基于背景知识的文本自动分类

卢　朋，曾隽芳，杨一平

（中科院自动化研究所，北京　100080）

摘　要： 文本自动分类是智能信息处理的一个重要分支，目前大量的研究都是基于统计的，采用的文本表示方法大都是基于向量空间模型的，这些方法存在着一些缺陷。本文在分析了这些缺陷的基础上，提出了一种新的分类方法——基于背景知识的文本自动分类方法。该方法模拟了人的分类过程，建立能够代表类别的背景知识——知识树，利用相应的分类算法对文本进行处理，试图让计算机在背景知识下具有人的认知能力。这种方法在已知知识树的前提下，考虑了文本的语义层次结构，把文本的结构与知识树相整合，以激活相应的树枝，最后得出文本类别的归属。试验表明，该方法具有较高的分类正确率和召回率。

关键词： 文本自动分类；知识树；认知能力

Automatic Text Classification Based on Background Knowledge

Lu Peng, Zeng Junfang, Yang Yiping

(Institute of Automation, Chinese Academy of Sciences, Beijing, 100080)

Abstract: Automatic text classification is one of important fields in intelligent information process. Most researchers focus on statistic method (Rocchio, SVM, KNN etc.) which is based on Vector Space Model (VSM) representing text. On the basis of analyzing their disadvantages, a new method —automatic text classification based on background knowledge is proposed in this paper. This method is to simulate the classification process of human being. And it includes background knowledge and classification algorithm in order to make computer cognitive ability. It combines text semantic structure and background knowledge to activate relative branches of knowledge tree and decide which classification it belongs to by reasoning. The experiment indicates that the model has higher classification precision and recall.

Keywords: Automatic Text Classification; Knowledge Tree; Cognitive Ability

1　引言

在当今信息爆炸的时代，人们面对的是多元海量的信息，仅仅依靠人来整理这些杂乱的信息是远远不够的。文本自动分类作为智能信息处理的一个重要分支，就是对大量的用自然语言写成的文本按照文本内容自动归属到预定的类别，现今文本自动分类已经应用到搜索引擎、邮件过滤等系统。

目前，国内外在文本自动分类方面的研究主要是统计的方法。基于统计方法的主要算法主要有：Rocchio[1]、简单向量距离法、朴素贝叶斯法[2]、KNN[3]、支持向量机[4]、神经网络[5]等。这些算法中，文本的表示采用

作者简介：卢朋（1980—），男，山东，博士，E-mail:Lu_Peng@mails.gucas.ac.cn;

了向量空间模型（Vector Space Model，VSM）。向量空间模型的基本思想是以向量来表示文本，把文本视为字、词或词组（即为向量空间模型特征）的序列，根据特征项（能够代表类别的特征词或短语）对文本的贡献赋以一定的权值，从而构成一个向量$(\omega_1,\omega_2,...\omega_n)$，其中$\omega_i$是第 i 个特征的权值，n 是特征总数[5]，然后通过计算向量之间的距离来判决文本的归属。

尽管基于统计的分类方法应用广泛，并取得了可喜的成果，但也有着诸多缺陷，比如没有考虑文本的语义结构。我们针对统计方法存在的问题，模拟了人的分类方式，提出了基于背景知识的文本分类方法，它是从人的认知角度对文本进行分类，目前的试验情况表明，已得到了期望的效果。

本文主要探讨了基于背景知识的文本分类系统的关键技术和系统实现，第 1 节为引言，第 2 节分析了统计的文本分类方法的一些缺陷，第 3 节给出了我们实现的基于背景知识的文本分类的思想起源和系统结构框架，第 4 节和第 5 节探讨了文本分类系统的关键技术，第 6 节是该系统的试验结果和试验分析，第 7 节是结束语。

2 统计文本分类方法的分析

统计文本分类是计算向量距离以判断文本的归属，而大都采用向量空间模型来表示文本，向量空间模型是把词或词组看成文本序列，用它来表示文本并不是十分完善，分以下几种情况说明：

