基于谱特征和图分割的图聚类算法

高 阳[1],李昌华[1],李智杰[1,2],崔欢欢[1]
GAO Yang[1], LI Changhua[1], LI Zhijie[1,2], CUI Huanhuan[1]

1.西安建筑科技大学 信息与控制工程学院,西安 710055
2.西安建筑科技大学 建筑学院,西安 710055
1.College of Information & Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
2.College of Architecture, Xi'an University of Architecture and Technology, Xi'an 710055, China

GAO Yang, LI Changhua, LI Zhijie, et al. Graph clustering algorithm based on spectrum and graph partition. Computer Engineering and Applications, 2017, 53(15):222-226.

Abstract: In order to analyze the structured data in database with valid clustering, the algorithm firstly mines the depth data characteristics of different graph sample, and constructs the association matrix including connection relations and nodes with hierarchy, completes the analysis of spectral characteristics combined with the Laplace matrix. Secondly, it uses the Gaussian kernel to build the similarity matrix, to facilitate post-processing the value of similarity matrix normalized in the range of 0-1. Finally, it combines *k*-means with graph partitioning algorithm to make the data *k*-partition, then gets *k* cluster of the database. The experimental results demonstrate that the improved Laplace has finer division of the internal structure in matrix, and improves the pre-process results of the sample. The minimum rate cut algorithm ensures the accuracy of the premise, and turns the NP-hard problem into a polynomial time to solve the problems and improve algorithm efficiency.

Key words: spectral decomposition; graph segmentation; similarity matrix; graph clustering

摘 要: 为了对图数据库中的结构化数据有效的聚类分析,首先对不同的图数据样本进行特征的深度挖掘,构造了包含节点间连接层次关系的关联度矩阵,与拉普拉斯矩阵结合共同完成谱特征分析;然后利用高斯核函数进行相似度矩阵的构建,将相似度归一化到0到1的范围内便于后期处理;最后结合图分割与*k*-means算法将相似度矩阵进行*k*分割,得到*k*个聚类。经过大量分析实验表明,改进的拉普拉斯矩阵对样本内部结构有更为精细的划分,提高了前期样本处理效果。最小比率割算法在保证精度的前提下,将NP难的问题转化为多项式时间内解决的问题,提高了算法的效率。

关键词: 谱特征分解; 图分割;相似度矩阵;图聚类
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1601-0047

1 引言

图的谱方法在图聚类上的应用主要是使用图转换成矩阵后的特征向量或者图谱特征的衍生特征来表示图的结构特征关系,然后再使用传统的聚类方法(比如*k*-means算法)实现图的聚类。Gibert等人提出向量空间在图中的嵌入成为了众多优秀算法在图中应用的桥梁[1],可以看出图的谱特征在聚类上的重要作用。Luo等提出的使用邻接矩阵的谱代表图的结构的方法[2],不同于传统基于树结构的算法,有一定的指导意义。常见的聚类算法有MST聚类算法、归一化之后的分割算法(SM)[3]及其改进算法(KVV)[4]、结合拉普拉斯矩阵谱实现的emeans算法[5]等。另外基于奇异值分解(SVD)的谱特征提取算法[6]以及邻接矩阵特征值分解(EVD)[7]都是基于邻接矩阵及其衍生性质的谱特征提取方法。

基金项目: 国家自然科学基金(No.61373112);陕西省自然科学基金(No.2016JM6078)。
作者简介: 高阳(1989—),男,硕士研究生,研究领域为图形图像处理,模式识别,E-mail:gaoyang615@163.com;李昌华(1963—),男,博士,教授,研究领域为数字建筑,模式识别;李智杰(1980—),男,博士,讲师,研究领域为数字建筑,模式识别。
收稿日期: 2016-01-05 **修回日期:** 2016-03-04 **文章编号:** 1002-8331(2017)15-0222-05
CNKI **网络优先出版:** 2016-03-25, http://www.cnki.net/kcms/detail/11.2127.TP.20160325.2047.096.html

不同于传统的聚类算法如k-means算法[8]，谱聚类算法能在任意形状上的样本空间实现聚类，并且不易陷入局部最优解的情况，具有较好的整体性能，并收敛于全局最优。谱聚类方法关注数据的大小和信息的主要特征，能够有效地对数据降维。所以，谱聚类是通过样本之间的相似度来计算特征向量的，与数据的维度无关。

