结合 LSA 的中文谱聚类算法研究

熊忠阳, 暴自强, 李智星, 张玉芳
(重庆大学 计算机学院, 重庆 400044)

摘 要: 传统的文本谱聚类需要的文本相似矩阵依赖于向量空间模型,忽略了词与词之间的语义关系,存在词频维数过高、计算代价高等问题。针对这些问题,提出了一种基于潜在语义分析(latent semantic analysis,LSA)的文本相似矩阵构造方法,利用奇异值分解(singular value decomposition,SVD)降维,在低维的语义空间表示文本,以此来提高同类文本间的语义相似度,并进行了相关对比实验。在该实验中,改进方法的聚类效果要好于传统的方法,从而验证了改进方法的有效性和可行性。

关键词: 文本聚类; 潜在语义分析; 奇异值分解; 谱聚类
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)03-0917-02
doi:10.3969/j.issn.1001-3695.2010.03.030

Research of Chinese spectral clustering with LSA

XIONG Zhong-yang, BAO Zi-qiang, LI Zhi-xing, ZHANG Yu-fang
(*College of Computer Science*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: Traditional text samples similarity matrix for spectral cluster heavily rely on the vector space model which ignores the semantic relationship among terms. It will give rise to problems such as curse of dimensionality, feature redundancy and high computing cost. To solve the problems above, this paper proposed a new method based on LSA to solve it, which used SVD to lowering rank of matrices. The experimental results turn out that the new method enhances the cluster accuracy and less the data-process elapsed time.

Key words: text clustering; LSA; SVD; spectral cluster

0 引言

聚类分析是模式识别和机器学习的重要研究领域。所谓聚类(clustering)就是将数据对象分组成为多个类或簇(cluster),使得在同一簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大。聚类分析在文本数据挖掘中有着非常重要的应用,被广泛地应用于文本数据挖掘和信息检索等领域,可以用来改进信息检索系统的查准率和查全率,也可用于查找最接近的文本,还可用于对 Web 上的文本进行分层次的聚类等[1]。传统的聚类算法,如 K-means、EM 等都是建立在凸球形的样本空间上,但当样本空间不满足凸形时,算法容易陷入局部最优。谱聚类算法(spectral clustering algorithm)避免了这个问题。该算法建立在图论中的谱图理论基础上,其本质是将聚类问题转换为图的最优划分问题。与传统的聚类算法相比,谱聚类算法将聚类转换为一个代数上的矩阵求解问题,具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点[2]。

由于谱聚类算法起源于谱图理论,当把它应用于文本聚类时,如何抽取文本特征来表示文本成为一个首要问题。传统的思路是使用基于关键词集的文档向量表示模型,但是这种方式忽略了词与词之间的语义关系,存在词频维数过高、聚类算法计算复杂度高等问题,往往并不能准确描述文档间的语义相关性。潜在语义分析(LSA)技术则可以在一个低维潜在概念语义空间重新描述自然语言文本,从而可以更好地反映文本之间的语义相似度[3]。

1 LSA 简介

文本数据的图表示模型对基于图的聚类算法最终的效果具有重要的影响。在传统的基于关键词集的向量空间模型中,文本间的相似性取决于文档间的词汇特征的共现率。然而,在自然语言文本中普遍存在着同义词和多义词的现象,多义词的现象导致两篇包含很多共有词汇的文本并不一定很相似,而同义词现象导致相似文本间可能并没有太多的共现词汇,这就造成了词条的语义意义上的大量冗余。这样,基于词条特征空间的文档表示有时并不能很好地反映文档之间的语义相关性,而且文本数据的向量空间表示模型中,样本都具有几千几万的维度,算法仅处理这些高维数据就需要花费大量时间。针对文本的这一特点,LSA[4]利用奇异值分解(SVD)技术对高维的词条—文档矩阵进行处理,在潜在的语义结构子空间中重新表示文本及文本间的相似度,同时达到降维的目的。其数学描述如下:

