# 关键词提取的 K-means 方法在设备分类中的运用①

陈 立

(浙江工商大学 实验室与设备管理处, 杭州 310018)
(浙江大学 计算机与科学学院, 杭州 310058)

**摘 要**: 利用文本分类技术对设备进行分类目前遇到的最大困难是,信息处理量的急剧增加造成分类过程中设备特征项维数的大幅增加, 使得对设备的分类变得愈加困难, 且效率愈来愈低. 而关键词提取是提高文本分类效率的常用方法. 根据设备文本描述的特点, 以预先假定的初始关键词及其特征项词频来构建向量空间模型(VSM), 在此基础上利用 K-means 算法将文本中的关键词提取出来. 实验表明, 基于 K-means 的关键词提取不仅大幅度地提高了设备分类效率, 且分类准确性也得到了提高.

**关键词**: 设备分类; K-means; 关键词; 提取; 降维

## Keyword Extraction Method Based on K-means in the Application of Equipment Classification

CHEN Li

(Laboratory and Equipment Department, Zhejiang Gongshang University, Hangzhou 310018, China)
(College of Computer Science, Zhejiang University, Hangzhou 310058, China)

**Abstract**: Expository is the most common form in text classification. With the continuous increase of information, how to improve the efficiency of expository text classification is a hotspot in researching. Keyword extraction is used to improve the efficiency in text classification. This paper suggests to extraction keywords in the text based on the use of K-means algorithm by constructing the Vector Space Model (VSM) subject to pre-assumed initial keyword frequency. The KNN text classification algorithm verifies keywords extracting based on the use of K-means algorithm not only makes the classification accuracy slightly increased, but also reduce the dimension, thus realizing the text classification efficiency.

**Key words**: equipment classification; K-means; keyword; extraction; dimension reduction

仪器设备分类的重要性在于能实现规模采购、精细管理、合理维修等目的, 如不进行科学合理的分类管理, 势必造成人力、物力和财力的巨大浪费. 就笔者所从事的设备管理工作来看, 目前的政府集中采购, 是先将用户所提交的设备进行集中, 然后再进一步进行分类, 当每一类达到一定规模后即可进行招标采购. 这一分类过程目前是采用计算机对设备进行文本分类[1,2]. 但随着设备数量及种类的增加, 以及分类算法中设备样本的增加, 文本分类中通常采用的向量空间模型(VSM)[3], 因向量维数可能涉及训练集中的所有词汇, 而使得计算机需要处理的数据量非常巨大, 从而限制了文本分类方法在设备采购中的应用. 由此也可以看出, 对向量维数实施降维是实现文本分类效率提升的必要前提[4].

为此, 本文提出一种基于 K-means 聚类的关键词提取方法. 由于关键词是对文档内容的高度概括, 能全面反映文档的内容和主题[5], 所以关键词提取可以有效地实现降维. 本文将以各部门工作中应用十分广泛的设备说明文本为例, 通过关键词提取来提高分类效率. 显然, 若将每一类文档的关键词挖掘出来, 会给计算机快速高效检索及有效组织资源等带来极大帮助[6].

① 基金项目:浙江省社会科学界联合会研究课题成果(2014Z052);浙江省教育厅科研项目(Y201432308)
收稿时间:2015-01-16;收到修改稿时间:2015-03-12

## 1 基于K-means聚类的关键词提取思路

典型的设备文本依照一定的顺序层次进行说明, 有的由现象到本质, 有的由主到次, 有的按事物的性质、功用、原理等顺序来说明. 包括设备名称、功能及参数等等. 如某部门提交的电脑说明书, 如图 1 所示, 就是一种典型的说明文体. 其中, “主板”、“处理器”、“内存”、“硬盘”、“显卡”、“声卡”、“网卡”、“显示器”等词能够很好地反映“计算机”这一设备的特征, 可将其视为图 1 所示文本中的关键词.

