

“天罗”个人信息代理系统的设计与实现

张小斌* 程学旗* 白 硕* 朱建民**

*国家智能计算机研究开发中心 北京 2704 信箱 100080

**中国人民解放军总后科研所

Email:cxq@ncic.ac.cn

摘要: 个人信息代理可以有效地解决人们在查找网上信息时的“信息迷向”和“信息拥塞”。本文介绍了“天罗”个人信息代理系统设计的思路和实现方法,包括用户信息需求的处理,文本内容的分析、信息分流和个性化用户反馈的实现等。在对实验数据分析的基础上,对以后的工作进行了展望。

关键词: 个人信息代理, 信息检索, 信息推送, 相关反馈

The Design and Implementation of TechRoute™ Personal Information Agent

Zhang xiaobing Cheng xueqi Bai shuo Zhu jianmin

National Research Center for Intelligent Computing systems(NCIC),beijing 2704, 100080

Email:cxq@ncic.ac.cn

Abstract: As the size of the Internet increases, users in hope of finding useful information are suffering from information disorientation and information congestion. Information agent technology can help to understand user's long-term information requirement in an intelligent way.

In this paper, we present the design and implementation of Tianluc Personal Information Agent, an active information push software. This includes the disposal of information requirements, analysis of hypertext content, information diffuence and relevancy feedback disposal of users. Based on the experiment result, some advanced work is discussed.

Keyword: Personal Information Agent, Information Retrieval, Information Push, Relevancy Feedback

1 网上信息服务的困境

随着 Internet/Intranet 的迅速发展,基于 WWW 的信息处理日益成为人们关注的焦点。越来越多的信息以电子化方式存放在网上,但是信息获取手段的提高并没有与信息的增长同步。虽然有许多信息检索和过滤工具被开发出来,然而,传统的信息服务系统没有有效

的手段理解用户准确的信息需求，这些系统也忽视用户的个性化要求，不能提供长期的主动的信息服务。“信息迷向”和“信息拥塞”困扰着寻找网上电子信息的人们。

信息代理技术可以智能化地理解用户的兴趣和长期的信息需求，使用信息检索、自动分类、机器学习和信息推送技术，为用户提供准确、可靠和方便的信息服务。目前，人们在信息检索、信息过滤、站点导游等方面展开了广泛的研究。

信息代理能够完成如下任务：

1. 充分了解所属用户的所有相关信息，建立并逐步优化用户的领域模型。
2. 将用户信息检索的需要，提交给 Agent，并完成用户信息检索的任务。
3. 根据用户的领域模型和历史记录，确认检索到的信息的相关性。
4. 自动记录、学习并调整用户的领域模型。
5. 代替用户完成信息的分析及其它处理任务。

2. 系统的设计

“天罗”个人信息代理系统是为克服信息过载以及信息迷向而设计的按照用户指定的个人信息需求描述，实行主动服务的软件实体。个人信息代理系统使得用户面对的是一个完全个性化的满足个人信息需求的虚拟 Internet。

系统设计的要求为每天能够处理信息采集器采集的数千至一万个页面，可以为大约 50 个左右的用户服务，每个用户可以提交 1 到 8 类兴趣。要求系统具有合适的处理速度和存储开销，同时有较高的准确率，用户界面友好。

系统由一个信息采集器和一个运行于个人信息代理服务器上的个人信息代理部件组成。信息采集器监视 Internet 上流动的信息，并将最新的和发生变化的信息采集到本地。而个人信息代理根据用户的信息需求，将符合用户信息需求的信息分发到相应的用户。系统可以同时为大量用户提供主动的信息服务。

服务器端主要由一个网络信息采集器和基于内容的缓存系统来管理网上的动态信息，系统使用传统的文本检索技术，并结合一定的 HTML 分析，利用自动分类(Classify)、信息过滤(Filter)、信息推送(Push)和机器学习技术(Machine Learning)为不同的用户整理和提交各类信息。并且根据用户的反馈信息进行自学习，来更新用户的信息需求模型。系统主要包括六个基本功能模块：