（1）特征项关系：向量空间模型是假设文章中词与词之间是独立的，没有关联的[6]。对于文本来说，文本结构都是有一些概念和概念之间的关系来组成的，而向量空间模型却没有体现出这种关系，在这个意义上来说，向量空间模型并不能表示文本。（2）特征项差别：对于体育、政治、经济、科技等这些大的类别来说，他们之间的重复特征项比较少，能够很好的区分，但是如果类别之间交叉现象比较严重的时候，他们之间的特征项重复的非常多，分类精度会大大下降。尤其是在多层分类体系中，子类之间的特征交叉更为严重，分类就很难达到需要的效果。（3）无用特征项：通常来说一篇文本会有成千上万个词，而有很多词是各个类别均可以出现的，比如，“我们”、“的”等词，并不能代表类别的特征。尽管可以建立了一个通用的停用词表（对文本类别不产生任何意义的词表），提取特征时去除这些停用词，但是剩余的特征项数也是庞大的，而对于文本来说具有关键作用的词才能判断文本的类别，这部分词往往并不是很多。由于这部分无用特征项的影响会导致向量距离计算的不准确。（4）其他：对于向量空间模型（VSM）来说，只考虑词形，并没有考虑词义，事实上自然语言中存在着大量的一词多义和一义多词的现象，以词形为单位的方法很难克服这个问题。比如“计算机”和“电脑”是同义，但是向量空间模型却看成不同的词；另外，对训练语料来说，还存在一个过学习和欠学习的过程，如果训练语料不全面，代表性不强，他们的特征项组成的向量就不能代表类别，会直接影响自动分类的精度，而训练语料过于多，可能有一些非本类的特征项也收集在其中，则也会影响分类的精度。

由此，我们提出了基于背景知识的文本自动分类方法。

3 思想起源与系统框架

3.1 思想起源

一般来说，人的分类过程是先通过阅读文章，以获得文章中语言符号的表层意义，即对文本的浅层理解，然后根据个人经验、文化背景知识对获得的文本信息进行加工，最后对文本做出判断。从认知心理学的角度来看，文本分类是一个复杂的认知心理过程，在这个认知心理过程中，涉及到两个非常重要的方面：认知能力和背景知识。模仿人的这种认知过程，我们提出了基于背景知识的文本自动分类方法，它首先是建立一个背景知识，然后根据背景知识对文本进行处理，最后整合这些信息以得出文本的分类信息。我们把这种基于背景知识的文本自动分类过程看作一个文本的语义结构与背景知识整合的过程。

3.2 系统框架图

根据以上的知识，我们构建文本自动分类系统分为两大过程，一、背景知识的构建，即认知心理学中的知识体系的构建，二、对文本的分类算法的实现，即认知能力的实现。其整体流程如图 1。

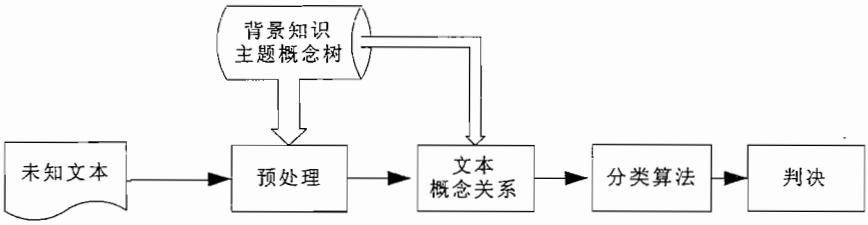


Fig 1 Schematic of the system architecture

图 1 基于背景知识的文本分类系统框架

4 背景知识的构建

对人来说，当输入的文本信息与记忆中贮存的有关信息相整合，才能得以对文本进行分类，如果缺乏有关的信息，或者未能激活记忆中的有关信息，那么就不能或难于对文本实现分类。比如，有的小孩指着田地里的"麦苗"叫"韭菜"，就是因为在他的背景知识里没有“麦苗”这样的背景知识，也就无从区分“麦苗”和“韭菜”。可见已有的背景知识对文本分类显得如此重要，以致可以把背景知识与文本信息之间的联系看作是文本分类的一种认知能力。根据认知心理学，人已有的知识构成一个认知结构（cognitive structure），即知识体系（knowledge structure），它包括三个基本组成部分：多个范畴系统、区分不同范畴的规则和各个范畴间相互关联的网络系统[7，8]。根据分类的需要把背景知识分为两个部分：每个类别中的概念和概念之间的不同关系（成员关系、父子关系和同义关系）。我们把这种背景知识以知识树的形式表示出来。如图 2 所示。

