基于KNN算法的医药信息文本分类系统的研究

许 幸,张启蕊
(广东药学院 医药信息工程学院,广东 广州 510006)

摘 要:针对目前医药信息文本分类领域的现状,设计并实现了一种基于KNN算法的医药信息文本分类系统。该系统充分利用了向量空间模型在表示方法上的优势和快速KNN算法的特点,并采用逆向最大匹配分词方法进行分词,可有效提高医药信息分类的准确性和信息处理效率。此外,构建了一个医药信息数据集,该数据集包含582篇医药类文本,其中训练文本433篇,测试文本149篇,并在该数据集上对医药信息文本分类系统进行了测试,得到了74.83%的F_1值。实验证明,该系统可以较好地实现医药信息文本分类。

关键词:医药信息;文本分类;向量空间模型;KNN算法
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2009)04-0206-04

Research of Medical Information Text Categorization Based on KNN Algorithm

XU Xing, ZHANG Qi-rui
(College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China)

Abstract: Designs and implements a system of medical information text categorization based on KNN algorithm. This system uses the vector space model to represent a text, uses the fast KNN algorithm to classify a text, and uses the reverse maximum match to segment the words. Therefore, it improves the accuracy of medical information classification and the efficiency of information processing. In addition, constructs a dataset of medical information including 582 medical documents, which is randomly divides into a training set including 433 documents and 149 documents. The system of medical information text classification is tested on our dataset and a F_1 score of 74.83% is obtained. The result shows the better classification performance on medical information.

Key words: medical information; text categorization; vector space model; KNN algorithm

0 引 言

医药信息历史悠久,累积了巨量的信息资源,大量传统的纸质信息转为电子文档形式保存,它容纳了医药海量的各种类别的原始信息。同时,在互联网上,电子文档医药信息每天都在急剧增加。如何在浩如烟海而又纷繁芜杂的医药信息文本中以最快的速度、最少的时间、掌握最有效的信息?据Forrest Research的统计资料指出,80%以上的数据以非结构化的形式存在[1]。因此,对非结构化数据的处理尤其显得重要。

面对海量信息,传统的做法是,对网上的信息进行人工分类,并加以组织和整理,为人们提供一种相对有效的信息获取手段。但这种人工分类的做法存在着许多弊端:一是耗费大量的人力、物力和精力;二是分类结果一致性不高。因此,自动文本分类成为处理海量数据的关键技术[2]。

文本分类在自然语言处理与理解、信息管理与组织、内容信息过滤等领域都有着广泛的应用。在文本自动分类中,著名的文本分类分类方法有支持向量机(Support Vector Machine, SVM)、K最近邻(K-Nearest Neighbor, KNN)、神经网络(Neural Network, NN)、线性最小二乘估计(LLSF)、贝叶斯算法(Bayes)和决策树等[3]。在这些方法中,KNN是一种简单、有效、非参数的方法,当训练样本数增加时,其分类时间将急剧增加,当词库增加,分类精度也会增加。同时,KNN方法也是一种基于实例的文本特征向量空间模型表示的分类方法[4]。

目前关于文本分类的系统基本上都是通用的文本分类系统,没有专门针对医药信息的文本分类系统。

收稿日期:2008-07-23
基金项目:广东省医学科研基金资助项目(B2008088);广东药学院科研基金资助项目(2007YGY01)
作者简介:许 幸(1984-),男,广东罗定人,助理工程师,研究方向为医药信息处理;张启蕊,博士,讲师,研究方向为信息处理、文本分类。

文中研究的基于 KNN 算法的医药信息文本分类系统,是针对医药信息文本自动分类而设计的系统,可以有效提高医药信息分类的准确性,大大提高信息处理效率,为医药信息搜索引擎提供基础。

1　KNN 算法

KNN(K－Nearest Neighbor)算法是机器学习领域的经典算法,其基本思想相当直观:把未知类别实例与训练集中的每个实例进行比较,找出最邻近的 k 个实例,通过选中的 k 个实例的类别来判断未知类别实例的类别[5]。

KNN 算法已在文本分类中得到了成功的应用,对给定的未知类别的文本,考虑在训练集中与该未知文本距离最近的 k 篇文本,根据这 k 篇文本所属的类别判定新文本所属的类别。

类别判断方法如下:

对找到的 k 篇文本,为每个类别打分,然后排序,只有分值超过指定阈值的类别才判定为文本 d 的类别。

KNN 算法在文本分类中的具体实现步骤如下:

① 根据特征项集合重新描述训练文本向量;

② 把待分类文本导入后,根据特征词分词该文本,确定待分类文本的向量表示;

