智能信息Agent的原理和实现方法

陈红英 李卫华 (广州广东工业大学计算机学院 510090)

摘要：本文介绍一个采用了多种机器学习技术的智能搜索Agent的原理和实现方法。

关键词：信息过滤 最大匹配法 中文分词 加权向量空间法 兴趣学习 ID3算法

1 前言

随着Internet的发展，网络上信息量增加，智能代理技术已成为计算机研究领域的一个崭新的课题，传统的网络搜索引擎例如Alta Vista、Excite、WebCrawler等引擎基本上不具有智能，其缺点如下：首先查询的准确性不高，返回的结果成千上万，使用户难于寻找到自己真正喜欢的信息。其次，它们不能主动从网络上发现和收集用户需要的信息，用户要查询同样的兴趣，只能再次搜索，以获得最新的网页的内容。浪费了用户大量的时间。探索智能化、知识化的搜索引擎已经成为十分迫切的要求。Agent原意是代理人，Michael Wooldridge对其下了一个定义：Agent是一个能够根据它对其环境的感知控制其自身的决策和行为的程序。Internet智能Agent,就是以 Internet这一规模庞大具有极高的异质、动态的软件环境为研究基础，提供对Internet网络信息的收集、搜索、分析、综合等高度体现智能行为的信息处理手段为目的的软件。它充分利用人工智能的技术成果，开拓了Internet方面的应用。目前介绍智能搜索引擎方面的文章也很多，但几乎都是针对某个具体的问题进行探讨，本文全面介绍了一个智能Agent的各个组成部分和实现思路，使读者对什么是智能搜索Agent，怎样实现一个智能搜索Agent有一个系统的了解。

2 Agent系统的组成

(1) 我们的Agent系统主要由三个子系统组成：信息搜索子系统、信息过滤子系统、兴趣学习子系统。

信息搜索子系统的目标是：

① 能够在较短的时间内，在指定的范围内搜寻所需信息。要求信息全面，速度快。

② 对其所覆盖的资料进行自动更新，对得到的数据进行加工处理。

兴趣过滤子系统的目标是：

① 能快速提取所处理网页的关键字内容。

② 过滤后的网页能够确实反映用户的兴趣。

用户兴趣学习和机器学习子系统的目标：

① 使Agent能够进行自我性能调整和改善，表现出学习和自适应能力。

② 这种学习技术要面向Internet网络信息环境，对Internet提供的各种信息有较强的领悟力和适应力。

③ 系统能根据掌握的知识向用户提供智能化的信息服务。

其系统实现流程图如图1，关系图如图3。

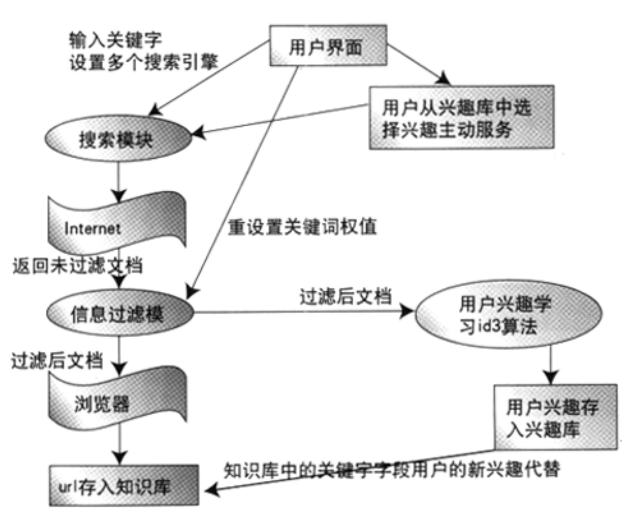

图1 系统流程图

3 所用机器学习技术

我们的系统在搜索机制、信息学习机制、过滤机制分别采用了当前最新的机器学习技术，如图 2。例如：在搜索机制中我们同时采用了组合引擎和概念检索两种策略，极大的提高了检索的查全率，这是一种较新的思路。中文分词机制和信息过滤机制中我们采用比较成熟的技术，中文分词机制我们采用最大匹配方法，它简单实用，可操作性强，同时为了提高分词效率我们对其进行了预处理，大大的提高了分词的效率。信息过滤机制中我们采用的是向量空间法，在其关键词权值方面和特征提取方面我们作了研究。在学习机制中我们综合了利用了启发式学习方法和决策树学习方法。另外我们增加了用户兴趣管理机制和主动服务机制，进一步提高了 Agent 的主动性功能，从以上的分析可见我们的Agent无论在智能方面、自学习方面和主动性服务方面都有较好的性能和突破。下面按照Agent工作的流程具体分析各个子系统的实现方法和思路：

