文章编号：2095-2163(2020)03-0241-05 中图分类号：TP-31 文献标志码：A

基于 K-prototype 聚类算法恐怖分子嫌疑度的划分

闫普虹，黄润才，姜 川，孙园园，孙刘成，王从澳
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：当今，恐怖分子作案的多样性和复杂性给相关机构的破案大大增加了难度，如何迅速简便地发现隐藏的恐怖分子，是安全机构最为关心的问题。本文基于 K-prototype 聚类算法，依据恐怖事件发生的数据特征，运用 SPSS 软件对此数据进行标准化处理，得出恐怖分子典型事件的嫌疑度样例的特征向量，通过 Python 进行聚类分析，得到五类别聚类中心分布图。实验结果验证了方法的可行性与有效性，为安全机构对恐怖分子嫌疑度的划分提供了一种分析方法。

关键词：k-mean++；嫌疑度；Python；K-prototype 聚类

Classification of terrorist suspects based on K-prototype clustering algorithm

YAN Puhong, HUANG Runcai, JIANG Chuan, SUN Yuanyuan, SUN Liucheng, WANG Cong´ao
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China))

【Abstract】Currently, the diversity and complexity of terrorist crimes have greatly increased the difficulty of solving relevant cases. How to quickly and easily discover hidden terrorists is the most concerned issue of security agencies. Based on the K-prototype clustering algorithm, this paper uses SPSS software to standardize the data according to the data characteristics of terrorist events, and obtains the feature vector of the suspected terrorist sample. The clustering analysis is performed by Python. Five categories of cluster center distribution map are given out. The experimental results verify the feasibility and effectiveness of the method, and provide an analysis method for the security agencies to divide the terrorists´ suspect degree.

【Key words】k-mean++; suspicion; Python; K-prototype cluster analysis

0 引 言

自美国“911”恐怖袭击以来，恐怖主义的危害性，以及恐怖袭击形式的多样化和复杂化的演变，引起了全社会的关注与重视。研究中发现因恐怖袭击者的由精心策划到“独狼式”随机游走暴动袭击，再加上网络化、全球化的发展，以及没有先验知识分类的影响恐怖袭击发生因素[1]，使许多恐怖案件的侦破变得更加棘手和困难，针对于这些没有事先的经验或一些国际、国内、行业标准的恐怖袭击案件，要对嫌疑程度进行划分和判别，如果直接分类便会显得随意和主观，不能得到科学合理的判断结果，对于海量数据上的处理也不现实。鉴于以上原因，研究可知聚类分析可以根据对象的内在属性，将其聚集成为不同的簇，每一个簇内部相似度高，簇之间差异度大。利用聚类分析的这种特点，可以对海量涉恐情报数据进行自动化、智能化的处理。通过引入以 K-means 改进的 K-prototype 聚类分析算法对标准化后的样本数据进行分析，发现内部高度相似的恐怖团伙，并在此基础上提炼恐怖团伙之间的关系，提升政府、公安机关分析反恐情报的能力和水平，进而提升打击恐怖主义的工作效能[2]，对于政府及公安机关尽早发现新生或隐藏的恐怖分子有着重要的意义与价值。

1 引入多层聚类算法划分恐怖分子嫌疑程度的背景

对于恐怖袭击者嫌疑程度的划分这一问题，现有的成果存在一定问题，缺乏科学方法对样本分类的预处理，只是单独地分析某一地区的情况，或在整体上缺乏对其中重点国家的关注，在新态势上，泛泛而谈者居多，没有对袭击者主题数据进行预处理，或以偏概全，未区分具体国家恐怖袭击事件的发生频度；在原因分析上，单项分析居多，缺乏整体性和完整性，恐怖袭击者嫌疑程度的区分是多种因素相互影响的一个结果，在分析中应该规避单一化或绝对化，以免得到错误的结论。划分聚类可用在对于一个包含 n 个多维对象的集合 D，划分出 k（$k \leqslant n$）个子集合，每个子集合就是一个簇。对于本文研究的主题而言，利用划分聚类可以有效地发现潜在的涉恐人员群体。集合 D 是公安机关掌握的人员的总体，集合中的每一个对象就是一个人员的信息。研究时要识别一个人是否是恐怖分子或者潜在的恐怖

