



(12) 发明专利申请

(10) 申请公布号 CN 104657472 A

(43) 申请公布日 2015. 05. 27

(21) 申请号 201510079140. 7

(22) 申请日 2015. 02. 13

(71) 申请人 南京邮电大学

地址 210023 江苏省南京市亚东新城区文苑
路 9 号

(72) 发明人 陈志 陈骏 岳文静

(74) 专利代理机构 南京经纬专利商标代理有限
公司 32200

代理人 叶连生

(51) Int. Cl.

G06F 17/30(2006. 01)

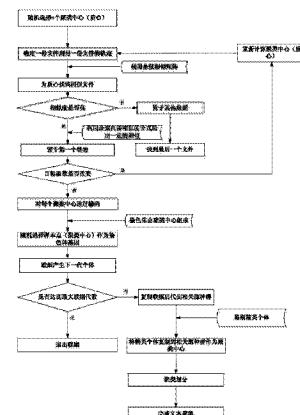
权利要求书2页 说明书5页 附图2页

(54) 发明名称

一种基于进化算法的英文文本聚类方法

(57) 摘要

本发明给出一种英文文本聚类方法，该方法首先将英文文本进行预处理成向量空间模型，然后在聚类过程中，第一步是随机选取 n 个聚类中心，对于聚类中心，利用欧氏距离进行聚类的划分，使同一类的文本归为一个聚类，这样完成得到一个局部最优的聚类划分；第二步是进行进化算法的处理，所用到的是联姻的思想以及基因交叉变异的过程进行新一代聚类中心的选择，通过与文本间距离最近的原则进行聚类划分从而达到全局最优。本发明能够对英文文本进行有效聚类，剔除不必要的聚类结果，使得聚类过程较快收敛。



1. 一种基于进化算法的英文文本聚类方法,其特征在于该方法包括以下步骤:

步骤 1) 将用户提供的多个英文文本拆分成单词,删除长度小于 2 的单词;删除停用词,将删除后的单词形成新的单词集合;所述停用词是由用户指定,一般为那些对文本标识没有太大作用的单词,主要功能是消除所有文本中出现频率都很高的词;

步骤 2) 统计用户所提供文本集中的文本总数、统计每个文本中删除后的单词总数;统计出在新的单词集合中每个单词出现在各文本中的数量,统计出新的单词集中每个单词所出现过的文本数;

步骤 3) 对所有的单词按照其权值从大到小排序,提取 4-6 个的权重较大的单词作为文本的特征表示;所述权值表示为 $f_i(d) * (\log \frac{1.1 * N}{n_i} + 0.5)$; $f_i(d)$ 为词频;所述词频是在该文本中,该单词的数量除以该文本中的总单词数;所述 N 为总文本数, n_i 为文本集合中含有该词的文本数;

步骤 4) 随机选取 2-4 个聚类中心即质心文本;利用欧氏距离

$Dis(d_p, d_q) = \sqrt{\sum_{k=1}^m (x_{pk} - x_{qk})^2}$ 进行文本的距离度量,为质心文本找到相似文本;

x_{pk}, x_{qk} 分别表示文本 d_p, d_q 的第 k 个文本特征的权重;利用余弦相似度公式,

$Sim(D_1, D_2) = \cos \theta = \frac{\sum_{k=1}^n W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^n W_{1k}^2)(\sum_{k=1}^n W_{2k}^2)}}$ 计算出非质心文本与质心的相似度,用户通过

给出阈值来判断相似度的高低,将相似度高的置于第一群集,相似度低的置于第二群集,直到找到最后一个文件为止;判断目标函数是否改变;所述 W_{1k} 表示第一个文本的权重, W_{2k} 表示第二个文本的权重,k 表示第 k 个特征项;所述目标函数是指所输入的文件与聚类中心的距离是否改变;重新计算聚类中心,重复步骤 4);

步骤 5) 对每个聚类中心进行编码,将聚类中心表示为染色体,随机选择样本点即聚

类中心作为染色体基因;确定适应度函数 $Fitness(I) = \frac{1}{1 + \sum_{\alpha=1}^m \sum_{X_{\alpha} \in C_{\beta}} \|X_{\alpha} - Z_{\beta}\|}$; 所述编