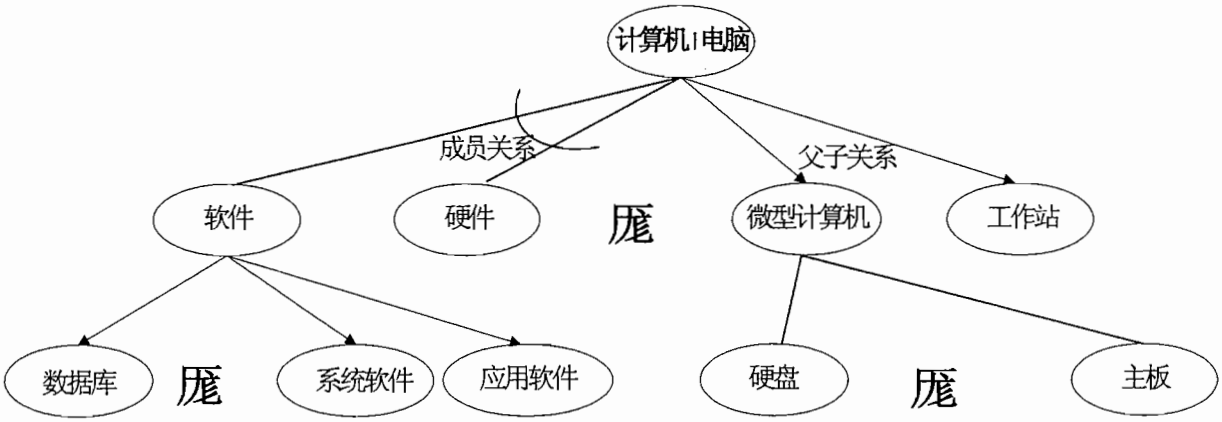


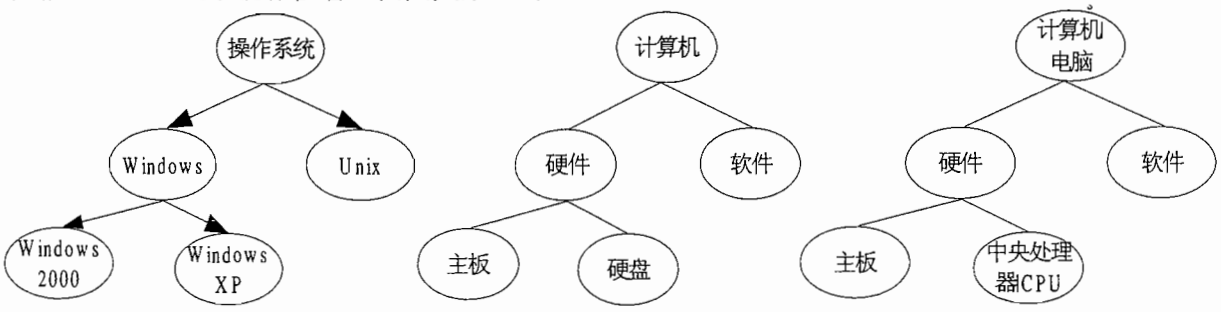
Fig.2 Knowledge Tree in Compute Science

图 2 计算机方面的部分知识树

总体来说，知识树的建立分以下二步：

（1）节点概念：节点概念是指知识树在其节点上的概念。概念是人类对客观世界认识的结果,是客观事物或现象在头脑中的反映[8]，在本质上是符号化的实体,它表示的是客观世界中的事物及其含义。在我们的系统中,概念是由节点概念表示的标识符。通常，人在阅读文本时可以在阅读若干行后甚至只阅读标题后，就可以较为准确的判断该文本的所属类别，而不必阅读整篇文本，这其中所依据的就是类别中的一些概念——主题概念，所谓主题概念是指能够代表本类以区分其他类别的词或短语。众所周知，《中国图书分类法》和《中国分类主题词表》是大部分图书分类的主要依据，他们实际上相当于专家知识，我们的知识树的节点概念就是参照《中国图书分类法》和《中国分类主题词表》来划分的。（2）概念关系：客观事物之间是普遍联系的，概念之间也存在着某种关系。要描述一篇文本，主要看文本所包含的概念信息和概念关系信息，到目前为止，我们把知识树上的节点概念关系分为三种 ：父子关系、成员关系和同义关系。父子关系具有继承特征，即父类的特征可以被子类所继承，在知识树上以“→”来表示，如“微型计算机”具有“计算机”的特征，如图 3（a）所示。成员关系即整体－部分关系是表达事物组成的一种关系，知识树上以“－”来表示，如图 3（b）描述了“微型计算机”的组成，成员关系属性具有传递的性质，如计算机包含硬件，硬件包含主板，主板包含中央处理器，则计算机也包含这中央处理器，但成员关系没有继承特征。同义关系，同一个客观事物和现象，常常可以用不同的概念去描述和修饰，