本文主要介绍图谱方法在聚类上的应用，使用IAM Graph Database Repository作为测试对象，IAM图库是一个测评性很高的公共测试图库，其中包含字符、蛋白质、WEB、指纹等结构信息。聚类问题可以转化为图分割问题，使用最优的方法实现图的划分也就解决了图的聚类问题。本文尝试使用拉普拉斯谱特征以及图分割方法来进行图的聚类，图数据之间的相似度矩阵以及加权矩阵来提高图聚类算法的精度。图的谱聚类流程如图1所示。

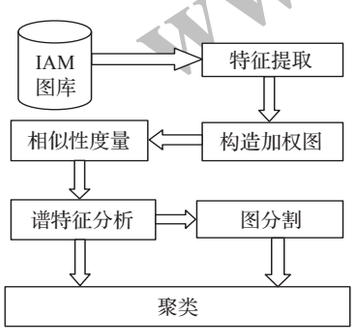

图1　图的谱聚类框图

2　图的谱特征选择与优化

有效地选择图的谱特征，又尽量减少大量的矩阵运算是图聚类问题的关键。图作为结构化的数据，具有节点、边、及对应的权值等属性，一般情况下使用邻接矩阵（记作A）及其对应的特征值和特征向量来描述图中顶点与边的关系，拉普拉斯矩阵（记作L）作为邻接矩阵的一种变形，具有良好的特性，更适用于图的匹配以及聚类问题的解决。矩阵之间的关系为$L=D-A$，其中D为对角阵，对角线上的数值等于A中行的和的绝对值，其他非对角元素为0。

拉普拉斯矩阵在图聚类上具有广泛的应用，Scott和Longuet-Higgins根据结构化学的概念，将谱方法应用于图像匹配[9]，即通过对亲和矩阵进行SVD分解求解特征值以及特征向量。王年等使用图的Laplace谱特征进行图像匹配[10]，首先规范化了LapLace矩阵，然后构造特征点匹配矩阵进行匹配。也有许多研究者通过图的路径来描述拓扑特征，如Bai等利用路径相似性实现结构骨架的匹配[11]，Ling等使用内部节点之间的距离来描述图结构的特征[12]。Gao等人通过直方图来描述图的内部特征实现相似性度量[13]，这种方法效果较好并且具有较低的复杂度。

2.1　改进的拉普拉斯谱特征

定义1（关联度矩阵）对无向图G，对应的邻接矩阵为A，w_{ij}为顶点v_i与v_j之间的连接权值，使用带深度的广度优先遍历算法进行关联度矩阵D_{ij}^*的计算。

关联度值的递归计算如公式(1)所示，其中已访问过的节点不再访问。

$$d_{ij}^*=d_{ij}^*+\left(\frac{1}{2}\right)^h\sum_{j=1}^{n}w_{ij},\ 0\leqslant h<n,\ i=j \tag{1}$$

其中h表示广度优先算法中遍历的深度。

D_{ij}^*的详细计算步骤：

(1)从节点v_i开始，令初始深度$h=0$，$i=0$，度值$d_{ij}^*=0$，度值计算深度m。

(2)如果$h<m$，访问与之相邻且为访问过的节点加入队列并计数m_k，已访问节点标记为$visited=1$；公式(1)计算度值；如果$h=m$，转到(4)。

(3)再依次访问与m_k个节点分别相邻的节点，如果存在，$h=h+1$，返回(2)；否则，继续。

(4)得到该节点的关联度值，$i=i+1$，转到步骤(1)，直到所有节点都被访问过。

在实际情况下，如果计算一个节点对所有节点的关联度，既降低了效率，又对结果无太大影响，因此设置关联度的计算深度m来提高度关联矩阵的计算效率。

传统度矩阵（对角阵）表示如式(2)所示：

$$D_{ij}=\begin{cases}\sum_{j=1}^{n}w_{ij},\ i=j\\0,\ i\neq j\end{cases} \tag{2}$$

w_{ij}表示无向图中两节点间的连接关系，$w_{ij}=1$表示两节点直接相连，$w_{ij}=0$表示不连接。

关联度矩阵D_{ij}^*与度矩阵D_{ij}不同，度矩阵D_{ij}的对角线元素仅表示了顶点v_{ij}与其他节点的直接连接关系，而关联度矩阵不仅含有图中顶点v_{ij}与其他顶点的直接连接关系，还包括间接连接关系。一个顶点与其他顶点的间接连接关系在图结构中具有重要的意义。如图2所示，表明关联度矩阵能够更好地层次化区分图结构中的关键节点。