码是指将特征值转化为二进制表示;所述适应度函数是基于欧氏距离进行的相似度度量, $Fitness(I)$ 表示个体 I 的适应度, X_{α} 为属于类 C_{β} 的样本点, Z_{β} 为第 β 个聚类中心;

步骤 6) 将染色体基因进行联姻产生下一代个体,并鉴别精英个体;然后采用轮盘赌法,由适应度函数对应的概率分布确定把当前群体中的第 α 个个体 I 按照选择概率

$P_s(I_{\alpha}) = \frac{Fitness(I_{\alpha})}{\sum_{\beta=1}^m Fitness(I_{\beta})}$ 抽出,并进行交叉和变异;所述联姻是仿照生物学中父代双亲结合产生下一代的过程;所述精英个体是指在聚类过程中总是获得较高相似度的聚类中心;

相似度的高低通过用户给出的阈值判断;所述轮盘赌是一种赌博方式,每一种选择的方式都是完全随机没有任何人为操作的;所述 $P_s(I_{\alpha})$ 表示第 α 个个体被选择的概率;

步骤 61) 从当前群体中按轮盘赌法选择两条染色体,随机选取交叉位置,将两条染色体从交叉点处分成两段,按概率 P_c 一次将两条染色体的右半段互换并重新连接,得到两

条新染色体；

步骤 62) 随机选择基因变异的位置,以用户指定的变异概率 P_m 对这些位置的基因进行变异,所述变异概率一般在 0.01 至 0.3 之间;

步骤 7) 复制联姻后代到上一代种群中,若精英个体优于联姻后代,则将精英个体复制到相关源种群作为新的聚类中心,否则仍使用联姻后代作为聚类中心;当达到联姻最大代数的时候停止联姻,确定聚类中心;在聚类中心确定的情况下,聚类划分采用与文本距离最近 $Dis(X_\alpha, Z_\beta) = \min_\gamma(Dis(X_\alpha, Z_\beta))$ 的原则确定, γ 为聚类数;最后使得聚类的划分更加准确,同一聚类的文本相似度更高。

一种基于进化算法的英文文本聚类方法

技术领域

[0001] 本发明涉及一种英文文本聚类方法,利用一种局部聚类的方法对文本进行聚类中心的选择,再利用一种进化算法进行全局聚类,属于机器学习、文本挖掘、统计分析、信息检索交叉技术应用领域。

背景技术

[0002] 随着数据库技术和互联网技术普及和发展,人们因为大量数据已经陷入了“数据丰富,知识贫乏”的尴尬境地。面对浩瀚的数据海洋,不知所措。信息量虽然巨大,但对于用户来说,所需要信息只是其中很小的一部分。如何从浩瀚的文本信息资源中准确获取所需信息,已成为信息处理的一个关键问题。文本挖掘指的是从大量的文本集合中发现潜在的模式和知识的过程。文本聚类是文本挖掘的主要技术之一。

[0003] 文本聚类是一种集成机器学习、模式识别、统计分析和信息检索技术于一体的文本挖掘方法,其特点是在不需要训练集和预定义类别的情况下,即可从给定的文档集合中找到合理的聚类划分。通过文本聚类,可将文档集合划分为若干簇,并使同一簇中的文档具有尽可能大的相似度,簇间文档保持尽可能小的相似度,为信息的查询和检索提供了较好的优化和分析方法。

[0004] 典型的文本聚类方法有很多种,其中 K-Means 算法因其简单和高效性,在文本聚类中占有重要地位。由于 K-Means 算法在聚类中心的计算过程中采用了启发式方法,因而有效地降低了算法复杂度,提高了运算速度。也因为如此,使得该算法对初始聚类中心的选择较为敏感,易于陷入局部最优解。