作者简介：闫普虹（1990-），男，硕士研究生，助理工程师，主要研究方向：基于智能算法的电力负荷预测；黄润才（1966-），男，博士，副教授，主要研究方向：信息系统、智能感知与控制。

收稿日期：2019-11-16

分子,仅仅根据单一的指标是无法做到的,必须要根据恐怖分子的历史数据,建立一个基于人员个人信息、活动轨迹、社会交往等多个方面多个指标构成的一个评价体系,因此文中要分析的每一个对象都是多维度的。

2 基于划分的聚类分析恐怖分子嫌疑程度模型建立与算法实现

2.1 基于 K-means 聚类恐怖袭击者嫌疑程度划分算法实现

基于划分的聚类算法可以说是一种基于原型的聚类方法,首先将恐怖袭击事件数据集的对象初始划分为 K 组,每一组表示一个簇,然后反复利用迭代重定位技术将反恐案件在各个簇中重新划分。其中,初始划分原则是:每个簇中至少有一个案件,每个案件只能属于一个簇。好的划分结果标准是:簇内案件特征尽量接近,簇间案件特征互相远离[3-4]。聚类分析法是一种探索性分析方法,能够分析事物的内在特点和规律,并根据相似性原则对事物进行分组,是数据挖掘中常用的一种技术。K-means 基本思想是:在数据集中随机选择一个样本点作为第一个初始化的聚类中心。选择出其余的聚类中心:计算样本中的每一个样本点与已经初始化的聚类中心之间的距离,并选择其中最短的距离,记为 $d-i$ 以概率选择距离最大的样本作为新的聚类中心,重复上述过程,直到 k 个聚类中心都被确定对 k 个初始化的聚类中心,利用 K-Means 算法计算最终的聚类中心[5]。综上可得,算法的整体描述见如下。

(1)为中心向量 c_1, c_2, …, c_k 初始化 k 个种子。

(2)分组:将样本分配给距离其最近的中心向量。由这些样本构造不相交(non-overlapping)的聚类,将样本分配给距离这些样本最近的中心向量,并使目标函数值最小,如式(1)所示:

$$\sum_{i=1}^{n} \min_{j\in\{1,2,\cdots,k\}} \| x_i - p_j \|^2, \tag{1}$$

(3)确定中心:用各个聚类的中心向量作为新的中心。亦须有助于减小目标函数值,原因见式(2):

$$\sum_{i=1}^{m} \| y_i - w \|^2 \leqslant \sum_{i=1}^{m} (\| y_i - \bar{y} \|^2 + \| \bar{y} - w \|^2), \tag{2}$$

等式成立的充要条件为:

$$w = \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i, \tag{3}$$

过程中,需重复分组和确定中心的步骤,直至算法收敛。

进一步可得,算法的具体过程可阐释分述如下。

Step 1 从数据集 $\{x_n\}_{n=1}^{N}$ 中任意选取 k 个赋给初始的聚类中心 $c_1, c_2, \cdots, c_k$。

Step 2 对数据集中的每个样本点 x_i, 计算其与各个聚类中心 c_j 的欧式距离并获取其类别标号:

$$label(i) = \text{argmin} \| x_i - c_j \|^2, i = 1,\cdots,N, j = 1,\cdots,k, \tag{4}$$

Step 3 重新计算 k 个聚类中心,计算公式为:

$$c_j = \frac{\sum_{s:label(s)=j} x_s}{N_j}, \ j = 1,2,\cdots k. \tag{5}$$

Step 4 重复 Step 2~Step 3,直至达到最大迭代次数为止。

2.2 基于 K-prototype 聚类算法在恐怖袭击者嫌疑程度的划分

虽然采用了经典 K-means 聚类算法,能够简单、快速地处理大数据集,而且该算法是相对可伸缩和高效的,但是影响恐怖袭击事件发生的因素来源于全球恐怖主义数据库(GTD),其中包含了 1997~2017 年世界各地恐怖袭击信息事件。与其他数据库不同的是,GTD 拥有 17 万例恐怖案例,针对如此庞大、观察值较多且不同种类的样本数据,K-means 聚类算法存在以下缺点:

在簇的平均值被定义的情况下才能使用,这对于处理符号属性的数据不适用,比如在对影响恐怖袭击发生因素中的一些样本数据:武器信息,附加来源等,K-means 算法的局限性较大,对“噪声”和孤立点数据是敏感的,少量的该类数据能够对平均数据产生极大的影响。对此,为了将恐怖组织或个人在不同时间、地点作案的若干事件进行归类,就需要提前指定 k, 因为 K-means 算法对初始化非常敏感。

因此,选择使用 K-prototype 聚类,K-prototype 聚类是处理混合属性聚类的典型算法,是综合了 k-mean 和 k-mode 算法的思想,并且加入了描述数据簇的原型和混合属性数据之间的相异度计算公式,对处理海量的影响恐怖袭击事件发生的样本数据可以快速有效地进行聚类分析,最终得到恐怖袭击者嫌疑划分的等级。

研究中,根据影响恐怖袭击因素的样本数据的不同特点, 可以按常规定义:$X = \{X_1, X_2, X_3, \cdots\cdots, X_n\}$ 表示数据集(含有 n 个数据),其中数据有 m 个属性。 数据 $X_i = \{X_{11}, X_{12}, X_{13}, \cdots\cdots, X_{1m}\}$,A_j 表示

属性 j，$dom(A_j)$ 表示属性 j 的值域；对于数值属性，值域 $dom(A_j)$ 表示取值范围；对于分类属性，值域 $dom(A_j)$ 表示集合 X_{ij} 中数据 i 的第 j 个属性。同样，数据 X_i 也可表示为：

$$x_i = (A_1 = x_{11}) \wedge (A_2 = x_{12}) \wedge (A_3 = x_{13}) \cdots\cdots \wedge (A_m = x_{1m}), \quad (6)$$

数据总共有 m 个属性，不妨设前 p 个属性为数值属性（r 代表），后 $m-r$ 个属性为分类属性（c 代表），K-prototype 算法是设定了一个目标函数，不断迭代，直到目标函数值不变。同时，K-prototype 算法提出了混合属性簇的原型，由此可以理解原型就是数值属性聚类的质心。混合属性中存在数值属性和分类属性，其原型的定义是数值属性原型用属性中所有属性取值的均值，分列属性原型是分类属性中选取属性值取值频率最高的属性。合起来就是原型。

此时，研究中还涉及到相异度距离。一般来说，数值属性的相异度一般选用欧式距离，在 K-prototype 算法中混合属性的相异度分为属性属性和分类属性，对其分别求值，然后相加。

对于影响恐怖袭击因素分类属性，使用海明威距离，即属性值相同为 0；属性值不同为 1。也就是：

$$d(x_{ij}, x_{pj}) = \begin{cases} 1, & \text{if } x_{ij} \,!\, = x_{pj}, \\ 0, & \text{if } x_{ij} == x_{pj}. \end{cases} \quad (7)$$

对于影响恐怖袭击因素数值属性，计算数值属性对应的欧式距离，则数据和簇的距离（相异度）可写为如下数学公式：

$$d(x_i, Q_l) = \sum_{j=1}^{p} (x_{ij}^r - q_{lj}^r)^2 + \mu_l \sum_{j=p+1}^{m} \delta(x_{ij}^c, q_{lj}^c). \quad (8)$$

其中，前 P 个是数值属性，后 m 个是分类属性，是簇 Q 的原型的 j 属性，u 是分类属性的权重因子。其恐怖分子嫌疑度的目标函数如式（9）所示：

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} d(x_i, Q_l). \quad (9)$$

至此，研究得到的算法步骤详述如下。

输入：聚类簇的个数 k，权重因子

输出：产生好的聚类

Step 1　从数据集中随机选取 k 个对象作为初始的 k 个簇的原型。

Step 2　遍历数据集中的每一个数据，计算数据与 k 个簇的相异度。再将该数据分配到相异度最小的对应的簇中，每次分配结束后，更新簇的原型，并计算目标函数。

Step 3　对比目标函数值是否改变，循环直到目标函数值不再变化为止。

3　实验与仿真分析

根据模式之间的相似性对模式进行分类，K-prototype 算法是一种非监督分类方法。相似性的含义为：有 n 个特征值则组成 n 维向量 $X = [x_1, x_2, \cdots, x_n]$，$X$ 称为该样本的特征向量。这相当于特征空间中的一个点，以特征空间中，点间的距离函数作为模式相似性的测量，以“距离”作为模式分类的依据，距离越小，越“相似”[6]。