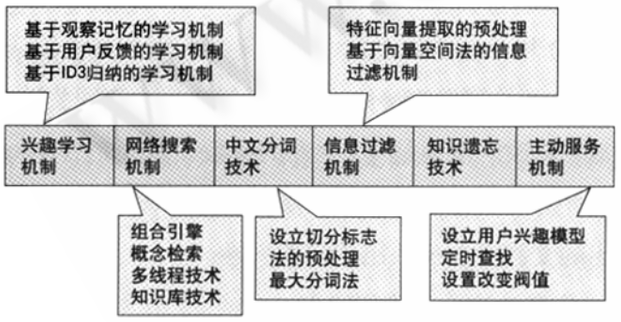

图 2 系统功能图和各部分采用的机器学习方法

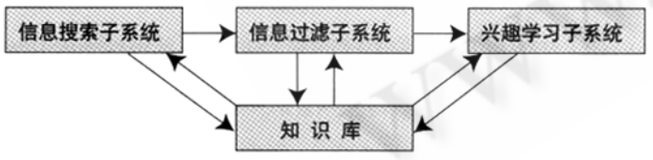

图 3 系统模块关系图

3.1 搜索子系统的实现技术

在搜索部分我们采用了如下技术：组合引擎，概念检索，多线程技术，知识库技术。采用组合引擎，概念检索技术提高了搜索的查全率，采用多线程技术和知识库技术则提高了检索的速度，为下面的信息过滤部分提供的充分的源文档。

(1) 组合引擎。现在网络上已经有很多搜索引擎，虽然它们不具有智能性，但我们可以利用它们查询，这样可充分利用各个搜索引擎的数据库，委托多个搜索引擎同时进行搜索，综合多个搜索引擎的返回结果。对用户的同一查询，不同的引擎返回的结果不同，采用组合引擎结合各引擎的返回，可以扩大检索范围，信息更加全面，更好的满足用户的需求，另外采用这种把查询工具与检索数据库分离，安置在客户端的方法，可以充分扩充客户端的信息过滤等智能性功能，大大加强返回信息的准确性和全面性。目前，Internet 上有些搜索工具（如 MetaCrawler , SavvySearch 等）采取的也是组合引擎，不过它们对各搜索引擎的返回结果没有充分的分析过滤，采用的是简单集成的方法。

(2) 概念检索。目前大多数搜索引擎都是基于关键词匹配的搜索，这种方法有它的缺陷，例如：用户想搜索“计算机”，引擎返回的都是含有“计算机”关键词的文档，但对只含有“电脑”和“微机”的文档它却放弃了。我们采用基于抽象的概念的搜索方式，采用同义词扩展方法。同义扩展采用的以知识库为基础的方法。将关键词提取它的抽象概念，在知识库中查找它的同义词，生成检索串，在上面的例子中我们将生成：“计算机 && 电脑 && 微机“这样的检索串。从而提高了访问的召回率，提高了访问的全面性。

(3) 多线程技术。我们的 agent 派生了几个线程，分别连接不同的搜索引擎，这样大大提高了检索效率，同时也防止了因为一个引擎的访问失败而导致整个访问的失败。

(4) 知识库技术。我们采用了知识库技术，对返回的页面进行管理。对由用户的兴趣返回的网页进行信息过滤，然后将其 URL 存入知识库，进行管理。这样，下次用户检索时我们可以先在本地知识库中查找，大大提高了检索效率。另外我们采用知识遗忘技术，对于每个 URL 设置了存活期和生命值，每次访问该 URL 则将它的生命值增加，否则它的生命值随时间的递减，当存活期到期时或生命值为0时，则将该知识抛弃，这样可以保证知识库的实时性，同时控制了知识库的规模，便于快速查找。