[0005] 遗传算法 (Genetic Algorithm) 是一类借鉴生物界的进化规律 (适者生存, 优胜劣汰遗传机制) 演化而来的随机化搜索方法。它是由美国的 J. Holland 教授 1975 年首先提出,其主要特点是直接对结构对象进行操作,不存在求导和函数连续性的限定;具有内在的隐并行性和更好的全局寻优能力;采用概率化的寻优方法,能自动获取和指导优化的搜索空间,自适应地调整搜索方向,不需要确定的规则。遗传算法的这些性质,已被人们广泛地应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域。它是现代有关智能计算中的关键技术。遗传算法也是计算机科学人工智能领域中用于解决最优化的一种搜索启发式算法,是进化算法的一种。这种启发式通常用来生成有用的解决方案来优化和搜索问题。进化算法最初是借鉴了进化生物学中的一些现象而发展起来的,这些现象包括遗传、突变、自然选择以及杂交等。

发明内容

[0006] 技术问题:本发明的目的是提供一种基于进化算法的英文文本聚类方法,将局部聚类算法和进化算法相结合对多个英文文本进行聚类,先将文本进行局部聚类以选出一批合适的聚类中心,以余弦相似度进行文本的相似度度量,而后利用进化算法进行源种群聚类中心的选取以及聚类的划分以达到全局聚类的效果,解决聚类中心的选取过于随机而无

法得到最优聚类、同一聚类文件相似度不高以及聚

[0007] 类次数过多而产生不必要的聚类结果使得聚类无法收敛等问题。

[0008] 技术方案：本发明所述的一种基于进化算法的英文文本聚类方法，将文本预先处理成为向量集，根据公式计算出单词的权重作为特征项来进行文本表示，然后随机选取聚类中心进行文本的局部聚类，最后通过进化算法中种群的联姻思想进行新一代聚类中心的选择，通过与文本间距离最近的原则进行聚类划分从而达到

[0009] 全局最优。

[0010] 本发明所述的英文文本聚类的方法包括以下步骤：

[0011] 步骤 1) 将用户提供的多个英文文本拆分成单词，删除长度小于 2 的单词；删除停用词，将删除后的单词形成新的单词集合；所述停用词是由用户指定，一般为那些对文本标识没有太大作用的单词，主要功能是消除所有文本中出现频率都很高的词；

[0012] 步骤 2) 统计用户所提供文本集中的文本总数、统计每个文本中删除后的单词总数；统计出在新的单词集合中每个单词出现在各文本中的数量，统计出新的单词集中每个单词所出现过的文本数；

[0013] 步骤 3) 对所有的单词按照其权值从大到小排序，提取 4-6 个的权重较大的单词作为文本的特征表示；所述权值表示为 $f_i(d) * (\log \left| \frac{1.1 * N}{n_i} \right| + 0.5)$ ； $f_i(d)$ 为词频；所述词频是在该文本中，该单词的数量除以该文本中的总单词数；所述 N 为总文本数， n_i 为文本集合中含有该词的文本数；

[0014] 步骤 4) 随机选取 2-4 个聚类中心（质心文本）；利用欧氏距离

$$Dis(d_p, d_q) = \sqrt{\sum_{k=1}^m (x_{pk} - x_{qk})^2}$$

进行文本的距离度量，为质心文本找到相似文本； x_{pk} , x_{qk} 分别表示文本 d_p , d_q 的第 k 个文本

特征的权重；利用余弦相似度公式， $Sim(D_1, D_2) = \cos \theta = \frac{\sum_{k=1}^n W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^n W_{1k}^2)(\sum_{k=1}^n W_{2k}^2)}}$ 计算出非

质心文本与质心的相似度，用户通过给出阈值来判断相似度的高低，将相似度高的置于第一群集，相似度低的置于第二群集，直到找到最后一个文件为止；判断目标函数是否改变；

所述 W_{1k} 表示第一个文本的权重， W_{2k} 表示第二个文本的权重， k 表示第 k 个特征项；所述目标函数是指所输入的文件与聚类中心的距离是否改变；重新计算聚类中心，重复步骤 4)；

[0015] 步骤 5) 对每个聚类中心进行编码，将聚类中心表示为染色体，随机选择样本点

(聚类中心) 作为染色体基因；确定适应度函数 $Fitness(I) = \frac{1}{1 + \sum_{\alpha=1}^m \sum_{X_{\alpha} \in C_{\beta}} \|X_{\alpha} - Z_{\beta}\|}$ ；所

