第 34 卷 第 18 期 计 算 机 工 程 2008 年 9 月
Vol.34 No.18 Computer Engineering September 2008
·软件技术与数据库· 文章编号：1000—3428(2008)18—0039—03 文献标识码：A 中图分类号：TP301.6

基于向量空间模型的文本聚类算法

姚清耘，刘功申，李 翔
(上海交通大学信息安全工程学院，上海 200240)

摘 要：文本聚类是聚类的一个重要研究分支，是聚类方法在文本处理领域的应用。该文探讨了基于向量空间模型的文本聚类方法，提出了一种文本聚类的改进算法——LP算法。同时，基于语料库的实际聚类效果，就维度确定、特征选择等方面提出优化方案。实验证明，LP算法有效地减少了聚类所消耗的时间，实用性和灵活性都较高。

关键词：向量空间模型；文本聚类；语料库

VSM–based Text Clustering Algorithm

YAO Qing-yun, LIU Gong-shen, LI Xiang
(School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240)

【Abstract】 Text clustering, one of the most important research braches of clustering, is the application of clustering algorithm in text processing. This paper discusses different Vector Space Model(VSM)-based clustering algorithms and presents an improved text clustering algorithm——Level-Panel(LP) algorithm. In addition, according to the effects of clustering for the corpus, it presents optimizations of clustering algorithm, including dimension determining, feature selection, etc. It is proved that LP algorithm can effectively reduce the time spending in clustering process. It is high in practicability and flexibility.

【Key words】 Vector Space Model(VSM); text clustering; corpus

1 文本聚类研究现状

Internet 已经发展为当今世界上最大的信息库和全球范围内传播信息最主要的渠道。随着 Internet 的大规模普及和企业信息化程度的提高，各种资源呈爆炸式增长。在中国互联网络信息中心(CNNIC)2007 年 1 月最新公布的中国互联网络发展状况统计报告中显示，70.2%的网络信息均以文本形式体现。对于这种半结构或无结构化数据，如何从中获取特定内容的信息和知识成为摆在人们面前的一道难题。近年来，文本挖掘、信息过滤和信息检索等方面的研究出现了前所未有的高潮。

作为一种无监督的机器学习方法，聚类技术可以将大量文本信息组成少数有意义的簇，并提供导航或浏览机制。

文本聚类的主要应用点包括：

(1)文本聚类可以作为多文档自动文摘等自然语言处理应用的预处理步骤。其中比较典型的例子是哥伦比亚大学开发的多文档自动文摘系统Newsblaster[1]。该系统将新闻进行聚类处理，并对同主题文档进行冗余消除、信息融合、文本生成等处理，从而生成一篇简明扼要的摘要文档。

(2)对搜索引擎返回的结果进行聚类，使用户迅速定位到所需要的信息。比较典型的系统有 Infonetware Real Term Search。Infonetware 具有强大的对搜索结果进行主题分类的功能。另外，由 Carrot Search 开发的基于 Java 的开源 Carrot2 搜索结果聚合聚类引擎 2.0 版也是这方面的利用，Carrot2 可以自动把自然的搜索结果归类(聚合聚类)到相应的语义类别中，提供基于层级的、同义的以及标签过滤的功能。

(3)改善文本分类的结果，如俄亥俄州立大学的Y.C.Fang等人的工作[2]。

(4)文档集合的自动整理。如Scatter/Gather[3]，它是一个基于聚类的文档浏览系统。

2 文本聚类过程

文本聚类主要依据聚类假设：同类的文档相似度较大，非同类的文档相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程、以及不需要预先对文档手工标注类别，因此具有较高的灵活性和自动化处理能力，成为对文本信息进行有效组织、摘要和导航的重要手段。文本聚类的具体过程如图 1 所示。

