一种基于 SimRank 得分的谱聚类算法

李鹏清[1] 李扬定[1] 邓雪莲[2] 李永钢[1] 方 月[1]
(广西师范大学广西多源信息挖掘与安全重点实验室 广西 桂林 541004)[1]
(广西中医药大学公共卫生与管理学院 南宁 530200)[2]

摘 要 传统的谱聚类算法在建立相似度矩阵时仅考虑数据点与点的距离,忽略了数据点之间隐含的内在联系。针对这一问题,提出了一种基于 SimRank 的谱聚类算法。该算法首先用无向图数据建立邻接矩阵,并计算出基于 SimRank 的相似度矩阵;然后根据相似度矩阵建立拉普拉斯矩阵表达式,对其进行归一化后再进行谱分解;最后对分解得到的特征向量进行 k-means 聚类。在 Zoo 等 UCI 标准数据集上的实验结果表明,所提算法在聚类精确度、标准互信息和纯度 3 个评价指标上均优于现有的 LRR(Low Rank Rrepresentation)等基于距离相似度的谱聚类算法。

关键词 谱聚类,相似度矩阵,SimRank 得分,邻接矩阵,拉普拉斯矩阵,k-均值聚类
中图法分类号 TP181 **文献标识码** A

Spectral Clustering Algorithm Based on SimRank Score

LI Peng-qing[1] LI Yang-ding[1] DENG Xue-lian[2] LI Yong-gang[1] FANG Yue[1]
(Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China)[1]
(School of Public Health and Management, Guangxi University of Chinese Medicine, Nanning 530200, China)[2]

Abstract Traditional spectral clustering algorithms only consider distance between data points, ignoring their intrinsic relation. To deal with this problem, a spectral clustering method based on SimRank score was proposed. Firstly, the method computes the adjacency matrix of the undirected graph data, and obtains the similarity matrix based on SimRank. Secondly, a Laplacian matrix expression is constructed based on similarity matrix, which is then normalized followed by spectral decomposition. Finally, a k-means clustering procedure is performed on the obtained eigenvectors to obtain the final clustering results. Experimental results on benchmark datasets from UCI data repository show that the proposed algorithm is superior to the existing spectral clustering algorithms based on distance similarity in terms of clustering accuracy, standard mutual information and purity.

Keywords Spectral clustering, Similarity matrix, SimRank score, Adjacency matrix, Laplace matrix, k-means clustering

1 引言

聚类分析是一种将实体或虚拟对象组成的集合分成多个类的科学的统计分析方法。在对数据对象进行分析时,要求根据每个对象的特点对其进行恰当的分类。该分类过程是无监督的,即在无先验知识的条件下完成聚类分析。当前聚类分析的研究主要集中在谱聚类算法[1]。谱聚类是利用数据集相似性矩阵的特征向量的性质对数据进行聚类,其基本思想是把样本中的对象看成节点,对象与对象之间通过边来连接,并对每条边赋予一定的权重,即相似度,将聚类问题转化成图的分割问题。该思想的目的是探索一种图的切割方式[2-3],使得类间相似度尽可能低,而类内相似度尽可能高。

谱聚类分析的应用范围非常广[4-7],例如信息科学、生物科学、经济学以及心理学分析等诸多领域都会用到这项技术[8]。在商业领域,基于谱聚类分析技术,运用买卖等方法发现不同客户的特点;在生物科学研究中,可以通过谱聚类分析探索动植物及基因结构并将它们分组,从而获得对相应种群结构特征的了解;谱聚类分析也被运用到电子商务的网站建立上,经过聚类分析技术得到浏览习惯相同的客户群,通过分析他们各自的特点,使得用户能为客户提供精准的服务。

目前具有代表性的谱聚类方法有 Liu 等[9]提出的用基于低秩表示的谱聚类算法(Low Rank Representation, LRR),它强制关联矩阵构造为低秩,保留数据点的整体信息。模型对噪声是鲁棒的,因为噪声将增加数据的秩。Elhamifar 等[10]

