K-Means聚类方法在简历数据中的应用

冯丽娜 周丽华
(云南大学信息学院计算机科学与工程系 昆明 650091)

摘 要 文本挖掘(Text Mining)是一个从非结构化文本信息中获取用户感兴趣或者有用的模式的过程。其中简历数据是一类内容以个人基本信息为主,语言精短、简约、明快,并且具有很强目的性的文本数据。针对简历数据区别于其他文本数据简短、明确的特点,采用聚类分析中的K-Means方法对其进行聚类分析。实验结果表明,将K-Means聚类算法应用在简历数据中,聚类结果符合用户需求,且具有实际意义并对人力资源效率的提高有一定帮助。

关键词 简历数据,特征提取,K-Means,聚类
中图法分类号 TP311 **文献标识码** A

Application of K-Means in the Curriculum Vitae Data

FENG Li-na ZHOU Li-hua
(Department of Computer Science and Engineering, Information College, Yunnan University, Kunming 650091, China)

Abstract Text Mining is a process to get an interesting or useful model from the unstructural text information. Curriculum vitae data is one kind of text data which mainly contains the personal information. The character of this data is visibly exact, short, simple and vivid. Besides, the motive of curriculum vitae is very intense and obvious. Aiming at the particular character of curriculum vitae data we choose the K-Means clustering method to make the clustering analysis. The experiment results show that the application of K-Means in the curriculum vitae data can provide an undersandable description of the discovered clusters and do some help with the enhance of working efficiency.
Keywords Curriculum vitae data, Feature extraction, K-Means, Clustering

1 引言

随着互联网技术的飞速发展,网络招聘已经成为各大公司企业实施招聘的主要渠道。就招聘管理流程而言,简历筛选环节无疑是最耗费人力的环节。招聘专员每天要收到上百甚至更多份简历,据不完全统计:人力资源从业者每天要花费至少1~3小时筛选不符合企业需求的简历,致使人力资源工作效率低下。

基于对这些问题的考虑,我们将K-Means聚类算法应用在简历自动评审中,解决工作效率低下的问题。首先,通过对简历数据自身特点的分析,选择适用于简历数据的特征提取方式,进而获得每份简历文档的特征向量表示,最后采用K-Means聚类算法完成聚类工作。对聚类后的简历数据可以根据类别分给各部门专员进行选择,这样在一定程度上提高了人力资源的工作效率。

本文的主要特点在于:

- 分析了简历数据区别于其他文本数据的独有特点;
- 总结了专门适用于简历数据的停用词列表;
- 算法简单易实现,且针对简历数据参数易确定。

本文第2节阐述相关工作;第3节给出处理简历数据的完整过程;实验结果在第4节给出;最后是结论与未来工作的展望。

2 相关工作

文档聚类主要是依据著名的聚类假设:同类的文档相似度较大,而不同类的文档相似度较小。目前,许多不同的文本聚类算法已经被提出,例如:Scatter/Gather[1],Suffix Tree Clustering[2]和二分K-Means[3]等,这些算法虽然有比较好的聚类结果,但是对某些类的描述却并不符合用户的本意,也就是说很多算法的应用是需要特定领域的数据分析为背景的,不同的文本数据采用同样的聚类算法效果不同。

在对文本的预处理过程中,特征函数也要根据

冯丽娜(1984-),女,硕士研究生,主要研究方向为数据挖掘,E-mail:Denaanny@126.com;周丽华 副教授,主要研究方向为数据挖掘与数据仓库。

具体领域数据的特点来选择，通常用到的特征评估函数有以下几种形式：文档频度、信息增益(IG)、互信息(MI)、χ^2 统计(CHI)及术语强度 TS 等，通过对中文文本语料分析，各有其优缺点[1]。本文选择文档频度方法作为特征评估函数，主要因为其算法简单、计算量小，且符合简历数据的特点。在简历数据中大多数低频词为噪音，而 TF-IDF 方法对低频词的处理效果较好，可以提高聚类精度。