述编码是指将特征值转化为二进制表示；所述适应度函数是基于欧氏距离进行的相似度度量， $Fitness(I)$ 表示个体 I 的适应度， X_{α} 为属于类 C_{β} 的样本点， Z_{β} 为第 β 个聚类中心；

[0016] 步骤 6) 将染色体基因进行联姻产生下一代个体，并鉴别精英个体；然后采用轮盘赌法，由适应度函数对应的概率分布确定把当前群体中的第 α 个个体 I 按照选择概率

$$P_s(I_\alpha) = \frac{Fitness(I_\alpha)}{\sum_{\beta=1}^m Fitness(I_\beta)}$$

抽出，并进行交叉和变异；所述联姻是仿照生物学中父代双亲结合产生下一代的过程；所述精英个体是指在聚类过程中总是获得较高相似度的聚类中心；相似度的高低通过用户给出的阈值判断；所述轮盘赌是一种赌博方式，每一种选择的方式都是完全随机没有任何认为操作的；所述 $P_s(I_\alpha)$ 表示第 α 个个体被选择的概率；

[0017] 步骤 61) 从当前群体中按轮盘赌法选择两条染色体，随机选取交叉位置，将两条染色体从交叉点处分成两段，按概率 P_c 一次将两条染色体的右半段互换并重新连接，得到两条新染色体；

[0018] 步骤 62) 随机选择基因变异的位置，以用户指定的变异概率 P_m 对这些位置的基因进行变异，所述变异概率一般在 0.01 至 0.3 之间；

[0019] 步骤 7) 复制联姻后代到上一代种群中，若精英个体优于联姻后代，则将精英个体复制到相关源种群作为新的聚类中心，否则仍使用联姻后代作为聚类中心；当达到联姻最大代数的时候停止联姻，确定聚类中心；在聚类中心确定的情况下，聚类划分采用与文本距离最近 $Dis(X_\alpha, Z_\beta) = \min_\gamma(Dis(X_\alpha, Z_\beta))$ 的原则确定， γ 为聚类数；最后使得聚类的划分更加准确，同一聚类的文本相似度更高。

[0020] 有益效果：

[0021] 1) 本发明提供一种英文文本聚类的方法，整个过程思路清晰，易于理解，首先借用的 K-means 算法的思想过程简单，使用者可以很快了解即可使用，而后的进化算法与之相结合，算法表示清楚，相关技术概念也较容易理解。

[0022] 2) 本发明所述聚类过程中，通过两种算法的结合，不断地优化聚类中心以及聚类的划分从而使得最终的结果由局部最优到全局最优。

[0023] 3) 本发明中所述的联姻过程，提供了整个联姻的方法已经下一代个体的选取法则，能够使得新的聚类中心更加合适，新的聚类划分更加准确，从而达到全局最优化。

[0024] 4) 本发明中所述的联姻方法中包含遗传算法的部分，因此搜索使用评价函数启发，过程简单；使用轮盘赌法进行迭代，具有随机性，具有可扩展性；比较容易和其他的算法相结合进行问题的处理。

附图说明

[0025] 图 1 文本预处理的方法流程图，

[0026] 图 2 一种英文文本聚类的方法流程图。

具体实施方式

[0027] 本发明以文本挖掘作为背景，对多个英文文本进行聚类，目的是根据类别的不同来获取更有价值的信息，本发明根据图 1 进行文本的预处理，向量化；根据图 2 进行文本间的聚类。具体实例如下所述：

[0028] 1. 将 4 个文本中的每一个文本拆分成单词，对每个文本中的单词进行长度分析；删除长度小于 2 的单词，删除停用词；

[0029] 2. 统计出 4 个文本的总单词数，每个文本中每个单词的数量，计算单词 a 在所在

文本中的词频 $f_i(d)$, 判断单词 a 是否在在该文本中出现, 出现过标记为 1 ;未出现过标记为 0, 统计出单词 a 所出现过的文本数 ;这里以一号文本 D1 作为例子 ,D1 的总单词数为 1000 个, 单词 a 出现在 3 个文本中, 单词 b 出现在 3 个文本中, 单词 c 出现在 4 个文本中, 单词 d 出现在 3 个文本中 ;各项参数如表 3 :