本文受国家重点研发计划项目(2016YFB1000905),国家自然科学基金(61363009,61672177,61573270,81701780),广西自然科学/青年基金(2015GXNSFCB139011,2017GXNSFBA198221),广西多源信息挖掘与安全重点实验室开放基金(16-A-01-01,16-A-01-02),广西研究生教育创新计划项目(XYCSZ2017064,XYCSZ2017067,YCSW2017065),广西研究生创新计划项目(YCSW2018094)资助。
李鹏清(1993—),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:1263647631@qq.com;**李扬定**(1986—),男,博士,主要研究方向为医学影像分析、数据挖掘;**邓雪莲**(1979—),女,硕士,主要研究方向为数据挖掘、复杂网络,E-mail:2183451435@qq.com(通信作者);**李永钢**(1989—),男,硕士,主要研究方向为数据挖掘、机器学习;**方 月**(1992—),男,硕士生,主要研究方向为数据挖掘、机器学习。

提出了稀疏子空间聚类(Sparse Subspace Clustering，SSC)，它强制自表征系数稀疏。Lu 等[11]提出了最小二乘回归(LSR)方法来构建关联矩阵。NCut[12]和 RCut[13]较相似，虽然 NCut 算法本身无法得到聚类的最优结果，但是结合某些离散方法进行聚类可以获得相对较好的聚类效果[14]。以上方法虽然都能获得较好的聚类效果，但是在构造相似度矩阵时由于都是基于距离的相似性度量方法，无法很好地处理高维数据。高维空间中数据点之间的距离相近，难以准确描述样本间的相似性关系。

谱聚类效果的好坏很大程度上取决于相似度矩阵的好坏。为了使谱聚类算法获得令人满意的结果，如何构造能够更好地描述数据之间相互关系的相似度矩阵是谱聚类算法的关键。本文通过计算样本间的 SimRank 得分得到相似度矩阵，用谱聚类方法分析相似度矩阵，并得到最终的聚类结果。

本文的创新点是运用 SimRank 得分来构造相似度矩阵，充分考虑了数据点之间的结构相似度，同时解决了传统谱聚类算法难以在高维数据上获得较好的聚类结果这一难题。该算法在计算相似度矩阵时不依赖于点与点之间的距离，而是根据其邻居信息来进行计算。该方法适用于高维数据分析。而传统的方法是运用距离公式计算得到相似度矩阵，如果数据点之间的距离小，则认为它们之间的相似度高。实际上，有些距离较远的数据点间的相似度也会很高；而距离较近的数据点的相似度却可能较低。针对这一问题，本文提出了一种高效的迭代优化方法，即基于 SimRank[15-16]得分的谱聚类算法(SimRank score based Spectral Clustering，SRCSC)。

2　相关研究背景

2.1　SimRank 相似度得分

在实际应用中需要测量不同对象之间的相似度，例如传统文本语料库或万维网中相似文档的查找。更一般地，人们希望能够利用相似性度量来聚类对象。例如，在推荐系统中，根据用户的喜好将相似的用户或项目进行分组。一般而言，相似的对象与类似的对象是相关的，例如，如果对象 o,q 分别和对象 g,h 相关，并且 g,h 是相似的，则对象 o,q 也是相似的，而且对象 g,o 之间的相似性与对象 h,q 之间的相似性很接近，则可以认为 h 是 g 的邻域，同样也可以认为 q 是 o 的邻域，即 g,o 之间的相似性得分可以用 h,q 之间的相似性得分来近似表示。基于以上考虑，两个对象之间的相似性得分可以由它们的邻居的相似性得分来表示，这是 SimRank 得分的基本思想。

2.2　图的分割

考虑社交网络中人与人之间的关系图，其中每个对象(object)被看作是这个网络中的一个节点，每两个对象之间通过边来连接。边代表对象之间是否有联系，而边的权重则表示对象之间关系的亲密程度，也称为权重，其值域为[0,1]，如图 1 所示。

