基于多 Agent 协作的自动分类知识库研究

李　萌　孙济庆

(华东理工大学科技信息研究所　上海　200237)(华东理工大学图书馆　上海　200237)

摘　要　提出一种基于多 Agent 协作架构的自动分类知识库更新思路，通过多 Agent 协作新文档与已有训练规则的匹配，有效地进行新类别的自动扩展和新分类规则的自动生成，同时为训练集的频繁维护问题提出了新的解决方案。

关键词　文本分类　多 Agent　分类知识库

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-8095(2009)05-0089-03

文本自动分类技术是利用计算机对文本集的知识特征按照一定的分类体系或标准进行自动分类，对具有相同知识特征的文本附以相同的类别标记。文本自动分类一般包括两个基本过程，即分类知识训练和新文献分类预测。分类知识训练是指应用分类算法从一个包含一定数量的已经进行分类标记的文献集合(即训练集)中获得分类知识;新文献分类预测则是运用训练得到的分类知识对训练集之外的文献进行分类预测。在文本自动分类中,训练集是相对固定的,一般不可能代表所有新文献包含的主题。人们一般采取定期或不定期更新训练集，再重新训练分类器的方法来解决这一问题。但是,这种方法不仅会增加训练时间，而且会造成分类知识的不一致性,从而导致整个分类系统中文献分类不一致,影响系统的检索性能。

本文结合当前新类别自动扩展的问题，采用多 Agent 技术,提出构建具有动态扩展功能的自动分类知识库系统。通过建立多个具有较高智能和自治能力的 Agent,在 Agent 之间通过共享资源、相互协作、相互服务,共同完成分类知识库更新任务,使得整个系统能够充分利用未标记文献,加强机器学习效果,逐步实现分类树的生长及更新。

1　自动分类知识库

1.1　自动分类知识库的构建

自动分类知识库的构建是自动分类的基础。分类知识库的构建实际上就是构建类特征词、类目、归属度算法。通过适当的相关度度量,形成特征词、类目、归属度三元组作为自动分类的基础和核心。用分类号控制类特征词,用类特征词表达类知识,实现三者的一对一、一对多的对应转换,并通过相应的算法确定文本的最终类别。在三元组构建的过程中,主要通过相关度的度量来确定关键词和类目之间的归属度。任何标引本质上都是一种知识概念的标识系统。由于标识及其组织方式的不同，于是形成了分别采用号码标识、受控语词标识、非控语词标识的分类语言、主题语言及自然语言。因此,以自然语言为基础的自动分类知识库存在着隐含的概念对应关系。通过等值对应、近似对应、包容对应等措施,就可使这种隐含的对应关系显现出来，实现与分类号之间互相控制和转换。

1.2　分类过程中的阈值

在给定待分类文献中，分别比较其知识特征与候选类别中每个类的相似度，并为其建立相对应的阈值,如果该文献的相似度大于相应阈值,该文献就归属于该类,否则就不属于该类。这种归类方法阈值的确定十分重要。有关相关度度量的方法有多种,如互信息、系数方法、Dice 系数、Cosine 系数、Jaccard 系数、开方统计、极大似然比估计等。这几种度量方法在不同的环境下各有优缺点。本文将最优截尾法进行简化,假定每个类相应阈值为相同值,计算出文档与各类别匹配的最大归属度,从而进行分析。

首先进行归属度计算,提取文本特征,用筛选后的(特征词、类目、归属度)三元组来构造特征词—类目矩阵 K-C(W_1,W_2,…,W_i),以此表示各特征词在每个类目中不同的归属度。其中 W_i 为特征词 i 在各类目中的归属度。$W_i=(V_{i1},V_{i2},\cdots,V_{ij})$,$V_{ij}$ 表示特征词 i 在类目 j 中的归属度。其次,将经过特征提取之后形成的若干个特征词,逐一在矩阵中匹配,并将匹配到的若干个向量存放到临时矩阵 TEMP 中。匹配结束后对临时矩阵 TEMP 进行归并，取最大值 V_{max} 作为分类的依据,将其与类别阈值进行比较,当其大于或等于阈值时则可判定该文档归属于某一类别。

$$V_{max}=\max\left(\sum_{i=1}^{i}W_i\right)=\max\left(\sum_{i=1}^{i}\sum_{j=1}^{j}V_{ij}\right)$$

