词聚类算法

1、背景

词的聚类我们不控制输出的类的数量，是类似基于密度的聚类。

机器学习中，常见的基于密度的聚类算法，多数将文本转化为数值向量后，通过计算向量之间的距离d，来进行聚类，比如：

DBSCAN ：不控制类的半径，通过密度可达的点的数量实现聚类；

meanshift： 控制类的有限半径，通过不停更新质心实现聚类。

对于近义词聚类算法，我们要实现的效果更像DBSCAN那样，是不控制类中最终的词的数量的（即不控制类半径）。

但是，由于近义词的识别过程中，我们使用了多个模型进行融合，其中既包括计算特征距离的相似算法（w2v_sim,tfidf_sim)，也包括计算空间距离的关联算法(relevance),

这样多个向量的距离阈值的控制都是不同的，且模型之间不是单纯的平行关系，融合条件很复杂，无法直接代入任何现有聚类模型。

遂，对于词的两两之间的距离计算并标识相似(相近）的算法比较复杂，无法在聚类过程中嵌入实现，但却可以最终输出两两词之间是相近关系的场景，这里设计了：

密度收敛可达的聚类算法 ：

类的延伸不是发散的，而是每次密度可达(相近)的新的点的数量，都要较上一次发生收敛，若出现发散或 连续无法收敛的次数达到阈值次数的情况，则剔除发散/无法收敛的点后，存余的点即可生成新类 。

区分DBSCAN：DBSACN是类的延伸应该是发散的，若发散的数量达不到阈值，则生成新类

2、算法设计及逻辑流

* 数据源：

已确定相似/相近的两两组合的词

* 收敛：

指下一次拓展出来的新点new_points较本次拓展出来的点current_points的数量差异，

new_points>current_points ：发散 ，触发发散，类在current收敛

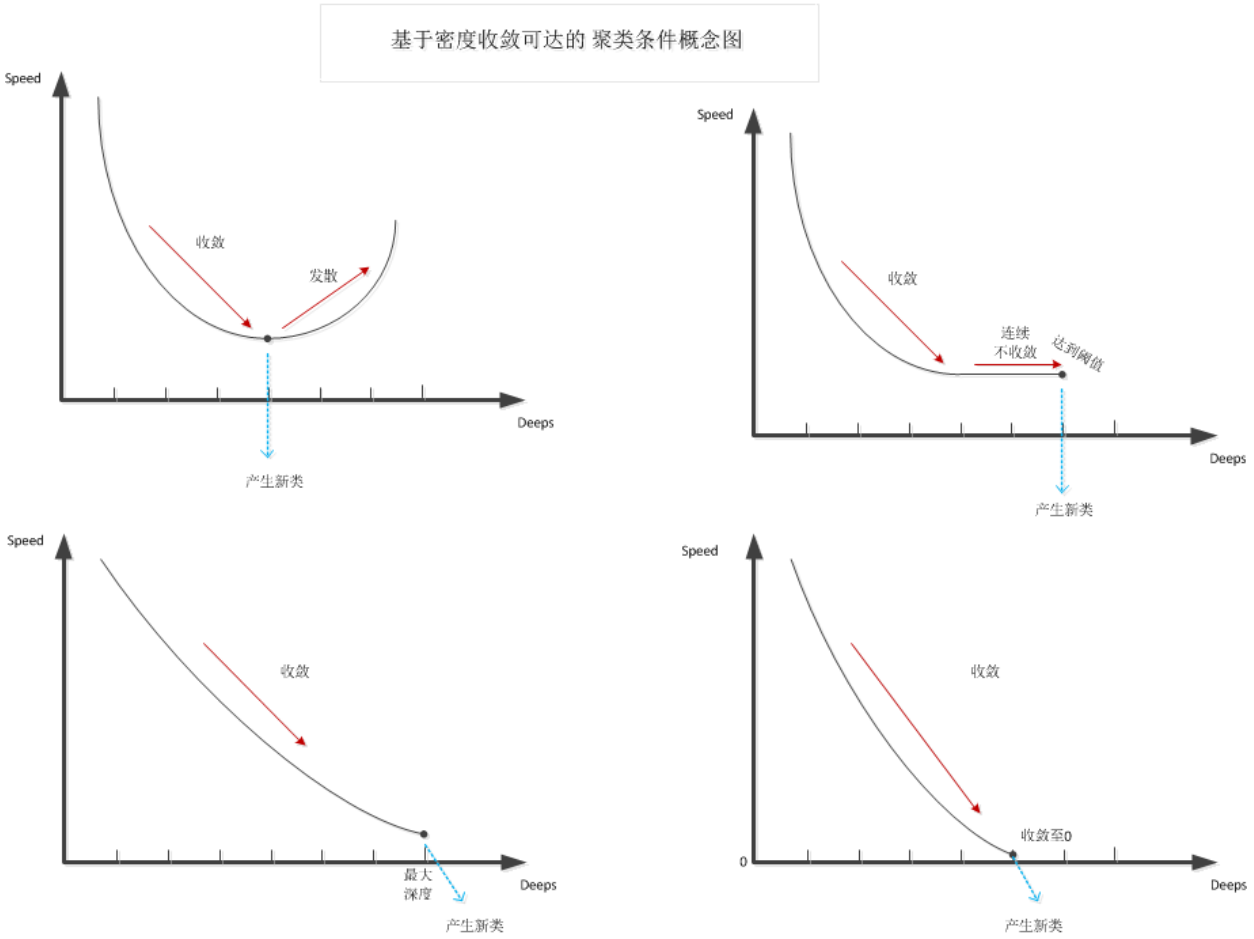
new_points=current_points ：不收敛，连续不收敛的次数存在阈值限制，超过阈值,收敛

