结构,以及在服务器端、客户端和代理端实现的利弊。

1.1 用户描述文档

在个性化服务系统来说,最重要的是用户的参与,有必要为每个用户建立一个用户描述文档(user profile)。

用户描述文档用来刻画用户特征,在制定用户描述文档之前,需考虑下面几个问题[11]:

(1)描述文档有没有现成的标准?

(2)收集什么数据?收集数据的目的?

(3)如何收集数据?根据什么信息源来收集?

(4)收集的数据如何组织?

(5)可自适应用户信息进行更新?

在收集用户的信息之前,首先需分析用户愿意提供什么信息。用户一般都很注意个人信息的保密性,调查显示,83%的用户愿意向 Web 站点提供自己的姓名、性别、年龄、教育背景和兴趣,但大多数用户不愿意提供私有、敏感的信息,如个人收入和信用卡号等。另一项调查显示,39%的用户愿意 Web 站点向其他 Web 站点共享自己的信息。

1.2 资源描述文档

个性化服务系统处理的资源是由其应用范围确定的[12]。Anatagonomy、SmartPush 应用的范围是报纸;GroupLens 应用的范围是 Usenet 新闻;CiteSeer 应用的范围是科技文档;FireFly 应用的范围是音乐和电影;Amazon.com、eBay 应用的范围是电子商务。还有一些个性化服务系统用于导航、推荐、帮助或搜索,但它们所处理的资源不太相同,如 Personal WebWatcher、WebWatcher、Letizia 处理 Web 页与链接;WebSIFT 处理 Web 访问日志;SiteSeer、PowerBookmark 处理 BookMark 和相关文档;Syskill & Webert、ProFusion 处理从其他搜索引擎返回的查询结果等。目前,个性化服务系统所处理的资源都是文本资源,FireFly 面向音乐和电影,其通过用户评价喜欢的音乐家和电影来进行协同过滤,所以仍然属于文本处理。资源的描述与用户的描述密切相关,通常是用同样的机制来表达用户和资源。资源描述文档可以用基于内容的方法和基于分类的方法来表示[13]。

(1)基于内容的方法

基于内容的方法是从资源本身抽取信息来表示资源,使用最广泛的方法是用加权关键词矢量。这种方法的关键的问题是特征选取,特征选取要达到 2 个目标:一是选取最好的词;二是选取最少的词。要抽取特征词条,需要利用停用词列表对文档进行词的切分,在完成词切分后,接着除去文档集中出现次数过少和过多的词。经过这些处理后,还需对特征进行进一步的选取,以降低特征的维数。常用的特征选择方法有:信息熵(entropy)、信息增益(IG)、互信息(MI)、X^2 统计方法(CHI)、TF-IDF 方法等。比较简单的做法就是计算每个特征的熵值,选取具有最大熵值的若干个特征,即信息量最大的若干个特征;也可以计算每个特征的信息增量,即每个特征在文档中出现前后的信息熵之差;还可以计算每个特征的互信息即每个特征和文档的相关性;还可使用 X^2 统计方法。以上的计算方法中,信息增量方法和 X^2 统计方法表现较好,但这两种方法的计算量比较大。在完成文档特征的选取后,还要计算每个特征的权值,使用最广泛的是 TFIDF 方法,其思想是出现次数越多且罕见的词汇越能代表文档内容,具有很好的类别区分能力,适合用来分类。对某一特征,TF 表示该特征在文档中出现的次数,某一特定词语的 IDF,可以由总文档数目除以包含该词语之文档的数目,再将得到的商取对数。矢量模型的代价是比较大的,有时为了加快处理速度,可以只考虑 TF 一项,矢量模型在只考虑 TF,IDF 以及没有考虑 TF 和 IDF 时都使效果显著下降。在资源描述中特征选择非常重要,一个合理、有效的特征选择方法可以在信息处理阶段去掉数据中的冗余,降低特征空间的维数,提高效率。

(2)基于分类的方法

基于分类的方法是利用类别来表示资源,在给定的分类体系下,根据内容确定所属类别的过程。对文档资源进行分类有利于将文档推荐给对该类文档感兴趣的用户。文本分类方法有多种,常用的有 K 近邻法(kNN),朴素贝叶斯(Naive-Bayes)和支持向量机(Support Vector Machines)等。K 近邻法是一种简单而常用的文本分类方法。该方法的思想是给定一个经过分类的训练文档集合,在对新文档进行分类时,首先从训练文档集合中找出与测试文档最相关的 k 篇文档,然后按照这 k 篇文档所属的类别来对该测试文档进行分类处理。朴素贝叶斯方法将概率模型应用于自动分类,是种简单而有效的分类方法。它的分类思想是使用贝叶斯公式通过先验概率和类别的条件概率来估计文档对类别的后验概率。支持向量机的理论基础是统计学习理论,作为一种相对较新的通用机器学习方法,最近十年来成为自动分类领域的研究和应用热点。资源的类别可以预先定义,也可以利用聚类技术自动产生。许多研究表明:聚类在精度上非常依赖于文档的数量,而且由聚类产生的类型可能对用户来说是毫无意义的,因此可以先使用手工选定的类型来分类文档,在没有对应的候选类型或需要进一步划分某类型时,才使用聚类产生的类型。

