计算机科学 2003Vol. 30№. 3

一种模仿人类的自动文本分类算法

王树梅[1] 戴保存[1] 黄河燕[2] 陈肇雄[2]
(南京理工大学计算机系 南京 210014)[1]
(中国科学院计算机语言信息工程研究中心 北京 100083)[2]

An Automatic Algorithm of Text Categorization Imitating Human's

WANG Shu-Mei[1] DAI Bao-Cun[1] HUANG He-Yan[2] CHEN Zhao-Xiong[2]
(Department of Computer, Nanjing University of Science and Technology, Nanjing 210014)[1]
(Computer Language Information Engineering Center, Chinese Academy of Sciences, Beijing 100083)[2]
E-mail: haomimi@mail.njust.edu.cn

Abstract An algorithm of text classification is given that imitates human's in this paper. On one hand, the algorithm enhances weight of theme when feature vector is processed, because of the assumption that the title of a document can project its content. On the other hand, a weight parameter ω vector is designed to simulate human's skimming and skipping behavior for calculating method of a document cluster center, and a weight of the feature that there are more positive examples than negative ones is enhanced. The experiment shows that the algorithm greatly improves the performance of a text classification system.

Keywords Text categorization, Corpus, Cluster center, Machine learning

1. 引言

Internet 上有着大量的且快速增长的文本，文本是信息和知识的宝贵资源。随着 Internet 的快速发展，不久的将来，人们所需要的大部分信息都可以在网上找到。Internet 正在成为人类的信息宝库，但是随着网上信息的爆炸性增长，人们想从这个信息宝库中获得自己所需要的信息已经变得日益困难，因此，如何快速有效地获得有用的信息已成为人们十分关心的问题。文本信息处理就成了人们研究的热门课题。文本的自动分类是文本信息处理的重要环节。

本文根据人类分类专家对文本进行分类时的一些行为特点设计了一个模仿人类的自动分类算法。该分类算法的重点是：在计算类别的几何文档聚类中心时，提升对类别有较强区分能力的特征项在特征向量中的权重，压缩对区分类别能力较弱的特征项在特征向量中的权重。经过对封闭语料和开放语料的测试结果表明这种分类器产生的结果提高了文本分类系统的性能。

2. 文本的表示

文本分类的任务是根据待分类的文本所描述的内容，来决定该文本所属的类别。从这个角度来看，文本分类应涉及到自然语言的篇章理解技术，由于该技术目前尚未到达实用化程度，本文采用信息检索中的自动标引技术对文本进行分析，用特征向量作为文本的内部表示。下面讨论特征向量生成之后，如何对特征向量进行分类。

· 设$\overline{d_1}$，$\overline{d_2}$是两篇文章的向量表示，即：

$\overline{d_1}=(d_{11}, d_{12}, \cdots, d_{1n})$；$\overline{d_2}=(d_{21}, d_{22}, \cdots, d_{2n})$；

则$\overline{d_1}$，$\overline{d_2}$的相似度为：

$$S(\overline{d_1},\overline{d_2})=\frac{\sum_{i=1}^{n} d_{1i} * d_{2i}}{|\overline{d_1}\| \overline{d_2}|} \quad (1)$$

· 设$\{\overline{d_1},\overline{d_2},\cdots,\overline{d_k}\}$是一个聚类，其中各文献的向量表示为：

$\overline{d_i}=(d_{i1},d_{i2},\cdots,d_{in})$ 其中 $1\leqslant i\leqslant k$

将聚类表示为

$$\overline{C}=\frac{1}{k}\sum_{i=1}^{k}\frac{\overline{d_i}}{|\overline{d_i}|}; \quad (2)$$

这种表示方法也称为质心法。

当要确定某待分类文献 d 的类别时，假设共有类别$(C_1, C_2, \cdots, C_k)$k 个类别，按式(2)分别计算 d 与 k 个类别的相似度，若把 d 归属到第 i 个类目，则必须满足下列条件：

$S(d,C_i)=\max_j(S(d,C_j))$；

· 正例集指训练语料中属于某个类别的所有文档构成的集合，反例集是训练集中不属于该类别的文档构成的集合。

3. 仿人分类的分类算法

人类分类专家对文本进行手工分类是一个比较复杂的智力活动，其中包括推理判断、总结概括等过程，计算机对文本进行分类，其实就是用计算机来模拟这一复杂的智力活动。充分探寻手工分类的实质与关键，将其更好地应用到自动分类中去，是建立一个好的文本分类系统的关键。

