●晏创业 张玉峰

机器学习在获取检索知识中的应用*

摘 要 将机器学习技术引入网络信息检索系统是提高其智能性的关键。机器学习方法包括:用户知识的学习,专家经验知识学习,网页分类知识学习。图 2。参考文献 8。

关键词 智能检索 检索知识 机器学习
分类号 G354.2

ABSTRACT It is a key to apply machine learning technology for the improvement of the intelligence of network information retrieval system. Methods of machine learning include the learning of user knowledge, the learning of expert experiences and the learning of website classified knowledge. 2 figs. 8 refs.

KEY WORDS Intelligent retrieval. Retrieval knowledge. Machine learining.
CLASS NUMBER G354.2

当传统的布尔逻辑检索和空间向量检索无法有效地获取网络知识时,人们便将注意力转向智能检索系统。智能检索的主要特性在于它能够通过与周围环境的交互学到新的知识,以提高其自身解决问题的能力,即将机器学习这样一种新型的人工智能技术成功地应用于信息检索系统,是提高其智能性的关键所在。

机器学习是人工智能的一个新的分支学科,从研究人类学习行为出发,研究一些基本方法(归纳、一般化、特殊化、类比等)去认识客观世界,获取各种知识和技能,以对人类的认识规律进行探索,模拟人类的各种学习过程,借助于计算机科学和技术原理建立各种学习模型,从而为计算机系统赋予学习能力[1]。机器学习的研究始于神经元模型研究,此后又经历了符号概念获取、知识强化学习研究阶段,至今已发展到连接学习和混合型学习研究阶段。

在网络信息检索系统中,信息用户、专家(包括知识领域专家和信息检索专家)和因特网资源是其重要的处理对象。检索系统的学习任务包括:向信息用户学习他们的信息偏好,以提高信息检索的个性化;向因特网学习网络信息的特征,以实现网络信息基于内容的分类与检索;向专家学习领域知识和检索经验,以提高系统解决问题的质量和效率。对于不同类型的检索知识,必须根据其具体属性,选择相应的机器学习方法。

1 用户知识的学习

1.1 用户偏好的学习算法

包括学习用户的长期信息偏好和短期信息偏好。长期信息偏好是在长时期内逐渐形成的,而且一旦形成,就相对比较稳定。因而改变长期偏好所花的时间与其形成时间成正比。短期信息偏好是动态的,比如用户关于热点话题的兴趣差不多每天都会变。并且用户对信息的短期偏好是现实生活中非常普遍的一种现象,是实现个性化信息检索必须关注的一个方面。如果系统能同时兼顾用户的长期和短期偏好,那么既可以向用户提供一般个性化的信息,也可向用户提供他们关心的热点信息。"基于多描述器的用户偏好算法"弥补了"基于单描述器算法"的不足,是较理想的、动态的用户偏好学习算法。这里以三描述器算法为例来介绍动态的用户偏好的学习算法。

在三描述器算法中,有一个描述器 LTD(Long-term Descriptor)用于用户长期信息偏好的学习。LTD 以用户过去的信息偏好为基准,综合考虑用户新近感兴趣的和厌恶的信息类型,通过对代表相关信息文档的关键词进行加权运算,动态地调整用户长期信息偏好。

用户短期偏好学习的主要任务是追踪用户最新

*本文是国家社科项目"基于学习的智能检索机制研究"(编号:01BTQ011)的研究论文。

反馈,以反映用户的暂时偏好,为"实时"个性化信息服务提供依据。三描述器算法中除了 LTD 外,剩下的两个描述器都用于用户短期信息偏好学习,我们将其称之为 STD (Short-term Descriptor)。其中的一个用于从非相关反馈信息(正例)中学习,叫做正描述器;另一个用于从非相关反馈信息(反例)中学习,叫做负描述器。采用两个描述器,是为了更有效地利用相关和非相关文档的信息,从而使这种可变因素更多的学习算法效率更高。每个描述器被赋予一定的权值,以表示该描述器在学习中所起的作用(比如,为了强调相关反馈学习的作用,可以给正描述器赋一个更大的权值)。无论用户偏好的变化多么快,这些权值都可用来控制正负描述器之间的交互。描述器对用户新偏好学习的快慢取决于学习率和用户对新信息偏好的信任水平,这种信任水平越高,新信息偏好所占据的主动权就越大,旧信息偏好消失的速度也就越快。