(5) 当从搜索引擎查询回来的网址还是不够多时，Agent系统可以启动自己的网络搜索算法，从现有的网址出发，利用有限区域深度广度优先算法进行搜索。

通过以上的方法我们的 Agent 在查全率和速度方面有了较大的提高，下一步是将返回的文章集首先通过中文分词后，再进行过滤，从而得到用户满意的文章。

信息搜索子系统特点：

① 查询速度快：系统优先调用其他引擎查询，只有在返回网址较少或不满足用户的要求时才使用有限深度-广度优先搜索算法在网络上搜索。

② 信息全面：采用了概念检索的方法，提高了查全率。

3.2 信息过滤机制

包括中文分词和信息过滤两部分

3.2.1 中文分词机制

由于中文和西文的区别，使中文引擎不同于西文引擎，主要区别在于分词部分，英语中词和词中间都由空格或标点符号隔开，汉语词与词则无明显的界限，这就影响了关键词的操作，(读取和匹配)，所以要正确的理解一篇文章首先必须分词。目前采用的分词方法主要有以下几种：最大匹配方法、反向最大匹配方法、逐词遍历法、设立切分标识法、最佳匹配法、有穷多层次列举法、二次扫描法、邻接约束方法、邻接知识约束方法、专家系统方法、最少分词词频选择方法、神经网络方法等等。其中词库匹配法是大势所趋，目前一些非词库的处理方法效果不能令人满意。

我们这里采用的是最大匹配法，并对其进行了预处理和改进，以提高其分词效率。方法如下：

(1) 预处理，主要是利用汉字的特点对其进行预处理，尽量在文章多设置“分词标志“，将长的汉字分成短的汉字，以便后面的分词。

具体方法是：

① 对文章进行两次扫描

· 在英文字符、标点符号、数字、其他非汉字符号的特殊字符的两侧分别插入空格，因为词语不可能跨越这些特殊符号存在。

· 利用汉语的特性，对有些字如“的”，“得”，“了”，“很”等等，这些字一般不能与其他字组成词语，所以我们把它们用空格代替。

② 对于中文词库，在使用前，将其排序，根据其中的与词对应的权值按降序排列，使其保证先从最高频的词开始匹配。这样可以使时间复杂度达到最小。

(2) 最大匹配法的思路。分词词典中的词有I个汉字组成，取汉字字符串序列中前I个汉字作为匹配字段，查分词词典，若能匹配，则将这个匹配字段切分出来，若不能匹配，则将匹配字段的最后一个词去掉，重复以上过程，直到匹配为止。

3.2.2 信息过滤

目前信息过滤的方法很多，有基于向量空间法、基于文章集的信息过滤方法、基于社会过滤的方法。我们采用向量空间法，向量空间法实现方法简单，可操作性强，其中特征提取是信息过滤智能体的重要组成部分。它实现从Internet网上的HTML文档到特征向量的转化，为用户兴趣的学习提供样本。

将文章进行中文分词后，我们采用向量空间法，提取特征值，将文章转化成一个向量，与用户兴趣的模板向量进行相似性比较，从而判断该文章是否满足用户的兴趣。为了加强向量空间法的过滤效率，为此在特征提取时我们进行了如下的改进：

(1) 含有大量URL的网页往往是一个引用页面，如果一个网页的URL超过某个阀值，则其相关度下降，甚至抛弃它。[3]

(2) 对HTML文档结构进行分析，根据关键字在文档中的作用进一步调整该关键字的权值。

例如：在标题中的关键字应增加其权值。

(3) 相关度高的文档中的关键字应增加其权值。

(4) 用户可以直接修改模板关键字串中不同关键字的权重，修改的范围是0-1之间的实数，修改完后，Agent重新计算各关键字的权值，使它们归一 [7]。这样使过滤出的文档更能反映出用户的兴趣。