信息搜集器：从不同的信息源获取信息，经过内容抽取，规范表达信息，为分流控制器使用。

用户输入代理：获取用户需求信息，规范和精化需求信息的描述，维护用户需求数据库。

判定树生成器：采用上述描述的算法，将需求信息构造成“高频属性优先分支”判定树。

信息分发器：根据用户的个人信息特点，确定信息分发的形式；

信息分流控制器：采用分流算法，对从信息搜集器来的信息分流到不同的需求槽内。

反馈信息收集器：收集用户的反馈信息，形成用户个人兴趣描述向量，并计算源信息与用户信息需求的相关度。

3 系统的实现

3.1 信息需求的处理

用户对于信息内容的要求和其它格式要求通过一个“高频属性优先”的判定树提交给系统。判定树在构造过程中，需要一个分解栈，一个需求掩模。分解栈是一个堆栈数据结构，掩模是一个包含当前全部部分兴趣编号的集合，掩模具有同判定条件倒排表中 idList 同样的数据结构。兴趣编号并不要求是连续的。

判定树的构造算法如下：

1. 根据原始的信息需求，构造属性表(L)：对每个属性构造对应的需求倒排链表，把“当前有效链长”置为链的全长，把“当前起始位置”置为0。把“需求掩模”置为所有编号的集合。初始化树的根节点，并将其置为当前节点(CurrentNode)。初始化分解栈(DivStack)；
2. 按固定排序选择第一个“当前有效链长”最长的需求链表，其对应的属性为 p，需求掩模分解为满足 p 的信息需求的编号集合(SET_p)和不满足 p 的编号集合($SET_{\bar{p}}$)。后者进栈，前者被置为当前需求掩模。生成新节点 N(p)，插入到树中，新节点成为当前节点。同时，修改 p 对应的需求链表中 BeginingIndex 的值；
3. 按照当前需求掩模调整整体属性表(L)；
4. 当属性表中所有需求倒排链表的 SET_SIZE 为零时，当前节点置为输出节点。如果 DivStack==NIL，成功退出。否则，出栈，修改需求掩模，当前节点指向其父节点，回到3；如果存在 SET_SIZE != 0，回到2。

3.2 信息分流

信息分流是分析并提取信息源的各种系统可以理解和处理的属性，并根据用户对信息的要求，进行判定，并将其推送给用户的过程。信息分流包括源文本信息的提取和通过判定树的过程。

页面分析首先使用的是传统的文本信息处理方法，包括如下一些步骤：

1. 从页面数据库中读取页面，过滤 HTML 文本中的标记，将其变成纯文本。
2. 如果是中文文本，使用最大词长匹配算法，进行汉语分词。如果是英文文本，使用 STEMMING 算法，进行词形还原和词根识别工作。
3. 过滤禁用词。
4. 记录关键词的位置信息，关键词的位置信息是指关键词在文本中出现的先后顺序，而不是从文本开始位置起的以字节计算的偏移值。
5. 统计关键词的词频等信息。关键词的权值是超文本格式信息对词频加权的結果。

根据信息需求判定树将源信息向用户个人推送，称之为信息的分流。信息分流操作在将源信息表达成一个属性向量之后，将这个属性向量与需求判定树匹配过滤。每满足一个信息需求节点时，就在信息分流库内添加一条推送记录。信息分流的算法描述如下：

1. 规整信息条目 I，形成带权属性集合；
2. 设当前节点 N 为树根节点；

3. 处理当前节点 N。在满足一定的阈值的前提下，结合权值，计算出当前节点在 I 中的表现值。标识路径；
4. 如果 N 是输出节点，输出；
5. 宽度优先遍历子树，产生的新节点为当前节点，没有新节点，则遍历结束并返回，否则到（3）。