用户短期偏好由 4 个值确定:STD = (POSD, NEGD, w_{POSD}, w_{NEGD}),其中,POSD 表示正描述器,NEGD 表示负描述器,w_{POSD}表示正描述器学习效果的权值,w_{NEGD}表示负描述器学习效果的值。可以得出用户对文档 D 的偏好度:

$$I_{STD(D)} = w_{POSD} \times SIM(POSD, D) - w_{NEGD} \times SIM(NEGD, D);$$

$SIM(POSD, D)$ 和 $SIM(NEGD, D)$ 运算表示取相似值。

至此,我们可以综合用户的长期和短期偏好,得出用户信息偏好的逻辑表示:

$$ITDR(D) = \eta I_{LTD(D)} + (1 - \eta) I_{STD(D)} \quad (0 \leq \eta \leq 1)。$$

1.2　基于三个描述器的用户知识学习模式

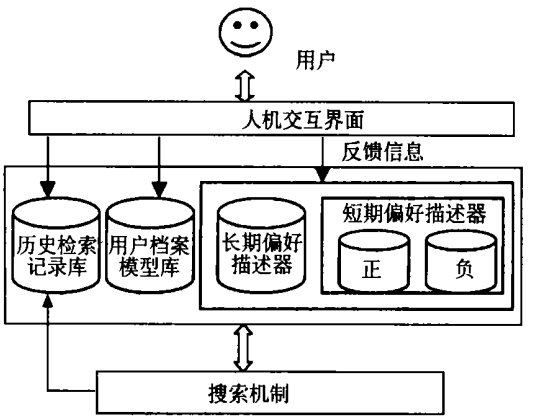

图 1　基于三描述器算法的学习模式

在对用户模型学习方法和任务分析的基础上,我们可以构建一个基于三描述器算法的用户知识学习模式(如图 1)。其中,历史记录库中的内容直接来源于检索过程;用户档案模型库中的内容通过用户填写在线表单的方式获取;用户信息偏好学习的内容来源于用户的信息反馈。3 个子模块相互协作共同实现用户知识的学习,为构建用户模型打下基础。

1.3　用户模型

用户模型是对用户知识进行归纳抽象的产物,是实现个性化信息检索的重要部件。用户模型的设计主要基于这样一种思想:若系统能准确地生成用户代表,那么它就能比较有把握地预言用户的行为和信息需求,从而使信息检索变得更为有效[2]。

从对用户信息偏好模型化的方式来看,主要有两种类型的用户模型:一类是面向检索式提炼的传统用户模型;另一类是具有描述性用户界面的交互式模型,该类用户模型非常关注用户搜索信息的过程。

著名的信息学家 Hearst 将第一类用户模型描述为一个反复提炼检索式的过程。即用户将自己的信息需求以表达式的方式提交给信息检索系统,系统根据表达式执行检索,然后用户审查检索结果,并结合自己的信息需求修改表达式,直至最终的检索结果令用户满意为止。这种传统的用户模型主要侧重于从系统中获取检索结果,而忽略了与用户之间的交互,或者说这种交互是低级的、被动的。

第二类用户模型非常注重用户与系统之间的交互,并力图在交互过程中学习用户的相关知识。加强用户与信息系统交流的主要原因在于用户信息需求的易变性,以及真正能满足用户信息需求的是用户获取若干有用信息的过程,而非一个孤立的文档集。O'day 和 Jeffries 认为:搜索的主要价值在于信息和经验的积累,而非最后得到的文档集。在信息检索系统中加强用户模型功能的建设,就是要强调用户的相关信息在检索中的重要作用。

2　专家经验知识学习

专家经验知识是人们在长期的生产实践中积累起来的财富,是领域知识与具体问题的解决方案相结合的产物,应该被人们共享。具体说来,使用专家经验知识主要有这样一些好处[3]:

(1)系统能够在无人监督的情况下,高效、准确、迅速地工作。

(2) 使人类专家的领域知识突破时间和空间的限制,专家系统程序可永久保存,并可复制任意多的副本或在网上供不同地区或不同部门的人们使用。

(3) 能够带来巨大的经济效益和社会效益。

从人们意识到专家经验知识的价值开始,人们一直为寻找理想的专家经验知识获取途径而努力,并先后尝试了手工获取、智能知识编辑和人工神经网络学习等多种方法。其中,借助神经网络可使系统进行自组织、自学习,不断地充实、丰富专家系统中原有的知识库,从而使知识获取问题得到很好解决。同时,可以采用归纳学习或者基于解释的机器学习方法,对通过神经网络获取的知识进行处理,并对学习的效果进行检测。

2.1　专家经验知识学习的一般步骤

虽然采取不同方法获取专家经验知识的具体操作不同,但从最初专家经验信息的输入,到最终专家经验知识的生成,信息流经历的主要环节大体一致。

(1) 从知识源抽取知识。抽取知识是指对知识元素的识别、理解、筛选、归纳等综合处理过程。因为事实性知识直接可以用文字的形式表示出来,所以对它的抽取相对比较容易,就是由知识工程师按照一定的领域规则和系统规则经过一定的分析、概括和筛选即可抽取出来。而启发性知识是抽象的、隐含的,必须在与领域专家交互的过程中才能得到。

(2) 知识输入。将从知识源中抽取的知识按照计算机可以识别的某种方法表示出来,然后输入系统的临时知识库即完成了知识输入工作。目前知识输入主要有两种方法:其一是利用计算机系统提供的编辑软件;其二是利用专门编制的知识编辑系统。

(3) 知识的融合学习。临时知识库中的知识是没有经过任何组织的"毛知识",要达到可以利用的目的,还要经过一定的学习推理,使其按照所解决问题的属性、规则等形成一系列的知识模型。目前用于知识学习的方法主要有:归纳学习 ——通过归纳推理,概括产生一般性的结论;解释学习 ——通过对具体事例的演绎推理,得到解决一般问题的知识模型;神经网络学习 ——通过学习、联想和自组织,形成知识模型。

(4) 专家经验知识的验证。对专家经验知识进行检验,首先因为它是专家经验信息经过一系列的自动化、半自动化处理而形成的,在这个过程中难免会有错;其次,系统对专家经验知识的要求非常高,是以专家经验知识的正确性为最低限度的;更重要的是专家经验知识经推广后,应该具有普遍意义,在正式投入应用之前,必须对其解决具体问题的能力作一般性验证。

2.2　专家经验知识学习模式

专家经验知识的获取,需综合应用多种学习方法,从多种角度实现复合学习。这里给出以人工神经网络为基础的专家经验知识学习模式,如图 2 所示。

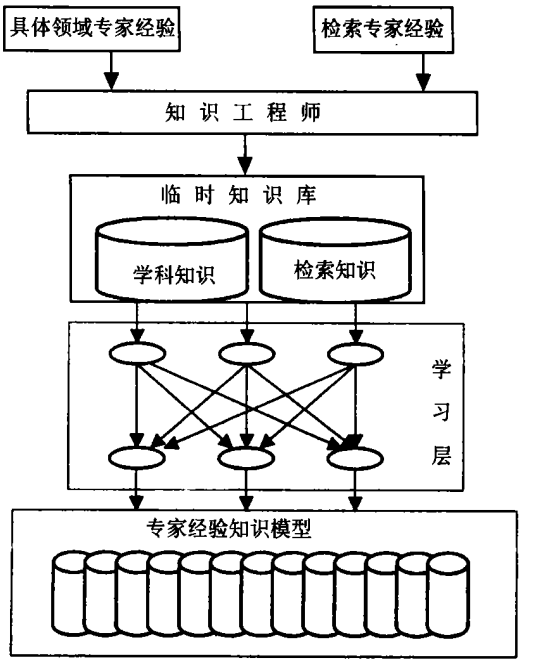

图 2　专家经验知识学习模式

以专家经验知识为基础的智能检索系统也可被看作一种知识系统,它不同于一般的专家系统,主要是因为网络信息检索是一个交叉领域,它涉及到计算机、信息管理等多领域的知识。而且根据网络信息类型的不同,与各具体的学科领域有关,所以智能检索系统需要获取多学科领域的专家经验知识。

智能检索系统的实现,依赖于检索专家的经验知识与应用领域专家经验知识的结合,即在系统中要形成融合了各方专家经验知识的知识模型。从给出的模式图可以看出,系统从知识工程师那里获取