首先在样本数据中筛选出任务二给出的恐怖分子关于典型事件的 10 个样例所对应的数据，运用 SPSS 软件对此数据进行标准化处理，然后进行聚类分析，最后得出各个事件相对应的特征向量。恐怖分子关于典型事件 10 个样例的特征向量如图 1 所示，嫌疑程度判断框图如图 2 所示。

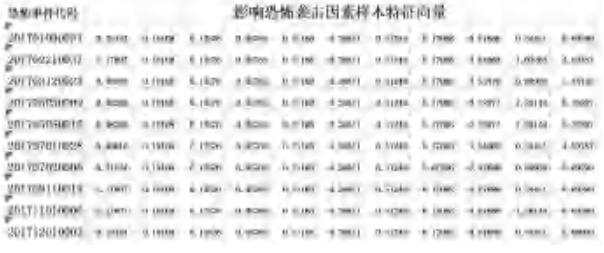

图 1　恐怖分子关于典型事件的嫌疑度样例的特征向量

Fig. 1　The eigenvectors of terrorist suspects for typical events

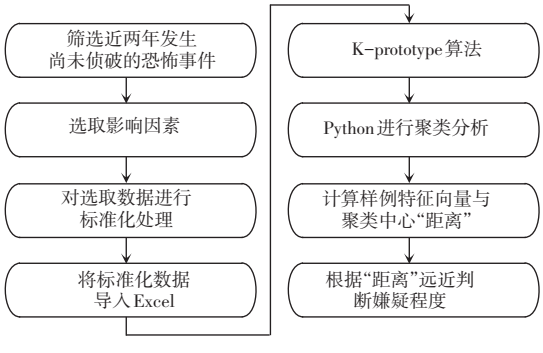

图 2　嫌疑程度判断框图

Fig. 2　Block diagram of suspect degree

在此基础上，筛选出近两年发生的、尚未有组织或个人宣称负责的恐怖袭击事件后，要选取影响恐怖分子关于嫌疑度的影响因素，这里选取的影响因素有：country，extended，crit1，crit2，crit3，doubtterr，success，suicide，attacktype1，targtype1，weaptype1。针对近两年发生的、尚未有组织或个人宣称负责的恐怖袭击事件在选取影响因素下的数据，将其在 SPSS 软件中进行标准化处理；而后将标准化数据导入 Excel 表格；基于 K-prototype 算法，用 Python 对

Excel 表格中数据进行聚类分析；$k = n$ 时，可将其聚为 n 类，但根据程序结果图形可知，将其聚为 5 类时效果最佳[7]。仿真生成的聚类图如图 3 所示。

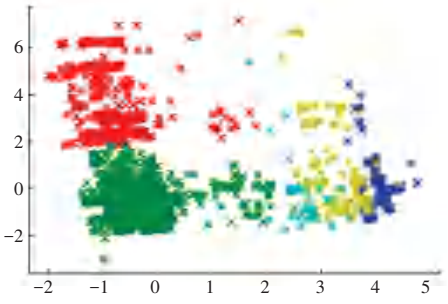

图 3 5 类别聚类中心分布图

Fig. 3 Cluster centers distribution of the five categories

在 Python 编程过程中，通过调用 K－prototype 算法函数，来计算 5 个类别的聚类中心。本文研究中计算出的 5 个类别的聚类中心详见如下。

(1)第一类聚类中心：

[3.005 289 69e−01 1.562 878 46e+00 1.357 577 76e−01
5.764 839 27e−02 1.185 770 04e−01 −1.395 417 48e−01
2.850 600 82e−01 −1.700 815 93e−01 2.521 767 70e+00
−3.411 997 49e−02 1.994 723 14e+00]

(2)第二类聚类中心：

[−6.239 664 53e−02 −1.822 817 61e−01 1.357 577 76e−01
5.764 839 27e−02 3.116 835 86e−01 −3.406 897 14e−01
−6.131 029 37e−02 −4.586 897 93e−03 −2.736 137 76e−01
9.940 345 92e−02 −2.117 192 45e−01]