**品名:**计算机(商用机)
**配置项目及要求:**⑴主板:英特尔 H81 Express 芯片组及以上, 板型结构:M-ATX, 支持系统电压、温度、风扇运行侦测, ⑵处理器:英特尔酷睿 i5, 核心数量:四核, 主频:3.2GHz 6 MB 三级高速缓存, ⑶内存:1*4G DDRIII 1600 内存及以上, ⑷硬盘: 1TB 7200 rpm SATA 6Gb/s, 类型: SATA 串行, 转速: 7200 转/分钟, ⑸显卡:芯片组:NVIDIA GeForce 705, 显存容量: 1G DDRIII, ⑹声卡: Realtek ALC887, 声道数:7.1 声道, ⑺网卡: 集成 10/100/1000 以太网卡+Dell 无线-N 1705 2.4GHz + 蓝牙 4.0, ⑻显示器:尺寸 19 寸非宽屏, 面板:液晶面板, 分辨率: 1600*900, 响应范围:5ms, 亮度:200cd/m2, LED 背光, ⑼保修:原厂三年免费上门保修服务, ⑽其它:提供系统备份恢复功能; 可同时保留系统出厂备份和用户自定义备份; 可设置管理员密码, 防止非法备份与恢复,

图 1 计算机类设备的说明文本

可以发现, 以上述计算机说明文本为例, 作为关键词, 都是一些频繁出现在各种计算机说明文本中、用于对计算机进行描述的词. 其特征表现在: ①在同类设备说明书中大量出现, 且词的用法变化较少. 相反, 非关键词, 由于使用人的不同, 以及计算机型号的不同, 其用法上差别也会相当大. 如图 1 文本中的“英特尔酷睿”这类描述设备参数的词, 在另一个电脑说明书中, 就可能被描述为“AMD A85X”. 也就是说, 图 1 文本中的关键词“处理器”、“内存”、“硬盘”等词在各种电脑说明书中呈现时, 其变化一般要小于图 1 文本中描述词部分的“酷睿”、“四核”等词. ②对关键词进行描述的词可作为关键词的特征词. 某些出现频率较高的特征词, 可以反映出某一类关键词. 而类似于图 1 文本第 8 项中的 “液晶”、“宽屏”、“亮度”、“响应”等词, 就会在对“显示器”这个关键词进行描述时出现的频率较高, 所以“液晶”、“宽屏”、“亮度”、“响应”等词可以称作显示器的特征词.

利用上述特征①的关键词与非关键词出现的频率不同,将高频词假定为初始关键词, 基于初始关键词构建某类文本的向量空间模型, 在此基础上利用特征②使用 K-Means 算法进一步聚类出关键词.

## 2 基于K-means聚类的关键词提取模型

### 2.1 K-Means 算法[7]

K-Means 算法是一种常用的聚类算法. 它是基于划分的一种算法, 基本思路为: 对给定的聚类数目 $k$, 首先随机创建一个初始划分, 然后采用迭代方法通过将聚类中心不断移动来尝试改进划分.

设数据集合 $X=\{x_1,x_2,\cdots,x_n\}$

$k$ 个聚类中心分别为 $z_1,z_2,\cdots,z_k$

用 $w_j(j=1,2,\cdots,k)$ 表示聚类的 $k$ 个类别. 有如下定义.

定义1 两个数据样例之间的欧式距离为:

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^T(x_i-x_j)}$$

定义2 属于同一类别数据样例的算术平均为:

$$z_j=\frac{1}{N}\sum_{x\in w_j}x$$

定义3 目标函数为:

$$J=\sum_{i=1}^{K}\sum_{j=1}^{n_j}d(x_j,z_i)$$

K-Means算法描述如下:

输入: 聚类个数 $k$ 以及包含 $n$ 个数据样例的数据集合; 输出: 满足目标函数值最小的 $k$ 个聚类算法流程:

① 从 $n$ 个数据样例中任意选择 $k$ 个样例作为初始聚类中心.

② 循环下述流程③到④, 直到目标函数 $J$ 取值不再变化.

③ 根据每个聚类样例的均值(中心样例), 计算每个样例与这些中心样例的距离, 并且根据最小距离重新对相应样例进行划分.

④ 重新计算每个聚类的均值(中心样例).