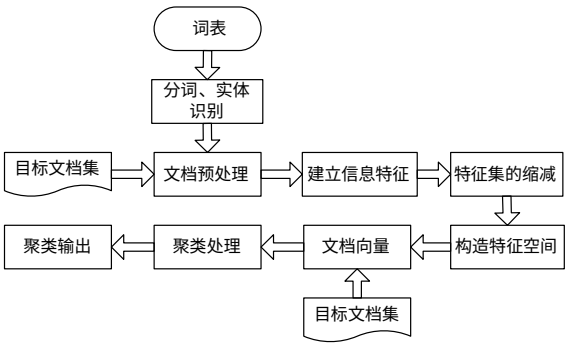

图 1 文本聚类过程

2.1 文本信息的预处理

文本聚类的首要问题是如何将文本内容表示成为数学上可分析处理的形式，即建立文本特征，以一定的特征项(如词条或描述)来代表目标文本信息。

要建立文本信息的文本特征，常用的方法是：对文本信

基金项目：国家自然科学基金资助项目(60502032, 60402019)；教育部新世纪优秀人才支持计划基金资助项目(NCET-06-0393)

作者简介：姚清耘(1981－)，女，硕士研究生，主研方向：文本挖掘；刘功申、李 翔，副教授

收稿日期：2007-10-20 **E-mail：**qyyao@sjtu.edu.cn

息进行预处理(词性标注、语义标注)，构建统计词典，对文本进行词条切分，完成文本信息的分词过程。

2.2 文本信息特征的建立

文本信息的特征表示模型有多种，常用的有布尔逻辑型、向量空间型、概率型以及混合型等。其中，向量空间模型(Vector Space Model, VSM)是近几年来应用较多且效果较好的方法之一[4]。1969 年，Gerard Salton提出了向量空间模型VSM，它是文档表示的一个统计模型。该模型的主要思想是：将每一文档都映射为由一组规范化正交词条矢量张成的向量空间中的一个点。对于所有的文档类和未知文档，都可以用此空间中的词条向量 $(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$ 来表示(其中，T_i 为特征向量词条；W_i 为 T_i 的权重)[5]。一般需要构造一个评价函数来表示词条权重，其计算的唯一准则就是要最大限度地区别不同文档。这种向量空间模型的表示方法最大的优点在于将非结构化和半结构化的文本表示为向量形式，使得各种数学处理成为可能。

2.3 文本信息特征集的缩减

VSM 将文本内容表示成数学上可分析处理的形式，但是存在的一个问题是文档特征向量具有惊人的维数。因此，在对文本进行聚类处理之前，应对文本信息特征集进行缩减。通常的方法是针对每个特征词条的权重排序，选取预定数目的最佳特征作为结果的特征子集。选取的数目以及采用的评价函数都要针对具体问题来分析决定。

降低文本特征向量维数的另一个方法是采用向量的稀疏表示方法。虽然文本信息特征集的向量维数非常大，但是对于单个文档，绝大多数向量元素都为零，这一特征也决定了单个文档的向量表示将是一个稀疏向量。为了节省内存占用空间，同时加快聚类处理速度，可以采用向量的稀疏表示方法。假设确定的特征向量词条的个数为 n，传统的表示方法为 $(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$，而稀疏表示方法为 $(D_1,W_1,D_2,W_2,\cdots,D_p,W_p,n)$ $(W_i\neq 0)$。其中，D_i 为权重不为零的特征向量词条；W_i 为其相应权重；n 为向量维度。这种表示方式大大减小了内存占用，提升了聚类效率，但是由于每个文本特征向量维数不一致，一定程度上增加了数学处理的难度。

2.4 文本聚类

在将文本内容表示成数学上可分析处理的形式后，接下来的工作就是在此数学形式的基础上，对文本进行聚类处理。

文本聚类主要有 2 种方法：基于概率[6]和基于距离[7]。基于概率的方法以贝叶斯概率理论为基础，用概率的分布方式描述聚类结果。基于距离的方法，就是以特征向量表示文档，将文档看成向量空间中的一个点，通过计算点之间的距离进行聚类。

目前，基于距离的文本聚类比较成熟的方法大致可以分为 2 种类型：层次凝聚法和平面划分法。

对于给定的文件集合 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_n\}$，层次凝聚法的具体过程如下：

