基于矢量空间模型的文本自动分类系统研究

Research of Automatic Text Categorization System Based on VSM

包 剑　冀常鹏　李义杰　（辽宁工程技术大学电子与信息工程系　123000）

摘要：随着互联网及其信息服务的快速增长，对于网络信息资源的获取越来越重要，而面向 Web 的文本挖掘技术的发展及应用对于解决这一问题将会起到极其重要的影响。通过分析文本自动分类的关键理论及技术，给出基于矢量空间模型的文本自动分类系统的执行过程，给出了系统的实现算法，提高了系统的精度和效率。

关键词：矢量空间模型 文本自动分类 特征抽取

1 引言

随着信息技术的不断发展，特别是 Internet 应用的日益广泛，人们已经从信息缺乏的时代过渡到了信息极为丰富的年代，其中 WWW 的发展最为迅速，成为包含多种信息资源、站点遍布全球的巨大信息服务网络，为用户提供了一个极具价值的信息源。而其中所包含的各种各样的信息情报、科技文献和新闻等需要管理，为有效地保留大的文本集合，耗费大量的时间和金钱。对文本进行有效管理的方法之一就是将它们进行系统分类。文本分类是自然语言处理的一个重要应用领域，在 80 年代，在文本分类方法占主导地位的是基于知识工程的分类方法，即由专业人员手工编写分类规则来指导分类，其分类过程是先由人类专家来将它们分类，然后被保存于适合的记录材料。在此期间需要大量工作，并且要求专业的分类人员具有较多经验和专门知识。然而分类质量有时还是得不到保证，且周期长费用高，效率低，不易满足人们的实际需要。为解决这些问题，提出网络文本信息处理，它包括信息检索、文本分类和信息过滤等，而文本自动分类则是其中一个重要的环节。

2 文本分类的关键技术

2.1 文本的表示

文档的内容是用自然语言描述的，计算机很难处理其语义，计算机只认识 0 和 1，所以必须将文本的内容特征转化计算机可以处理的格式。根据“贝叶斯假设”，假定组成文本的字或词在确定文本类别的作用上相互独立，这样，可以就使用文本中出现的字或词的集合来代替文本。而这将丢失大量关于文章内容的信息。但是这种假设可以使文本的表示和处理形式化，并且可以在文本分类中取得较好的效果。

近年来，矢量空间模型是信息检索领域应用广泛且效果较好的模型。在该模型中，文档 d 被看作一系列无序词条的集合，对每个词条加上一个对应的权值，矢量空间模型以矢量表示文本：$(\omega_1,\omega_2,\cdots,\omega_n)$，其中 ω_i 为第 i 个特征项的权重。要将文本表示为矢量空间中的一个矢量，就首先要将文本分词，由这些词作为向量的维数来表示文本。最初的矢量表示完全是 0、1 形式，当文本中出现了该词，那么文本向量的该词为 1，否则为 0。这种方法无法体现这个词在文本中的作用程度，逐渐被更精确的词频代替，词频分为绝对词频和相对词频。绝对词频即使用词在文本中出现的频率表示文本；相对词频为归一化的词频。矢量空间模型将文档映射为一个特征矢量 $V(d)=(t_1,\omega_1(d);\cdots;t_n,\omega_n(d))$，其中 t_i 为词条项，$\omega_i(d)$ 为 t_i 在 d 中的权值。$\omega_i(d)$ 一般被定义为 t_i 在 d 中出现频率 $tf_i(d)$ 的函数，即 $\omega_i(d)=\psi(tf_i(d))$。在信息检索中常用的词条权值计算方法为 TF－IDF 函数 $\psi=tf_i(d)\times\log(\frac{N}{n_i})$，其中 N 为所有文档的数目，$n_i$ 为含有词条 t_i 的文档数目。TF－IDF 公式有很多变种，下面是一个常用的 TF－IDF 公式：

$$\omega_{ik}(d)=\frac{tf_{ik}(d)\log(\frac{N}{n_k}+0.01)}{\sqrt{\sum_{k=1}^{n}(tf_{ik}(d))^2\times\log^2(\frac{N}{n_k}+0.01)}}$$

其中，$tf_{ik}(d)$ 表示词条 t_k 在文档 d 中出现的频率，N 表示全部样本文档的总数，n_k 表示包含词条 t_k 的文档数。

根据 TF－IDF 公式，文档集中包含某一词条的文档越多，说明它区分文档类别属性的能力越低，其权值越小；另一方面，某一文档中某一词条出现的频率越高，说明它区分文档内容属性的能力越强，其权值越大。

