文档自动分类技术及其实现

邹涛 孙赛 (南京大学多媒体计算机研究所 210093)

摘要:文档自动分类是信息处理领域中的一项重要研究课题,也是一项重要的应用技术。本文介绍了实现文档自动分类中的几项关键技术,并给出了实现文档自动分类的一般方法。

关键词:文本分类 分类模型 VSM Naive Bayes

一、引言

随着信息技术的发展,特别是 Internet 应用的普及,人们已经从信息缺乏的时代过渡到了信息极大丰富的时代,如何自动处理大量的数字化文本为了一项重要的研究课题。

文档分类是指根据文档的内容或属性,将大量的文档归到一个或多个类别的过程。文档自动分类是一项重要的信息处理技术,在邮件分类、电子会议、信息过滤等方面得到了较为广泛的应用。例如麻省理工学院(MIT)为白宫开发的邮件分类系统就承担了白宫几乎所有的电子信件的分拣与处理工作。本文就介绍一下文档自动分类中的几项关键技术及实现文档分类系统的一般方法。

二、关键技术

文档自动分类的关键问题是如何构造一个分类函数或分类模型(也称为分类器),并利用此分类模型将未知文档映射到给定的类别空间。分类器的构造方法有多种,主要有统计方法、机器学习方法、神经网络方法等。向量空间模型(Vector Space Model, VSM)与 Naive Bayes 模型是近些年应用较多且分类效果较好的两种文档分类模型。

1. 向量空间模型

在向量空间模型 VSM 中,将文档看作为是由相互独立的词条组(T1, T2, ...; Tn)构成,对于每一词条 Ti,都根据其在文档中的重要程度赋以一定的权值 Wi,并将 T1, T2, ... Tn 看成一个 n 维坐标系中的坐标轴,W1, W2...; Wn 为对应的坐标值。这样由(T1, T2, ...; Tn)分解而得的正交词条矢量组就张成了一个文档向量空间,文档则映射成为空间中的一个点。对于所有文档和文档类都可映射到此文本向量空间,用词条矢量(T1, W1; T2, W2; ...; Tn, Wn)来表示,从而将文档类的匹配问题转化为向量空间中的向量匹配问题。假设已知文档类为 Q,未知文档为 D,两者的相似程度可用向量之间的夹角来度量,夹角越小说明相似度越高,相似度计算公式如下:

$$Sim(Q,D)=\cos(Q,D)=\frac{\sum_{k=1}^{n}W_{qk}\cdot W_{dk}}{\sqrt{\sum_{k=1}^{n}W_{qk}^2}\cdot\sqrt{\sum_{k=1}^{n}W_{dk}^2}}$$

2. Naive Bayes 模型

Naive Bayes 分类模型是一种基于概率的分类方法,虽然对文本的分类处理作了很多的简化,但 Naive Bayes 法仍然能得到较高的分类正确率。Naive Bayes 分类法是基于所有词条在文档中的出现概率是相对独立的假设之上的。假设集合 C 为文本类的集合,判断一个文档 d' 是否属于某个类别 Ci 可通过计算 P(Ci|d')的概率完成,即给定文档 d',它属于文档类 Ci 的概率是多少。Naive Bayes 法的判别原则就是将 d'指定到使 P(Ci|d')达到最大概率的 Ci 类中,即求解 arg max P(Ci|d')。P(Ci|d')可根据文档的长度进行分解:

$$P(C_i\mid d')=\sum_{l=1}^{\infty}P(C_i\mid d',l)\cdot P(l\mid d')$$

根据 Bayes 定律可得:

$$P(C_i\mid d')=\frac{P(d'\mid C_i,l')\cdot P(C_i\mid l')}{\sum_{C'\in C}P(d'\mid C',l')\cdot P(C'\mid l')}$$

3. 中文信息的处理

中文与英文不同，句子中各词条间没有分隔符(空格)，因此在利用以上分类模型进行文档分类时，需要先对中文文档进行词条切分，以划分出每个词条。文档分类系统对分词的准确度没有很高的要求，可以采用较为简单的基于词表的机械分词法。词表分词法需要建立分词词典，其分词原理就是根据分词词典中的词条来划分出文档中的所有词。在切分词条时，可先根据标点进行粗切分，然后再分别使用正向和逆向最大匹配法进行细切分。如果切分结果相同，则认为切分正确；如果不相同，则在不同之处取包含两部分的最小长度串，作为词典候补词条，这样可不断发现新词汇并将其添加入分词词典中。分词词典一般都较为庞大(包含几万词条)，对文档进行分词处理需要较长的时间，降低了文档分类效率。我们可对分词处理作一些简化以提高分类效率：将每个汉字都作为一个词条。这样可以取消庞大的分词词典和耗时的分词过程。实验表明，经此简化后的分词系统仍能达到较高的分词准确度。

4. 禁用词表

在自然语言文档中，一般都包含了大量的介词、冠词、连词(如 in、the、and、在)等无具体含义的虚词词汇，这些虚词在几乎所有都的文档中都有很高的出现频率并且没有具体含义，对文档没有区分作用，还会对其它一些实词起到抑制作用，降低了分类系统的处理效率与准确度，应予以虑除。因此还需要建立禁用词表，虑除那些无助于分类的词条。

三、分类系统的实现

下面以采用向量空间模型进行文档分类为例，简述文档自动分类系统的组成与工作流程(如图 1 所示)。

1. 系统组成

文档自动分类系统一般由语料库、词典、特征提取、文档分类等四部分组成：

(1)语料库模块：管理、维护用于算法学习和特征提取的训练文档集。

(2)词典模块：管理、维护切分词典、禁用词表和同义词典、蕴含词典等辅助词典。

(3)特征提取模块：对训练文档进行词条切分和词频统计，并根据词频分布提取出代表文档类的特征项集及相应权值，生成特征向量表。

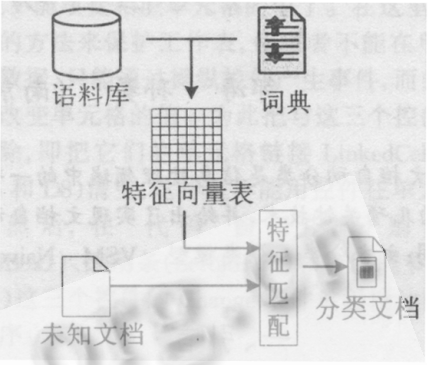

图 1 系统组成与工作流程

(4)文档分类模块：根据词频分布，提取出待分类文档的代表向量，并计算与各文档类特征向量的相似度，将符合一定的阈值条件的文档归属到相应的类别。

2. 工作流程

(1)利用经过人工分类的文档建立训练语料库；

(2)建立切分、统计词典和禁用词表；

(3)对训练文档进行词条切分和词频统计，并生成各文档类的特征向量和初始阈值；

(4)读入待分类文档，并提取特征向量；

(5)计算待分文档向量与各文档类向量的相似度，根据阈值条件生成输出结果。

四、结束语

根据上述分类模型与实现方法，我们已经在 Windows 98 环境下实现了一个针对中文技术文档的自动分类系统，经过对词典与阈值条件的适当调整，达到了较好的分类效果，平均分类正确率达到了 88%。

参考文献

[1] G. Salton, A. Wong and C. S. Yang. A Vector Space Model for Automatic Indexing, Communications of ACM, Vol. 18, 1975, pp613 - 620.

[2] 吴立德 大规模中文文本处理 复旦大学出版社 1997.7

(来稿时间：1999 年 3 月)