目前对于简历数据的聚类方法研究不足，鉴于此种情况，本文采用经典的 K-Means 算法对简历数据进行了聚类处理。

3 K-Means 聚类算法在简历数据中的应用

简历数据的主要处理流程如图 1 所示。

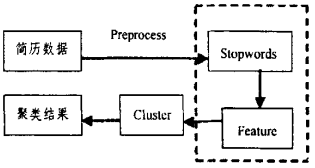

图 1 简历数据处理流程

方法：

(1)获取文档输入；

(2)初始化分词器，移除停用词；

(3)初始化 TF-IDF 测量器，生成每个文档的 TF-IDF 权重；

(4)初始化 K-Means 算法(data,K)；

(5)迭代；

(6)获取聚类结果，输出。

3.1 简历数据特点

经过分析发现简历数据具有以下一些特点：内容简短，每份简历文档字数在 500～800 之间；内容目的性较强，并不存在大量的虚词，相反，某些能标识文本特性的实词在简历数据中反而成了不具实际意义的词；相似文档中包含的特征词比较相近，且代表性较强；低频词偏多，且低频词对主题文档的代表性较低。

文本数据中的停用词是指文本中出现频率很高，但对确定文本特性没有具体意义的词，主要指连词、介词、语气词等。因此，对这些词或词组的去除不仅不影响整篇文档特征向量的提取，反而会提高关键词密度、简化计算、提高搜索效率。通常在文本数据中使用频率较高的中文停用词有 800 多个，如："啊"，"其他"，"但是"等等。但针对于简历数据，由于它自身具有以上那些特点，传统的中文停用词表对它并不适用。

在对简历文本预处理过程中，我们归纳并总结了适用于简历数据的停用词包。它与通用的中文停用词有较大的区别：首先在内容上，例如，在简历数据当中出现频率较高的字段"姓名"，"民族"，"出生年月"，"联系地址"等等均被列入停用词行列，但在大多数的文本数据当中，这些词在很大程度上都是具有重要意义的实词。其次是在词量上，简历数据内容精短且常用字段比较集中，所以它的停用词量相对于通用的中文停用词量要小很多。

3.2 特征提取

在去除停用词之后，为了更精确地表示文档，进一步降低向量空间维数，防止过分拟合，对剩余的词还需进行特征的提取。这是文本聚类过程中至关重要的环节，特征提取的好坏直接影响着聚类结果的精确度。特征提取的主要过程是通过构造一个特征评估函数，把测量空间的数据投影到特征空间，得到在特征空间的值，然后根据特征空间中的值对每个特征进行评估，选择值最高的若干个特征作为文本的特征向量[5]。

本文采用 Salton 提出的 TF-IDF 值计算文档特征权重，其中 $tf(t,d)$ 表示词 t 在文本 d 中的绝对频度；$idf(t)=\log(N/n(t))$ 表示反比文档频率，其中 N 为文档总数，$n(t)$ 为词 t 出现的不同文档的总数。则某词 t 在文档 d 中的权重如式(1)所示：

$$w_i(d)=tf(t,d)\cdot idf(t) \tag{1}$$

3.3 简历数据的向量表示及聚类算法

我们采用一种比较常用的文本特征表示模型——向量空间模型(Vector Space Model,VSM)。在该模型中，文本空间被视为一组向量组成的向量空间，而每一份文档被表示成高维空间中的一个点。为了便于对不同长度的文档进行比较，获取特征集合后，文档的特征向量在 VSM 中会被归一化成统一的长度，它的具体形式为：

$$d_i=(w_1,w_2,w_3,\cdots,w_n)$$

其中，n 为特征空间的维数，w_i 表示当前文档在第 i 维上所占的权重。

根据词频矩阵中词频的相似相关程度计算文档间的相似度，并使用欧式距离作为文档相似的衡量尺度。式(2)如下：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(|x_i-y_i|)^2} \tag{2}$$

