个性化智能信息提取中的用户兴趣发现

Discovering User's Interests in Personalized Intelligent Information Retrieval System

欧 洁 林守勋
(中国科学院计算技术研究所 北京100080)
刘桂林
(郑州大学北区新开普电子技术有限公司 郑州450002)

Abstract The increasing stream of electronic information available makes it ever more time consuming to find interesting information. Personalized intelligent information retrieval system holds the promise of relieving the user's burden by learning a model of the user's interests and using this model to find interesting documents. A method that can fast and accurately discover user's interest in a personalized intelligent information retrieval system is described in this paper. Firstly, user provides his interests by the means of answering problems, so system can get the initial personal vector of user from the answers. After that, system alters the persona vector of user according to the feedback information and Server Log Data of user.

Keywords Vector space model, Personalized intelligent information retrieval, Web usage mining

1 引言

1990年,WWW(World Wide Web)出现,在随后的几年中它获得了空前的发展,Internet 上的信息量以指数形式飞速增长,现在,Internet 已成为一个浩瀚的海量信息源。但由于 Internet 是一个具有开放性、动态性和异构性的全球分布式网络,资源分布很分散,且没有统一的管理和结构,这就导致了信息获取的困难。如何快速、准确地从浩瀚的信息资源中寻找到所需的信息已经成为困扰用户的一大难题。

为了实现信息提取的智能化,人们将人工智能技术引入到信息提取中,已经研究出了各种智能信息提取方法,提出了一些智能型检索系统模型。智能信息提取系统将人工智能技术应用于 WWW 信息处理领域来解决网上信息有效获取的问题,是帮助人们快速获取信息的有效手段。下面介绍几种典型的智能信息提取系统。

文[2]所介绍的 WebWatcher 是一个非常著名的导航器,它帮助用户在网上导航,同时该系统通过为用户选择"链路"或站点跟踪学习,改善了导航的质量。其学习算法属于一种强化学习算法。

文[8]所介绍的系统是一个主动推送网页的系统。它每天提供一些可能会让用户感兴趣的网页,用户根据自己的兴趣来评价这些页面,系统根据这个评价信息自我调整、从而改善推送性能。系统的实现算法如下:

①搜索 WWW 中的网页。

②选择最符合用户兴趣的 p 个页面给用户。

③从用户那儿得到关于每一个推送页面的评价。

④根据反馈信息更新智能选择算法。

⑤继续从第①步开始执行。

文[5]所介绍的 IDGS 系统(Information discovering and gathering system)是为了在 WWW 上自动进行中英文技术资料的搜集而设计开发的,能够根据用户提交的挖掘目标样本,在 WWW 上自动查找用户所需的信息。IDGS 系统采用了向量空间模型(VSM)和基于词频统计的权值评价技术,由特征提取、源站点查询、文档采集、模式匹配等4部分组成(图1)。IDGS 系统的工作流程如下:

①特征提取:对用户提交的挖掘目标样本进行特征提取,生成挖掘目标的特征矢量。

②站点查询:在特征矢量中取权值最大的3~10个特征项作为查询关键字,向多个资源索引系统发送查询请求,将返回的结果 URL 作为文档采集的起点。

③信息采集:运行 Robot 程序从查询到的源 URL 开始进行文档采集。

欧 洁 博士生,研究方向为知识发现、数据挖掘;林守勋 研究员,博士生导师,研究方向为 CSCW、数据库,

④模式匹配：提取出源文档的特征矢量，并进行特征匹配，把符合阈值条件的文档提交给用户。

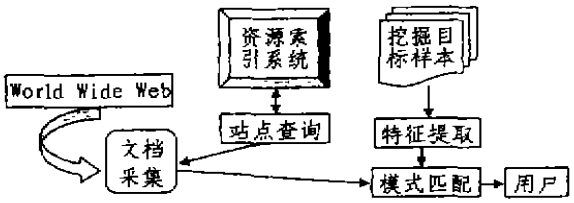

图 1

2 现有系统的缺陷和解决方案

个性化智能信息提取系统是帮助人们快速获取信息的有效手段。然而，现在系统仍然存在如下一些缺陷或不足。

(1)这些系统发现用户兴趣的方式通常为：①由用户以关键字方式提供自己的兴趣。②从一类文档中提取。以关键字的方式提供自己的兴趣的缺陷为：用户经常不能准确地表达自己的兴趣，从一类文档中提取用户兴趣的缺陷为：不能全面发现用户的兴趣。