其中,i 为矩阵中词的个数，j 为矩阵中类目的个数。

1.3　自动分类知识库的更新

在层次型文本分类(HTC)中,其类别分布呈树状结构(见图 1)。其中每个类别都可以看作是树上的一个节点，每个类别的上位类是其代表节点的父

收稿日期:2008-07-05

作者简介:李萌(1982—),女,情报学 2006 级硕士研究生。

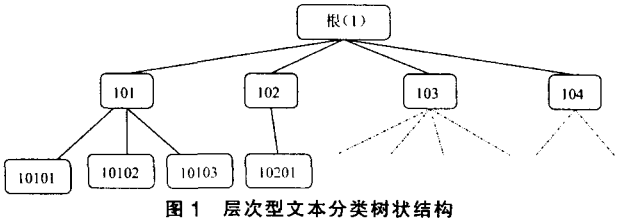

图 1　层次型文本分类树状结构

节点,下位类是其代表节点的子节点,同位类是指代表节点的兄弟节点。通常而言,HTC 采用一种基于层次的自顶向下的策略对文档进行分类。这种分类方式首先判断一个文档是否属于一个分类树的根节点类别,如果属于,接下来判断是否属于分类树中的一个或多个子类别。重复该过程,直到这个文档不能被归属到任何的下一层子节点类别或到达叶子类别。

基于多 Agent 协作的自动分类知识库系统设定,当特征匹配结果值 u(取 V_{max})大于某一阈值(设为 u_0) 时,认为该文本属于某一类别。若某文档的 V_{max} 小于阈值,则无法将该文档归属到底层叶子类别,即无法将该文本归属至某一类别,此时将对该未标记文献的文本特征进行分析处理,就需将该文档归为最相近的父节点,提取特征项并将其特征暂存,累积到数据库中。例如,某一文本:

当 $u_{101}>u_0$ 时,文档可归为 101 大类;继续判断,若 u_{10101}、u_{10102}、u_{10103} 均小于 u_0,则无法将该文本归属至 101 下层某一节点。对该文本进行特征词记忆,当该类文本的特征累积量达到某一阈值时,则生长出 101 的新的子节点,10101、10102、10103 的兄弟节点 10104 。将待分类文档存入数据库进行特征累积,从而最终实现知识库的更新。

2　系统设计

2.1　系统构成

基于 Agent 的自动分类知识库系统结构如图 2 所示,具有三大主体功能,即分类知识库搭建、分类特征构建和分类类表自动更新。系统主要由预处理 Agent、信息搜集 Agent、特征提取 Agent、特征匹配 Agent、类别扩展 Agent、词典扩展 Agent、模块调度 Agent 等几个智能体相互作用构成。

分类知识库是由通过信息搜集 Agent 在网络中搜集相关训练文档,经预处理、特征提取等 Agent 处理后形成自动分类的类特征三元组构成的。

当有新的分类知识特征时,由词典扩展 Agent 发出协作请求,信息搜集 Agent 搜集类似文档,然后进入系统先由预处理 Agent 对文本进行分词和词性标注等进行前期处理,建立词汇表,并过滤掉虚词及统计得出的高频禁用词。文档经预处理 Agent 处理后,特征提取 Agent 对该内容词表运用相应的算法进行选取,并通过统计词频,计算出权重,获得文本的特征词表。然后交由特征匹配 Agent,构造关键词—类目矩阵,计算各关键词与类别的归属度,取出