③ 在训练文本集中选出与待分类文本最相近(相似)的 K 个文本,计算公式为:

$$d_{pq} = 2\sqrt{\sum_{h=1}^{n} w_h (a_{ph} - a_{qh})^2} \qquad (1)$$

其中,d_{pq} 是待分类文本 p 和训练样本 q 距离;n 是属性总数,a_{ph} 是待分类文本 p 中的第 h 个属性;w_h 是第 h 个属性的权重。

④ 在待分类文本的 K 个最相近(相似)的训练文本中,依次计算每类的权重,计算公式为:

$$p(x, c_p) = \sum_{p=1}^{n} \mathrm{sim}(a_p, x)\, p_a(a_p, c_q) \qquad (2)$$

其中,$\sum \mathrm{sim}(a_p, x)$ 是 x 的 k 个最近邻中的样本 a_p 和 x 之间的相似度。$p_a(a_p, c_q) = 1$,a_p 是类别 c_q 的样本;$p_a(a_p, c_q) = 0$,a_p 不是类别 c_q 的样本。

⑤ 根据公式(1)计算待分类文本 X 和每个训练样本的距离,选择与待分类样本距离最小的 K 个样本作为 X 的 K 个最近邻。根据公式(2)计算待分类文本与 X 的 K 个最近邻样本的权重。把各类的权重进行比较,把文本分类到权重最大的那个样本所属的类别中。

2　医药自动文本分类系统

文中设计的医药信息文本分类系统的训练和分类模型如图 1 所示。该模型包括两个模块:训练模块和分类模块。训练模块由预处理、文本表示、特征降维、分类器和性能评价五个部分组成,分类模块由预处理、文本表示和分类器三个部分组成。

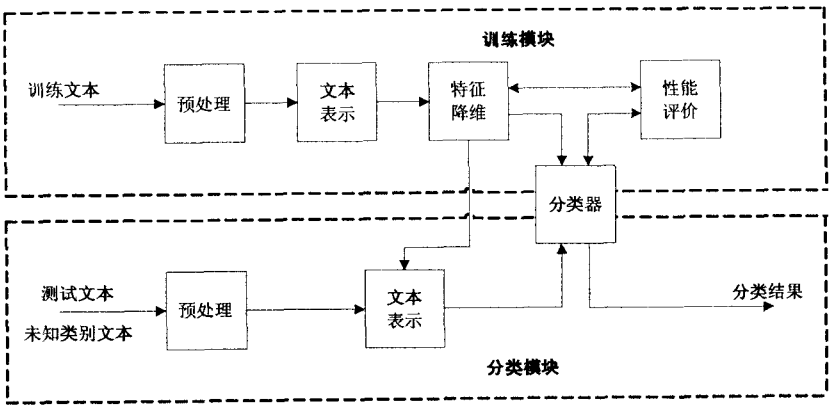

图 1　文本分类过程图

下面对该系统中关键模块采用的技术进行介绍。

2.1　技术路线

中文文本分类领域,预处理主要完成标点符号的去除、词的切分、停用词的删除等功能。中文分词是自然语言处理领域的一个研究热点和难题,常用的几种机械分词方法有正向最大匹配分词方法、逆向最大匹配分词方法、最少切分方法。一般说来,逆向匹配的切分精度略高于正向匹配,遇到的歧义现象也比较少。有研究显示,正向最大匹配的错误率为 1/169 左右,逆向最大匹配的错误率为 1/245 左右[6]。因此逆向最大匹配分词方法可以达到较好的分词效果。因此,文中设计的医药信息分类系统采用了逆向最大匹配法对医药信息文本进行分词。

医药信息文本分类的预处理主要包括分词和停用词处理两部分内容。分词是利用特定词典(通用集和医药信息专用集结合)进行分词,停用词处理则是利用禁用词集去除文档中的语义虚泛的禁用词,例如:"的","地","得"等。

为了配合本系统的逆向最大匹配分词方法,在建立医药信息分词字典时,采用长词优先的法则进行建立。即在收集医药信息词里,尽量把词语的长度也计算在内,如:"风湿性关节炎"收集为一个词,同时"风湿"与"关节炎"也收集在字典里。在建立字典时,把医药信息的词语与通用的词语相结合,这对医药信息文献的文本训练与分类提供更多的特征词的提取。查阅大量的医药类的书籍,经过分析,建立字典如下:吖啶、吖啶橙、吖啶黄、吖啶黄素、阿克拉霉素、阿霉素、阿米巴、阿米巴病、阿米巴痢疾、阿米卡星、阿米替林、阿米