(2)用户与系统间的交互方式比较单调。现有系统普遍采用相关反馈技术作为用户和系统进行交互的主要手段，根据用户浏览页面的信息来自动更新用户的个性化向量是目前现有系统所缺少的。

本文所介绍的方法是：以回答问题的方式让用户提供兴趣，并用相关反馈和用户访问信息挖掘相结合的方法来更新用户兴趣库，从而有效地利用了以回答问题的方式让用户提供兴趣的优点：能快速地获得用户兴趣，以及将用户浏览页面的信息与用户对检索结果的反馈信息结合起来更新用户兴趣库，全面地利用了用户留在服务器方的信息，因而能自适应用户兴趣的变化。因此，该方法能快速、精确地发现用户的兴趣。

3 个性化智能信息提取

个性化智能信息提取即根据用户的兴趣过滤查询结果并将信息主动推荐给用户。它的实现过程包括二个阶段：(1)发现用户兴趣。(2)将 WWW 上的信息与用户兴趣相比较，把与用户兴趣相符合的信息提供给用户。因为这里所指的个性化智能信息提取是由计算机来完成的，所以信息和用户兴趣都必须以一定的方式在计算机内表示出来，这就是目标表示。

3.1 目标表示与特征提取

目标表示是指从目标信息(如用户兴趣和 WWW 上的信息)中选取一些特征项(如词条等)，用这些特征项及其在目标信息中的重要性来代表目标信息。目标表示模型有多种，常用的有布尔逻辑型、向量空间型、概率型等。近年来应用较多的模型是向量空间模型(vector space model，VSM)法。

用 VSM 实现目标表示的方法为：(1)选取一组适合于表示所有目标信息的词条($T_1, T_2, \cdots, T_n$)，(2)对于目标信息 F，根据词条 T_i 在 F 中的重要程度求出权值 W_i，$i=1, \cdots, n$，这样就可以把目标信息 F 用词条特征矢量($W_1, W_2, \cdots, W_n$)表示出来，W_i 表示 T_i 在 F 中的权重，$i=1, \cdots, n$。词条和权值的选择方法称为特征提取。词条 T_i 在文章 F 中对应的权重 W_i 可以用以下公式计算。

$$W_i=\frac{(0.5+0.5\frac{tf(i)}{tf_{max}})(\log\frac{n}{df(i)})}{\sqrt{\sum_{T_j\in F}((0.5+0.5\frac{tf(j)}{tf_{max}})^2(\log\frac{n}{df(j)})^2)}}$$

tf(i)为词条 T_i 在文章 F 中的出现次数，df(i)是包含词条 T_i 的文章数，n 是全部文章数目，tf_{max} 是文章 F 中出现最多的词条的出现次数。具体请参阅文[8]。

用以上方法可以把文档和用户兴趣用向量的方式表示出来，从而将文档信息和用户兴趣的匹配问题转化为向量空间中的向量匹配问题来进行处理。假设用户兴趣的词条特征矢量为 $U=(u_1, u_2 \cdots, u_n)$，未知文档的词条特征矢量为 $V=(v_1, v_2 \cdots, v_n)$，两者的相似程度可用向量之间的夹角来度量，夹角越小说明相似度越高，相似度计算公式如下：

$$Sim(V,U)=\cos(V,U)$$
$$=\sum_{k=1}^{n}u_k\cdot v_k/\sqrt{\sum_{k=1}^{n}u_k^2}\cdot\sqrt{\sum_{k=1}^{n}v_k^2}$$