[0030] 3. 计算每个单词的权重 $f_i(d) * (\log \left| \frac{1.1 * N}{n_i} \right| + 0.5)$; , 由大到小依次选择 5 个单词

作为文档的特征项, 权重作为这 5 个单词的特征值来进行文本表示, 1 号文本 D1 的特征项为 a, b, c, d, 权重分别为 30, 20, 20, 10 ;2 号文本 D2 的特征项为 a, c, d, e, 权重分别为 40, 30, 20, 10 ;3 号文本 D3 特征项为 b, c, d, e, 权重分别为 30, 20, 10, 10 ;4 号文本 D4 特征项为 a, b, c, e ;权重分别为 40, 20, 10, 10 ;分别用向量表示为 D1(30, 20, 20, 10, 0) , D2(40, 0, 30, 20, 10) , D3(0, 30, 20, 10, 10) , D4(40, 20, 10, 0, 10) ;

[0031] 4. 随机选取 2 个聚类中心, 利用欧氏距离 $Dis(d_p, d_q) = \sqrt{\sum_{k=1}^m (x_{pk} - x_{qk})^2}$ 进行文本

的距离度量, 利用余弦相似度公式 $Sim(D_1, D_2) = \cos \theta = \frac{\sum_{k=1}^n W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^n W_{1k}^2)(\sum_{k=1}^n W_{2k}^2)}}$ 计算出文件

与质心的相似度 ,D1、D4 作为聚类中心, 则通过公式计算 D2 与 D1 相似度为 0.86 ;D3 与 D1 相似度为 0.67 ;D2 与 D4 相似度为 0.78 ;D3 与 D4 相似度为 0.49 ;显然 D2, D3 更适合归为 D1 一类 ;

[0032] 5. 目标函数有改变, 重新计算聚类中心, 经过几轮聚类之后得出文本相似度如表 2 ;选择 D1, D2 作为聚类中心可以得到相对最优的聚类划分, 即 D3, D4, D1 为第一聚类, D2 为一个聚类, 可以看出聚类中心无论选取哪两个都不能得出最好的聚类, 目标函数无法取得最小值 ;只能做到局部最优 ;

[0033] 6. 对每个聚类中心进行二进制编码, 聚类中心表示为染色体, 选择样本点 D1, D2, 把它们作为染色体基因, D1(011110, 010100, 010100, 001010, 0) , D2(101000, 0, 011110, 010100, 001010) , 确定适应度函数

$$Fitness(I) = \frac{1}{1 + \sum_{\alpha=1}^m \sum_{X_{\alpha} \in C_{\beta}} \|X_{\alpha} - Z_{\beta}\|},;$$

[0034] 7. 将 2 个染色体基因进行两两联姻产生下一代个体, 通过单点交叉和变异产生两个新的聚类中心, 并进行精英个体的鉴别, 交叉我们选择最末两位进行交叉, 交叉出来得到的新的个体 D1, D2 为分别为 :D1(011100, 010100, 010110, 001000, 000010) , D2 进行交叉之后再让第四个特征项的第三位进行变异为 (101010, 0, 011100, 010010, 001000) ;则对应的特征值分别为 D1(28, 20, 22, 8, 2) , D2(42, 0, 28, 18, 8) ;

[0035] 8. 在聚类中心改变的情况下进行文本间聚类, 联姻之后的聚类相似度如表 3 ;由此看出, D3, D4 与 D1 的相似度非常高, 分别达到了 0.86 和 0.9, 所以 D1 则可以作为精英个体进行保存 ;再次进行下一步联姻 ;

[0036] 9. 当达到联姻最大代数的时候停止联姻, 确定聚类中心 ;在聚类中心确定的情况

下,聚类划分采用与文本距离最近 $Dis(X_\alpha, Z_\beta) = \min_\gamma(Dis(X_\alpha, Z_\beta))$ 的原则确定;所述 γ 为聚类数,最后使得聚类的划分更加准确,同一聚类的文件相似度更高。

[0037] 表 1 一号文本各项参数表

[0038]

单词	数量	频率	所出现过的文本数
a	450	0.45	3
b	300	0.3	3
c	370	0.37	4
d	150	0.15	3
e	0	0	0

[0039] 表 2 聚类算法迭代完成时各文本间的相似度表

[0040]