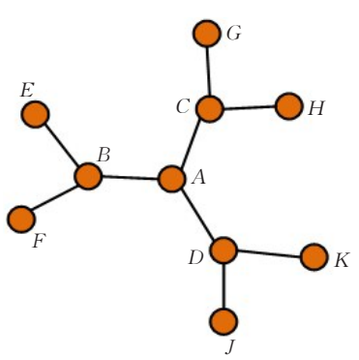

图2　图结构示例

传统度矩阵关于度值的计算中，A、B、C、D 四个连接点的度值都是3，但在图结构中，A 点显然是中心点，其作用远远大于其他三点，在本文提出的关联度矩阵中 A 点的关联度为6，而 B、C、D 三点的关联度为5，以此类推，可以得到 E、F、G、H、J、K 的关联度值为3。使用关联度矩阵能够清晰地将各节点的作用量化。表1表示两种度矩阵表示方式下，各节点的度值。

表1 度矩阵与关联度矩阵下的值

权值	A	B	C	D	E	F	G	H	J	K
D_{ij}	3	3	3	3	1	1	1	1	1	1
D^*_{ij}	6	5	5	5	3	3	3	3	3	3

可以看出，在关联度矩阵中比传统度矩阵具有更好的层次区分效果，各点具有明显的区分度，实验证明，使用关联度矩阵，能够提高图之间的匹配准确率。改进的拉普拉斯矩阵定义形式如下：

$$L=D^*-A \tag{3}$$

特征值以及特征向量具有如下关系：

$$Ly=\lambda y \tag{4}$$

其中 D 是对角矩阵，λ 表示特征值，y 表示特征值对应的特征向量。拉普拉斯矩阵是对称的半正定矩阵，这些性质有利于将原来的NP难问题转化为多项式时间内可以解决的问题。

2.2 相似度矩阵的构造

实验证明，仅仅使用拉普拉斯矩阵的特征值以及特征向量，作为聚类的标准，聚类效果往往较差，拉普拉斯矩阵忽略了原始结构图中的空间信息，主要关注节点与节点之间的连接信息，无权值的邻接矩阵不关注节点之间的权重关系。节点之间的权值代表了样本之间的关联程度，因此相似度矩阵的构造在图聚类过程中十分重要。相似度矩阵的构造方式有多种，代表性的有基于元素的方法、基于结构的方法、基于实例的方法。

基于元素的方法注重实体的属性相似度，容易导致局部最优的结果；基于结构的方法注重样本内部的关系结构，恰当地提取关系结构能够提高聚类效率；基于实例的方法强调整体的表象特征，缺点是计算复杂度及较高。本文采用的高斯核函数模型属于结构型构造方法，并且参数 δ 的动态调节能够适应不同的应用场景。

高斯核函数计算相似度矩阵，距离越大，代表其相似度越小。

$$s(x_i,x_j)=\mathrm{e}^{-\|x_i-x_j\|^2/(2\delta^2)} \tag{5}$$

参数 δ 决定着平滑程度，δ 越大，相似度曲线平滑程度越好。在聚类过程中，δ 过大或者过小都会导致聚类效果较差，一般根据图数据集的实际情况，动态调节参数的大小。样本之间的相似度在高斯核函数的约束下，取值范围规范化在0到1之间，有利于数据的统一与后期运算处理。

$$\boldsymbol{S}=[s_1,s_2,\cdots,s_n]^{\mathrm{T}} \tag{6}$$

其中 s_k 是相似度矩阵中第 k 行数据，每一行数据代表其中某一样本与其余所有样本之间的相似度。在实验过程中由于样本只可能属于一类，与其他类之间的相似度必然很低，在实际计算过程中造成大量计算资源的浪费。因此采取 k 近邻算法（k-Nearest Neighbor），为了提高计算效率，仅保留相似度矩阵中除自身外前 k 个较大的相似度值。