这些不同概念之间的关系就存在着“同义关系”。在知识树上以“l”来表示，如图3（c）所示。



(a) Hyponymy Relation
(a) 上下位关系

(b) Member Relation
(b) 成员关系

(c) Synonymy Relation
(c) 同义关系

Fig.3 Concept Relation

图 3 概念关系

知识树是层次结构的语义组织，不同的层次概念和层次关系描绘了整个背景知识。我们的背景知识是通过Stias系统[1]来实现的。此系统为我们建立知识树提供了便利的工具。

5 分类算法

人的信息加工需要利用一定的策略，表现出人的心理活动的智慧性。文本分类也是如此。人在已有的知识和经验的基础上，常应用各种策略，如语义策略、次序策略和句法策略等等，来加工语言信息。已有背景知识对文本分类的作用不仅表现在策略运用上，而且还表现在信息整合上，即背景知识与文本信息之间的整合，只有激活了背景知识中贮存的这些知识，才能使文本所传达的信息与已贮存的有关信息得以实现整合，从而达到对文本进行浅层理解或者能够准确地理解，最终对文本进行分类。

尽管文本中的谴词造句会随着作者的风格和习惯而不同，但描述同一主题的文章所涉及的概念都是有关联的。我们在考察一篇文本所描述的内容时，主要是看其中表达被描述对象的概念及关系，以确定被描述对象内概念及其概念关系的语义结构，然后再与背景知识相整合。

由此，文本自动分类实际上是文本中主题概念与知识树上节点概念的匹配，也就是计算待分文本与知识树节点主题概念 S_j 的关联概率 $R(S_j)$，文本自动分到概率最大的节点。分类算法大致如下：

（1）节点概念的初始关联系数

节点概念的初始关联系数是指知识树上节点概念 S_j 在文本中频率，即 S_j 在文本中重复出现的次数 N_i 除以文本中所有概念总数（记重复）M。对汉语这种没有区别性分词标志的语言来说，要得到文本的所有概念，即需要对文本进行分词（segment）。节点概念 S_j 的初始关联系数为：

$$R_0 = N_i / M \ ,(N_i \qquad 0 \square CM > 0) \tag{1}$$

(2)节点概念关联系数的计算

节点概念关联系数是指知识树上的节点概念与文本中的概念的相关性，这种相关性主要包括节点概念的初始关联度和下位节点对上位节点的贡献。节点概念的初始相关度即为节点概念初始关联系数 R_0。本系统只针对下位节点对上位节点的贡献，即是考虑知识树上的层次概念之间的语义关系，目前我们定义节点概念之间是存在三种关系：父子关系、成员关系和同义关系。只考虑下位节点对上位节点的影响，对于叶节点没有下位节点，也就没有其他节点概念的贡献了，而非叶节点概念 S_j 的关联系数由三个部分组成：它自身的初始关联系数 R_0，来自父子关系继承的关联系数 R_h，来自成员关系的关联系数 R_c。

父子关系继承的关联系数：

[1]科技情报信息分析系统(Science Technology Information Analysis System, Stias)是中科院自动化研究所综合信息研究中心建立的一套科技情报收集、加工、分析系统。

$$R_h = \sum_n (a_n \times R_n) \quad (2)$$

其中 R_n 为与节点概念 S_j 具有父子关系的下位节点概念的关联系数，加权系数 a_n 应满足 $0 \le a_n < 1$，并且 $\sum_n a_n < 1$，如果我们认为当一个主题概念所有子类主题都被激活，就相当于该主题也被激活，并且每个子类主题对父类主题概念的贡献都相同，这时

$$R_h = \sum_n \left(R_n / N_h \right) \quad (3)$$

其中 N_h 为父子关系下子元素的总数。

同理可推出成员关系的关联系数 R_c：

$$R_c = \sum_m \left(R_m / N_M \right) \quad (4)$$

其中 N_M 为成员关系下子成员元素总数,R_m 为与 S_j 具有成员关系的下位节点概念的关联系数。

对于同义关系来说，其表现在知识树上的同一个主题概念上，影响的是概念初始概率 R_0，影响后的 R_0 为：

$$R'_0 = (N_i + N_s) / M \quad (5)$$

其中 N_i 为节点概念 S_i 在文本中出现次数，N_s 为与 S_j 具有同义关系的概念在文本中出现次数。

最后，我们利用概率统计中的独立概率公式：

$$P(XYYYZ) = P(X) + P(Y) + P(Z) - P(X) \times P(Y) - P(Y) \times P(Z) - P(Z) \times P(X) + P(X) \times P(Y) \times P(Z)$$