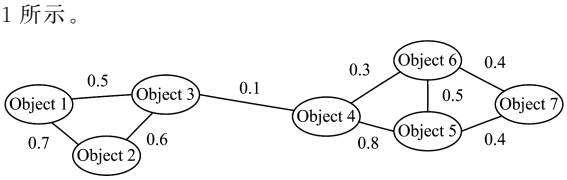

图 1　节点之间关系的亲密度网络图

由图 1 可以看出 Object 1，Object 2，Object 3 的关系较紧密，同时 Object 4 至 Object 7 之间的关系亦相对紧密，而这两个子图是通过 Object 3 和 Object 4 连接起来的。以现实生活为例，甲公司有员工 3 人，分别为员工 1、员工 2、员工 3；乙公司有员工 4 人，分别为员工 4、员工 5、员工 6、员工 7。如果员工 3 和员工 4 刚好相识，但是关系普通，那么就可以把连接甲公司的员工 3 和乙公司的员工 4 之间的边切断，得到两个相互独立的子图，从而实现对这个关系图的一个合适的分割。

2.3　图的拉普拉斯矩阵

给定一个无向图 G(包含 n 个节点)，G 的相似度矩阵 $W\in R^{n\times n}$ 的定义如下：若第 j 个节点与第 i 个节点之间有边相连，则 $W_{ij}=1$；否则，$W_{ij}=0$。所有与第 i 个节点相连的边数即为该节点的度，定义为 $d_i=\sum_j W_{ij}$。令 $D=\mathrm{diag}(d_1,\cdots,d_n)$ 表示无向图 G 的对角矩阵，定义无向图 G 的拉普拉斯矩阵(图论中图的矩阵表示方法)为 $L=D-W$。

3　算法描述

本文提出的算法包括两个步骤：1)建立基于 SimRank 得分的相似度矩阵 W；2)使用谱聚类算法[17-20]得到最终的聚类结果。其中，步骤 2)是根据步骤 1)计算得到的相似度矩阵 W 来建立拉普拉斯矩阵 L，并对 L 进行归一化谱分解，最后运用 k-means[21]聚类方法对谱分解得到的特征向量进行聚类，从而得到最终的结果。

3.1　相似度矩阵的建立

对于给定的图[22] $G=(V,E)$，$V=[x_1,x_2,\cdots,x_n]$ 中的节点代表样本中的对象，E 表示对象之间的关系，节点 x_i 和 x_j 的 SimRank[23]相似度得分为 x_i,x_j 的邻居相似度得分的平均值，其中 $x_i,x_j\in V$ 且 $i,j\in[1,\cdots,n]$。

例如，在图 2 中，m,a,c,d,b,e,f,p 表示数据样本中的 8 个节点，图中节点 m 的相邻节点为 a,c,d，节点 p 的相邻节点为 b,c,e,f。根据 SimRank 得分的定义，m 和 p 的相似性等于 a,c,d 分别与 b,c,e,f 的相似得分的和再除以 m 与 p 邻居个数的乘积。

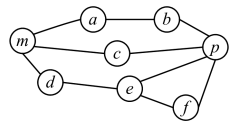

图 2　示例节点关系图

本文用 $s(m,p)\in[0,1]$ 来表示对象 m 和 p 的相似度。根据上文所述，为 $s(m,p)$ 建立基本递归方程：

$$s(m,p)=\frac{C}{|I(m)||I(p)|}\sum_{i=1}^{|I(m)|}\sum_{j=1}^{|I(p)|}s(I_i(m),I_j(p)) \quad (1)$$

其中，$I_i(m)$，$I_j(p)$ 分别代表对象 m 的第 i 个邻居和对象 p 的第 j 个邻居，如果 $m=p$，则 $s(m,p)=1$。对于极端情况，即当 m 或 p 没有邻居时，无法计算出 m 和 p 之间的相似度。因此，当 $I(m)=\emptyset$ 或 $I(p)=\emptyset$ 时，本文将式(1)中的结果定义为 0，而 C 是 0 和 1 之间的常数。根据式(1)可以看出，SimRank 的得分是对称的，即 $s(m,p)=s(p,m)$。