### 2.2 关键词提取算法步骤

#### 2.2.1 以初始关键词构建文本空间向量模型

① 初始关键词的给定

所谓的初始关键词是指在关键词提取出来前假定的一些关键词. 对于初始关键词的给定, 首先将同属

一类的所有文本合并为一个大文本，如，现有计算机类文本 102 篇、打印机类文本 87 篇、交换机类文本 76 篇、……，将上述计算机类文本、打印机类文本和交换机类文本按照类别各自合成一个大文本，最终各类文本各有一篇大文本，有几大类就有几篇文本.

其次，将文本的词频求出. 目前很多文本分类研究都采用经典的 TF_IDF 方法来求词频[8]，该方法使得在越少类别文档中出现的特征词，其权重越大，这样就可以将那些不具备类别区分能力的词，如“的”、“地”，以及英文的“a” “an”等词排除.

再次，按照词频的大小由高到低排序. 我们发现大部分类别的文本中，词频最高的 8%～10%的词中含有了 80%～90%的关键词，图 2 证实了这一点. 该图为人工统计的结果，横轴表示词频最高的词占整个文本的比例，纵轴表示所选出的词中关键词占整个文本的关键词的比例. 所以我们将文本中词频最高的前 10%的词作为初始关键词.

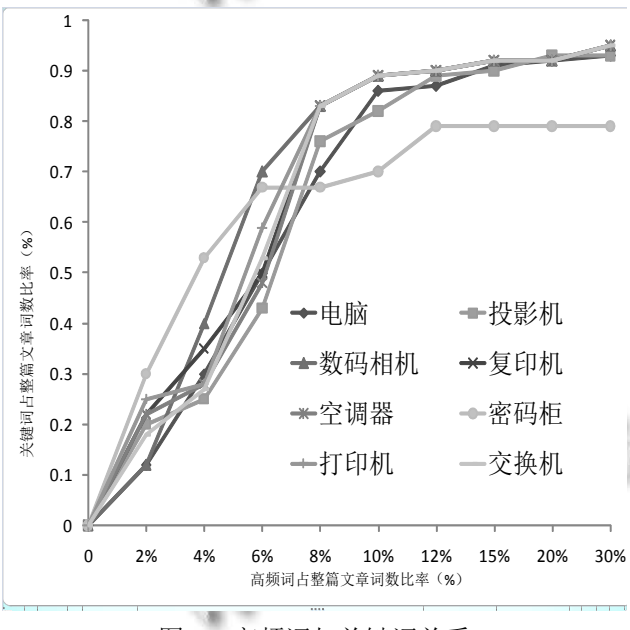


图 2 高频词与关键词关系

② 向量空间模型的给定

在信息检索领域中，向量空间模型(VSM)[3]最为简洁有效，影响力也最大. 不论是关键词提取还是文本分类，都需要首先建立文本的向量空间模型. 在某些情况下,对关键词进行描述的、被称为特征项的词，其自身也有可能是该类文本的关键词. 那么，如何寻找这种关键词的特征项呢？事实上，对关键词进行描述的特征项既可能紧随关键词之后，也可能出现在关键词之前，表现为和关键词之间的固定搭配关系. 如图 1 所示的“宽屏”这个词，其前面的“尺寸”就是“宽屏”这个关键词的特征词. 因此，为了将关键词尽可能提取出来，我们将以每一个初始关键词为基础，将其周围词作为特征词. 在某一个初始关键词周围出现的特征词频率越高，该初始关键词就越具备某类关键词特征. 我们将这些特征词的词频作为空间向量的项，这样就形成了该类文本的空间向量模型. 初始关键词的确定和向量空间模型的给定具体如下.

对某个类的所有文档进行计算，以每一个初始关键词为中心计算出其前后初始关键词、之间的词，也就是特征项出现的频率，其中为第个初始关键词，为初始关键词周围的第个词. 据此每类文档构成如下的二维矩阵,一个类文本集可以表示成一个的词. 文本矩阵:

$$A=(W_{ij})_{m\times n}\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ & & \cdots & \\ W_{m1} & W_{m2} & \cdots & W_{mn} \end{bmatrix}=(X_1,X_2,...,X_n)=\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} \quad (3)$$

其中每一列 $X_j$ 代表一个关键词及其周围特征词的一个组合，每一行 $Y_i$ 代表特征词在各个初始关键词中的权值，$W_{ij}$ 表示第 $i$ 个词在第 $j$ 个初始关键词周围出现的权重.