设词条—文档矩阵 X 是个 $m \times n$ 矩阵,其中 m 为词条数,n 为文本集的文档数。令 $k << \min(m,n)$, $\text{rank}(A) = r, k < r$。经过奇异值分解,矩阵 X 可表示为三个矩阵的乘积,如下:

$$X = SVD^{T} \quad (1)$$

收稿日期: 2009-08-10; **修回日期**: 2009-09-07

作者简介:熊忠阳(1964-),男,博导,主要研究方向为数据挖掘、数据库、并行计算、网络信息处理;暴自强(1983-),男,硕士研究生,主要研究方向为文本分类、互联网应用技术(baoziqiang1234@163.com);李智星(1985-),男,博士研究生,主要研究方向为文本分类、互联网应用技术;张玉芳(1967-),女,硕导,主要研究方向为数据挖掘、数据库、并行计算、网络信息处理.

其中：S、D 是 $m \times r$ 和 $n \times r$ 的正交矩阵，分别称为矩阵 X 的左右奇异向量矩阵；V 是 $r \times r$ 的对角矩阵，称为矩阵 X 的奇异标准形，其对角元素为矩阵 X 的奇异值。

$$V = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r) \quad \sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0 \tag{2}$$

矩阵 X 的奇异值按递减排列成对角矩阵 V，取 V 的前 k 个最大奇异值构成 $k \times k$ 的 V，取 S 和 D 的前 k 列构成 $m \times k$ 的 S_k 和 $n \times k$ 的 D_k，构建 X 的 k-秩近似矩阵 X_k。

$$X_k = S_k V_k D_k^{\mathrm{T}} \tag{3}$$

其中：S_k 和 D_k 中的行向量分别作为词条向量和文档向量；k 是降维后的维数。实际应用中 k 值常常取到几百，极大地减小了文本向量的维数。

2　谱聚类算法介绍

谱聚类算法的思想来源于谱图划分理论。假定将每个数据样本看做图中的顶点 V，根据样本间的相似度将顶点间的边 E 赋权重值 W，这样就得到一个基于样本相似度的无向加权图 $G = (V, E)$。那么在图 G 中，就可将聚类问题转换为在图 G 上的图划分问题。基于图论的最优划分准则就是使划分成的两个子图内部相似度最大，子图之间的相似度最小。划分准则的好坏直接影响到聚类结果的优劣。常见的划分准则有 Minimum cut、Normalized cut、Ratio cut 等。其中 2000 年 Shi 等人[5]提出的 Normalized cut 算法，由于较好地解决了划分的倾斜问题，研究表明效果最好。本文采用了这一算法，使用 Normalized cut的谱聚类算法数学表达如下：

a）给定或建立表示样本集的相似矩阵 W 以及聚类数 S。

b）计算拉普拉斯矩阵 $L = D - W$。

c）计算 $D^{-1}L$ 的第一最小特征值到第 n 最小特征值分别对应的特征向量 $v_1, v_2, \cdots, v_s$，令 $V = [v_1, v_2, \cdots, v_s]$，则 $V \in R^{n \times s}$。令 $V = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$，得到一个 k 维数据集合 $\{y_1, y_2, \cdots, y_n\}$。

d）对于数据点 $y_1, y_2, \cdots, y_n$，利用 K-均值算法聚成 s 个类。

3　结合 LSA 的谱聚类算法

针对上述谱聚类算法存在的不足，本文提出一种结合 LSA 的谱聚类新方法。该方法的主要思路是：针对传统的高维词条文本矩阵，首先对其进行 LSA 处理，将其映射到一个低维度的语义空间，然后在此基础上生成谱聚类算法需要的相似矩阵，进而在此基础上运行谱聚类算法。该方法有两个好处：a）通过 LSA 降维，可以减少数据处理时间，一般而言，可以将文本表示从上万维降低到几百维；b）将文本映射到潜在语义空间，增强了文本间的语义相关性，提高了谱聚类算法的精度。在中文文本数据集上的实验表明，该方法达到了以上效果。