本文通过迭代得到一个固定值，该值即为图 G 的 SimRank 方程的解。假设图 G 中的节点个数为 n，对每对节点迭代 k 次，可以得到 n^2 个 $R_k(*,*)$ 的值，其中 $R_k(m,p)$ 表示

迭代 k 次后 m，p 之间的相似度得分。本文利用迭代法，通过 $R_k(*,*)$来求得 $R_{k+1}(*,*)$。从 $R_0(*,*)$开始计算，其中 $R_0(m,p)$表示实际 $s(m,p)$的 SimRank 得分的初始值，即

$$R_{k+1}(m,p)=\begin{cases}\dfrac{C}{|I(m)||I(p)|}\sum_{i=1}^{|I(m)|}\sum_{j=1}^{|I(p)|}R_k(I_i(m),I_j(p)), & m\neq p\\ 1, & m=p\end{cases} \tag{2}$$

其中，$R_k(*,*)$为 k 的单调不减函数，$\lim_{k\to\infty}R_k(m,p)=s(m,p)$。实验中发现 $R_k(*,*)$的收敛速度较快，因此迭代次数 k 不应设置得太大。根据式(2)可以计算出图 G 的相似度矩阵 W，其中 W 是对称矩阵。

3.2 SRCSC 算法

由 3.1 节得到图 G 的相似度矩阵(即关联矩阵)W，$W\in R^{n\times n}$。根据相似度矩阵，可以通过计算得出无向图 G 的拉普拉斯矩阵 L(Laplacian)：

$$L=D-W \tag{3}$$

其中，矩阵 D 是对称的，$D=diag(d_1,\cdots,d_n)$，$d_i=\sum_j W_{ij}$，d_i 表示矩阵 W 中第 i 行所有元素的和。对矩阵 L 进行归一化谱分解，得到其前 k 个最小特征值所对应的特征向量，并将其组成一个 $k*n$ 的矩阵 P，然后对矩阵 P 进行 k-means 聚类得到向量 F。显然，相似度矩阵中每一列对象所属的类别分别对应于 n 维向量 F，从而获得最后的聚类结果。

本文提出的 SRCSC 算法具有以下优点：

1)该算法在计算相似性矩阵时运用了基于 SimRank 相似性得分的方法，比传统的基于距离的方法具有更明显的优点，该算法不依赖于传统的距离度量。

2)在本文提出的算法中，两个数据点之间的相似度并不是由它们距离来决定的，而是由它们邻居之间的相似程度来共同决定的，从而提高了相似性度量的鲁棒性。

SRCSC 算法的伪代码如算法 1 所示。

算法 1 SRCSC 算法伪代码

输入：训练样本 $X\in R^{d\times n}$，控制参数 C

输出：聚类准确率(Accuracy，ACC)，标准互信息(Normalized Mutual Information，NMI)和纯度(Purity)

1. 利用 knn 方法对数据初始化，得到其邻接矩阵；
2. 利用式(2)计算得到 SimRank 相似度矩阵 W，然后根据 W 建立归一化后的拉普拉斯矩阵 L；
3. 对 L 进行谱分解，得到前 k 个最小的特征值所对应的特征向量，并按列放置形成矩阵 P；
4. 对矩阵 P 进行 k-means 聚类；
5. 计算 ACC，NMI 和 Purity。

4 实验结果及分析

为了验证所提出的 SRCSC 算法的有效性，基于 MATLAB2014a 对其进行了实现，并在安装有 Windows7 操作系统、3.6 GHz CPU、16 GB 内存、1024 GB 硬盘的台式机上进行实验。实验采用了 UCI 上的 6 个标准数据集，包括动物数据集 Zoo[24]、镜头数据集 Lenses、电离层数据集 Ionosphere、澳大利亚手语标志数据集 Australian、汽车数据集 Cars、大肠杆菌数据集 Ecoli。实验数据集的简要介绍如表 1 所列。