(3)第三类聚类中心：

[1.348 017 64e−01 −1.384 136 91e−01 1.357 577 76e−01
5.764 839 27e−02 −3.208 122 75e+00 2.728 185 13e+00
2.014 265 40e−01 2.703 895 89e−01 −2.911 168 71e−01
−8.892 568 29e−01 −2.535 869 82e−01]

(4)第四类聚类中心：

[3.242 600 13e−01 2.932 263 11e−01 −7.365 465 42e+00
5.764 839 27e−02 3.116 835 86e−01 2.731 192 22e+00
1.934 173 23e−01 −1.430 765 14e−01 −1.293 261 27e−01
−2.639 028 88e−01 −2.215 299 37e−01]

(5)第五类聚类中心：

[4.839 322 97e−01 −5.877 853 05e−02 1.357 577 76e−01
−1.734 513 59e+01 3.116 835 86e−01 2.731 192 22e+00
1.518 345 07e−01 −2.254 165 21e−02 −3.157 306 05e−01
7.788 116 21e−01 −1.909 595 81e−01]

运用 Python 编程计算样例的特征向量与聚类中心“距离”，并以“距离”的远近判断嫌疑程度。根据程序输出结果，第 $i(i = 1,2,\cdots,10)$ 个样例与第 $j(j = 1,2,\cdots,5)$ 个嫌疑人的相似度见表 1。

表 1 嫌疑人的相似度数值表

Tab. 1 Suspect similarity table

样例	聚类中心	嫌疑相似度
1	1	3.868 758 463 425 298 3
	2	0.802 482 829 361 951 8
	3	4.953 168 023 995 85
	4	8.197 745 832 732 444
	5	17.699 772 540 001 36
2	1	3.861 117 513 106 249
	2	6.552 496 930 888 028
	3	8.037 481 979 880 727
	4	10.393 498 119 099 558
	5	18.943 644 864 732 47
3	1	8.938 912 476 602 331
	2	9.653 088 683 072 356
	3	10.526 076 676 222 468
	4	12.281 298 624 442 131
	5	19.945 970 088 673 008
4	1	9.923 191 600 858 871
	2	9.167 942 248 672 4
	3	10.277 998 925 278 538
	4	12.006 858 994 250 342
	5	19.636 621 359 191 757
5	1	9.923 191 600 858 871
	2	9.167 942 248 672 4
	3	10.277 998 925 278 538
	4	12.006 858 994 250 342
	5	19.636 621 359 191 757
6	1	3.643 786 451 912 132
	2	6.415 282 611 751 556
	3	8.111 227 864 735 142
	4	10.348 947 554 955 686
	5	18.828 927 765 970 377
7	1	7.205 272 550 072 988
	2	6.061 983 403 659 567
	3	7.352 213 375 401 506
	4	10.154 577 266 643 98
	5	18.718 585 480 273 262
8	1	4.104 007 112 748 831
	2	1.359 589 429 812 961 8
	3	5.109 110 818 171 969 5
	4	8.313 994 203 271 236
	5	17.762 225 572 207 008
9	1	4.264 554 377 541 797
	2	1.732 221 476 670 083 6
	3	5.354 683 200 024 032 5
	4	8.414 006 477 781 632
	5	17.767 235 305 072 788
10	1	3.868 758 463 425 298 3
	2	0.802 482 829 361 951 8
	3	4.953 168 023 995 85
	4	8.197 745 832 732 444
	5	17.699 772 540 001 36

根据上述第 $i(i=1,2,\cdots,10)$ 个样例与第 $j(j=1,2,\cdots,5)$ 个嫌疑人的相似度的数据，对任务二中所给出的样例恐袭事件，按嫌疑程度对 5 个嫌疑人排序的结果见表 2。

表 2　恐怖分子关于典型事件的嫌疑度分类表

Tab. 2　Classification of suspects´ suspicion on typical events

	1 号	2 号	3 号	4 号	5 号
样例 XX	4	3	1	2	5
201701090031	2	1	3	4	5
201702210037	1	2	3	4	5
201703120023	1	2	3	4	5
201705050009	2	1	3	4	5
201705050010	2	1	3	4	5
201707010028	1	2	3	4	5
201707020006	2	1	3	4	5
201708110018	2	1	3	4	5
201711010006	2	1	3	4	5
201712010003	2	1	3	4	5