妥、阿奇霉素、阿奇霉素分散片、阿奇霉素片、阿司咪唑、阿司匹林等。

另外，文本表示方法采用向量空间模型，特征选择方法使用 χ^2 方法，分类算法采用 KNN 算法。

2.2 数据集

根据国家的药品分类管理办法，药品的分类体系包括如：处方药与非处方药[7]。根据资料整理，现在得出以下分类：中药（安神药、补虚药、活血化瘀药、理气药、清热药等）、西药（维生素、抗病毒药、减肥药、解毒药、抗疟药、平喘药、抗真菌药等）、保健品（美容祛斑、营养强化、抗疲劳、调节血脂等）、医疗器械（护理设备、能量治疗器械、医用敷料、植入器械等）、仪器设备（包装设备、分析和检测仪器、粉碎机械、饮片机械等）等 125 类。文中根据实际情况，构建的数据集包括理气药、清热药、抗肿瘤药、化痰止咳平喘药和影响血液及造血系统的药物共五个类别，训练样本与测试样本的数量分布如表 1 所示。训练集和测试集彼此之间不重叠，不包括任何重复的文本。

表 1 数据集的各类别文本分布

主题类	理气药	清热药	抗肿瘤药	化痰止咳平喘药	影响血液及造血系统的药物
训练样本集	72	109	105	55	92
测试文本集	24	45	35	20	25

为了测试特征集规模对分类效果的影响，在选择理气药的特征词时总量与其它类别相对较少，而其他类别的特征词总量大体相当。然后对此语料库进行训练，最后使用测试集文本进行测试并进一步分析实验结果。

2.3 性能评价

系统评价采用经典的指标查准率、查全率、F_1 值进行评价，各指标定义如下：

准确率是某类别中所有判为该类的文本中分类正确的文本所占的比率，其计算公式为：

$$准确率=\frac{该类分类正确的文本数}{实际分到该类的文本数} \tag{3}$$

召回率是某类别中所有应分为该类的文本中分类正确的文本所占的比率，其计算公式为：

$$召回率=\frac{该类分类正确的文本数}{该类所有参与分类文本数} \tag{4}$$

准确率与召回率反映了分类质量的两个不同方面，二者必须综合考虑，不可偏废，因此，现在在分类系统中，存在一种综合评估指标，即为 F_1 测试值，其计算公式为：

$$F_1\text{测试值}=\frac{准确率\times 召回率\times 2}{准确率+召回率} \tag{5}$$

最后通过准确性、召回率、宏平均 F_1 值来衡量该分类系统的性能。

2.4 系统实现

文中设计的医药信息文本分类系统考虑了系统的可扩展性、用户的习惯以及系统的保密性等问题，包含了添加类别、用户管理等模块，系统功能模块图如图 2 所示。下面对主要模块分别进行介绍。

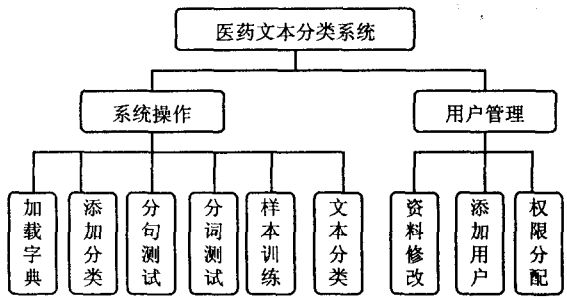

图 2 系统功能模块图

(1) 词典管理。

① 添加词典：把建立的词典加载记录在一个临时的哈希表里。

② 清除词典：把记录词典临时哈希表里的信息清除并删除这个临时的哈希表。

(2) 类别管理。

① 添加类别：增加训练样本的类别，并记录在类别.XML 里。

② 清空特征：清空该类别的特征表示，把该类别的 XML 表还原成原始状态。

(3) 训练管理。

① 加载样本：加载属于同一类的训练样本，并在 Textbox 中显示出来。

② 样本训练：把该类别的训练进行预处理，去掉停用词与干扰词，然后采用逆向最大匹配分词算法对每一个训练样本进行提取特征表示，并存储在该类别的 XML 文件中。

③ 继续学习：当该类别的训练样本增加时，不必重新训练，只需要把再学习的 Checkbox 勾上，然后把增加的样本加载，就可以进行再学习。

(4) 分类管理。

① 加载文本：把待分类文本加载并在 Textbox 中显示。

② 文本分类：把每一个待分类文本先进行预处理，然后进行文本特征提取，并记录在一个临时的哈希表，通过公式(1)计算每一个待分类文本与训练样本的距离，然后再通过公式(2)计算权重，把待分类文本分到权重最大的那一个样本所属的类别中，并显示结果。