3.2 发现用户兴趣

通常有三种方法用于发现用户的兴趣：1. 通过用户主动提供自己的兴趣来得到用户的个性化向量。2. 在用户没有明确参与的情况下，系统通过观察用户行为来得到用户的兴趣，从而得到用户的个性化向量。3. 通过用户的反馈信息来更新用户的个性化向量。

实现第一种方法有两种方式：1. 用户将自己感兴趣的信息类或在线文档分类后提供给系统，系统从这些文档或信息类中发现用户的兴趣[7]。2. 用户提供自己的研究方向和其它阅读爱好等信息，系统从这些信息中发现用户的兴趣。本文主要讨论第二种方式。实现它的一种方法为：让用户回答一些问题，如：(1)从事的专业、研究兴趣和研究方向。(2)最近参加的项目以及用一两句话描述这些项目。(3)除了自己的专业外经常阅读哪些专业和研究方向的资料。(4)是否对音乐、体育感兴趣？(5)对哪些种类的音乐和体育感兴趣？(6)经常读哪些非本专业的书？等等。对用户输入的答案进行目标表示后，对其进行聚类。一般来说，不同问题的答案形成不同的类，表示用户的各种兴趣。这样，在根据

用户兴趣推送页面或进行信息过滤时，可以根据用户的不同兴趣来推送或过滤，从而有效避免将各种不同兴趣表示成一个用户兴趣向量的缺陷，如只与某一项兴趣有关的文档与用户的个性化向量的相关性小等[1]。

这种方法的实现无须训练，因此可以快速地得到用户的兴趣，而且实现简单，但是，由于用户的兴趣并不是一成不变的，让用户每次访问网站时都输入这些内容会使用户觉得很繁琐，而且用户一般不可能提供所有的兴趣以及感兴趣的程度。

第二种方法是通过用户已访问过的页面去挖掘用户兴趣。一般来说，用户只会访问自己感兴趣的页面，除非用户判断失误，点击了自己并不感兴趣的页面，但这种情况很少发生。所以，我们可以认为用户访问过的页面中包括用户感兴趣的信息，对这些页面上的信息进行分析就可以发现用户的兴趣。从用户已访问过的页面去挖掘用户兴趣要用到页面访问挖掘。

这种方法能自适应用户兴趣的变化，而且不要求用户输入任何信息。但是，如果要达到一定的精确度的话，一般需要经过比较长的训练时间。

第三种方法为：根据用户对推送页面的评价信息来更新用户的个性化向量。这种方法实现简单，但需要用户对推送页面做出评价。

为了快速、精确地发现用户兴趣，本文提出采用这三种方法相结合的方法。首先，它要求用户通过回答问题的方式来提供自己的兴趣，从而得到用户的初始个性化向量。然后，通过用户的反馈信息和挖掘用户在访问 Web 时在服务器方留下的访问记录来修改用户的个性化向量。前面已经介绍了如何实现通过让用户回答问题的方式来得到用户的初始个性化向量的方法，下面介绍根据用户的页面访问信息和用户的反馈信息更新用户的个性化向量的方法。

为了根据用户的页面访问信息来更新用户的个性化向量，首先要发现用户的访问信息，所以下面介绍从日志文件中发现用户的访问信息的实现。

3.3 从日志文件中发现用户的访问信息

用户与 Web 服务器间交互的过程信息（包括用户的登录信息、用户的浏览记录）都以日志文件的形式存在，通过分析这些日志文件可以发现用户已浏览过的页面集和浏览这些页面的时间长度，这个过程是页面访问挖掘[3]中的一个重要组成部分。这个过程要完成的任务为：将日志文件中的信息按用户进行分类整理，得到每一个用户的访问情况库，库中的记录表示页面的访问信息，包括的内容有：页面的 URL 地址、页面的访问日期和访问页面的时间长度等。这些记录按访问时间的先后进行排序。访问情况库的内容可用以下三元组来描述：

$$t=(ip_i,uid_i,\{(P_1.url,P_1.time,P_1.length),\cdots,(P_m.url,P_m.time,P_m.length)\})$$

