中国农业大学学报 1999,4(4):74～77
Journal of China Agricultural University

文本信息自动分类系统 ITC98(Ⅰ)①
——ITC98 总体结构与编码子系统

陶 兰 申军霞
(中国农业大学电子电力工程学院)

摘 要 介绍文本信息自动分类系统 ITC98(Intelligent Text Classify)的总体结构及特点,对借鉴人类文本分类方式设计的启发式编码子系统作了讨论。

关键词 文本分类;编码;Kohonen 网络
分类号 TP 391;TP 387

Intelligent Text Classifying System(Ⅰ):
Overall Architecture and Encoding Subsystem of ITC98

Tao Lan Shen Junxia
(College of Electronic and Electric Power Engineering,CAU)

Abstract The overall architecture and characteristics of intelligent text classifying system (ITC98) are introduced. Heuristic Encoding Subsystem of ITC98 designed with artificial intelligent text classifying method is described in detail.

Key words intelligent text classify; code; KNN

目前,Internet 上存在着海量的多媒体信息,但数量及信息含量最大的当属文本类信息,而网上信息的发布、检索和提取等信息处理过程都涉及基于信息内容的信息分类技术。文献[1]借鉴人类对文本分类的思维模式和过程,构造并实现了一个基于人工神经网络(ANN)的文本信息自动分类系统 ITC98。下面概述 ITC98 的总体结构及特点,重点介绍采用启发式编码方法建造的编码子系统。

1 ITC98 的总体结构及特点

ITC98 的总体结构如下:

"自动分词子系统",完成中文文献的词语切分处理,为下一步文本编码作准备。

"特征信息标注模块",完成对已切分文献中特征信息的标注。它依据各类别不同级别的特征信息库,标注特征信息的所属类别及相应级别。

"依据文本标题分类模块",在特征信息标注的基础上,计算文献标题中各类别权重,若标题中某一类别权重大于既定阈值,则系统调用采用标题特征信息加权统计的分类方法,完成文献分类。

收稿日期:1999-01-19
①清华大学智能技术与系统国家重点实验室开放课题基金资助项目
②陶 兰,北京清华东路 17 号 中国农业大学(东校区)64 信箱,100083

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

“文献编码子系统”,对采用 ANN 依正文内容完成分类的文献进行编码。

“基于 BP 网络的文本分类子系统”,将编码后的文献集输入 BP 分类器,完成学习与文本分类工作。

ITC98 具有以下特点:

1)借鉴人类文本分类的工作机制,针对不同文献特点分别采用标题特征信息加权统计和 BP 分类器 2 种不同的分类方法,不仅提高了系统效率而且保证了分类精度。

2)对待分类文献的格式、类型、系统运行环境,以及所用语料库均无特殊要求,适合在微机上完成各种文献的分类处理,实用性较好。

3)将分类处理后的文献以 html 文件格式输出保存,这样便于与其他网上信息处理系统联接,对网上信息做进一步的深层处理,如信息抽取、信息管理等。

4)用户可以依据分类需求定义文献类别主题,并建立相应类别特征信息库,系统会依据用户定义的类别主题完成文献分类。系统灵活性高,适用面广,易于推广。

5)对于有摘要的文献,ITC98 会主要依据摘要完成类别判定,以进一步提高效率。

2　ITC98 的编码子系统

ITC98 的文本分类功能主要由 ANN 中的 BP 网络实现。ANN 用于自然语言处理要解决的首要问题是,如何将语言信息编码成 ANN 可以处理的量化信息,因而,信息编码子系统的设计是整个文本信息处理系统中必不可少且至关重要的部分。笔者借鉴人类在文本分类过程中所采用的思维方式,设计了一种本文中称为启发式的编码子系统。

2.1　启发式编码子系统

启发式编码子系统的结构如下:

1)类别特征信息库及信息库维护更新模块。

人类进行文本信息分类,是以日常积累形成的类别判据为基础的。这些类别判据可能是类别特征词、类别特征结构,或类别特征概念点。启发式编码子系统首先以系统建造者或系统用户的类别判据知识为基础,建立类别特征信息库。系统管理员可随时更新特征信息库,以不断提高分类精度。

在启发式编码子系统中,根据分类精度的不同需求,将特征词库划分成不同的级别,为了实现量化,分别用 1,2,3,…表示特征词的级别,其中“1”表示最高级别,其余类推。

2)类别特征信息标注模块。

此模块以类别特征信息库为依据,对已切分文本进行标注,标注的内容包括特征信息所属类别以及重要级别。现举例说明如下:

|Dell|计算机|公司|5|月底|宣布[C1][P2]|,|将[C2]|于|今年|晚|些|时候|发布[P1]|工作|站|计划[C2]|。

上例中:*|计划[C2]|*,表示*计划*一词为 C 类 2 级特征词,而*|宣布[C1][P1]|*,表示*宣布*一词同时为 C 类和 P 类 1 级特征词。

3)特征信息编码模块。

在确定了标注词的编码类别和级别后,采用下式确定它们相应的特征信息编码:

$$W_{C_j}=2/(1+x_j)$$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

式中：W_{Cj}表示C类第j个特征词信息编码结果；x_j表示第j个特征词信息在此类中所属级别。例如上例中/*发布[P1]*/的编码结果为2/(1+1)=1，/*计划[C2]*/的编码结果为2/(1+2)=0.667。

