K-means 聚类算法的研究和应用

熊志斌，朱剑锋，王冬
（琼州学院电子信息工程学院，海南 三亚 572022）

摘　要：介绍了 K-means 算法的思想，分析了在文档聚类中运用 K-means 算法的步骤。以开源的机器学习软件 Weka 为平台，详细论述在 Weka 上进行文档聚类的前端处理过程，利用搜狗语料库中的文档在 Weka 上进行了 K-means 算法的聚类测试。实验结果表明，K-means 算法在 Web 文档聚类中表现出较好的效果。根据实验结果，分析了 K-means 算法存在的不足和聚类分析中特征选择的重要性。

关键词：K-means 算法；聚类分析；Web 文档

Research and Application of K-means Clustering Algorithm

XIONG Zhi-bin，ZHU Jian-feng，WANG Dong
(College of Electronics and Information Engineering Qiongzhou University, Hainan Sanya 572022 , China)

Abstract: This paper introduced the principal of the K-means algorithm, and presented the method by which K-means algorithm is applied to document clustering. The front-end process of documents clustering was discussed in detail, based on Weka that is an open source machine learning software. The K-means clustering algorithm was tested on Weka with documents corpus from Sogou Corporation, and results from experiments show that the K-means algorithm is effective in Web document clustering. According to experimental data, the defect of K-means algorithm was discussed, and importance of feature selection was discussed in unsupervised clustering.

Key words: K-means algorithm ; Clustering analysis ; Web documents

1　引言

聚类就是对数据集中的数据分组，使得组内数据具有高度的相似性，而组间的数据有较大的相异性。聚类分析是一种重要的数据分析方法，广泛应用于数据挖掘、模式识别、机器学习等领域。

K-means 算法是一种简单快速的聚类算法。在文本聚类领域，K-means 聚类算法已经成为一种基本算法[1]。

Weka（Waikato Environment for Knowledge Analysis）是一款开源的基于 Java 语言的机器学习和数据挖掘软件[2]，由新西兰 Waikato 大学开发。以 Weka 平台为基础，分析了 K-means 算法在文档聚类中的应用，在 Weka 平台实现文档聚类。首先用文档向量空间模型[3]表示文档，对 Web 文档进行数量化的描述。在文档向量化的过程中，特征词的选取至关重要，前人在这方面做了一些研究工作。研究表明[4]，特征词过多导致高维数据，不仅增大了空间开销，还加重机器学习的负担，特征词过少，影响聚类的效果。最后利用 Weka 平台中对 Web 文档进行聚类测试。实验结果表明 K-means 聚类算法对 Web 文档聚类有较好的效果。

2　K-means 算法

2.1　算法思想

1967 年 Macqueen 提出了 K-means 算法[10]，基本思想是把数据集中的数据点随机生成 k 组，把每组的均值作为中心点。重新计算每个数据点与各组的中心点的相似性，根据数据点相似性的度量准则，把每个数据点重新分组，计算每组新的均值作为中心点。不断重复上述过程，直到中心点的均值收敛，停止迭代过程。

K-means 算法是一种比较快速的聚类方法，时间复杂度为 O（nkt），其中 n 是数据点的数目，k 是分组数目，t 是迭代次数。K-means 算法也存在不足，最大问题要指定分组数目并且在运行过程中容易导致局部最优。

2.2　算法在文档聚类中的应用

K-means 算法在不同的领域都有成功的运用。运用 K-means 算法进行文档聚类，首先需要对文档建立文档表示模型，VSM (vector space model) 模型是一种常用的文档表示模型。VSM 模型用向量表示文档，文档转换成向量数据，可以利用 K-means 算法实现文档聚类。算法流程如下：