其中 ip_i 表示用户的 IP 地址，uid_i 表示用户的 ID，对于任意的 $0<k<m+1$，P_k 表示用户在该事件中访问的第 k 个页面，$P_k.url$ 表示页面 P_k 的 URL 地址，$P_k.time$ 表示用户访问 P_k 的日期，$P_k.length$ 表示用户访问 P_k 的时间长度，$P_1.time<P_2.time<\cdots<P_m.time$。

3.4 发现内容页面

用户浏览 Internet 上的信息的主要方法为：打开信息所在站点的主页面，点击一系列链接来打开信息所在的页面。这样，用户为了访问到自己需要的信息所在的页面（内容页面），一般都需要经过一系列其它的页面（辅助页面）。为了发现用户的兴趣，只需要对内容页面进行挖掘，所以要区分用户访问的页面哪些是内容页面，哪些是辅助页面，区分的方法有二种：①根据访问时间的长短来进行区分，实现方法为：设定一个阈值 a，访问时间超过 a 的认为是内容页面，访问时间小于 a 的认为是辅助页面。②最大向前规则，这种方法的根据是：用户访问完内容页面后，用户将退回已访问过的页面。因此该方法认为内容页面是用户按“后退”键前所访问的页面。如访问路径为：ABCDCEFECBG，则认为内容页面为 D，F，G，实验证明第一种方法比第二种方法略好些。

在求出了用户访问的内容页面集后，采用上节所述的 VSM 方法对这些页面进行特征提取和目标表示。

目标表示的实现方法为：把汉语中适合描述页面特征的词条形成数据库，分别求出这些词条在页面中的出现次数。对每一个页面，取出现次数最多的二十个词条作为其特征，求出每一个词条的权重，这样就将页面在计算机内用目标表示的方法描述了出来。

3.5 根据用户的页面访问信息更新用户库的方法

我们提出了如下根据用户的页面访问信息更新用户库的方法。

设 $M_1,\cdots,M_n$ 表示对用户主动提供的兴趣或感兴趣的文档进行聚类后形成的所有个性化向量，向量 $V_1,\cdots,V_m$ 表示用户已访问过的所有页面。那么，如果页面 V_i 与用户的某一项兴趣 M_j 非常接近，即 $V_i\cdot M_j$ 的值大于阈值 a，则用该页面对应的向量加强到用户的兴趣 M_j 中，即：

$$M_j=M_j+\frac{m^2*V_i*VI(i)*T(i)}{\sum_{k=1}^{m}VI(K)*\sum_{k=1}^{m}T(k)}$$

$VI(i)$ 表示用户访问页面 V_i 的频度，$T(i)$ 表示用户访

问页面 V_i 的时间，选择上式的主要依据是：以用户访问页面 V 的访问频度余以平均访问频度和用户访问页面 V_i 的访问时间除以平均访问时间为标准衡量用户对页面 V_i 的感兴趣程度。通过上式可知，用户访问页面的信息加强了用户的相应兴趣。如果页面 V_i 不与用户的任何一项兴趣接近的话，则将分量值对应为

$$M_j = M_j + \frac{m^2 * V_i * VI(i) * T(i)}{\sum_{k=1}^{n} VI(k) * \sum_{k=1}^{n} T(k)}$$

的向量作为用户的新兴趣。当然这样会导致用户的兴趣种类大量增加，所以设定一个阈值 b，如果用户的某一项兴趣对应向量的模小于 b 时，在推送相关页面时不考虑该项兴趣。