可计算出主题概念 S_j 的关联系数为

$$R(S_j) = R'_0 + R_h + R_c - R'_0 \times R_h - R_h \times R_c - R'_0 \times R_c + R_0 \times R_h \times R_c \quad (6)$$

（3）潜在激发条件

对于非知识树类别的文章来说，节点概念也可能出现在文章中，但是这些节点概念在文章是分散且互不关联的，为此我们可以预先设定一个规则，去除这些节点概念，以区分非知识树类别的文章。我们设定的规则如下：

（a）知识树上的节点概念 S_j 的初始关联系数 R_0<0.01。(b)S_j 的兄弟节点概念和同树枝节点概念并没有出现。

符合以上两个条件我们就认为此节点概念 S_j 是与文本不相关的，自动去除 S_j。

6 实验结果与分析

到目前为止，我们建立了自动化、计算机、机器人三个领域的比较健全的知识树，暂时只对自动化、计算机、机器人的语料进行测试，对于其他类别的语料本系统只能识别出是不是这三个类别。为了验证本算法的合理性，我们在互联网上找了计算机、自动化、机器人、体育四类语料对系统进行测试。通常，对系统效果的评价主要使用准确率，查全率，F1 综合指标三个指标。准确率和查全率反映了分类质量的两个不同方面，两者必须综合考虑，采用 F1 综合指标。

为了与统计方法性能进行比较，我们采用了 KNN 分类器。两种方法的试验结果如表 1 所示。

从试验结果可以看出：

（1）基于背景知识的文本分类可以达到预期的效果。（2）对于体育领域的测试综合指数是最好的，机器人语料中由于涉及到一些计算机和自动化领域的概念，效果稍微差点。

总体来说，基于背景知识的文本自动分类能够达到我们的要求，能够应用到实际的分类系统中去。但是到目前为止还有一些不完善的地方，如暂时还没有对文本标题和文本关键字进行处理等。

表 1 试验结果

Tab.1 Result

评价 / 类别	基于背景知识			基于统计的（KNN）		
	正确率	召回率	F1 综合	正确率	召回率	F1 综合
计算机	96.9%	96.9%	96.9%	85.3%	86.1%	62.4%
自动化	95%	94%	89.7%	80.2%	82%	81.1%
机器人	92.6%	92.2%	92.4%	79.6%	81%	80.3%
体育	98.2%	99%	98.6%	95.8%	97%	96.4%

7 结束语

在分析了基于统计的文本自动分类的基础上，我们模仿人的分类系统，提出了基于背景知识的文本自动分类，它模拟了人的认知能力，考虑到了背景知识和文本的语义结构，避免了采用向量空间模型对文本分类时的缺陷。从试验结果来看，取得了良好的效果,这证明了基于背景知识的方法对进一步提高分类系统的效果是可行的。尽管现在只是考虑了三类的情况，但知识树有好的扩展性，我们可以方便的利用 Stias 平台建立其他领域的背景知识。今后进一步的工作：在理论研究上，对知识树上的复合概念进行处理；在实践上，健全知识树，对更多、更大规模的文本进行基于背景知识的文本自动分类实验研究。

参考文献

[1] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997, pp.143--151.

[2] Susana Eyheramendy, David D. Lewis, David Madigan, On the Naive Bayes Model for Text Categorization,Ninth International Workshop on Artificial Intelligence and Statistics,2003

[3] Gongde Guo, Hui Wang, Using KNN Model-based Approach for Automatic Text, 2003

[4] T. Joachims. Text categorization with Support vector machines. In Proc. 10th European Conference on Machine Learning (ECML'98), pages 137--142, Chemnitz, Germany, 1998.

[5] 庞剑锋，卜东波，白硕，基于向量空间模型的文本自动分类系统的研究与实现，计算机应用研究， 2001

[6] 李晓黎，刘继敏，史忠植，概念推理网及其在文本分类中的应用，计算机研究与应用，2000

[7] Kazuhiro Morita , El-Sayed Atlam , Masao Fuketra , Kazuhiko Tsuda , etc., Word classification and hierarchy using co-occurrence word information, Information Processing and Management: an International Journal,2004

[8] 王甦，汪安圣，认知心理学，背景大学出版社，1992，P240～241，P261～275