(1)将 D 中的每个文件 d_i 看成一个具有单个成员的簇 $c_i=\{d_i\}$，这些簇构成了 D 的一个聚类 $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$；

(2)计算 C 中每对簇 (c_i,c_j) 之间的相似度 $sim(c_i,c_j)$；

(3)选取具有最大相似度的簇对 $sim(c_i,c_j)$，将 c_i 和 c_j 合并为一个新的簇 $c_k=sim\,c_i\cup c_j$，从而构成了 D 的一个新的聚类 $C=(c_1,c_2,\cdots,c_{n-1})$；

(4)重复上述步骤，直至 C 中剩下一个簇为止。

该过程构造出一棵生成树，其中包含了簇的层次信息以及所有簇内和簇间的相似度。

对于给定的文件集合 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_n\}$，平面划分法的具体过程如下：

(1)确定要生成簇的数目 k；

(2)按照某种原则生成 k 个聚类中心作为聚类的种子 $S=(s_1,s_2,\cdots,s_j,\cdots,s_k)$；

(3)对 D 中的每个文件 d_i，依次计算它与各个种子 s_j 的相似度 $sim(d_i,s_j)$；

(4)选取具有最大相似度的种子 $\underset{c_i,c_j\in C}{\operatorname{argmax}}\,sim(d_i,s_j)$，将 d_i 归入以 s_j 为聚类中心的簇 c_j，从而得到 D 的一个聚类 $C=\{c_1,c_2\}$；

(5)重复此步骤若干次，以得到较为稳定的聚类结果。

这 2 种类型各有优缺点。层次凝聚法能够生成层次化的嵌套簇，准确度较高。但在每次合并时，需要全局地比较所有簇之间的相似度，并选出最佳的 2 个簇，因此执行速度较慢，不适合大量文件的集合。而平面划分法相对来说速度较快，但是必须事先确定 k 的取值，且种子选取的好坏对群集结果有较大影响。

综合考虑这 2 种聚类类型的优缺点，本文提出了一种基于向量空间模型的文本聚类的改进方法——LP 算法。具体过程如下：

对于给定的文件集合 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_n\}$：

(1)将 D 中的每个文件 d_i 看作是一个具有单个成员的簇 $c_i=\{d_i\}$；

(2)任选其中一单个成员簇 c_i 作为聚类的起点；

(3)在其余未聚类的样本中，找到与 c_i 距离满足条件的 d_j (可以是与 c_i 距离最近的点，即相似度 $sim(c_i,d_j)$ 最大的 d_j，也可以是与 c_i 距离不超过阈值 d 的点，即相似度 $sim(c_i,d_j)\geq d$ 的任意 d_j)。将 d_j 归入 c_i 形成一个新的簇 $c_k=sim\,c_i\cup d_j$；

(4)重复步骤(3)，直至与 c_i 距离最近的 d_k 与 c_i 之间的距离超过阈值 d，此时认为已经聚完了一类；

(5)选择一个未聚类的单个成员簇，重复步骤(3)和步骤(4)，开始新的一轮聚类，直至所有的单个成员簇 c_i 都参与了聚类。

LP 算法不需要比较所有簇之间的相似度，执行速度较快，适合大量文件的集合，实用性更高。同时，在聚类过程中不需要事先确定 k 的取值，降低了与领域知识的依赖性，提高了灵活性。

3 实验设计

本文采用搜狐研发中心搜狗实验室的互联网语料链接关系库 SOGOU-T。该关系库提供了一个大规模互联网链接关系对应表，用于验证各种链接关系分析算法的有效性与可行性。

语料关系库中的数据分为 10 大类(C000007 汽车，C000008 财经，C000010 IT，C000013 健康，C000014 体育，C000016 旅游，C000020 教育，C000022 招聘，C000023 文化，C000024 军事)。

