

自动分类技术的发展与展望

[作者] 杨建良, 王永成

[单位] 上海交通大学计算机科学与工程系

[摘要] 随着信息化浪潮席卷全球,手工分类索引已经不适用于大规模信息的处理了,自动分类的研究随着时代的需要蓬勃发展了起来。本文首先介绍了自动分类技术的背景和历史发展,然后着重介绍了目前应用最广泛的几种自动分类方法;其后,结合实际研究课题介绍了基于仿人思想的自动分类方法,并对自动分类技术未来发展的方向作了展望。

[关键词] 自动分类, 信息检索

1、背景介绍

自上个世纪 80 年代以来,信息化的浪潮席卷全球,信息技术迅速地渗透到社会经济的各个领域。信息的来源是多方面的,比如报纸、电视、广播等等。近几年来,随着 Internet 的普及和网络技术的不断完善,Internet 已经成为了全球最庞大最丰富的信息资源库。由于 Internet 的开放性,各类信息都能在第一时间发布在 Internet 上。然而,Internet 的这种开放性也导致了 Internet 上信息的杂乱性和冗余性。因此,自动分类技术随着时代的需求而蓬勃发展了起来。作为一种有效的信息处理方法,自动分类技术将各类信息按照一定的分类体系进行分类整理,从而大大提高了用户搜集情报的效率。

自动分类技术是在手工分类技术的基础上发展起来的。传统的信息手工分类技术已经相当成熟,但却不适用于对 Internet 上时刻更新的信息进行处理。因为它不具有实时性,另外查全率和分类的一致性也受到一定的制约[1]。世界著名搜索引擎 Yahoo 长期以来集中了大量人力进行手工分类,并且曾经因此获得了巨大的成功,但这种成功的背后已潜伏着落后的危机。最近,Yahoo 宣布同 Google 合作,开发自动分类技术以取代手工分类——自动分类技术已经成为大势所趋。

2、自动分类历史

自动分类技术的研究始于 20 世纪 50 年代末,IBM 公司的 H. P. Luhn 在这一领域进行了开创性的研究。1960 年,Maron 在 Journal of ACM 上发表了有关自动分类的第一篇论文 On Relevance, Probabilistic Indexing and Information Retrieval,随后许多著名的情报学家如 K. Sparch、G. Salton 及 R. M. Needham 等都在这一领域进行了卓有成效的研究。相对于国外的情况,我国开展自动分类的研究起步稍晚一些。80 年代中期开始,我国的一些大学、图书馆和文献工作单位开展了档案、文献或图书的辅助或自动分类研究,并陆续研制出一批计算机辅助分类系统和自动分类系统,这些系统主要集中在中文处理领域。

3、自动分类算法简介

目前,世界上已研究出多种具有一定效率的自动分类算法,而以下 4 种算法是最常用的:

(1) KNN 法(K-Nearest Neighbor)

KNN 法即 K 最近邻法,最初由 Cover 和 Hart 于 1968 年提出的,是一个理论上比较成熟的方法。该方法的思路非常简单直观:如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

KNN 方法虽然从原理上也依赖于极限定理,但在类别决策时,只与极少量的相邻样本有关。因此,采用这种方法可以较好地避免样本的不平衡问题[1]。另外,由于 KNN 方法主要靠周围有限的邻近的样本,而不是靠判别类域的方法来确定所属类别的,因此对于类域的交叉或重叠较多的待分样本集来说,KNN 方法较其他方法更为适合。

该方法的不足之处是计算量较大,因为对每一个待分类的文本都要计算它到全体已知样本的距离,才能求得它的 K 个最近邻点。目前常用的解决方法是事先对已知样本点进行剪辑,事先去除对分类作用不大的样本。另外还有一种 Reverse KNN 法,能降低 KNN 算法的计算复杂度,提高分类的效率。

该算法比较适用于样本容量比较大的类域的自动分类,而那些样本容量较小的类域采用这种算法比较容易产生误分。

(2) SVM 法

SVM 法即支持向量机(Support Vector Machine)法,由 Vapnik 等人于 1995 年提出,具有相对优良的性能指标。该方法是建立在统计学习理论基础上的机器学习方法。通过学习算法,SVM 可以自动寻找出那些对分类有较好区分能力的支持向量,由此构造出的分类器可以最大化类与类的间隔,因而有较好的适应能力和较高的分准率。该方法只需要由各类域的边界样本的类别来决定最后的分类结果。

从图我们可以更直观地了解 SVM 算法:图中的“+”和“-”表示空间中样本,点划线表示的是类域的边界,类域边界上的样本对最终的决策面的确定起着决定性的作用,称之为支持向量(SV)。支持向量机算法的目的在于寻找一个超平面 $H(d)$,该超平面可以将训练集中的数据分开,且与类域边界的沿垂直于该超平面方向的距离最大,故 SVM 法亦被称为最大边缘(maximum margin)算法[2]。

从对图 1 的分析可以发现,待分样本集中的大部分样本不是支持向量,移去或者减少这些样本对分类结果没有影响[3]。因此,SVM 法对小样本情况下的自动分类有着较好的分类结果。

(3) VSM 法

VSM 法即向量空间模型(Vector Space Model)法,由 Salton 等人于 60 年代末提出。这是最早也是最出名的信息检索方面的数学模型。其基本思想是将文档表示为加权的特征向量: $D=D(T_1, W_1; T_2, W_2; \dots; T_n, W_n)$,然后通过计算文本相似度的方法来确定待分样本的类别。当文本被表示为空间向量模型的时候,文本的相似度就可以借助特征向量之间的内积来表示。