2.2.2 $k$ 值的设置

空间向量模型(VSM)[3]构建好后，在进行K-means聚类算法前需要先确定聚类数 $k$ [9]. K-means聚类对初始聚类中心极为敏感，如果初始聚类中心选择不当，算法很容易陷入局部最优解，而非全局最优解.

对于 $k$ 值得多少，我们采用经验法，聚类的个数 $k$ 一般小于样本个数的平方根[10]，因此，一般可以取初始点为文本总数的平方根. 在这里我们选择初始关键词数的平方根作为 $k$ 值.

2.2.3 利用 K-means 聚类算法求出关键词

初始关键词中包含了许多非关键词，我们采用 K-Means 算法进一步排除一些非关键词，以达到提取关键词的目的. 利用 K-means 聚类算法得到聚类中心点,去除相同的词后，即获得文本的关键词. 图 3 显示了在一次实验中利用 K-Means 算法对计算机类文本进行关键词提取后的文档.

主板 英特尔 芯片 支持 系统 Windows 中文 运行 CPU 处理器 数量 主频 GHz MB 三级 高速 缓存 内存DDRIII 内存 以上 硬盘 TB 类型 SATA 串行 转速 独立 显卡 显存 容量 声卡 网卡 集成 显示器 寸 液晶 LED 标 键盘 无线 键鼠 电池 保修 原厂 服务 备份 恢复

图 3　计算机类文档提取的关键词

## 3　实验设置

### 3.1 实验设计

为了验证关键词提取方法在文本分类中的降维效果，我们采用 KNN 算法进行验证. KNN(K Nearest Neighbor)算法是向量空间模型中最好的分类算法之一[11,12],具有分类精度高、结构简单的优点，但也存在不足，如分类时间过长、每次分类都需要训练[13]，所以导致该算法对文本的维度很敏感，当维度增加时效率会大幅下降. 正因为如此，KNN 算法可以很好地反映出本文关键词提取在文本分类中的降维效果. 在本次实验中，我们使用以 K-means 算法提取的关键词来取代所对应类别在训练集当中的文本，用以构建在 KNN 算法中所使用的向量空间模型，并与不采用 K-means 算法进行比较. KNN 算法首先在已知类别样本中寻找与待分类样本最相似的个样本. 文本样本之间的相似性可以通过文本向量之间的余弦来度量，表达如下:

$$Sim(d_i,d_j)=\cos\theta_{ij}=\frac{d_i^T\cdot d_j}{\|d_i\|\|d_j\|}=\frac{\sum_{k=1}^{m}w_{ki}w_{kj}}{\sqrt{\sum_{k=1}^{m}w_{ki}^2}\sqrt{\sum_{k=1}^{m}w_{kj}^2}} \quad (2)$$

其中，$d_i$ 为测试文本的特征向量，$d_j$ 为第 $j$ 类的中心向量，$m$ 为特征向量的维数，$k$ 的值确定须根据试验测试结果进行调整，$w_{ki}$ 和 $w_{kj}$ 是表示文本 $d_i$ 和 $d_j$ 文本向量中相应特征项的权重.

KNN算法可基于这 $k$ 个已知类别样本的类别属性对未知样本的类别做出预测,将未知样本的类别预测为在这个 $k$ 最近邻样本中包含最多实例的类别. 即:

$$C(X)=\arg\max_i\{n(d_j,c_i)|j\in[1,k]\} \quad (3)$$

式中 $n(d_j,c_i)$ 表示最近邻中属于 $c_i$ 类别的样本个数，$k$ 表示 $X$ 的最近邻的个数，一般认为 $k$ 的值为总文本数的2%左右时分类性能最好[14]. 本文选取KNN的 $k$ 值为总文本数的2%.