1.3 个性化推荐

个性化推荐目前大多采用基于规则的技术、基于内容过滤的技术和协同过滤技术。

(1)基于规则的技术

规则可以由用户定制,也可以使用基于关联规则的挖掘技术来发现,基于规则的推荐信息依赖于规则的质量和数量,其缺点是随着规则的数量增多,系统将变得越来越难以管理。规则本质上是一个 If-Then 语句,它可以利用用户静态属性或动态信息来建立。为了能够使用规则来推荐资源,用户描述文档和资源描述文档需用相同的关键词集合进行描述。信息推荐时的工作过程为:首先根据当前用户阅读过的感兴趣的内容,通过规则推算出用户还没有阅读过的感兴趣的内容,然后根据规则的支持度,对这些内容排序并展现给用户。基于规则的系统组成如图 1 所示。

关键词层提供上层描述所需的关键词,并定义关键

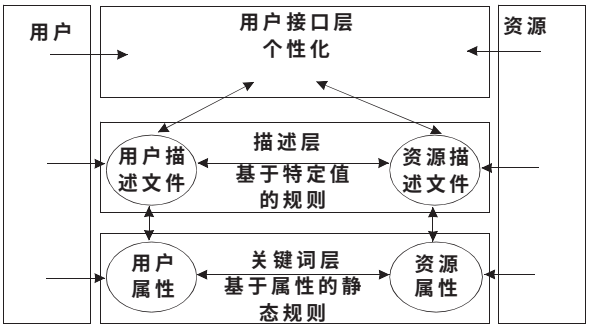

图 1 基于规则的技术关键词层、描述层和用户接口层

词间的依赖关系，该层可以定义静态属性的个性化规则。描述层定义用户描述和资源描述，由于描述层是针对具体的用户和资源，所以描述层的个性化规则是动态变化的。用户接口层提供个性化服务，根据下面 2 层定义的个性化规则将满足规则的资源推荐给用户。

(2)信息过滤技术

信息过滤技术可分为基于内容过滤的技术和协同过滤技术[15-16]。基于内容过滤的系统，如 Personal Web-Watcher、Letizia 等，它们利用资源与用户兴趣的相似性来过滤信息。其优点是简单、有效，缺点是难以区分资源内容的品质和风格，而且不能为用户发现新的感兴趣的资源，只能发现和用户已有兴趣相似的资源。协同过滤系统，如 WebWatcher、GroupLens 等，它们利用用户之间的相似性来过滤信息，优点是能为用户发现新的感兴趣的信息，缺点是存在 2 个很难解决的问题。一个是稀疏性，即在系统使用初期由于系统资源还未获得足够多的评价，系统很难利用这些评价来发现相似的用户；另一个是可扩展性，即随着系统用户和资源的增多，系统的性能会越来越低。还有一些个性化服务系统，如 WebSIFT、FAB 等，同时采用了基于内容过滤和协同过滤这 2 种技术。这 2 种过滤技术结合可以克服各自的一些缺点，利用用户浏览预期用户对其他资源的评价，增加资源评价的密度克服协同过滤的稀疏性问题，利用这些评价再进行协同过滤，从而提高协同过滤的性能。用户与资源的关系数据可以用 $m \times n$ 阶矩阵来表示，m 代表用户数量，n 代表项目数量，R_{ij} 代表第 i 个用户对项目 j 的评分。用户资源信息建模如表 1。

表 1 用户资源信息建模

Item / User	$Item_1$	…	$Item_j$	…	$Item_n$
$User_1$	$R_{1,1}$	…	$R_{1,j}$	…	$R_{1,n}$
…	…	…	…	…	…
$User_i$	$R_{i,1}$	…	$R_{i,j}$	…	$R_{i,n}$
…	…	…	…	…	…
$User_m$	$R_{m,1}$	…	$R_{m,j}$	…	$R_{m,n}$

1.4 个性化服务体系结构

基于 Web 的个性化服务体系结构与用户描述文档分布的位置有很大关系。用户描述文档可以存储在服务器端、客户端、代理端[17-18]。大部分个性化服务系统的用户描述文档都存放在服务器端，如 Syskill&Webert、Letizia、GroupLens、Anatagonomy 等。优点是可以避免用户描述文档的传输，除了支持基于内容的过滤，还可以支持协同过滤；缺点是用户描述文档不能在不同的 Web 应用之间共享。也有一些系统的用户描述文档是存储在客户端的，如 PointCast Network，这种体系的个性化服务可以在服务器端实现，也可以在客户端实现。它的优点是用户描述文档可以在不同的应用之间共享，缺点是只能进行基于内容的过滤。还有一些系统的用户描述文档是存储在代理上的，如 Personal WebWatcher、PVA 等，这种体系的个性化服务可以在服务器端实现，也可以在代理上实现。其优点是不仅可以支持基于内容的过滤和协同过滤，还支持用户描述文档在不同 Web 应用之间的共享；缺点是可能需要传输用户描述文档。