对分流到用户的文本还要计算文本与用户信息需求的相关性评分，并根据词间距信息进行加权。系统将分流得到的信息以电子邮件、频道和预留空间等多种方式提交给用户。

以上的算法更具体的描述可以参照 [程 98]。

3.3 个性化的相关反馈处理

用户的信息需求常常与系统推送结果存在着偏差，表现在 1：用户无法用准确的语言表达自己的信息需求；2：系统返回的结果虽然与信息需求相关，但并不合用户的兴趣。在这里兴趣指 Web 页面与用户长期信息需求目标的相关度。可以用相关反馈来解决这个问题，对于用户先前访问过的一组站点或者页面的反馈信息可以用来生成用户兴趣描述向量，来对未访问过的页面和站点进行预测。目前最为常用的用户兴趣模型是向量特征表示方法。

为了使系统尽可能实用，系统选择了折中的设计。在实现中，提出了以下一些要求：

1. 反馈功能为可选项，用户完全可以只选择判定树的输出直接作为推送信息。这是为了系统减少时间开销，同时，许多用户并不情愿提供反馈信息。
2. 用户对通过判定树判定的文本进行反馈。
3. 用户的反馈信息并不要求一次完全得到，而是添加式。系统记录用户历次的反馈信息，并进行合并。
4. 用户初始反馈页面通过提取特征项，形成该用户对对应兴趣的带权值的兴趣描述向量。
5. 用户以后的反馈信息通过算法合并到已有兴趣向量中，因为系统面对的是对用户信息需求的不完全的训练，因此，在两个向量合并中，应给予后来反馈的向量较高的权值；
6. 用户认为以前的反馈与信息需求发生偏差时，可以重置兴趣描述文件(profile)，即将该向量置空；
7. 为了减少计算量，只有通过判定树判定的页面才需要通过相应用户的兴趣向量的判定，决定是否推送给用户。这样一方面保证了系统推送的页面是与信息需求相关的，另一方面减少了计算量。判定需要计算输入页面与用户个性化向量的余弦距离；
8. 用户可以自己调节阈值，来得到合适的输出。

在系统中，用户可以在浏览时，向系统提交反馈信息。反馈信息在下一次信息推送时处理。反馈信息被处理后，清空反馈信息数据库，以便记录下一次的用户反馈信息。

系统对用户兴趣描述的处理提供四个方法：

初始化：如果是用户首次提交的反馈信息，形成用户的兴趣描述文件；

修正：对已经存在用户兴趣描述的用户反馈信息，需要将新的反馈信息加入到原有

的用户兴趣描述文件中。

清除： 当根据用户兴趣描述文件而得到的页面与用户兴趣相似度偏离用户真正的信息需求时，用户可以方便地清除原来的兴趣描述文件，从而产生新的描述文件。

判定： 用户兴趣描述文件与页面的相似度判定。即对输入的页面，需要计算该页面与用户兴趣的相似度，只有高于阈值的页面才向用户进行推送。

4. 系统运行结果

用户数：50	兴趣词条数：733
兴趣关键词：581	判定树节点数：639个
信息源：59,548K	站点数：15
有效页面（HTML 格式文件）：8232 页	
分流时间：228 秒（不包括反馈功能）	
每个推送条目：平均 280 字节	

系统在三分多钟时间有效地为 50 个用户进行了信息分流，共处理了 8232 个页面。由于系统通过判定树的时间很短，因此，可以认为，系统在时间和空间开销上基本满足应用的要求。

系统目前存放了信息采集器采集得到的页面 9 万多个，目前系统抓取的站点全部是国内高校的站点，共有 465 个站点，其中一个主机提供多个虚拟主机的 WWW 服务被视为不同的主机。

图 1 系统性能参数

系统使用前两万个页面进行测试：

1. 使用 10000 到 20000 之间的页面得到没有反馈信息的推送输出。见表 1 的 I 部分
2. 系统使用 1 到 10000 之间的推送结果，用户根据自己的兴趣对这些推送页面提供反馈信息（将符合需求的页面反馈回个人代理系统服务器）。见表 1 的 II 部分
3. 在处理了反馈信息之后，重新分流 10000 到 20000 之间的页面。见表 1 的 III 部分。
4. 用时情况见表 2。