语料关系库可供下载的共有 3 个版本：Mini 版，精简版，

完整版。本文使用前 2 个版本进行实验。

语料库的组织方式如下 :为 10 个大类各建立 1 个文件夹，在每个文件夹中，每 1 份语料自成 1 个.txt 文件。

实验过程如下：

(1)将所有文件夹下的.txt 文件随机连结成一个大的完整文件，同时保留.txt 文件的所属类别(本实验保留了类别的最后 2 位：07, 08, …)。

(2)采用中国科学院计算技术研究所数字化室&软件室发布的中文自然语言处理开放平台汉语词法分析系统 ICTCLAS。利用 ICTCLAS_Win，将(1)中的文件进行一级标注的词语切分。

(3)统计标注好的切分词语的词频。

(4)按照权重(词频)的大小整理切分词语，并保留权重超过一定限定值(阈值)的特征项。(本实验保留了词频大于 100 的词语作为特征项)同时，根据汉语的特点，在实验中设计了 2 种情况，以分析比较词性对于聚类效果的影响：

1)所有类型的词语都参与聚类；

2)只保留被标注为名词的词语。

(5)根据(4)中确定的切分词语构造空间向量的基向量，同时确定空间向量的维数等参数。

(6)将语料库中的每一份语料文件(.txt 文件)都表示为一个空间向量。在实验过程中，采用了如下 2 种表示方法：

1)传统的空间向量表示方法：$(T_1,W_1,T_2,W_2,\cdots,T_n,W_n)$；

2)稀疏的空间向量表示方法：$(D_1,W_1,D_2,W_2,\cdots,D_p,W_p,n)$。

(7)聚类：聚类过程是实验的重点，也是目标所在。

1)在开始聚类前，首先对(6)中已经表示好的文本空间向量做归一化处理。向量归一化在模式识别中是很重要的一环，其目的是把事件的统计分布概率统一归纳在 0-1 灰色聚类的隶属性上，这样，聚类过程对于每一个空间向量的敏感度都是一样的。

传统空间向量：$\boldsymbol{X}=(T_1,\frac{W_1}{d_1(\boldsymbol{X})},T_2,\frac{W_2}{d_1(\boldsymbol{X})},\cdots,T_n,\frac{W_n}{d_1(\boldsymbol{X})})$，其中 $d_1(\boldsymbol{X})=sqrt(W_1^2+W_2^2+\cdots+W_n^2)$；

稀疏空间向量：$\boldsymbol{X}=(D_1,\frac{W_1}{d_2(\boldsymbol{X})},D_2,\frac{W_2}{d_2(\boldsymbol{X})},\cdots,D_p,\frac{W_p}{d_2(\boldsymbol{X})})$，其中 $d_2(\boldsymbol{X})=sqrt(W_1^2+W_2^2+\cdots+W_p^2)$。

2)在实验中，采用欧几里德距离来表示任意 2 个文本向量之间的距离。

传统空间向量：令 $\boldsymbol{X}=(T_1,x_1,T_2,x_2,\cdots,T_n,x_n)$，$\boldsymbol{Y}=(T_1,y_1,T_2,y_2,\cdots,T_n,y_n)$，则 $d_1(\boldsymbol{X},\boldsymbol{Y})=sqrt(\sum_{i=1}^{n}(x_i-y_i)^2)$；

稀疏空间向量：计算方法与传统空间向量类似，计算相同词条之间距离平方和的算术平方根。

3) LP 算法要求预先确定阈值。实验中，采取的阈值策略是：制定初始阈值(即针对单个成员簇的阈值，此阈值根据实验效果多次调整)，当 2 个簇合并为 1 个簇时，新簇的阈值由合并算法根据被合并簇的聚类特征求出。

2 个簇进行合并，其特征向量分别为 $\boldsymbol{X}=(T_1,x_1,T_2,x_2,\cdots,T_n,x_n)$，$\boldsymbol{Y}=(T_1,y_1,T_2,y_2,\cdots,T_n,y_n)$，则组成的新簇的特征向量为

$$\boldsymbol{Z}=(T_1,\frac{(x_1+y_1)}{2},T_2,\frac{(x_2+y_2)}{2},\cdots,T_n,\frac{(x_n+y_n)}{2})=(T_1,z_1,T_2,z_2,\cdots,T_n,z_n)$$