### 3.2 数据来源

本次实验的设备语料库，是笔者所在高校以往所采购的设备文本,经过人工整理，共计 32 类设备 1567 个文本. 我们挑选出其中占比最大的 7 类设备进行最终实验效果的比较. 这 7 类设备分别是电脑、投影机、数码相机、复印机、空调器、密码柜、打印机、交换机，共计 561 个文本，每类文本平均词数分别为: 电脑(265)、投影机(298)、数码相机(262)、复印机(179)、空调器(74)、密码柜(97)、打印机(176)、交换机(367). 分词过程采用开源的ICTCLAS汉语分词系统，并去除相关停用词.

### 3.3 评价方法

信息检索、分类等领域两个最基本指标是召回率(Recall Rate)和准确率(Precision Rate)，召回率也叫查全率，准确率也叫查准率[15]，公式表达如下:

召回率(Recall) =系统检索到的相关文件/系统所有相关的文件总数

准确率(Precision) =系统检索到的相关文件/系统所有检索到的文件总数

查全率和查准率反映了分类质量的两个不同方面，两者必须综合考虑. 因此,综合考虑查准率和查全率得到一个新的评估指标 F1 测试值，其计算公式如下:

F1 测试值=(查准率＊查全率＊2)/(查准率+查全率).

## 4　实验结果及分析

### 4.1 关键词提取结果

我们使用关键词提取准确率及关键词筛选的遗漏率，作为判断k-means算法关键词提取效果的指标. 其中，关键词提取的准确率=实际的关键词数/经 K-means 算法提取的关键词数; 遗漏率=K-means 算法未能提取出的关键词数/实际的关键词数. 在上述两个表达式中，“实际的关键词”是指 K-means 算法所提取的关键词中，由人工判断去除其中的非关键词，所剩下的就是“实际的关键词”; 而“K-means 算法未能提取出的关键词”是指，单纯由人工去筛选的关键词中，K-means 算法提取所未能涉及的关键词. 显然，准确率越高及遗漏率越小，都表明 K-means 算法提取关键词的效果越好.

如图 4 所示，实验结果为,关键词准确率约为 78.21%,关键词筛选的遗漏率约为 17.33%. 而提取的

关键词数占原文本词数比例约为 25.11%.

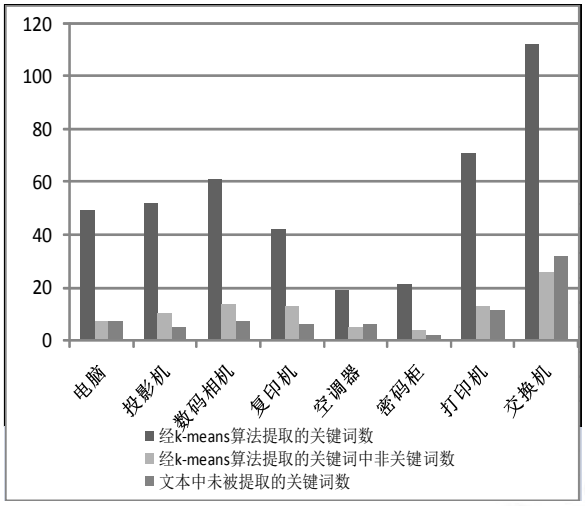


图 4 关键词提取结果分析

通过分析发现, 造成关键词提取不够完全准确的原因主要有三点: ①在实验中使用的工具误差影响, 如分词系统中对某些词语造成过度切分, 像“操作系统”可能会被拆分为“操作”和“系统”两词. ②文本中对关键词的描述太过随意, 如在“密码柜”类文本的关键词提取中, 因为对关键词描述太过随意, 使得在使用聚类算法时聚类过于分散, 导致其遗漏率就较高. ③文本中对设备的描述不够单一, 如在“打印机”类和“交换机”类文本的关键词提取中, 因为这些产品品种较多、差别较大, 像打印机分为针式打印机、喷墨打印机、激光打印机等, 由于这些设备的工作原理、制造方式相差较大, 文本描述时使用的关键词差别也较大, 若无差异地将这些设备放在一起, 则关键词提取效果变差, 若能将打印机进一步细化分类, 则效果可变得较好.