4　结束语

实验证明，K-prototype 聚类克服了对初始化非常敏感和只能对单一数值属性聚类的缺点，对处理海量的影响恐怖袭击事件发生的样本数据可以进行快速有效的聚类分析，最终得到恐怖袭击者嫌疑划分的等级依据，即簇内案件特征尽量接近，簇间案件特征互相远离的标准，通过 Python 进行聚类分析，得到 5 类聚类中心图[8-9]，又绘制出恐怖分子关于典型事件嫌疑度的直观描述图形，并按个人的危害性从大到小选出其中的前 5 个进行嫌疑程度排序，给相关安全机构统一组织侦查和提高破案率提供了一种技术支持。

参考文献

[1] 陈安，陈宁，周龙骧. 数据挖掘技术与应用[M]. 北京：科学出版社，2006.

[2] 夏颖，王哲，程琳. 聚类分析在犯罪数据分析中的应用[J]. 合肥工业大学学报(自然科学版)，2009，32 (12)：1924.

[3] 马立平. 聚类分析法[J]. 北京统计，2000(5)：36.

[4] 王千，王成，冯振元，等. K-means 聚类算法研究综述 [J]. 电子设计工程 ，2012，20(7)：21.

[5] 杨文雅. 聚类分析算法理论研究综述[J]. 华章，2012(23)：305.

[6] OLUKANMI P O，TWALA B. K－means－sharp：Modified centroid update for outlier-robust k-means clustering[C]// 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics(PRASA－Rob Mech)，Bloemfontein：IEEE，2017：14.

[7] 沈艳，余冬华，王昊雷. 粒子群 K-means 聚类算法的改进 [J]. 计算机工程与应用 ，2014，50(21)：125.

[8] 陈磊磊. 不同距离测度的 K-Means 文本聚类研究 [J]. 软件 ，2015，36(1)：56.

[9] 陈小雪，尉永清，任敏，等. 基于萤火虫优化的加权 K-means 算法[J]. 计算机应用研究 ，2018，35(2)：466.

[10] 向培素. 聚类算法综述[J]. 西南民族大学学报(自然科学版)，2011(S1) ：112.

[11] 贾瑞玉，李玉功. 类簇数目和初始中心点自确定的 K-means 算法 [J]. 计算机工程与应用 ，2018，54(7)：152.

[12] RODRIGUEZ A，LAIO A. Clustering by fast search and find of density peaks[J]. Science，2014，344(6191)：1492.

[13] GU Lei. A novel locality sensitive k－means clustering algorithm based on subtractive clustering[C]// 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). Beijing，China：IEEE，2017：836.

[14] XUE Wei，YANG Rongli，HONG Xiaoyu，et al. A novel k－means based on spatial density similarity measurement[C]//2017 29th Chinese Control and Decision Conference (CCDC). Chongqing，China：IEEE，2017：7782.

[15] GANESH S H，PREMKUMAR M S. A median based external initial centroid selection method for K－Means clustering[C]// World Congress on Computing and Communication Technologies (WCCCT). Tamil Nadu，India ：IEEE Computer Society，2017：143.

[16] SINGH J P，BOUGUILA N. Proportional data clustering using K－means algorithm：A comparison of different distances[C]// 2017 IEEE International Conference on Industrial Technology (ICIT).Toronto，ON，Canada：IEEE，2017：1048.

(上接第 240 页)

[2] HUANG L C，YEN T J，CHOU S C T. Community detection in dynamic social networks：A random walk approach [C]// International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011). Kaohsiung，Taiwan：[s.n.]，2011：110.

[3] 刘阳，季新生，刘彩霞. 网络社区发现优化：基于随机游走的边权预处理方法[J]. 电子与信息学报，2013，35(10)：2335.

[4] 杨海陆，张健沛，杨静. 利用 2-hop 随机游走进行异质网络社区发现[J]. 哈尔滨工程大学学报，2015，36(12)：1626.

[5] 辛宇，杨静，谢志强. 基于随机游走的语义重叠社区发现算法[J]. 计算机研究与发展，2015，52(2)：499.

[6] XIN Y，XIE Z Q，YANG J，et al. An adaptive random walk sampling method on dynamic community detection[J]. Expert Systems With Applications，2016，58：10.

[7] OKUDA M，SATOH S，SATO Y，et al. Community detection using restrained random-walk similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019：1－1(Early Access).