3 基于相似度矩阵的图分割算法

图聚类的直观解释就是根据样本间的相似度，依据一种或者多种属性，完成聚类。图分割是聚类问题的一个中间操作，谱聚类的思想是将样本间的相互关系用图理论来表达，使用图论的知识来解决聚类问题。因此，数据的聚类问题可以转换为图的分割问题。数据样本与图结构的转化是根据样本之间的相似度来完成的，样本作为节点，相似度定义了样本之间的连接权值。汤进等将提取的特征点构成Delaunay图[14]，计算其特征向量进行聚类，也取得了比较好的效果。文中关联度矩阵细致地描述了图的内部结构特征，结合图分割算法，对数据具有降维作用，相对提高聚类的精度和效率。

算法步骤：

（1）计算样本之间的相似度矩阵。

（2）构造关联度矩阵以及拉普拉斯矩阵 L。

（3）计算 L 的前 k 个最小的特征向量。

（4）k 个特征向量组成 $n\times k$ 的矩阵，k-means聚类。

3.1 图分割算法

图分割算法主要由以下几种，最小分割算法（Minimum cut）[15]、最小比率分割算法（Minimum ratio cut）[16]、最小规范化分割算法（Minimum normalized cut）[17]、最小最大分割算法（Min-max cut）[18]等。图分割的主要目的是使同组之间的权重最高，而不同组别之间的权重尽可能得低。权重越高，表示相似度越大，权重太低的边就要舍去。

经过处理的拉普拉斯矩阵谱对图的分割具有良好的性质。其中，图的二分可以采用基于Fiedler向量[19]的方法，图的 k 分采用基于多个主要特征向量的方法来完成[20]。本文主要考虑图的 k 分问题。

图分割的主要目的就是消除相关性、影响性最小的边，达到将原始图分为 k 个子图的目的，来完成聚类。子集之间的权值可以表示如式（7）：

$$W(C_r,C_t)=\sum_{v_i\in C_r,v_j\in C_t} w_{ij} \tag{7}$$

C_r,C_t 是不同的聚类组别，$W(C_r,C_t)$ 是各组之间的权重之和。v_i,v_j 分别是 C_r,C_t 中样本点，图 G 的所有可能的子图集合使用 π^k 表示，$\overline{C_i}$ 是 C_i 的补集。最小分割算法：

$$\min cut(\pi^k)=\frac{1}{2}\sum_{i=1}^{k}W(C_i,\overline{C_i}) \tag{8}$$

最小分割算法的目的是让不同类别集合中点之间的权值之和最小。然而在许多测试图集中，在各个集合中经常会出现孤立点，导致聚类结果不理想。对于非规范化数据，此目标函数会出现偏向最小分割的结果。

最小比率分割算法：

$$\min ratiocut(\pi^k)=\frac{1}{2}\sum_{i=1}^{k}\frac{W(C_i,\overline{C_i})}{|C_i|} \tag{9}$$

其中 $|C_i|$ 表示 C_i 组中包含的样本数目。最小比率分割算法中含有元素总数量因子，在一定程度上避免出现孤立点的情况，解决了分组过小问题。本文采用最小比率割作为分割的依据。

3.2　最小比率割算法

RatioCut的求解属于NP难问题，可以做如下转化，设图集 G 可以分割为 k 个子集，分别为 $C_1,C_2,\cdots,C_k$。定义向量：

$$\boldsymbol{h}_j=(h_{1,j},h_{2,j},\cdots,h_{n,j})^{\mathrm{T}} \tag{10}$$

$h_{i,j}$ 表示如公式(11)所示：

$$h_{i,j}=\begin{cases}1/\sqrt{|C_j|},v_i\in C_j\\0,v_i\notin C_j\end{cases} \tag{11}$$