$R_0(m,p)=\begin{cases}0 & m\neq p\\ 1 & m=p\end{cases}$。

从 $R_k(*,*)$计算出 $R_{k+1}(m,p)$：

表 1 数据集概况

数据集	样本数	属性数	类别数
Zoo	101	16	7
Lenses	24	4	3
Ionosphere	351	34	2
Australian	690	14	2
Cars	392	8	3
Ecoli	336	343	8

本文选用了 5 个对比算法与本文提出的 SRCSC 算法进行比较，分别是 SSC，LSR(Least Squares Regression)，LRR，RCut(Ratio cut)和 NCut(Normalized Cut)。

5 评价方法及实验结果分析

本文采用最为常用的聚类评价指标 ACC，NMI，Purity 来评价算法的聚类性能。

ACC 是目前最主要的聚类性能评价指标之一，其定义如下：

$$ACC(p,r)=\frac{\sum_{i=1}^{n}\delta(r_i,map(p_i))}{n} \tag{4}$$

其中，$\delta(a,b)=\begin{cases}1, & a=b\\ 0, & \text{其他}\end{cases}$，$map(p_i)$为最佳映射函数。可以将样本聚类标签映射为与之相互等价的样本真实标签，聚类所得到的标签与样本真实标签之间为一一映射关系。

NMI 表示数据的归一化互信息。对于已有的两个随机变量 R 和 S，两者之间的 NMI 可表示成如下形式：

$$NMI(R,S)=\frac{I(R,S)}{\sqrt{H(R)H(S)}} \tag{5}$$

其中，$I(R,S)$为变量 R 和 S 间的互信息，$H(\cdot)$表示信息熵。

Purity 表示数据的纯度，可通过求解以下公式得到：

$$Purity=\max\frac{\sum_{i=1}^{k}\sum_{j=1}^{s}w_{ij}x_{ij}}{n} \tag{6}$$

其中，$\sum_{j=1}^{s}x_{ij}=1,i=1,\cdots,k;\sum_{i=1}^{k}x_{ij}=1,i=1,\cdots,k;x_{ij}=0$ or 1，$i=1,\cdots,k,j=1,\cdots,s;n$ 为样本个数。实际上，Purity 就是每一行的最大值之和除以样本总数。

评价指标 ACC，NMI，Purity 的值越大，表明聚类性能越好，评价指标的取值范围都为[0,1]。各数据集上的实验结果如表 2－表 7 所列。

表 2 Zoo 动物数据集上的实验结果

(单位：%)

算法	*ACC*	*NMI*	*Purity*
SRCSC	76.24	80.20	65.21
LRR	53.47	68.32	46.96
LSR	55.45	72.28	49.65
NCut	56.44	78.28	62.54
RCut	60.40	82.18	67.31
SSC	51.49	69.31	45.52

表 3　Lenses 镜头数据集上的实验结果

（单位：%）

算法	*ACC*	*NMI*	*Purity*
SRCSC	62.50	75.00	35.32
LRR	50.00	62.50	10.37
LSR	50.00	66.67	14.69
NCut	41.67	62.50	7.74
RCut	45.83	62.50	14.99
SSC	54.17	62.50	18.10

表 4　Ionosphere 电离层数据集上的实验结果

（单位：%）

算法	*ACC*	*NMI*	*Purity*
SRCSC	87.18	87.18	41.40
LRR	63.82	64.10	3.57
LSR	55.56	64.10	11.73
NCut	69.23	69.23	10.85
RCut	69.23	69.23	10.85
SSC	53.85	64.10	6.66

表 5　Australian 澳大利亚手语标志数据集上的实验结果

（单位：%）

算法	*ACC*	*NMI*	*Purity*
SRCSC	66.52	66.52	8.45
LRR	65.22	65.22	6.05
LSR	64.49	64.49	6.08
NCut	66.67	66.67	8.54
RCut	66.67	66.67	8.54
SSC	55.65	55.65	0.17