2 关键问题

(1)综合各系统优点

基于规则的系统其优点是简单、直接，缺点是规则质量很难保证，不能动态更新。此外，随着规则的数量增多，系统将变得越来越难以管理。基于内容过滤的系统其优点是简单、有效，缺点是难以区分资源内容的品质和风格，而且不能为用户发现新的感兴趣的资源，只能发现和用户已有兴趣相似的资源。基于协同过滤系统的优点是能为用户发现新的感兴趣的信息，缺点是它的稀疏性和可扩展性，结合基于内容的过滤和协同过滤这两种技术可以克服各自的缺点，可以利用用户的浏览记录推测用户对其他资源的评价，增加资源评价的密度，来克服协同过滤的稀疏性问题。利用这些评价进行协同过滤，从而提高协同过滤的性能。

(2)用户描述文档的存放

用户描述文档可以存放在服务器端、客户端、代理端，大部分个性化服务系统的用户描述文档都存放在服务器端。其优点是可以避免用户描述文档的传输，支持基于内容的过滤和协同过滤；缺点是用户描述文档不能在不同的 Web 应用之间共享。用户描述文档存储在客户端，优点是用户描述文档可以在不同的 Web 应用之间共享，缺点是只能进行基于内容的过滤。用户描述文档存储在代理上，优点是不仅可以支持基于内容的过滤和协同过滤，还支持用户描述文档在不同的 Web 应用之间共享，缺点是需要传输用户描述文档。

(3)为推荐文档评分和排名

个性化服务技术是目前非常流行的一种技术，本文分析了各种具有代表性的个性化服务系统，并在此基础上详细描述了建立个性化服务的关键技术。要将个性化服务引入搜索引擎中，面对日益增长的 Web 信息，要满足不同背景、不同目的和不同时期的查询请求，必须针对不同用户提供不同的服务才能真正解决这个问题。个性化服务技术在搜索引擎上仍有很多值得研究和探讨的问题。

参考文献

[1] PRETSCHNER A. Ontology based personalized search[MS. Thesis]. Lawrence, KS: University of Kansas, 1999.

[2] LEE D L, CHUANG H, SEAMONS K. Document ranking

and the vector-space model. IEEE Softwar, 1997,14(2).

[3] XUE Gui Rong, LIN Chen Xi. Scalable collaborative filtering using cluster-based. In: Hong Kong University of Science and Technology, ACM, 2005.

[4] Jun Wang, Arjen P. de Vries, Marce J. T. Reinders1: unifying user-based and item-based collaborative filtering approaches by similarity fusion. Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, ACM, 2006.

[5] 曾春,邢春晓,周立柱.基于内容过滤的个性化搜索算法[J].软件学报,2003,14(5):999-1004.

[6] 曾春,邢春晓,周立柱.个性化服务技术综述[J].软件学报,2002,13(10):1952-1961.

[7] 刘均,李人厚.一种面向个性化协同学习的任务生成方法[J].软件学报,2006,17(1).

[8] 陈健,印鉴.基于影响集的协同过滤推荐算法.软件学报[J],2007,18(7).

[9] 韩立新,陈贵海,谢立.一个面向 Internet 的个性化信息检索系统模型[J].电子学报,2002,30(2).

[10] 邢春晓,高凤荣,战思南,等.适应用户兴趣变化的协同过滤推荐算法[J].计算机研究与发展,2007,44(2).

[11] 张锋.使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J].计算机研究与发展,2006,43(4).

[12] SRIVASTAVA J, COOLEY R, DESHPANDE M, et al. Web usage mining: discovery and applications of usage patterns from Web data. In:Fayyad, U., ed. Proceedings of the ACM SIGKDD Explorations. New York: ACM Press, 2000,1(2):12-23.

[13] 赵亮,胡乃静,张守志.个性化推荐算法设计[J].计算机研究与发展,2002,39(8):986-991.

[14] 汪晓岩,胡庆生.面向 Internet 的个性化智能信息检索[J].计算机研究与发展,1999,36(9):1039-1046.

[15] VOLOKH E. Personalization and privacy [J]. Communications of the ACM, 2000,43(8):84-88.

[16] WU Y H, CHEN Y C, CHEN A L P. Enabling personalized recommendation on the web based on user interests and behaviors. In:Klas, W., ed. Proceedings of the 11th International Workshop on Research Issues in Data Engineering. Los Alamitos, CA: IEEE CS Press, 2001.

[17] 沈云斐,沈国强,蒋丽华,等.基于时效性的 Web 页面个性化推荐模型的研究[J].计算机工程,2006,32(13):80-81,99.

[18] 秦国,杜小勇.基于用户层次信息的协同推荐算法[J].计算机科学,2004,31(10):138-140.

(收稿日期:2010-02-02)

作者简介:

欧建斌,男,1984 年生,硕士研究生,主要研究方向:数据挖掘技术。