4　实验

4.1　实验数据集

实验选取复旦大学李荣陆的中文文本语料库，共 10 个类，删除重复文档后有中文文档 2 126 篇，分词后产生 51 363 个词条，词条权重使用如下的正则化 TF * IDF 公式表示：

$$\mathrm{weight}_{\mathrm{tfidf}}(T_{ik}) = \mathrm{tf}(T_{ik}) \times \mathrm{idf}(T_k) / \sqrt{\sum_{j=1}^{n} (\mathrm{tf}(T_{ik}) \times \mathrm{idf}(T_k))^2} \tag{4}$$

4.2　实验评价标准

实验采用正则互信息来度量聚类效果，CAT（category label）与 CLS（cluster label）之间的正则互信息定义为

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{i,j} \log(n \cdot n_{i,j} / n_i \cdot n_j)}{\sqrt{(\sum_i n_i \log n_i / n)(\sum_j n_j \log n_j / n)}} \tag{5}$$

其中：n 是数据集文本数量；n_i 和 n_j 分别表示实际类别 i 与聚类类别 j 中的文本数量；$n_{i,j}$ 代表同时在实际类别 i 与聚类类别 j 中的文本数量。当 NMI = 1 时，聚类与实际类别完全匹配；NMI = 0 时，表明文本完全随机散布，NMI 越高表明聚类效果越好。

4.3　实验结果与分析

实验针对进行过 LSA 处理和未经过 LSA 处理后的相似矩阵两种情况进行对比实验。实验发现，当 k（LSA 降维后的维度）取 150 时，改进后的实验效果最好。LSA 降维处理后，生成相似矩阵时间由几分钟降低到不到 1 s，时间复杂度大大降低。另外，实际应用谱聚类算法时，为加快运算速度，常常需要对词条特征矩阵 X 使用 KNN 方法处理成稀疏矩阵，不同的 K 值会对实验结果造成影响。本文实验在进行 LSA 处理和未经过 LSA 处理两种情况下，针对 K 分别取 10、20、30、50、80、100、150、200 值时的实验结果进行了分析比较，结果如图 1 所示。

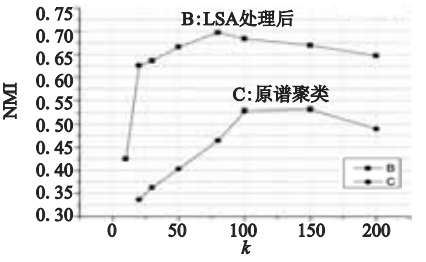

图1　LSA前后实验对比

图中折线 B 代表结合 LSA 后的谱聚类实验效果，折线 C 代表未结合 LSA 的谱聚类实验结果。由图中对比结果可以看出，改进算法的 NMI 值基本在 0.65 左右，相对于原始算法提高了 15% ~20%，这表明本文的算法不仅缩短了数据处理运行时间，而且提高了谱聚类的精度。

5　结束语

为提高谱聚类在文本聚类上的应用，本文引入了 LSA 技术。实验结果表明，使用 LSA 处理后聚类效果有较大提高，而且使得谱聚类表现更稳定。谱聚类算法作为一个新方向，还有许多东西有待进一步发掘，笔者将结合领域知识来进一步改进谱聚类效果。

参考文献：

[1] HAN J, KAMBER M. Data mining: concept and techniques[M]. San Fransisco: Morgan Kaufmann, 2001.

[2] 蔡晓妍，戴冠中，杨黎斌. 谱聚类算法综述[J]. 计算机科学，2008，35(7)：14-18.

[3] 戴新宇，田宝明，周俊生. 一种基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT[J]. 电子学报，2008，36(8)：1628-1630.

[4] DEERWESTER S, DUMAIS S, FUMAS G W, *et al*. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[5] SHI Jian-bo, MALIK J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.