信息过滤子系统的特点：

· 中文分词考虑了中文的书写习惯，使分词的效率大大提高。选择了基于词库的最大匹配法，算法实现简单有效。

· 过滤部分采用加权向量空间法，充分考虑了web文档具有元标对称的特点及各标签重要程度，改善了过滤的效果。

其原理图如图4：

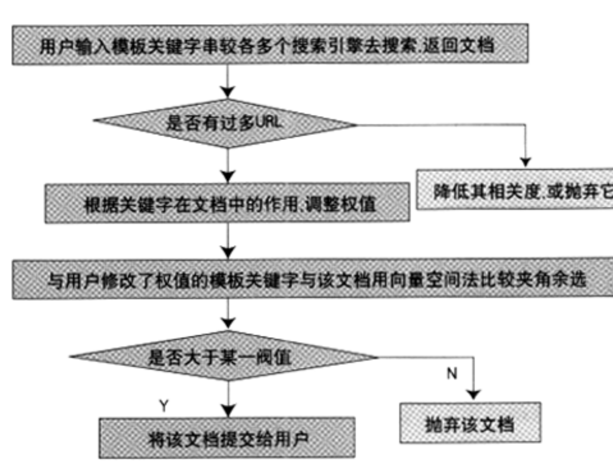

图 4 过滤算法流程图

4 兴趣学习机制

对过滤后的文档集，我们一方面可以直接返回给用户，另一方面，我们可以进一步根据用户的反馈进行用户兴趣的学习，只有学习到用户的真正兴趣，生成更加有效的关键词串，才能更有效的进行搜索。对于用户兴趣的学习我们采用启发式兴趣学习和决策树学习方法。启发式学习实现简单，可以缩短学习的周期时间，提高效率。决策树学习则可以更加精确的学习到用户的兴趣，从而提高了兴趣学习的质量。

4.1 启发式兴趣学习

观察用户在文档上停留的时间（由存取日志得到），我们可以认为如果用户在某个文档上停留的时间比较长，说明该用户对此文档感兴趣。然后通过：

(1) 提取并分析次文档的关键词，得到用户的兴趣。用户浏览一篇文章后，对该文章进行评价：满意、一般、不满意等。通判别的方法。首先找出最有判别力的因素，把数据分成多个子集，每个子集又选择最有判别力的因素进行递归划分，一直进行到所有子集仅包含同一类型的数据为止，从而得到一棵决策树。

方法是通过用户浏览网页后，对一篇网页进行评价，如果认为符合他的兴趣，则把该文章记为正例，否则记为反例，同时程序通过启发式短语提取方法，提取网页中的关键短语，提取文档关键短语后，当提取文档数超过一定数目（例10篇）时，通过改进的ID3算法学习用户的兴趣。

4.2 用户兴趣管理机制和主动服务机制

Agent为每个用户建立用户模型，用户模型纪录用户提交的信息查询任务、用户的爱好。用户可以指出某些文档或兴趣应当保持最新，选择某个URL或查询任务，由Agent去网上定时搜索（或选择系统空闲时）。

(1) 对某个具体的URL，将搜索回来的网页进行后台分析，与原文档进行分析，如果改变超过一定阀值，则提醒用户进行阅读，这样可以防止只有一点小改动就提醒用户，提高了效率。

(2) 根据某种兴趣搜索回来的网页，按信息过滤方法进行过滤后，选择相关度较高的文档，将其URL加入用户的网址库。

兴趣学习子系统的特点：

(1) 采用了启发式学习和改进的ID3算法，满足了Agent学习用户兴趣的要求。

(2) 采用用户兴趣管理机制和主动服务机制。

5 总结

我们用一组数据测试我们的搜索引擎，采用如下方法：

	相关文档A	不相关文档!A
已检索B	A ∩ B	!A ∩ B
未检索!B	A ∩!B	A ∩!B

对一组给定的文章集我们已知其中相关文档集合为A,不相关文档集合为！A，检索过滤后返回的文档集为B，则查全率=| A ∩ B|/|A|，查准率为=| A ∩ B|/|B|，通过实验证明搜索引擎查准率较高，学习效果很好。本课题由广东省自然科学基金资助，在此表示感谢。■

参考文献

1 《并行式Meta Search 系统的设计与实现》，解冲锋、 李星， 计算机工程与应用， 1999。2。

2 《个性化网上信息过滤智能体的实现》，傅忠谦、 王新跃 等，计算机应用 ，2000。3。

3 一种基于智能体的Web文档预取模式， 梁意文、 曹霞、董红斌，计算机工程与应用， 2001。4。

4 《基于信息Agent通知站点内容的有价值变化》，路海明、卢增祥、李衍达，计算机科学， 2000 vol.27.9。

5 《基于Agent技术的个性化主动信息服务》，路海明、卢增祥、徐晋晖， 计算机工程与应用 ，1999。6。

6 Moukas A. Amalthaea: Information Discovery And Filtering Using A Multiagent Evolving Ecosystem . Applied Artificial Intelligence ,1997(11):437–457。

7 Caglayan A, et al. Learn Sesame ———A Learing Agent Engine. 1997。

8 M. Balabanovic等, VISITOR–HOSTOR: Towards An Intelligent=Electronic Secretary. CIKM'94, Maryland, USA。