输入：文档向量集 $D=\{d_1, d_2, \cdots, d_n\}$，聚类个数 k

输出：k 个聚类

s1：从文档向量集 D 中随机取 k 个向量作为 k 个聚类的中心点 C_1，C_2，…，C_k。

s2：遍历文档向量集 D 中的向量 d_i，计算 d_i 与 C_j (j=1, 2, …, k) 的相似度，把 d_i 重新分配到最相似的组。

s3：根据 s2 得到新的 k 个聚类，重新计算每个聚类中的向量的均值作为中心点 C_1，C_2，…，C_k。

s4：比较聚类的中心点。

s5：如果中心点位置不再变化，则停止迭代；否则，转入 s2。

作者简介：熊志斌（1973-），男，讲师，硕士，研究方向：人工智能；朱剑锋，副教授，硕士；王冬，副教授，硕士。

收稿日期：2014-01-12

算法中需要建立文档相似性函数和聚类效果评价函数。文档的相似性表征了文档之间的相关程度。度量文档的相似性普遍采用两类方法，一种是基于距离的度量方法，一种是基于相似系数的度量方法[9]。基于距离的度量方法包括欧式距离、曼哈顿距离、明考斯基距离。基于相似系数的度量方法包括余弦相似系数、Jaccard 系数。在文档聚类中，通常采用余弦相似系数作为文档相似的度量值，公式如下：

$$Sim(d_i,d_j)=\frac{\sum_k d_{i,k}\times d_{j,k}}{\sqrt{(\sum_k d_{i,k}^2)\times(\sum_k d_{j,k}^2)}} \quad (1)$$

公式中 d_i 表示聚类中第 i 个文档向量，$d_{i,k}$ 表示向量 d_i 中的第 k 个分量。计算值越大，文档相似度越高。

聚类效果评价函数通常采用平方误差和，公式如下：

$$V=\sum_{i=1}^{k}\sum_{d_i\in S_i}(d_i-c_i)^2 \quad (2)$$

公式中 S_i 表示第 i 个聚类，d_i 表示聚类 S_i 中的向量，c_i 表示聚类 S_i 的均值。函数值越小，表明聚类内部越紧密。

3 基于 Weka 平台的 K-means 聚类

K-means 算法可以用来实现 Web 文档的聚类。Weka 平台实现了包括 K-means 算法在内的一些聚类算法。利用 Weka 平台实现文档聚类只需要做一些文档的前端处理工作，生成指定格式的文件，再调用 Weka 中的 K-means 算法，即可以完成文档的聚类分析。文档的前端处理工作包括文档分词化、去停用词、生成文档集词语表，根据词语表统计每篇文档的词频，建立词频矩阵；选择特征词生成特征向量；对每篇文档计算特征词的权重，完成文档的向量化。具体流程如图 1 所示。

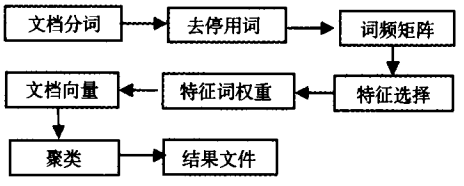

图 1 文档聚类流程

3.1 文档预处理

文档预处理包括文档分词化、去掉停用词、生成词频矩阵。首先对文档分词化，分词软件采用中科院张华平老师开发的 NLPIR 汉语分词系统 (又名 ICTCLAS2014 版) [11]。所有的文档被切分成词语，去掉对聚类分析无意义的停用词。停用词采用哈工大停用词表。在去停用词过程中，把单字词全部作为停用词处理。经过分词、去停用词处理后，得到文档集词表。

词频是指词语在一篇文档中出现的次数。对文档集中的每篇文档的进行统计，统计文档集词表中的词语在每篇文档中的出现次数，得到一个词频矩阵。

3.2 特征选择

K-means 聚类属于无监督的机器学习，由于事先不知道类别的信息，文档的特征词只能采用无监督的特征选择算法。常见的无监督的特征选择算法包括 3 种[5]，文档频数 (document frequency, DF)，单词权 (term strength, TS)，单词熵 (entropy-based feature ranking, EN)。

文档频数 (document frequency, DF)，词语的文档频数是指文档集中包含该词的文档数。文档频数的假设前提是，出现次数过多或过少的词语对区分类别没有贡献，删除这些词语可能有助于提高聚类结果。该算法时间复杂度 O (n)，适合海量数据处理。在实际应用中，文档频数的上限值和下限值的设定没有可靠的理论依据，可以根据实验结果做调整。