在实际应用中,VSM 法一般事先依据语料库中的训练样本和分类体系建立类别向量空间。当需要对一篇待分样本进行分类的时候,只需要计算待分样本和每一个类别向量的相似度即内积,然后选取相似度最大的类别作为该待分样本所对应的类别。

由于 VSM 法中需要事先计算类别的空间向量,而该空间向量的建立又很大程度的依赖于该类别向量中所包含的特征项。根据研究发现,类别中所包含的非零特征项越多,其包含的每个特征项对于类别的表达能力的越弱。因此,VSM 法相对其他分类方法而言,更适合于专业文献的分类。

(4) Bayes 法

Bayes 法是一种在已知先验概率与类条件概率的情况下的模式分类方法,待分样本的分类结果取决于各类域中样本的全体。

设训练样本集分为 M 类, 记为 $C = \{c_1, \dots, c_i, \dots, c_M\}$, 每类的先验概率为 $P(c_i)$, $i = 1, 2, \dots, M$ 。当样本集非常大时, 可以认为 $P(c_i) = c_i$ 类样本数/总样本数。对于一个待分样本 X , 其归于 c_j 类的类条件概率是 $P(X/c_i)$, 则根据 Bayes 定理, 可得到 c_j 类的后验概率 $P(c_i/X)$:

$$P(c_i/X) = P(X/c_i) \cdot P(c_i) / P(X) \quad (\text{式 1-1})$$

若 $P(c_i/X) = \max_j P(c_j/X)$, $i = 1, 2, \dots, M$, $j = 1, 2, \dots, M$, 则有 $x \in c_i$ (式 1-2)

式(1-2)是最大后验概率判决准则, 将式(1-1)代入式(1-2), 则有:

若 $P(X/c_i)P(c_i) = \max_j [P(X/c_j)P(c_j)]$, $i = 1, 2, \dots, M$, $j = 1, 2, \dots, M$, 则 $x \in c_i$

这就是常用到的 Bayes 分类判决准则。经过长期的研究, Bayes 分类方法在理论上论证得比较充分, 在应用上也是非常广泛的。

Bayes 方法的薄弱环节在于实际情况下, 类别总体的概率分布和各类样本的概率分布函数(或密度函数)常常是不知道的。为了获得它们, 就要求样本足够大。另外, Bayes 法要求表达文本的主题词相互独立, 这样的条件在实际文本中一般很难满足, 因此该方法往往在效果上难以达到理论上的最大值。

4、基于仿人思想的自动分类算法

目前, 国际上研究并投入实际应用的自动分类算法已有很多, 但没有一种算法能够对于各种类型的样本信息进行自动分类时取得较高的分类指标。产生这一情况的原因是多方面的, 但归根结底在于目前的自然语言处理技术还远远不能适应自然语言的复杂性, 这种情况在短时期内很难从根本上加以解决。

上海交通大学王永成教授在总结长期从事语言信息处理工作的基础上, 提出了“在计算机还不能完全像人类一样实现思维的情况下, 让计算机最大程度地模仿人的思维, 是实现人工智能的最有效手段”[1]。基于上述仿人思想, 并结合国际上现有的多种成熟的自动分类算法, 我们开发了一个 Internet 新闻自动分类系统。利用该系统对中国资讯网 (<http://www.chinainfobank.com/>) 提供的 10 万个样本进行了测试, 在其包含 194 个小类的分类体系框架下, 自动分类系统的查全率和查准率均达到了 95% 以上。

5、自动分类技术的展望

自动分类技术发展至今, 已经有 50 多年的历史了。目前看来, 未来自动分类系统的发展方向应该主要聚集于以下 3 个方面:

立体性

所谓立体性, 指的是文本的内容可以从不同角度或不同侧面进行考察, 从而挖掘出不同偏重的信息。具体而言, 目前的自动分类系统都是适应于某一个特定的分类体系的, 两个不同的分类体系间转换并非易事。而自动分类技术中立体性的发展目标就是要建立一个全面的分类系统, 其中可以同时包含多种分类体系, 各种分类体系之间可以方便地进行转换。

动态性

所谓动态性, 指的是分类法可以动态地随信息内容概率分布的变化进行变化, 力求分类法的树型结构是一个平衡结构, 使分类法更利于快速检索 [1]

信息的增长速度是不平衡的, 这样的不平衡包括空间上的不平衡和时间上的不平衡。特别是空间上的不平衡, 导致分类体系下某些类别所包含的信息剧增而某些类别长期以来无法得到新的信息。这样的不平衡性破坏了信息均匀分布以便于快速检索的原则。因此, 未来的

分类技术应该能够自动检测到这样的不平衡性的出现,并随着信息的不平衡增长而动态地进行调整,从而最大限度地保持类别体系的平衡性,以保证信息的快速检索。

面向用户性

所谓面向用户性,指的是分类系统的实时调节能力。不同用户有着不同的分类要求,同一用户在不同场合也可能有着不同的分类要求。因此,未来的自动分类系统应该更多的考虑增强学习功能,能够在用户的指导下对分类体系及分类法做出个性化的调整,以满足用户的需求。

6、结束语

物质、能源和信息是人类社会的三项基本资源。在信息爆炸的今天,对信息资源的挖掘和提炼显得更为重要。自动分类技术的出现大大的提高了信息的利用效率,为人类更好的利用信息资源提供了很好的一个途径,因此必将有着广阔的发展前景。

参考文献

- 1、尹中航. 网络新闻智能分类技术的研究与实现, 上海交通大学博士学位论文, 2002 年月
- 2、陶卿. 一种新的机器学习算法: Support Vector Machines, 模式识别与人工智能, 2000 年 9 月, Vol. 13, No. 3
- 3、尹中航、王永成. 应用支持向量机进行网上信息自动分类, 高技术通讯, 2000 年 11 月, 107 ~ 110