的临时知识是零散的知识单元(学科知识和检索知识相分离),经人工神经网络学习处理后,二者有机地结合起来,形成专家经验知识模型。

3　网页分类知识的学习

从网络技术发展的初期至今,人们就网络信息分类问题提出多项解决方案。比如,在早期的搜索引擎中,人们利用标引器浏览整个网页,计算每个词在网页中出现的频率,将这些词按频率从大到小的顺序排列,并用出现频率最高的词标引网页。随后,为了提高查全率,人们借鉴字索引的基本思想,仅将网页中起语法作用的公用词滤去,把剩下的所有词都作为索引项。随着统计和空间向量等技术在网络信息检索领域的应用,网络信息分类技术朝多极化方向发展:Apte 用决策树技术来构造分类器[4],Yang 采用了一种基于余弦距离计算的邻近算法进行分类[5],Lewis 构造了一个线性分类器[6],Cohen 设计了一种建立在权值更新基础上休眠专家算法[7]。总的来说,不断更新的网络信息分类技术在提高分类质量方面确实起到很大的作用。但与此同时,网络信息数量的剧增又给网络信息分类提出更高的要求,人们期待着一种高效、智能的分类方案。机器归纳学习方法利用归纳推理向实例学习,不断提高其自身解决问题的能力,受到人们的普遍关注。

机器归纳学习以训练实例集作为自己学习的依据,如何选择训练集是学习的关键。在网络信息分类中应用机器归纳学习,其训练集有两种设置方法:人们根据一些基本的分类规则,人为地设置训练集;学习机制自动选择训练集。在第 1 种方法中,人的主观因素起了很大的作用,而且学习机制缺乏自我更新训练集的能力,一定程度上影响了学习效果。在第 2 种方法中,学习机制根据网络"原信息",自动选择训练集,并能在工作中不断更新训练集,提高分类的质量,具有很高的智能性。

以机器归纳学习为基础,带有优秀训练集的相关词库分类规则,例如,REPGER(RElevant term Pool with Good training Example classifying Rule),是一种新的在线网络信息分类学习算法。REPGER 有 3 个特点:带有一个具有很强预测能力的词库;采用了新的机制以选择好的训练例子;能动态学习,以确定更新训练集的合适阀值。REPGER 最初根据权威的分类体系确立知识类别,并结合应用领域为每个知识类赋予原始训练实例集。对于一个待分类的网页,我们不仅要根据对它的预测相关度来确定其所属类别,还要根据网页的新特征(这些特征描述该类事物更具有代表性)判断它是否有被作为新的训练例子的价值。

4　结束语

采用机器学习技术能大大提高信息检索的智能性,这点是毋庸置疑的。然而,网络信息检索强调的是一种系统功能,即并非将检索技术和机器学习技术简单地相加就可达到预想的功能。必须将二者有机地结合起来,即根据具体的知识类型采用合适的多种学习方法。机器学习等人工智能技术与信息检索技术的融合,是实现基于知识的智能检索、面向用户的个性化主动性检索的有效途径。

参考文献

1　王万森. 人工智能原理及其应用. 北京:电子工业出版社,2000
2　张玉峰,晏创业. 基于 Agent 的个性化信息服务模型研究. 情报学报,2001(5)
3　张玉峰. 智能情报系统. 武汉:武汉大学出版社,1991
4　Apte c, Damerau F, Weiss S. Automated learning of decision rules for text categorization. ACM Transactions on Information System, 1994,12(3)
5　Yang Y. Expert network: Effective and efficient learning from human decision in text categorization and retrieval. Proc Seventeenth International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin,1994
6　Lewis D D, Schapone R E, Callan J P, Papka R. Training algorithms for linear text classifiers. Proc Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, 1996
7　Cohen W W, Singer Y. Context-sensitive Learning Methods for Text Categorization. Proc Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, 1996
8　范焱,郑诚等. 用 Naive Bayes 方法协调分类 Web 网页. 软件学报,2001(9)

晏创业　北京大学信息管理系博士生。通讯地址:北京。邮编 100871。
张玉峰　武汉大学信息管理学院教授。通讯地址:武汉。邮编 430072。

(来稿时间:2002-07-12)