单词权 (term strength, TS) 该方法的理论假设是，词语在相关的文本中出现的频率越高，在不相关的文本中出现的频率越低，该词的对类别区分越重要。单词权的计算不依赖类信息，适合用于无监督的文本聚类。计算单词权，首先必须计算所有文本对之间的相似度，该算法的时间复杂度较高，至少是 O (n^2)。

单词熵 (entropy-based feature ranking, EN) 是专门用于聚类问题的特征选择方法。该方法的理论假设是，不同的词语对数据的结构或分布的影响是不同的，单词越重要对数据的结构影响也就越大，不重要的词语对数据的结构几乎没有贡献。词语熵的时间复杂度为 O (mn^2)，不适合海量数据的处理。

在第 4 节的实验中，采用文档频数 (DF) 方法选择特征词。根据第 3.1 中的词频矩阵，把文档集词表中的高频词和低频词删除掉，得到文档集的特征词向量。

3.3 文档向量化

VSM 模型是 G.Soltn 等在 1975 年提出来的一种文档表示模型，最先用在文档检索的过程中。VSM 模型可以运用到文档聚类领域里，设 D 是文档集合，对于 $d_j\in D$，则文档 d_j 可以用向量表示成：

$d_j = (w_{1,j}, w_{2,j}, \cdots, w_{n,j})$

公式中 $w_{i,j}$ 表示第 i 个特征词在第 j 篇文档中的权重。通过计算向量之间的相似度就可以判断文档是否属于同一类别。

建立向量模型的要点是在特征词的选取和特征词权重的计算。特征词权重的计算最基本的模型是 TF-IDF (term frequency-inverse document frequency) 模型[12]。

TF 表示某个特征词在某篇文档中出现的次数。计算公式如下：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \quad (3)$$

公式中 $n_{i,j}$ 表示第 i 个特征词在第 j 篇文档中的出现次数，而分母则是在第 j 篇文档中所有特征词的出现次数之和。

IDF 是一个特征词普遍重要性的度量。计算公式如下：

$$idf_i=\log\frac{N}{n_i} \quad (4)$$

公式中 N 表示文档总数，n_i 表示是包含特征词 i 的文档数量。特征词权重 $w_{i,j}$ 计算公式如下：

$$w_{i,j}=tf_{i,j}\times idf_i \quad (5)$$

利用 3.2 节中所选的特征词向量、3.1 节中的词频矩阵和本节的公式 (5)，对文档集合 D 中的每篇文档计算特征词向

量的权重，完成文档的向量化，所有的文档形成一个向量集。

3.4 调用 Weka 中的 K-means 算法

Weka 平台已经实现了一些基本的聚类算法，其中包括 K-means 算法。利用 Weka 平台完成文档聚类，要求把文档向量集写成 arff (Attribute-Relation File Format) 格式的文件，作为 Weka 的输入数据。关于 arff 格式的规范，在 Weka 的联机文档和官方网站都详细的介绍。把文档向量集转换成 arff 格式文件后，把生成的 arff 格式文件加载到 Weka 平台，利用平台的可视化界面，按聚类过程操作步骤，设置聚类的类别数目和种子数目，完成文档的聚类分析。通过调用 Weka 的 visulize cluster assignments 功能，图形化地观察聚类结果，然后保存聚类结果，以便程序分析文档的聚类效果。在利用 Weka 完成聚类分析时，也可以在 Java 语言编写的程序中直接调用 Weka 软件包中相关类，得到聚类结果。

4 实验与结果分析

在文档聚类分析中，查准率和查全率是对聚类效果进行评价的最基本的最常用的指标。查准率和查全率的计算公式如下：

$$查准率 P = \frac{n_i}{S_i} \tag{6}$$

$$查全率 R = \frac{n_i}{N_i} \tag{7}$$

公式中 n_i 表示聚类后形成的第 i 个聚类中与类别相关的文档数目，S_i 表示聚类后形成的第 i 个聚类中的全部文档数目，N_i 表示第 i 个聚类中与类别相关的全部文档数目。