其中，x,y 为两个空间向量，n 为空间向量的维度。

具体的聚类过程如下：

算法 1 K-Means

输入：

K：聚类个数；

V：文档特征向量集；

输出：聚类结果。

方法：

[1] 从 V 中任选 K 个数据对象为初识簇中心；

[2] Repeat

[3] 根据簇中对象均值，将每个特征向量归到最相似

的簇中；

[4] 更新簇均值；

[5] Until 不再变化。

4 实验

根据国内各大院校开设的课程，所修专业涵盖36个大的类别和方向，例如：电子信息类、经济学类、哲学类、工商管理类等等。而每个类别当中细分了很多专业，如电子信息类包括：电子信息技术、微电子学、电子商务等等。类别中专业与专业之间关联是紧密的，因为开设的课程相关度度较高，而不同类别之间开设的课程相关程度非常小。因此我们期望对简历数据的聚类结果与专业类别的划分是一致的。

本文的实验数据从网络上获得，下载了5组类别相异的求职简历。每组包含8份简历文档，共计40份简历文档。所有文档经过分词和停用词处理，分词采用了正向最大匹配的方法，所用的部分停用词列表如图2所示，停用词处理结果对比如图3所示。

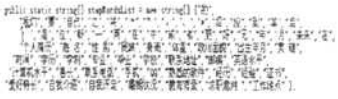

图2 简历数据的停用词列表

停用词处理前

停用词处理后

图3 停用词处理前后数据对比

经过TF-IDF方法选取的类别及类别部分特征词如表1所列。

表1 类别及类别部分特征词

类别	特征词列表
1	女 男 工商管理 本科 北京 调查 行政 会计 技术 财务 经济 文化 产业 企业 责任心
2	女 公共管理 政治学 行政 管理 公共 社会学 事业 劳动 保障 政策 团体 科研 事务 服务
3	女 汉 公关 文秘 大专 中文系 写作 公共 档案 办公 事务 管理 责任心 服务 自信 自律 礼仪
4	电子电路 电路 设计 实践 通信 技术 物理 基础 经验 信心 设施

通过K-Means聚类算法对40篇简历数据的特征向量进行聚类，结果如图4所示。

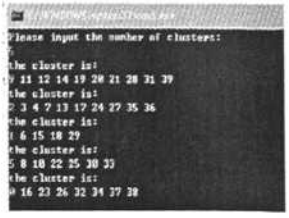

图4 聚类结果

类别数 K 取值为5，每一类所得到的文档数分别为：10,10,5,7,8。对聚类结果进行分析发现：1）类别间存在相似专业的类，聚类结果有一定偏差，像第一类工商管理类、第二类公共管理类和第三类文学类别中，管理学专业、行政管理专业和文学专业某些课程或实践经历相近，实验所得相似度较大，导致数据聚类不够准确；2）类别间差别较大的专业聚类结果较为准确，比如：第四类电子信息类与第五类医学类；3）因简历数据选取有限，并没有进行更加广泛全面的测试，还需要更进一步的改进。

结束语 本文分析了简历数据区别于其他文本数据简短、明确、目的性强的独特特点，并针对简历数据的这些特点，总结出适用于简历数据的停用词表，选择TF-IDF特征提取方式进行简历文档特征向量的建立，最后通过K-Means聚类方法对简历数据进行聚类分析。实验结果表明，将K-Means聚类方法应用在简历数据当中是切实可行的，且聚类结果提供了可理解的实际意义。

在未来的工作当中，我们希望可以采用语义图来进行特征的提取，并通过K-Means聚类方法对聚类结果进行分析、比较，完善K-Means算法在简历数据中的应用效果。

参 考 文 献

[1] Cutting D R, Karger D R, Pedersen J O, et al. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collection[C]// Proc. ACM SIGIR 92. 1992: 318-329

[2] Zamir O, Etzioni O. Web Document Clustering: A Feasability Demonstration[C]// Proc. ACM SIGIR 98. 1998: 46-54

[3] Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques[C] // Proc. Text Mining Workshop. KDD 2000, 2000

[4] Hong S, Baoliang L, Masao U, et al. Comparison and improments of feature extrction methods for text categorization. Computer Emulation, March 2006

[5] Lewis D D. Feature selection and feature extraction for text categorization[A] // Proceedings of Speech and Natural Language Workshop[C]. February 1992: 213-216

K-Means聚类方法在简历数据中的应用

作者： 冯丽娜， 周丽华
作者单位： 云南大学信息学院计算机科学与工程系 昆明 650091