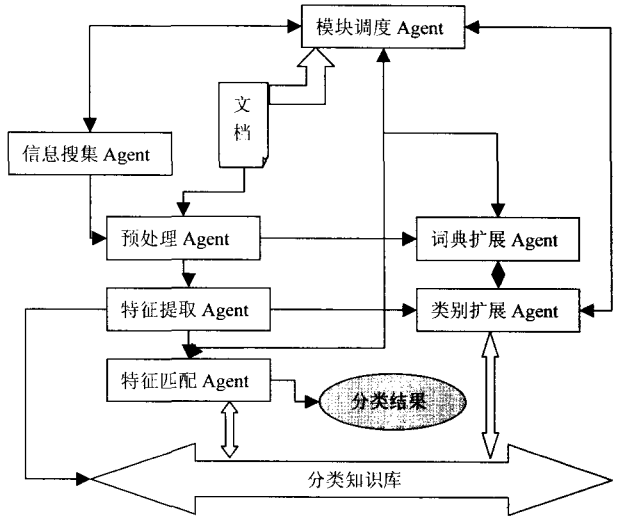

图 2　基于多 Agent 协作的自动分类知识库结构图

最大值 V_{max},进行阈值大小比较,形成分类结果。如果待分类文档特征提取后不能在已有知识库中找到匹配的叶子类别,则特征匹配 Agent 将会把相关信息传递给调度 Agent,由调度 Agent 将该文档的特征项传送给类别扩展 Agent 进行相关分析,与分类知识库进行交互,将该待分类文档存入语料库中进行文档备案及机器学习。若此时语料库中该类别的特征累积尚未达到阈值,则特征匹配 Agent 将该待分类文档相匹配的父类别反馈为分类结果。在两者之间有一个黑板结构的公共数据区用来存放待分现象、特征样本及预分结论,实现数据共享,供各 Agent 调用。若此时语料库中该类别的特征累积已达到相应阈值,分类知识库就在类表该父类别下生长出新的叶子类别,由类别扩展 Agent 将该新类别反馈给模块调度 Agent,再由模块调度 Agent 传送至特征匹配 Agent 得到新的分类结果,从而实现分类知识库的更新。

2.2　类别扩展功能

知识查询及操作语言(KQML)是目前最主要的 Agent 通信语言之一。KQML 分为内容层、消息层和通信层。内容层说明消息的内容、知识含义;消息层定义消息格式、语言行为类型;通信层定义通信体制、通信参数。由 KQML 承担各 Agent 之间的通信与协作。当一篇新文献在进行分类预测时,它与所有类别的匹配度均小于系统规定阈值时,新类别就应该自动产生。但是,新文献分类失败的原因可能是由分类异常或该文献的个体异常等造成。因此本文定义:当某类未分类文献特征累积值 w 达到某一程度 (设阈值为 W_0)时,则产生新类别及其相应的分类规则。另外,新类别的增加,容易引起子父节点之间隶属关系的不一致性。增加新类别时要尽可能地减少子父节点之间的差异,而且不能影响同级路径上其他节点与子父节点之间的关系。因而需要由多个 Agent

相互协作才能完成。

分类及类别扩展模型运行如下:

输入:新文献及其关键词

输出:新类别和分类规则

```
Do {
计算新文献与各类别的归属度 //特征匹配 Agent
找到与新文献归属度最大的类别，定位节点 L //特征匹配 Agent
}
While(u>=u0) //将新文献逐层归类直至无法进行下去
If(L 为叶节点)
{得到新文献最终分类结果为 L,输出分类结果
}
Else
{定位文献的父类别为 L
将文献导入数据库进行特征累积 //类别扩展 Agent
  If (w>=W0)
  {产生 L 新的子节点,即新类别
  生成分类规则 //类别扩展 Agent
  }
}
Return 新节点及其相应分类规则
```

从该模型可以看出,当一篇文献输入后,根据判定条件，决定是否产生新类别。如果需要创建新类别,则必须同时为该类别生成分类规则。当一篇文献加入时,产生了新的类别,此时将该文献的关键词作为分类规则的条件属性,而将该类别作为决策属性,生成一条新的分类规则。之后,如果有新文献加入该类别,需要根据每个新文献的特征项,生成相应的分类规则。

总之，理论分析和测试结果表明系统具有较好的分类知识库构建效果。笔者下一步将在本系统的基础上,深入地结合分类知识库的构建、自然语言处理等理论知识,尝试其他不同的算法,进一步提高分类效率和分类精度。

参考文献

1　张雪英. 文本自动分类中的动态类别扩展研究. 计算机应用研究,2007,24(5)

2　刘壁松,李春平. 一个可扩展的文本分类系统的设计与实现. 计算机工程与应用,2004(30)

3　杨为民,李龙澍. 基于 Agent 的文本分类系统. 计算机技术与发展,2007,17(2)

4　侯汉清,薛鹏军. 中文信息自动分类用知识库的设计与构建. 情报学报,2003,22(6)

5　苏金树. 基于机器学习的文本分类技术研究进展. 软件学报,2006(9)

(责任编辑:黄　建)