2.2 文本的特征抽取

利用上面的方法表示文档时，文档的特征矢量会达到上万维甚至数十万维的大小，因此必须进行维数压缩。在文本自动分类中，文本的特征抽取应该注意词条项所在的区域。文本的标题、副标题以及关键字表中的词条项包含了有关文档类别的重要信息，所以应该把其中的特征项作为一类重要特征保留；其次摘要中的特征词条对于分类的贡献也很大，但是仅仅这些特征信息是不够的，还需从正文内容中抽取特征信息。

对于正文内容中特征的抽取可以构造一个评估函数，对特征集中的每个特征进行独立的评估，每个特征都获得一个评估分，然后对所有的特征按照其评估分的大小排序，选取预定数目的最佳特征作为文本的特征集。常用的评估函数有词频、信息增益、期望交叉熵、文本证据权等。

3 自动分类的实现方法

根据分类知识的获取方法不同，可以将文本自动分类系统分为两种类型：基于知识工程的分类系统和基于统计的分类系统。基于知识工程的方法主要依赖语言学知识，需要编制大量的推理规则作为分类知识，实现相当复杂，而且其开发费用相当昂贵。现在应用比较多的是基于统计的自动分类系统，它忽略文本的语言学结构，将文本作为特征项集合来看，利用加权特征项构成矢量进行文本表示，利用词频信息对文本特征进行加权。它实现起来比较简单，并且分类准确度也高，能够满足一般应用的要求。矢量空间模型是基于统计的分类系统中广泛采用的文本计算模型。矢量空间模型可以将给定的文本转换成一个维数很高的矢量。矢量空间模型最突出的特点是可以方便的计算出两个矢量的相似度，即矢量所对应的文本的相似性。自动分类的实现过程如图 1 所示。

3.1 矢量空间模型方法

矢量空间模型（Vector Space Models）由 Salton G. 等人于 60 年代末提出，它是近些年来所研究的信息检索方法的一个重要分支。由于矢量空间模型是建立在规范的数学模型基础上，所以该模型在信息检索领域中的应用最为广泛。其中最为著名的应用该模型的检索系统是 Smart 系统。矢量空间模型用特征项及其相应权值代表文档信息，其应用前提是一篇文档的中心涵义能通过其中的词汇信息（即特征项）体现出来。在进行信息检索时，文档与查询请求之间的相关程度是通过矢量运算来描述的。如果将文献过滤中的新文档和用户兴趣模型也用矢量形式表示出来，那么矢量空间模型将同样适用于信息过滤领域。

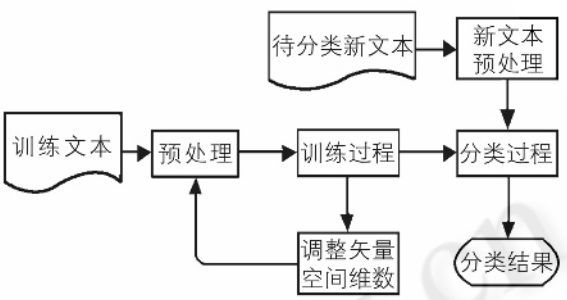

图 1　文本自动分类的实现过程

在矢量空间模型中，文本 D（Document）泛指一般的文献或文献中的片断。特征项 T（Term）是当文档的内容被简单地看成是它含有的基本语言单位（字、词、词组或短语等）所组成的集合时，这些基本的语言单位统称为特征项，即文档可以用项集（Term List）表示为 $D(t_1,t_2,\cdots t_i,\cdots,t_n)$，其中 t_i 是第 i 个特征项，$1\leqslant i\leqslant n$。文本可以用特征项的权值（Term Weight）表示，对于含有 n 个项的文档 $D(t_1,t_2,\cdots t_i,\cdots,t_n)$，特征项 t_i 常常被赋予一定的权值 w_i，表示其在文档中的重要程度，即 $D=(<t_1,w_1>,<t_2,w_2>,\cdots,<t_i,w_i>,\cdots,<t_n,w_n>)$，为描述方便，将文档简记为 $D=(w_1,w_2,\cdots,w_i,\cdots,w_n)$。同理，用户的信息需求也可以用矢量形式表示出来，如下所示 $Q=(q_1,q_2,\cdots,q_i,\cdots,q_n)$，$q_i$ 是用户查询请求中第 i 特征项的权值，$1\leqslant i\leqslant n$。对于给定一个文档 $D=(<t_1,w_1>,<t_2,w_2>,\cdots,<t_i,w_i>,\cdots,<t_n,w_n>)$ 由于 t_i 在文档中既可以重复出现又应该有