3.1 算法描述

(1)对训练语料中的正例集和自例集文本分别对标题和正文进行词频统计，得到两个向量"Head"，"Body"：

$Head=(h_{i1},h_{i2},\cdots,h_{ij},\cdots,h_{im})$

$Body=(b_{i1},b_{i2},\cdots,b_{ij},\cdots,b_{im})$

(2)生成特征项在正例集和反例集中的文档频率，向量 Pos和 N_{eg}

正例集文档频率向量：$Pos=(P_1, P_2, \cdots, P_i, \cdots, P_m)$

反例集文档频率向量：$N_{eg}=(N_1, N_2, \cdots, N_i, \cdots, N_m)$

$$P_i=\sum_{k=1}^{s_1}(((h_{ki}+b_{ki})>0)?\ 1:0) \quad (3)$$

$$N_i=\sum_{k=1}^{S_2}(((h_{ki}+b_{ki})>0)?\ 1:0) \quad (4)$$

其中，S_1 为正例集中文档的总数，S_2 为反例集中文档的总数。在式(3)中，h_{ki}，b_{ki}表示在正例集中第 k 篇文档的第 i 个特征项在标题向量和正文向量中的权重；在式(4)中，h_{ki}，b_{ki}表示在反例集中第 k 篇文档的第 i 个特征项在标题向量和正文向量中的权重；

(3)计算正例集中文档 d_i 的特征向量 $d_i=(f_{i1},f_{i2},\cdots,f_{ij},\cdots,f_{im})$

$$f_{ij}=ah_{ij}+b_{ij}$$

(4)对正例集中文档 d_i 的特征向量进行归一化处理得到向量 V_i，V_i 作为文档的内部唯一描述

$$V_i=(V_{i1},V_{i2},\cdots,V_{ij},\cdots,V_{im})$$

其中 $V_{ij}=f_{ij}/\sqrt{\sum_{j=1}^{m}f_{ij}^2}$

(5)计算类别的一个一维加权因子向量 $\omega=(\omega_1,\omega_2,\cdots,\omega_i,\cdots,\omega_m)$

$$\omega_i=\begin{cases}0.01 & \frac{p_i}{p_i+n_i}\leqslant 0.25\\ \exp(1+\frac{p_i}{p_i+n_i})-\exp(1+0.25)+0.01 & \frac{p_i}{p_i+n_i}>0.25\end{cases}$$

(6)计算文档的聚类中心 C。设类别有 N 个正例集，每个正例集用 V_i 表示，V_i 乘以加权因子向量 ω，得到中间向量 $V'_i=(V_{i1}\omega_1,V_{i2}\omega_2,\cdots,V_{ij}\omega_j,\cdots,V_{im}\omega_m)$，再对 V_i 进行归一化处理，得到向量 V_i^*

$$V_i^*=(t_{i1},t_{i2},\cdots,t_{ij},\cdots,t_{im})$$

其中 $t_{ij}=V_{ij}*\tilde{\omega}_j/\sqrt{\sum_{i=1}^{m}V_{ij}*\tilde{\omega}_j}$

生成中间向量 $C'=(C'_1,C'_2,\cdots,C'_j,\cdots,C'_m)$，其中 $C'_j=\sum_{i=1}^{N}t_{ij}$

对 C′进行归一化处理，得到文档的聚类中心 C：

$$C=(C_1,C_2,\cdots,C_j,\cdots,C_m)$$

其中 $C_j=C'_j/\sqrt{\sum_{j=1}^{m}C_j'^2}$；

(7)分类时计算文档向量 V_i 和文档聚类几何中心 C 之间的相似度 $Sim(V_i,C)$

$$Sim(V_i,C)=\sum_{j=1}^{m}V_{ij}*C_j$$

然后根据一定的阈值来判定该文档是否属于这个类别。

3.2 算法分析

根据相关的研究表明，自然科学期刊论文的篇名与其内容不符合的(占统计的 0.23%)极少，而绝大多数(占统计的 89.2%)都能较好地反映内容(另有 10.57%部分符合)。对《文汇报》这样的新闻报刊研究统计表明，文章标题也有 95%左右的切题性，80%以上的标题关键词都可以反映文章的主题。虽然研究是对中文语料而言的，但这说明了文章的题目在一定程度上反映了文章的主题，所以对从文章标题之中抽取的特征项应给予较高的权值，因此在算法的第 3 步，对标题向量乘以了一个加权系数。由于在训练阶段，训练语料的文本长度是变化的，为了考虑不同长度的训练文本在训练过程中对文档聚类中心的贡献作用是相同的，本算法对向量进行了归一化处理，步骤 4 就是完成这方面工作的。