(5) 后台管理。

① 系统管理：系统管理员对系统进行管理，包括系统的用户管理、添加删除与分配权限。

② 权限设置:本系统的权限为二级管理权限,系统管理员级别最高,可以进行任何操作,系统操作员只能对自己的信息进行维护并拥有对分类器进行操作的权限。

③ 修改密码:管理员与操作员都可以对自己的密码进行修改。

3　实验结果与分析

3.1　实验结果

使用文中构建的训练集对设计的系统进行训练,使用构建的测试集对医药信息分类系统进行测试,测试结果如表 2 所示,结果图如图 3 所示,图中 L 为理气药,Q 为清热药,K 为抗肿瘤药,H 为化痰止咳平喘药,Y 为影响血液及造血系统的药物,F 为宏平均值。

表 2　医药信息分类系统的分类结果

类别	理气药	清热药	抗肿瘤药	化痰止咳平喘药	影响血液及造血系统的药物	宏平均值
准确率	73.91%	74%	76.47%	63.16%	70.83%	71.67%
召回率	70.83%	73.33%	74.29%	60%	68%	69.29%
F_1 值	74.32%	73.66%	74.87%	61.54%	69.39%	70.46%

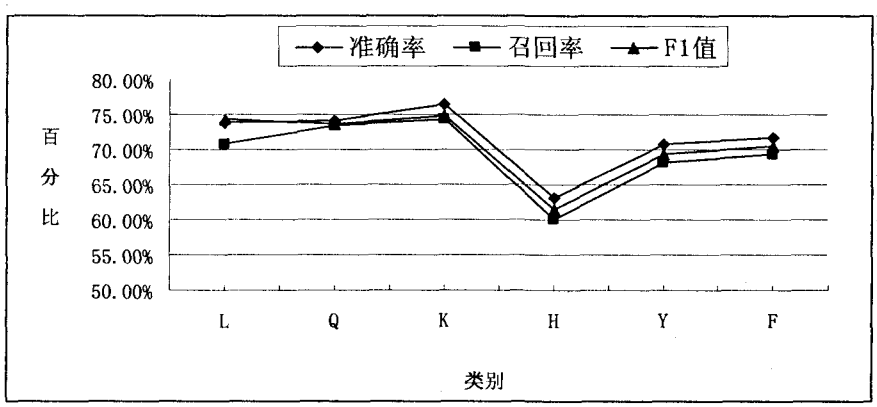

图 3　医药信息分类系统的分类结果图

3.2　结果分析

实验过程与结果显示,该系统的分类处理速度比较慢,反应了算法上还有待改进与优化,词典集的词量不够丰富,需要不断增加,分词的方法还存在着两个缺点:一是限制了词的长度,二是每次分词都有若干次无效循环(要一直循环到最大长度为止)。并且该系统对于类别差异性较大的类别具有较高的分类性能,而对于类别差异不大的类别的识别能力还有待提高。这说明文本的特征表示是影响分类系统性能的主要原因。

在实验中,发现特征词的选择与语料库情况关系密切,所以训练语料要反映一定的广度。在本系统进行设计的时候,一度认为每一类的训练库集大,提取的特征词越多越好,但经过测试后发现这种认识是有错误的。抽取的特征词太多,也会不利于文本的自动分类。因为有些特征词对于相对权重小的文本的分类会有干扰的作用。所以语料库的大小,需要经过大量的测试才可以确定。

4　结束语

构建了一个实验用医药语料库,并结合医药信息的特点利用 KNN 算法实现了医药信息文本分类系统。实验结果显示,在构建的数据集上该系统可以获得 74.83% 的 F_1 值。因此,该系统较好地实现了医药信息的自动分类,有效提高了医药信息分类的处理效率。但是,这还仅仅在于对医药信息自动分类的初步成功尝试,在接下来的工作中,将重点在构建更为丰富的医药数据集以及相应的分类算法方面进行深入的研究。

参考文献:

[1] 唐　菁,沈记全,杨炳儒. 基于 Web 的文本挖掘系统的研究与实现[J]. 计算机科学,2003,30(1):60-63

[2] 张启蕊,张　凌,董守斌,等. 基于免疫算法的文本分类研究[J]. 微计算机信息,2007,23(8):210-212.

[3] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.

[4] 王　煜,白　石,王正欧. 用于 Web 文本分类的快速 KNN 算法[J]. 情报学报,2007,26(1):60-64.

[5] 印　鉴,谭焕云. 基于 X2 统计量的 KNN 文本分类算法[J]. 小型微型计算机系统,2007,28(6):1094-1097.

[6] 杨　超. 分词技术研究报告[R/OL]. 2008-03. 教学资源网,计算机网络专栏, http://www.tingko.com/Lunwen/86083.html.

[7] 国家食品药品监督管理局. 处方药与非处方药分类管理办法[S/OL]. 1999-06-11. http://www.sda.gov.cn/WS01/CL0288/24524.html.