### 4.2 分类结果

表 1 分类结果

| 类别 | 文本篇数 | 不采用关键词提取的KNN算法 | | | 采用关键词提取的KNN算法 | | |
|---|---|---|---|---|---|---|---|
| | | 准确率 | 召回率 | F1值 | 准确率 | 召回率 | F1值 |
| 电脑 | 102 | 82.9 | 79.6 | 81.2 | 85.4 | 82.2 | 83.8 |
| 投影机 | 62 | 89.1 | 82.3 | 85.6 | 95.3 | 93.8 | 94.5 |
| 数码相机 | 56 | 78.8 | 72.4 | 75.5 | 78.2 | 77.9 | 78.0 |
| 复印机 | 32 | 89.2 | 84.0 | 86.5 | 93.7 | 86.2 | 89.8 |
| 空调器 | 78 | 90.4 | 88.9 | 89.6 | 91.6 | 93.9 | 92.7 |
| 密码柜 | 68 | 68.2 | 62.4 | 65.2 | 71.5 | 59.2 | 64.8 |

显然, 采用关键词提取的KNN算法与不采用关键词提取的KNN算法相比, 前者能减少 70%～80%的特征向量, 其分类评价指标也表现出了更好的效果.

## 5 结语

本文提出了一种针对设备文本的新的关键词提取方法. 用初始关键词周围的词作为特征项来构建文本的向量空间模型, 在此基础上用 K-means 算法对初始关键词进行聚类, 将聚类出的中心点作为某类设备的关键词. 从实验的结果来看, 关键词准确率较高. 同时, 与传统的分类方法相比, 该方法降维效果明显, 分类效果略有提高. 可见, 该方法具有较强的实际应用意义.

不过, 虽然基于 K-means 的关键词提取方法其降维效果较明显, 但这种方法的适应性还有一定局限, 如其对文本内容要求规范, 且相对单一等. 本文的后续研究将逐步解决上述问题, 进一步研究提高设备分类的准确性.

## 参考文献

1 Sebastiani F. Maehine learning in automated text eategorization. ACM Computing surveys, 2002, 34(l):l–47.

2 陈立.基于贝叶斯分类的高校设备批量集中采购.实验技术与管理,2014,31(5):265–268.

3 陈治纲,何丕廉,孙越恒,郑小慎.基于向量空间模型的文本分类系统的研究与实现.中文信息学报,2004,19(1):36–41.

4 Seo YW, Sycara K.Text clustering for topic detection. CMU-RI-TR-04-03. Pittsburgh: Robotics Institute, Carnegie Mellon University, 2004.

5 刘俊,邹东升,邢欣来,李英豪.基于主题特征的关键词抽取.计算机应用研究,2012,29(11):4224–4227.

6 罗杰,陈力,夏德麟,王凯.基于新的关键词提取方法的快速

文本分类系统.计算机应用研究,2006,31(4):32–34.

7 MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press.1967: 281–297.

8 Sebastiani F. Machine learning in automated text categorization .ACM Computing Surveys, 2002, 34(1):1–47.

9 张健沛,杨悦,杨静,张泽宝.基于最优划分的初始聚类中心选取算法.系统仿真学报,2009,21(9):2586–2590.

10 Vesanto J, Alhoniemi E. Clustering ofthe self-organizing map. IEEE Trans. on Neural Networks, 2000, 11(3): 586 – 600.

11 Yang Y, Liu X. A re-examination of text categorization methods. Proc. of the 22nd Annual International ACMSIGIR Conference on Research and Development in InformationRet rieval(SIGIR.99). Berkeley: ACM Press, 1999: 42–49.

12 HeJ, Tan AH, Tan CL. A comparative study on Chinese text categorization methods. Proc. of the International Workshop on Text and Web Mining. Singapore: Melbourne, 2000. 24–35.

13 张晓辉,李莹,王华勇,赵宏.应用特征聚合进行中文文本分类的改进 KNN 算法.东北大学学报自然科学版,2003,24(3):229–232.

14 于瑞萍.中文文本分类相关算法的研究与实现[硕士学位论文].西安:西北大学,2007.

15 Van Rijsbergen K. Information Retrieval. Butterworths, London,1979.