表中“站点”表示推送站点数、“页面”表示推送页面数、“符合”表示符合兴趣页面数、“反馈”表示用户反馈的页面数。其中系统向用户推送前 n 个相关度最高的输出，实验中选取 n 为 40。四种需求分别是：研究生教育、局域网、信息检索、人工智能。

兴趣	I			II			III		
	站点	页面	符合	站点	页面	反馈	站点	页面	符合
研究生教育	20	40	15	12	40	12	19	40	19
局域网	16	40	21	11	40	13	19	40	27
信息检索	17	40	12	13	40	4	16	40	18
人工智能	15	40	3	11	40	3	16	40	4

表 1 试验结果

条目	数目	时间（秒）
处理反馈页面	3 1	6
计算页面与用户兴趣描述相关度	7 3 2	2

表 2 系统一些处理的时间开销

5. 结果分析及进一步的工作

系统使用了判定树来组织用户的信息需求，通过简单的 HTML 分析，特征项权值计算关键词页面内局部密度加权和阈值过滤，并且过滤评分教底的页面，保证了向用户推送的页面内容的正确性，达到了比信息检索系统高的信息准确率。通过相关反馈进一步提高了推送信息的准确率。从速度和服务性能上达到了较好程度。

然而，系统仍然存在着一些不足：系统仍然主要依赖于简单信息检索技术。判定树总体上属于布尔判定，虽然增加了 HTML 分析，文本内局部权重，文本内词间距加权评分，但评分的核心仍然是文本内词频。由于文本内词频与文本内容之间存在关联的误差，会造成部分推送内容的不准确。

如果使用 TF/IDF 来对文本进行相关性评分，结果会更合理一些。但是，由于文档频数是与语料相关的，而系统只有在对当前源信息分流完毕之后，才能得到全局分布信息，造成这一信息事实上不可用。

反馈功能的引入可以使系统的推送结果更加符合用户个性化的需要，增加了系统的用户满意度。但是，大量反馈信息的处理和页面与个性化向量的计算必将占用许多时间。同时，个性化向量的获得主要是根据反馈页面的词频信息来取舍的，必将在个性化向量中引入不相关的成分。

目前，我们希望将基于概念词典的研究作为系统改进的一个重要方向。概念词典可以改进传统的基于统计分析的方法的不足，将基于内容分析有效地引入信息服务系统中来。其他的改进方法有待进一步研究。

参考文献

- [BS98] Eric W. Brown, Alan F. Smeaton, "Hypertext Information Retrieval for the web", *SIGIR'98 on Hypertext Information Retrieval for the web*, pp8-13;
- [程 98] 程学旗等, "一种基于判定树的信息分流机制", *ICCIP'98*;
- [郭 99] 郭宏蕾, "WWW 信息智能检索技术研究", *北京航空航天大学博士后研究报告*, 1999. 5;
- [JFM97] Thorsten Joachims, Dayne Freitag, Tom Mitchell, "Web Watcher: A Tour Guide for the World Wide Web", *IJCAI-97*, pp770-777, NAGOYA, JAPAN, August 23-29, 1997;
- [Kir98] Steve Kirsch, "Infoseek's experience searching the internet", *SIGIR'98 on Hypertext Information Retrieval for the web*, SIGIR Forum, Fall 1998;
- [LK95] Lang, K. "NewsWeeder: Learning to Filter Netnews", *Proceedings of the 12th International Machine Conference (ML95)*, pp331-339, San Francisco, CA: Morgan Kaufmann;
- [PMB96] Michael Pazzani, Jack Muramatsu, Daniel Billsus, "Syskill & Webert: Identifying interesting web sites", *AAAI 96*, pp56-61, August 4-8, 1996;
- [Sod97] Stephen Soderland, "Learning to Extract Text-based Information from the World Wide Web", *AAAI 97*.