这样，用户访问页面的情况也反映在用户的兴趣向量中，从而可以对发现用户兴趣的前二种方法扬长避短、快速、精确地发现用户的兴趣。

3.6 根据用户反馈的信息更新用户库的方法

如果用户对被推荐页面进行了评价的话，可以根据用户的评价信息来更新用户库。例如用户对页面的评价标准可以设为：正好符合要求、相关、无关，对应的用于更新用户个性化向量的评价分数为：正好符合要求：+2，相关：+1，无关：-2，设 $M_1, \cdots, M_n$ 表示对用户主动提供的兴趣进行聚类后形成的所有个性化向量，向量 $V_1, \cdots, V_m$ 表示用户评价过的所有页面。那么，如果页面 V_i 与用户的某一项兴趣 M_j 非常接近，即 $V_i \cdot M_j$ 的值大于阈值 a，则将该页面对应的向量乘上用户对它的评价分数后累加到用户的兴趣 M_j 中，即：$M_j = M_j + e_i \cdot v_i$，$e_i$ 表示用户 M 对页面 i 的评价。通过上式可知，用户对页面的评价反应到了用户的相应兴趣中。如果页面 V_i 不与用户的任何一项兴趣接近的话，则将分量值对应为 $e_i \cdot v_i$ 的向量作为用户的新兴趣。

3.7 通过用户兴趣预送页面

(1)用户一进入站点就将其可能会感兴趣的页面推荐给他。根据用户的兴趣和文档的相关程度来确定。

(2)根据用户进入站点后已访问的页面来预测他下一步可能会访问的页面，并将其推荐给用户。应用时序规则发现的方法来实现。

4 方法与分析

本文所介绍的方法是：以回答问题的方式让用户提供兴趣，并用相关反馈和用户访问信息挖掘相结合的方法来更新用户兴趣库，从而有效地利用了以回答问题的方式让用户提供兴趣的优点，能快速地获得用户兴趣，并且采用的问题能使用户比较全面地提供自己的兴趣。而且该方法将用户浏览页面的信息与用户对检索结果的反馈信息结合起来更新用户兴趣库，全面地利用了用户留在服务器方的信息，因而能准确地自适应用户兴趣的变化。因此，该方法能快速、精确地发现用户的兴趣。

结束语 本文介绍了 WWW 上个性化智能信息提取中发现用户兴趣的方法。该研究内容属目前智能信息检索领域的重要研究课题，具有很强的理论意义和实际意义，目前国内并没有将个性化智能信息提取和页面使用挖掘结合起来的系统，因此本文对于如何实现一个优秀的个性化智能信息提取系统有着重要的意义。

参 考 文 献

1 De Kroon H C M, Mitchell T M, Kerckhoffs E J H. Improving Learning Accuracy in Information Filtering. Available at: http//WWW. cs. cmu. edu

2 Joachims T, Freitag D, Mitchell T. Web Watcher: A Tour Guide for the WWW. Available at: http://WWW. cs. cmu. edu

3 Mobasher B, et al. Web Mining, Pattern Discovery from World Wide Web Transactions. Available at: http://www cs. cmu. edu/

4 汪晓岩，胡庆生，李斌，庄镇泉. 面向 Internet 的个性化智能信息提取. 计算机研究与发展，1999，36(9)：1039～1046

5 邹涛，王继成，朱华宇，金翔宇，张福炎. WWW 上的信息挖掘技术及实现. 计算机研究与发展，1999，36(8)：1019～1024

6 Chen L, Sycara K. WebMate: A Personal Agent for Browsing and Searching Available at: http//www. cs. cmu edu

7 Krulwich B, Burkoyc. The InfoFinder Agent: Learning user interests through heuristic phrase extraction. IEEE Expert, 1997, 12(5): 22～27

8 Balabanovic M, Shoham Y, Yun Y. An Adaptive Agent for Automated Web Browsing. Available at: http://www-diglib. stanford. edu/cgi-bin/get/SIDL-WP-1995/0023.