本算法认真分析了手工分类的基本程序和手工分类人员的思维过程，进一步提炼出手工分类中的一些技术关键与人工分类的经验和规律，从而形成了本算法的第 5 步和第 6 步，而且通过本文的分类实验证明这套算法对文献自动分类是可行的，下面解释第 5 步和第 6 步是如何模仿人工分类的。人工分类人员在进行手工分类时，他并不需要通读整篇文档之后，才能判断该文档属于哪一个类别，而实际情况是，他们只要略读和跳读待分类的文档，就可以判断该文档属于哪一个类别。当然在这种跳读和略读的背后人的大脑进行了推理、概括、总结等一系列复杂的、高智能的活动。手工分类人员在跳读或略读过程中，发现了一些有用关键字之后，经过大脑的判定就可以确定文档的类别了。那么手工分类人员到底是看到什么样的关键字之后马上就可以确定文档的类别呢？本文认为应该是那种和类别有较高相关度的关键字，也就是能够揭示文档内容的关键字，同时也是那种区分相关文档和非相关文档能力较强的关键字，这样的关键字直观地说也就是那些在正例集中大量出现的而在反例集中出现较少的关键字。

本文设计的加权因子向量 ω 就是用来模仿人手工分类时的跳读和略读过程的。对那些在正例集中大量出现而在反例集中出现较少的特征项，计算文档的聚类中心时，该特征项权重就得到了提升；相反那些在反例集中大量出现而在正例集中出现较少的特征项的权重就得到压缩。对特征项在文档聚类中心的权重的提升，并不是线性提升，而是非线性提升。

例如一篇文档 d 中出现了某一个关键字 i，如果手工分类，工作人员只要看到关键字 i 就可以立即确定该文档的类别 C_k，如果该文档 d 经过本算法进行分类，再计算文档和聚类中心之间的相似度：

$$Sim(d,C_k)=V_1*C_1+V_2*C_2+\cdots+V_i*C_i+\cdots+V_m*C_m$$

因为在计算文档聚类中心时，关键字 i 在文档聚类中心的权重得到了提升，因此该相似度就会得到提升，那么把该文档归类到该类别的可能性就会大。

4. 实验结果

实验用的语料库是路透社(Reuters-21578 Ver 1.0)语料库，它是 Reuters-22173 语料库的改进版。整个语料库收集的是路透社关于金融方面的新闻报道，语料库大约有 340 万单词，约 23MB，有 135 个关于金融方面的话题。本实验从路透社语料库中选取了 6699 篇作为训练文档集，5667 篇作为测试文档集。实验将语料库中的 Mergers/Acqusitions 话题作为研究对象。

4.1 文本分类器的性能指标

列联表(Contingency table)是统计分析中采样数据的一种频数分布，采样数据用两个以上因子进行分类，其中每个因子又具有两种或两种以上的类型。假设一个系统需要进行 n 次判断，每次判断都给出一个“yes”或“no”的判定，n 次判定可用下面的列联表表 1 表示。

表 1 n 次判定的列联表表示

	Yes 是正确	No 是正确	
判定为 Yes	a	b	a+b
判定为 No	c	d	c+d
	a+c	b+d	a+b+c+d=n

根据列联表可以计算出系统有效性的二个指标，例如表

(下转第 53 页)

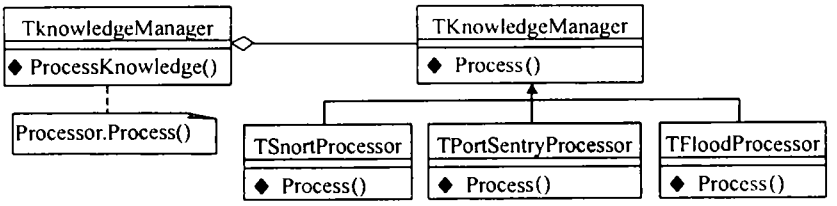

图 8　不同知识类型的统一处理模式

dgeProcessor 子类的实现可以用动态链接库，通过运行时加载，可以动态更新系统。如果用 Java 实现则是一种最自然的 OO 构建方式。以下用伪代码示(Object Pascal-like)示意模式中的关键部分：

```
TKnowledgeManager. ProcessKnowledge(Packet: String);
Begin
  Decode packet to get K_Type and the payload of knowledge;
  For I := 0 to Processors. Count - 1 do //Processors is a list of instances of subclass of TKnowledgeProcessor.
    Processors[I]. Process(K_Type, Knowledge);//different subclasses of TKnowledgeProcessor override method of Process.
End;
```