4)文本信息编码模块。

在对类别特征进行编码的基础上，进一步对文本信息进行编码。为了压缩文本信息编码的维数，且突出文本类别特征，ITC98采用下式对文本信息进行编码：

$$T_i=\exp\sum_{j=1}^{N}W_{Cj}^{-1}$$

式中：T_i表示文本信息编码的第i维编码，它在编码中代表了某类特征信息；N表示C类特征信息的总个数。

2.2 基于Kohonen网络的编码测试与调整

为了测试并改进编码系统的性能，采用Kohonen网络进行了实验。实验的目的有2个：一是测试编码的结果能否使文本同类相聚；二是研究不能正确聚类的原因，并对编码方法进行相应的改进和调整。实验是在主频为166，内存为64 M的微机上进行的。

Kohonen网络具有自组织有序特征映射的功能，可利用这一特性来测试编码结果是否能使同类样本相聚而异类样本分离。

本实验选择网上《计算机世界报》国内外行业动态中的90篇文本作为实验样本，样本长短不等，主要以新闻题材为主。实验选择了3类样本，类别分别为“新产品介绍”、“企业动态”和“市场行情”，每个类别各有文本30篇左右。由于本实验所选择的3种样本类型内容比较接近，都属于计算机行业动态，所以分类难度相对较大。

采用Kohonen网络进行测试实验，首先选择网络输入为6个神经元，分别对应3类实验样本的2级编码，输出平面由400个神经元组成，其中x，y方向各20个神经元。

调用启发式编码子系统，完成实验样本的编码，编码结果举例如表1所示。将编码结果输入Kohonen网络，经自动聚类后，样本在输出神经元上的空间分布如图1所示。

表1 系统编码结果

文本编号	T_{C1}	T_{C2}	T_{P1}	T_{P2}	T_{M1}	T_{M2}
1	0.367 879	0.388 275	3.472 652	0.381 354	0.367 879	0.367 879
2	0.606 531	0.381 354	0.685 545	0.367 879	0.367 879	0.367 879
3	0.367 879	0.374 556	1.130 273	0.392 748	0.367 879	0.385 747
⋮						
86	0.367 879	0.367 879	0.367 879	0.395 322	4.436 374	0.422 045
87	0.685 545	0.374 556	1.000 000	0.367 879	15.405 980	0.419 297
88	0.685 545	0.374 556	0.685 545	0.381 354	8.267 206	0.469 382

说明：T_{C1}，T_{C2}，T_{P1}，T_{P2}，T_{M1}，T_{M2}分别表示3类实验样本的2级编码结果。

对图1显示的样本聚类分布情况可作如下分析。

1)3种不同图形，基本按类别聚集在一起，说明3类实验样本的编码基本可以按同类相聚，编码结果基本符合分类要求。

2)代表“新产品介绍”和“企业动态”的2类样本之间有重叠现象，这与此2类文献的类别特征容易混淆的客观情况相一致。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3)个别样本出现聚类错误。经查这些样本有其特殊情况,需进行个别调整。

a. 对于 P 类与 C 类样本交叉的情况,一方面要修改特征信息库,减少特征词之间的重叠,增加具有类别特征的动宾短语,尽量加大 2 类样本的距离;另一方面由于 2 类样本本身存在着相似性,难以将其准确分离,所以在实现分类时,可以用隶属度模糊表示分类结果,或者使此类文本同时出现在 2 类文献集中。

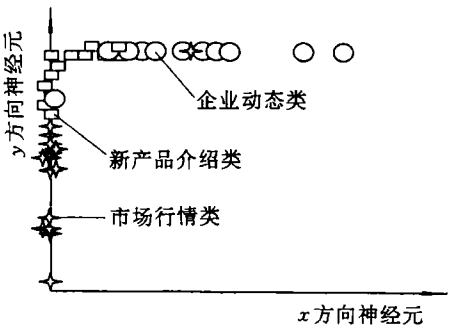

图 1　与输出样本对应的输出神经元分布

b. 对于个别出现错误分类的样本,视具体情况作如下调整:

样本中几乎没有类别特征信息存在时,可提取一些类别特征字或短语,加入特征信息库中,以增加类别判据;

样本中某 2 类特征信息均很多,难以判断类别时,要加大 1 级类别特征信息的编码权重,具体操作中可以将原 1 级特征信息编码权重加倍,再做累加计算。

对调整后的编码进行测试,问题可得到改善。

3　结束语

设计并实现了一种本文称之为启发式信息编码的子系统。实例测试结果表明,此系统的性能满足中文文本分类要求。它与 ITC98 的其他子系统共同实现了中文文本信息的自动分类。

参　考　文　献

1　申军霞. 文本信息自动分类系统及其应用研究:[学位论文]. 北京:中国农业大学,1998. 14～17,28～40

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net