表 6　Cars 汽车数据集上的实验结果

（单位：%）

算法	*ACC*	*NMI*	*Purity*
SRCSC	52.18	69.13	20.99
LRR	48.47	71.43	24.27
LSR	51.02	62.50	13.33
NCut	48.21	69.12	19.74
RCut	48.21	69.12	19.74
SSC	54.59	64.54	20.60

表 7　Ecoli 大肠杆菌数据集上的实验结果

（单位：%）

算法	*ACC*	*NMI*	*Purity*
SRCSC	63.99	82.44	56.14
LRR	37.20	57.44	20.74
LSR	52.68	76.79	44.88
NCut	57.44	83.33	54.56
RCut	58.04	83.33	54.61
SSC	44.05	75.60	37.20

从表 2 动物数据集（Zoo）上的实验结果可以看出，SRCSC 的精确度比其他算法提高了 15.84%～24.75%；SRCSC 的互信息和纯度略低于 NCut，但是要高于其他比较算法。在表 3 所列的镜头数据集（Lenses）上，对于精确度、互信息和纯度而言，SRCSC 都要优于对比算法，其中精确度提高了 8.33%～20.83%，互信息提高了 4.17%～12.5%，纯度提高了 16.95%～27.58%。由表 4 电离层数据集（Ionosphere）结果可以看出，本文实验结果明显优于其他对比算法，精确度提高了 15.95%～33.33%，互信息提高 15.95%～23.08%，纯度提高了 28.28%～37.83%。由表 5 澳大利亚手语标志数据集（Australian）可以看出，SRCSC 算法的精确度和互信息分别低于 NCut，RCut 0.15%，但是要高于其他几个对比算法。另一方面，SRCSC 的纯度和 NCut，RCut 相同，但是要高于其他对比算法 2.4%～8.28%。对于表 6 汽车数据集（Cars）而言，SRCSC 的精确度要稍低于 SSC 算法，但是高于其他对比算法；SRCSC 的互信息和纯度略低于 LRR，但是高于其他对比算法。由表 7 可知，SRCSC 在大肠杆菌数据集（Ecoli）上取得的精确度高于其他几类算法 5.95%～26.79%，而纯度提高了 1.53%～35.4%；虽然 SRCSC 的互信息比 NCut 和 RCut 低 0.89%，但是高出其他算法约 2.98%～25%。

综上可知，无论在聚类准确率、标准互信息或纯度上，SRCSC 算法都要优于其他对比算法。

将 Zoo 数据集嵌入到欧几里德空间中（第二小的和第三小的特征值对应的特征向量作为横坐标和纵坐标），SRCSC 及其他 5 种对比算法的聚类效果如图 3－图 8 所示。