new_point=0 ：收敛至0，即不再可达新的点，收敛完成。

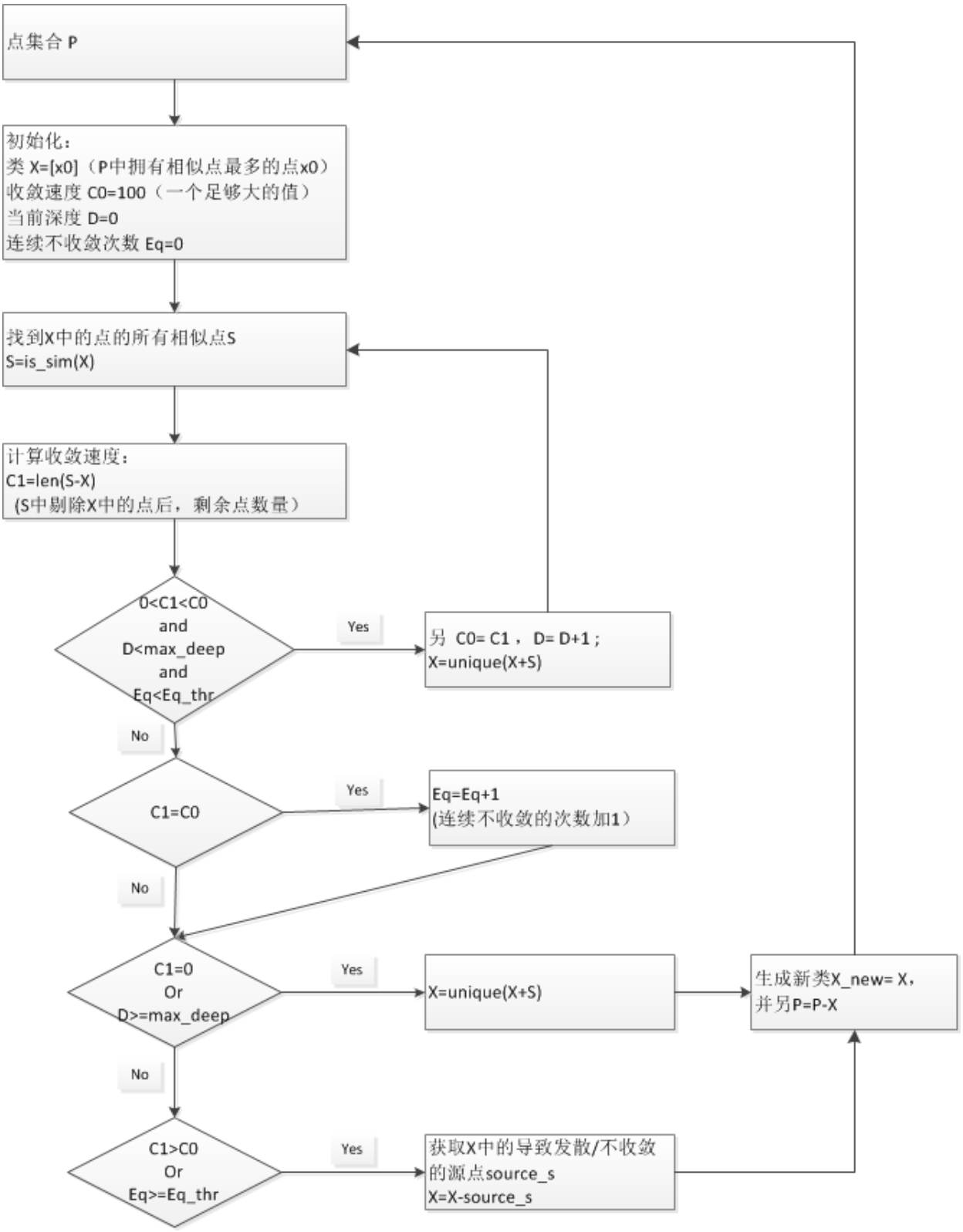
* 最大深度：

聚类过程中，进行类拓展的最大次数，即每个类的最大迭代次数。

* 聚类条件：


基于密度收敛可达的 聚类条件概念图
Speed
收敛
发散
Deeps
产生新类
Speed
收敛
连续
不收敛
达到阈值
Deeps
产生新类
Speed
收敛
最大
深度
Deeps
产生新类
Speed
收敛
0
收敛至0
Deeps
产生新类

* 聚类流程：


点集合 P
初始化：
类 X=[x0]（P中拥有相似点最多的点x0）
收敛速度 C0=100（一个足够大的值）
当前深度 D=0
连续不收敛次数 Eq=0
找到X中的点的所有相似点S
S=is_sim(X)
计算收敛速度：
C1=len(S-X)
(S中剔除X中的点后，剩余点数量）
0<C1<C0
and
D<max_deep
and
Eq<Eq_thr
Yes
另 C0= C1 ， D= D+1 ;
X=unique(X+S)
No
C1=C0
Yes
Eq=Eq+1
(连续不收敛的次数加1）
No
C1=0
Or
D>=max_deep
Yes
X=unique(X+S)
生成新类X_new= X，
并另P=P-X
No
C1>C0
Or
Eq>=Eq_thr
Yes
获取X中的导致发散/不收敛
的源点source_s
X=X-source_s