合并定理：假定对 2 个簇进行合并，合并后的簇的阈值表示为

$$d'=\max(dist(\boldsymbol{Z},\boldsymbol{X}+d_x),dist(\boldsymbol{Z},\boldsymbol{Y}+d_y))$$

其中，*dist* 指 2 个特征向量之间的距离。

4 数据分析

实验中对于本文提到的 3 种聚类方式都有涉及，对于它们的优劣在实验层面上做了研究比对。

A：所有类型的词语都用于构建空间向量；

B：只采用名词构建空间向量；

C：采用传统的空间向量表示方法；

D：采用稀疏的空间向量表示方法。

Mini 版(SogouC.mini.20061102)：共 100 篇文档，每个类别 10 篇。

精简版(SogouC.reduced.20061102)：共 10 020 篇文档，每个类别 1 002 篇。

表 1 是实验结果。其中，*t*(time)表示聚类消耗时间，单位为 ms；*a*(accuracy)表示聚类准确度。聚类消耗的时间依赖于执行的具体状况，因而有一定的差异。表中所取的数据是排除突变数据(即坏数据)之后，多次实验结果的平均值。

表 1 聚类实验效果

聚类方式	Mini 版				精简版			
	A		B		A		B	
	C	D	C	D	C	D	C	D
层次聚类法	*t*:5 634 *a*:82%	*t*:1 078 *a*:82%	*t*:4 614 *a*: 87%	*t*:920 *a*:87%	*t*:560 684 *a*: 83%	*t*:107 824 *a*:83%	*t*:477 894 *a*: 89%	*t*:91 920 *a*:89%
平面划分法	*t*:2 770 *a*:64%	*t*:532 *a*:64%	*t*:2 371 *a*:71%	*t*:456 *a*:71%	*t*:284 867 *a*:66%	*t*:53 787 *a*:66%	*t*:232 896 *a*:72%	*t*:44 788 *a*:72%
LP 算法	*t*:4 202 *a*:77%	*t*:797 *a*:77%	*t*:3 487 *a*:82%	*t*:675 *a*:82%	*t*:415 662 *a*:77%	*t*:79 934 *a*:77%	*t*:335 790 *a*:83%	*t*:67 165 *a*:83%

对实验结果进行分析，可以总结出以下 5 点：

(1)对于精简版的聚类，3 种方法的效果都优于 Mini 版。这是因为，精简版的基础数据量较大，个别的突变数据对于聚类效果的影响就相对较小。

(2)采用稀疏向量表示法之后，聚类的时间消耗减少了约 4/5，表明对于高维向量采用其稀疏表示可以有效地节省内存占用空间，加快聚类处理速度。

(3)相较于层次聚类，LP 算法在时间消耗上下降了约 30%，因此，对于数据量较大，实时性要求较高的场合，由于有效地减少了消耗时间，LP 算法还是显示出了它的优势。

(4)相较于平面划分法，LP 算法在聚类的准确性上提高了 11%~13%，达到了 77%~83%，从而保证了聚类的准确度在可接受的范围之内。

(5)本次实验中，LP 算法在聚类准确性上略逊于层次法，笔者认为这主要是因为：层次法的主要思想是全局最优，每次聚为一个簇的 2 个成员之间的相似度都是最大的，而在 LP 算法中，决定将 2 个成员归为一类的唯一衡量就是阈值 *d*。阈值选取的好坏对于实验效果的影响非常大。因此，如何选取阈值的初始值以及在聚类过程中如何动态地调整阈值是下一步的主要工作。

5 结束语

文本聚类在文本模式识别中占有重要的地位，这也是本文研究的价值所在。本文分析了基于距离的文本聚类中比较成熟的 2 种方法：层次凝聚法和平面划分法，并提出了一种改进方法 LP。从实验效果上看，LP 算法速度更快，灵活性也更高。在后续的工作中，还将进一步在实验的基础上对算法进行反复修正和拓展，以达到更好的聚类和实用效果。

(下转第 44 页)