实验中，使用搜狗语料库的精简版作为测试数据来源。搜狗语料库是搜狗实验室从因特网搜集的文档，进行人工分类后的文档集，从精简版选择了 5 大类，每类 50 篇文章，作为测试数据集，对每篇文章的文件名作了类别标注，以便程序计算实验结果查准率和查全率。

测试步骤如下：

(1) 生成 arff 格式的文件。

(2) 在 Weka 中进行聚类分析，保存分析结果。

(3) 计算分析结果的查准率、查全率。

测试过程中，开始设定种子数为 10，聚类数为 5。反复测试了 6 次，每次测试种子数增加 1，每次测试结果不一样。在 Weka 平台，以平方误差和 (sum of squared errors) 作为聚类评价指标，该值越小表明聚类效果越好。选取平方误差和最小的一次，实验结果如表 1 所示。

实验数据表明 K-means 算法在 Web 文档聚类中具有较好的聚类效果。6 次测试中每次结果不一样，表明聚类的结果不稳定，与种子的数目和选择有关，但实验数据上也没有呈现出种子数目越多平方误差和越小的趋势。为了到达较好的聚类效果，需要反复测试几次，选取平方误差和最小的一次作为实验结果。

表 1 实验结果

	军事	体育	IT	财经	教育
查准率 P	0.516	0.708	0.655	0.689	0.825
查全率 R	0.64	0.68	0.72	0.62	0.65

5 结语

研究了 K-means 聚类算法，以 Weka 平台为基础，把 K-means 聚类算法用到 Web 文档聚类中。传统的 K-means 聚类算法要求事先指定聚类数目，在 Web 文档聚类的实际应用中是无法事先知道合适的类别数目的。如何让机器自动确定聚类数目是下一步的研究方向。特征词的选择直接关系到聚类的准确性和算法时间空间效率。聚类是个无监督的学习过程，事先不知到类别信息，因此特征词选择也是无监督的。在无监督的聚类过程，如何选择高区分度的特征词也是以后要开展的工作。Weka 已经实现包括 K-means 聚类算法在内的大量的与数据挖掘相关的算法，其源代码开放，可以充分利用 Weka 平台开放性，在其基础上做二次开发，改进 K-means 算法，扩聚类算法。

参考文献

[1] 吴夙慧，成颖，郑彦宁，潘云涛. K-means 算法研究综述 [J]. 现代图书情报技术, 2011, (5): 28-35.

[2] Weka Data Mining Software in Java [EB/OL].

[3] 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 [J]. 计算机应用研究，2001, 18 (9): 17-26.

[4] 刘海峰，吴功宜，陈正. 一种高效的用于文本聚类的无监督特征选择算法 [J]. 计算机研究与发展，2005，42 (3): 381-386.

[5] 龚静，曾建一. 文本聚类中的特征选择方法 [J]. 吉首大学学报 (自然科学版)，2008，29(2).

[6] 王正，刘培玉，刘克非. 一种用于 Web 文本聚类的特征选择方法 [J]. 计算机应用与软件，2007，24 (1): 154-156.

[7] Yan, Pedersen J O. A comparative study on feature selection ext categorization [C] //ICML. 1997, 97: 412-420.

[8] 李匹，满自斌. 自适应特征权重的 K—means 聚类算法 [J]. 计算机技术与发展, 2013, 6: 027.

[9] 樊东 基于文本聚类的特征选择算法研究 [D]. 西北师范大 2012.

[10] Maeen, James. " Some methods for classification and anal of multivariate observations." Proceedings of the fifth berkeley symposium on mathematical statistics and probity. Vol. 1. No. 281-297. 1967.

[11] NLP汉语分词系统 [EB/OL].

[12] Marg C D, Raghavan P, Schütze H. Introduction to information retrieval [M]. Cambridge: Cambridge university press, 2008.