知识类型的 ID(K_Type)必须唯一，我们考虑可以采取两种方式：分散式管理(如 COM 中的 GUID)和集中式管理(如域名管理)。两种方式各有优缺点。

结论　在安全系统中安全知识(防火墙规则、入侵检测规则等)的及时更新是系统安全的重要保证。人工方式有诸多缺点，如果能以自优化的方式，形成一个知识社区(管理域节点形成的树)，则每一个社区中被管理的子系统都可以自动在社区中通告其新发现的知识(可能是管理员新加入的)。每一个子节点上的知识发现将在社区全局内形成共享。如此，系统将不断自优化和进化，保证了安全系统知识规则的及时有效更新。我们提出了入侵检测知识自优化的概念，给出了一个开放的、高伸缩性的自优化框架，并实现了一个实验性系统 Vortex，类似的工作在入侵检测领域尚未见报道。该框架实际上是一个通用的知识自优化框架，安全领域其它类型的知识，比如防火墙的规则也可纳入其中。未来的工作是在框架中加入评估，使得一个管理域可以评估新提交知识的有效性和重要性，让系统能以更有序的方式进化。

参 考 文 献

1　Waltz E. Information Warfare principles and operations. Artech House Boston . London

2　Kephart O, et al. Biologically inspired defenses against computer viruses. In: Proc. of IJCAI '95, 985-996, Montreal, Aug. 1995

3　www. gecities. com/researchtriangle/3372/firewall_farm. html

4　Kephart J O. A biologically inspired immune system for computers. In: R. A. Brooks, P. Maes, eds. Artificial Life IV. Proc. of the 4th Intl. Workshop on the Synthesis and Simulation of Living Systems, MIT Press 1994. 130～139

5　Communication in the Common Intrusion Detection Framework. http: // www. isi. edu/gost/cidf/ drafts/communication. txt

6　Gamma E, Ralph Johnson Design Patterns: Elements of Reusable Object-Oriented Software

(上接第45页)

中的 a 表示系统给出判定为"yes"，而"yes"为正确判定的次数。

$$查全率=\frac{a}{a+c}\quad 查准率=\frac{a}{a+b}$$

对文本分类系统来说 a+c 就是人工分入该类的文档数量，在封闭测试时，就是训练集中属于该类的文档数，成为机器归入文档数；列联表中的 a 表示机器正确地分入该类的文档数，成为正确归入文档数，因此对于文本分类系统来说查全率和查准率可定义为如下表示：

$$查准率=\frac{正确归入文档数}{机器归入文档数}*100\%$$

$$查全率=\frac{正确归入文档数}{应有文档数}*100\%$$

查准率和查全率这两个指标是相互矛盾的，有时为了提高系统的查准率就会使系统的查全率下降；为了提高系统的查全率，就会使系统的查准率下降，一个好的系统应该很好地兼顾这两个指标。

4.2　实验的设计及其描述

本文实验主要想验证仿人算法中所设计的计算聚类中心的新方法对分类器性能的影响。

实验的基准是：特征向量的特征项完全由单词组成，不包括任何短语，理想文献的计算方法用2中的式(2)，不对训练语料中的文本进行聚类分析；计算文本和理想文献的相似程度使用2中的公式(1)，本文所做的设计对分类器的影响通过查准率和查全率反映出来。

下表(表2)列出了本文对封闭语料和开放语料的实验结果：

表2　对封闭语料和开放语料的实验结果

	封闭语料		开放语料	
	查准率	查全率	查准率	查全率
基准	0.50	0.59	0.57	0.56
加入短语搭配之后	0.57	0.67	0.64	0.62
仿人算法	0.66	0.74	0.70	0.72

从实验的结果可以看出，把短语加入特征向量中之后，可以提高特征向量的聚类特性。本文设计的模仿人类专家的略读和跳读行为的文献聚类中心的计算方法，对系统性能有极大的提高，这给我们的一个启示是：探索人工分类的实质并把其应用到文献自动分类中去是一个系统是否能成功的关键。

参 考 文 献

1　康耀红著. 现代情报检索理论. 科学技术出版社，1990

2　朱兰娟. 中文文献的自动分类:[学位论文]. 上海交通大学，1987

3　曹素青，曾伏虎，等. 一个中文文本自动分类数学模型. 情报学报，1999，18(1)

4　张玥杰，姚天顺. 基于特征相关性的汉语文本自动分类模型的研究. 小型微型计算机系统，1998，19(8)

5　王永成，张坤. 中文文献自动分类研究. 情报学报，1997，16(5)

6　成颖，史九林. 自动分类研究现状与展望. 情报学报，1999，18(1)

7　Cohen, William W. Text Categorization and Relational Learning. Machine Learning. In: Proc. of the Twelfth Intl. Conf. (ML95)