先后次序的关系,所以分析起来有一定的难度。为简化分析,可以暂不考虑 ti 在文档中的先后顺序并要求 t_i 互异。这时可把 $t_1, t_2, \cdots, t_i, \cdots, tn$ 看成一个 n 维的坐标系,$w_1, w_2, \cdots, w_i, \cdots, w_n$ 为相应的坐标值,则 $D = (w_1, w_2, \cdots, w_i, \cdots, w_n)$ 可以看成是 n 维空间(特征项文档空间,即 TD 空间)中的一个矢量。

矢量空间模型的相似度(Similarity)是用来度量文档之间或用户的信息需求之间的(内容)相关程度。文档与查询矢量之间的相似度使用以下公式来计算:

$$Sim(D,Q) = \sum_{k=1}^{n} w_k \times q_k$$

或用矢量的夹角余弦值来表示: $Sim(D,Q) = \cos\theta = \dfrac{\sum_{k=1}^{n} w_k \times q_k}{\sqrt{(\sum_{k=1}^{n} w_k^2)(\sum_{k=1}^{n} q_k^2)}}$

利用此方法进行信息检索或信息过滤时,先是将文档表示成能单独加权和操作的特征项集合,再在 TD 空间上计算文档矢量和用户信息需求矢量之间的相似渡,最后提供给用户一组按相似度降序排列的文档列表。在自动归类中可以利用类似的方法来计算待归类文档和某类目的相关度。例如文本 D_1 的特征项为 t_1, t_2, t_3, t_4,权值分别为 30,20,20,10,类目 G_1 的特征项为 t_1, t_3, t_4, t_5,权值分别为 40,30,20,10,则 D_1 的矢量表示为 D_1(30,20,20,10,0),G_1 的矢量表示为 G_1(40,0,30,20,10),则根据上式计算出来的文本 D_1 与类目 G_1 相关度是 0.9036。

3.2 评估方法

一个文本检索系统按一定查询格式的输入检索出了一组文档,文本分类系统的评估指标根据文本检索的度量来定义,即查准率(precision)和查全率(recall)。查准率 p 是指分类器判定的属于类别 c_i 的所有文档中与实际相符的文档所占的比例(即反映正确性)。查全率 r 是指专家判定的属于类别 c_i 的文档中,分类器做出同样判定的文档所占比例。

3.3 实验结果与分析

本系统针对计算机控制类技术文档进行分类,对从 670 篇文档进行训练和测试。所有这些文档分为 4 个类别:组态软件、工控机、控制网络、PLC,取出 170 篇作为测试集,另外 500 篇又分为两部分:360 篇作为训练文档集(Training set);余下 140 篇作为 Validation set,用于调整矢量维度。测试结果如表 1 所示。

表 1　测试结果 %

类别	组态软件	工控机	CAN 总线	PLC	平均
p	91.5	100.0	77.0	92.2	90.2
r	82.0	83.5	85.0	76.2	81.7

由测试结果看出,本系统达到了较好的分类效果。另外,从算法的时间复杂度考虑,若训练文档集有 m 篇,矢量维数为 n,类别数为 k,则训练算法复杂度为 O(mn),分类算法复杂度为 O(kn)。

4 结束语

随着互联网及其信息服务的快速增长,网络信息资源的获取越来越重要,而面向 Web 的文本挖掘技术的发展及应用对于解决这一问题将会起到极其重要的影响。今后,还将在矢量空间模型的基础上,对层次分类体系及算法进行研究,以进一步提高分类效率和分类精度。

参考文献

1 Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization [J], In Journal of Information Retrieval, 1999(1):67 – 88.

2 Mena J. Data mining your website [M], USA:Digital Press,1999.

3 HWANG Kai, XU Zhiwei. Scalable Parallel Computing: Technology Architecture Programming. San Francisco:Mcgraw – Hill,1998.

4 Baker M, Buyya R. Cluster Computing: The Commodity Supercomputer. Softw. Pract. Exper., 1999, 29(6) 551 – 576.

5 李晓黎、刘继敏、史忠植,基于矢量机和无监督聚类相结合的中文网页分类器[J],计算机学报,2001(1):62 – 68。

6 刘开瑛、薛翠芳等,中文文本中抽取特征信息的区域与技术[J],中文信息学报,1998(2):1 – 7。