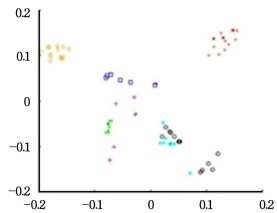

图 3　SRCSC 聚类效果图

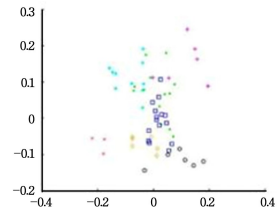

图 4　LRR 聚类效果图

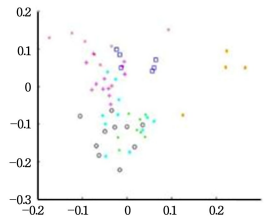

图 5　LSR 聚类效果图

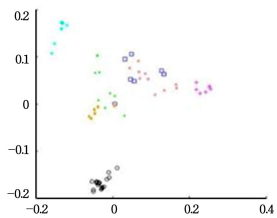

图 6　NCut 聚类效果图

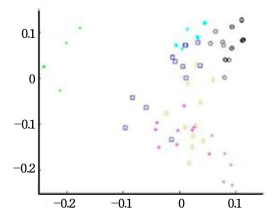

图 7　RCut 聚类效果图

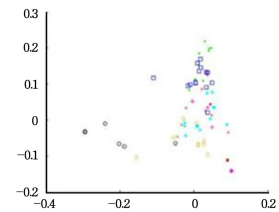

图 8　SSC 聚类效果图

由图 3－图 8 可以看出，SRCSC 的聚类效果比 LRR 等对比算法的聚类效果更好。

通过以上实验结果及相关算法的对比表明，本文提出的 SRCSC 算法优于 SSC 等聚类算法。

结束语　本文通过引入一种基于 SimRank 相似性得分的计算方法，建立基于 SimRank 得分的相似度矩阵来代替传统的基于距离的相似度矩阵；然后建立与所得相似度矩阵对应的归一化拉普拉斯矩阵，并对其进行谱分解；最后进行 k-means 聚类得到最终的结果。通过详细的实验结果验证了本文提出的 SRCSC 算法的有效性。后续将进一步研究如何降低 SRCSC 算法的时间复杂度和空间复杂度，同时提高其聚类准确率。

参 考 文 献

[1]　刘紫涵，吴鹏海，吴艳兰，等. 三种谱聚类算法及其应用研究[J]. 计算机应用研究，2017，34(4)：1026-1031.

[2]　MIGUEL C. On the diameter of the commuting graph of the matrix ring over a centrally finite division ring[J]. Linear Algebra & Its Applications，2016，509：276-285.

[3]　LI X，DU Y，WEI Y，et al. The research of concept context graph layer division based on six degrees of separation theory [J]. Journal of Computational Information Systems，2013，9(22)：9219-9226.

（下转第 467 页）

tion: scalable online collaborative filtering[C]// International Conference on World Wide Web. ACM, 2007: 271-280.

[3] BILLSUS D, PAZZANI M J, CHEN J. A learning agent for wireless news access[C]// Proceedings of the 5th International Conference on Intelligent user Interfaces. 2000: 33-36.

[4] IVÁN C, CASTELLS P. Ontology-Based Personalised and Context-Aware Recommendations of News Items[C]// Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2008: 562-565.

[5] IJNTEMA W, GOOSSEN F, FRASINCAR F, et al. Ontology-based news recommendation[C]// Edbt/icdt Workshops. ACM, 2010: 16.

[6] CANTADOR I, BELLOGÍN A, CASTELLS P. A multilayer ontology-based hybrid recommendation model[J]. Ai Communications, 2008, 21(2-3): 203-210.

[7] 杨武，唐瑞，卢玲. 基于内容的推荐与协同过滤融合的新闻推荐方法[J]. 计算机应用，2016，36(2)：414-418.

[8] 孟祥武，陈诚，张玉洁. 移动新闻推荐技术及其应用研究综述[J]. 计算机学报，2016，39(4)：685-703.

[9] 陶永才，李俊艳，石磊，等. 基于地理位置的个性化新闻混合推荐研究[J]. 小型微型计算机系统，2016，37(5)：943-947.

[10] SON J W, KIM A Y, PARK S B. A location-based news article recommendation with explicit localized semantic analysis[C]// International ACM SIGIR Conference on Reserach and Development in Information Retrieval. ACM, 2013: 293-302.

[11] YOON H G, SONG H J, PARK S B, et al. A personalized news recommendation using user location and news contents[J]. Applied Mathematics & Information Sciences, 2015, 9(2): 439-449.

[12] 张晓艳，王挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005，32(4)：44-48.

[13] 郭喜跃，何婷婷. 信息抽取研究综述[J]. 计算机科学，2015，42(2)：14-17.

[14] 俞鸿魁，张华平，刘群，等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报，2006，27(2)：87-94.

[15] 鞠久朋，张伟伟，宁建军，等. CRF 与规则相结合的地理空间命名实体识别[J]. 计算机工程，2011，37(7)：210-212.

[16] 姚清耘. 基于向量空间模型的中文文本聚类方法的研究[D]. 上海：上海交通大学，2008：27.