其中每列数据都是正交的特征向量，由 k 个特征向量 $\boldsymbol{h}_j$ 组成的 H 是标准正交的矩阵，即 $H'H=I$，则

$$\boldsymbol{h}_i'L\boldsymbol{h}_i=\frac{cut(C_i,\overline{C}_i)}{|C_i|} \tag{12}$$

$$\boldsymbol{h}_i'L\boldsymbol{h}_i=(H'LH)_{ii} \tag{13}$$

综合以上可以得到：

$$RatioCut(A_1,A_2,\cdots,A_k)=\sum_{i=1}^{k}\boldsymbol{h}_i'L\boldsymbol{h}_i=\sum_{i=1}^{k}(H'LH)_{ii} \tag{14}$$

所以NP难问题就转化为求解 $H'H=I$，H 的解即为拉普拉斯矩阵 L 的前 k 个特征向量组成的矩阵。H 矩阵的每一行可以看做一个样本点进行聚类分析。

4　实验结果

为了验证优化后的拉普拉斯矩阵以及图分割算法应用后的图聚类的实际效果，实验采用IAM图库中的字符图库作为数据样本进行测试。包含A、E、F、H、I、K、L、M、N、T、V、W、X、Y、Z的结构信息。原始数据使用XML文档存储，每个字符有150种结构表示，为了简化，随机使用其中每个字符集中的50个字符作为实验样本。部分字符可视化后结果如图3所示。

实验首先从每种字符集中随机抽取50个样本，因为一共有15种字符，所以一共有750个样本参与测试。公式(9)，(14)中的参数 k 设置为15。对测试样本首先

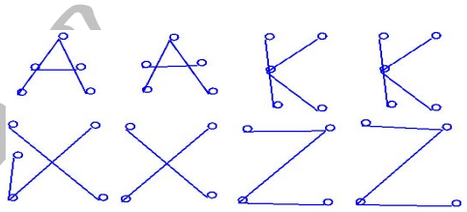

图3　字符的可视化图形

进行特征提取，充分考虑到图的每个节点在整体结构中的作影响因子，构造关联度矩阵，计算其特征向量。利用高斯核函数计算样本与样本之间的相似度，利用KNN算法获取前 m 个相似度最大的保留下来，形成相似度矩阵；最后结合图分割与 k-means算法将具有较大相似度的样本分别聚类，通过调节参数，得到比较理想的结果。对样本整体进行聚类分析结果如表2所示。

表2　关联度矩阵对聚类算法的影响

矩阵	正确率/%	
	k-means	比率分割算法
度矩阵	82.51	84.72
关联度矩阵	84.35	86.93

可以看出，原始样本的特征提取对于比率分割算法的影响较大，从比率分割算法对于改进的拉普拉斯矩阵的聚类效果来看，使用关联度矩阵的Laplace矩阵，明显地提高了聚类准确率，依赖 k-means算法强大的聚类能力，具有很高的稳定性，使用关联度矩阵后，准确率有了一定的提高。

对逐个单个字符集进行聚类分析，如表3所示。行字符是正确的聚类标识，列字符是待聚类的字符集合。表格内数据表示识别成为一种字符的数目。比如(A，A)=45，(A，E)=4，则表示A识别为A和E的字符个数分别为45个和4个。

可以看出不同字符间结构相差较小的一般容易产生错误识别的情况。如表中I与L，V与M，由于在书写过程总存在字符轮廓不明显等问题，导致节点提取存在误差。平均识别率在86%以上，比Munkres的KM算法有所提高，原因是基于关联度矩阵的相似度矩阵构造方法，能够更精细地区分不同位置的节点作用，对节点的层次化区分更明确，对后期聚类提供了良好的前提条件。

5　结束语

本文所使用优化后的Laplace矩阵，把图谱分解理论和图分割方法结合起来的方法，提高了图聚类的准确率，而且简化了计算，提高了识别效率。另外，本文对于聚类过程中，各特征点之间的相对位置关系与其他算法相似，没有在相似度矩阵中体现位置关系特征，如果加入到后期聚类中，可能会得到更理想的效果，这也是后期工作需要完善改进的地方。聚类的结果对字符提取