[17] 李佳珊. 个性化新闻推荐引擎中新闻分组聚类技术的研究与实现[D]. 北京：北京邮电大学，2013：20-29.

[18] STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[C]// World Text Mining Conference. 2000.

[19] KRISHNA B S V, PROFESSOR S, ENGINEERING M C O, et al. Comparative study of K-means and Bisecting k-means techniques in wordnet based on document clustering[J]. Human Movement, 2012, 13(2): 127-131.

[20] 项亮. 推荐系统实践[M]. 北京：人民邮电出版社，2012：44-59.

（上接第 461 页）

[4] ZHANG J M, SHEN Y X. Review on spectral methods for clustering[C]// Control Conference. IEEE, 2015: 3791-3796.

[5] CHE W F, FENG G C. Spectral clustering: A semi-supervised approach[J]. Neuro Computing, 2012, 77(1): 119-228.

[6] ZHAO Y C, ZHANG S C. Generalized Dimension-Reduction Framework for Recent-Biased Time Series Analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 231-244.

[7] LANGONE R, MALL R, ALZATE C, et al. Kernel Spectral Clustering and Applications[M]// Unsupervised Learning Algorithms. Springer International Publishing, 2016.

[8] 李瑞琳，赵永华，黄小磊. 一种基于 MPI 的稀疏化局部尺度并行谱聚类算法的研究与实现[J]. 计算机工程与科学，2016，38(5)：839-847.

[9] LIU G, LIN Z, YAN S, et al. Robust recovery of subspace structures by low-rank representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[10] ELHAMIFAR E, VIDAL R. Sparse subspace clustering [C]// CVPR. 2009: 2790-2797.

[11] LU C Y, MIN H, ZHAO Z Q, et al. Robust and efficient subspace segmentation via least squares regression[C]// ECCV. 2012: 347-360.

[12] 邹小林，冯国灿. 基于正则割(Ncut)的多阈值图像分割方法[J]. 计算机工程与应用，2012，48(19)：174-178.

[13] WANG S, SISKIND J M. Image Segmentation with Ratio Cut [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(6): 675-690.

[14] SRINIVASARAO P, SURESH K, RAVI K B. Image Segmentation using Clustering Algorithms[J]. International Journal of Computer Applications, 2015, 120: 36-38.

[15] 刘萍，黄纯万. 基于 SimRank 的作者相似度计算[J]. 情报理论与实践，2015，38(6)：109-114.

[16] ZHENG W, ZOU L, CHEN L, et al. Efficient SimRank-Based Similarity Join[J]. Acm Transactions on Database Systems, 2017, 42(3): 16.

[17] CHEN W F, FENG G C. Spectral clustering with discriminate cuts[J]. Knowledge-Based Systems, 2012, 28(7): 27-37.

[18] BOOBALAN M P, LOPEZ D, GAO X Z. Graph clustering using k-Neighbourhood Attribute Structural similarity [J]. Applied Soft Computing, 2016, 47: 216-223.

[19] ALZATE C, SUYKENS J A. Hierarchical kernel spectral clustering[J]. Neural Networks, 2012, 35(2): 21-30.

[20] 刘敏，韩宾，郭有倩. 一种改进的基于 K-means 的信息聚类算法研究[J]. 信息通信，2015(9)：35-36.

[21] FANG R, POUYANFAR S, YANG Y, et al. Computational Health Informatics in the Big Data Age: A Survey[J]. ACM Computing Surveys, 2016, 49(1): 12.

[22] ZHU X F, LI X L, ZHANG S C. Block-Row Sparse Multiview Multilabel Learning for Image Classification[J]. IEEE Transactions on Cybernetics, 2016, 46(2): 450-461.

[23] 李翠平. 一种基于 SimRank 的结点相似度计算方法：CN104933312 A[P]. 2015.

[24] GAO Y, WANG M, TAO D C, et al. 3-D object retrieval and recognition with hypergraph analysis [J]. IEEE Transactions on Image Processing a Publication of the IEEE Signal Processing Society, 2012, 21(9): 4290-4303.