表3 字符聚类准确率

	A	E	F	H	I	K	L	M	N	T	V	W	X	Y	Z
A	44	3	0	0	0	0	0	0	0	0	0	0	0	0	0
E	4	42	3	4	0	0	0	0	0	0	0	0	0	0	0
F	0	2	45	2	0	0	0	0	0	1	0	0	0	0	0
H	2	3	1	43	0	3	0	0	0	0	0	0	1	0	2
I	0	0	0	0	43	0	5	0	0	3	0	0	0	1	0
K	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0
L	0	0	0	0	6	2	42	0	0	2	0	0	0	0	0
M	0	0	0	0	0	0	0	45	2	0	2	3	0	0	5
N	0	0	1	1	0	0	0	2	42	0	0	0	0	0	1
T	0	0	0	0	1	0	0	0	0	43	0	1	0	3	0
V	0	0	0	0	0	0	2	2	3	0	45	1	3	3	0
W	0	0	0	0	0	0	0	0	1	0	0	43	2	0	0
X	0	0	0	0	0	0	0	0	0	0	2	0	44	0	0
Y	0	0	0	0	0	0	0	1	0	1	0	2	0	43	0
Z	0	0	0	0	0	0	1	0	2	0	1	0	0	0	42
准确率	0.88	0.84	0.90	0.88	0.86	0.90	0.84	0.90	0.84	0.86	0.90	0.86	0.88	0.86	0.84

的特征有较大的依赖,下一步工作重点也可以放在特征的深度挖掘上,结合其他领域的优秀思想,提高聚类算法的准确率。

参考文献:

[1] Gibert J,Valveny E,Bunke H.Graph embedding in vector spaces by node attribute statistics[J].Pattern Recognition,2012,45(9):3072-3083.

[2] Luo B,Wilson R,Hancock E.Spectral embedding of graphs[J].Pattern Recognition,2003,36(10):2213-2230.

[3] Shi J,Malik J.Normalized cuts and image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.

[4] Kannan R,Vempala S,Veta A.On clusterings-good,bad and spectral[J].Journal of the ACM,2004,51(3):497-515.

[5] Ng A,Jordan M,Weiss Y.On spectral clustering:Analysis and an algorithm[J].Neural Information Processing Systems,2002.

[6] Pilu M.A direct method for stereo correspondence based on singular value decomposition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.USA:IEEE Computer Society,1997:261-266.

[7] Shapiro L S,Brady J M.Feature-based correspondence:an eigenvector approach[J].Image and Vision Computing,1992,10(5):283-288.

[8] MacQueen J B.Some methods for classification and analysis of multivariate observations[C]//Proc Fifth Berkeley Symposium Mathematic Statistics and Probability.Berkeley,Calif:University of California Press,1967:281-297.

[9] ScottG L,Longuet-Higgins H C.An algorithm for associating the features of two images[J].Proceedings of Royal Society of London,1991,244:21-26.

[10] 王年,范益政,韦穗,等.基于图的Laplace谱的特征匹配[J].中国图象图形学报,2006,11(3):332-336.

[11] Bai X,Latecki L J.Path similarity skeleton graph matching[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2008,30(7):1282-1292.

[12] Ling H B,Jacobs D W J.Shape classification using the inner-distance[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(2):286-298.

[13] Gao X,Xiao B.Image categorization:Graph edit distance+edge direction histogram[J].Pattern Recognition,2008,41(10):3179-3191.

[14] 汤进,张春燕,罗斌.基于图谱分解和概率神经网络的图像分类[J].中国图象图形学报,2006(5):630-634.

[15] Wu Z,Leathy R.An optimal graph theoretic approach to data clustering theory and its application to image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1993,15(11):1101-1113.

[16] Wang S,Siskind J.Image segmentation with ratio cut[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2003,25(6):675-690.

[17] Shi J,Malik J.Normalized cuts and image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.

[18] Ding C,Ren X F,Zha H,et al.Spectral min-max cut for graph partitioning and data clustering[C]//Proceedings of the IEEE International Conference on Data Mining.USA:IEEE CS,2001:107-114.

[19] Fiedler M.Algebraic connectivity of graphs[J].Czechoslovak Mathematical Journal,1973,23(2):298-305.

[20] Dhillon I.Co-clustering documents and words using bipartite spectral graph partitioning[J].Knowledge Discovery and Data Mining,2001.