	D1	D2	D3	D4
D1	1	0.86	0.67	0.90
D2	0.86	1	0.42	0.78
D3	0.67	0.42	1	0.49
D4	0.90	0.78	0.49	1

[0041] 表 3 联姻完成后各文本间的相似度表

[0042]

	D1	D2	D3	D4
D1	1	0.86	0.71	0.90
D2	0.86	1	0.39	0.80
D3	0.71	0.39	1	0.49
D4	0.90	0.80	0.49	1

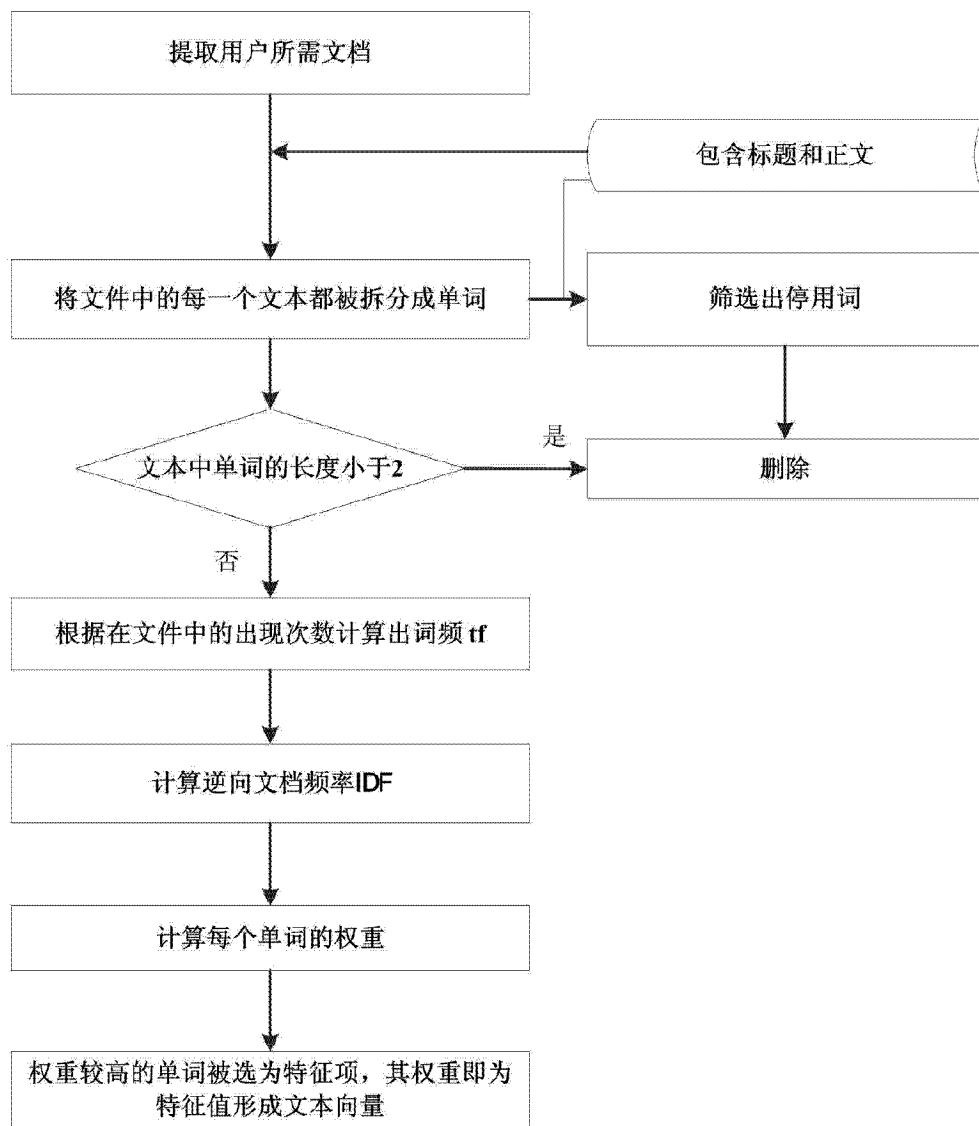


图 1

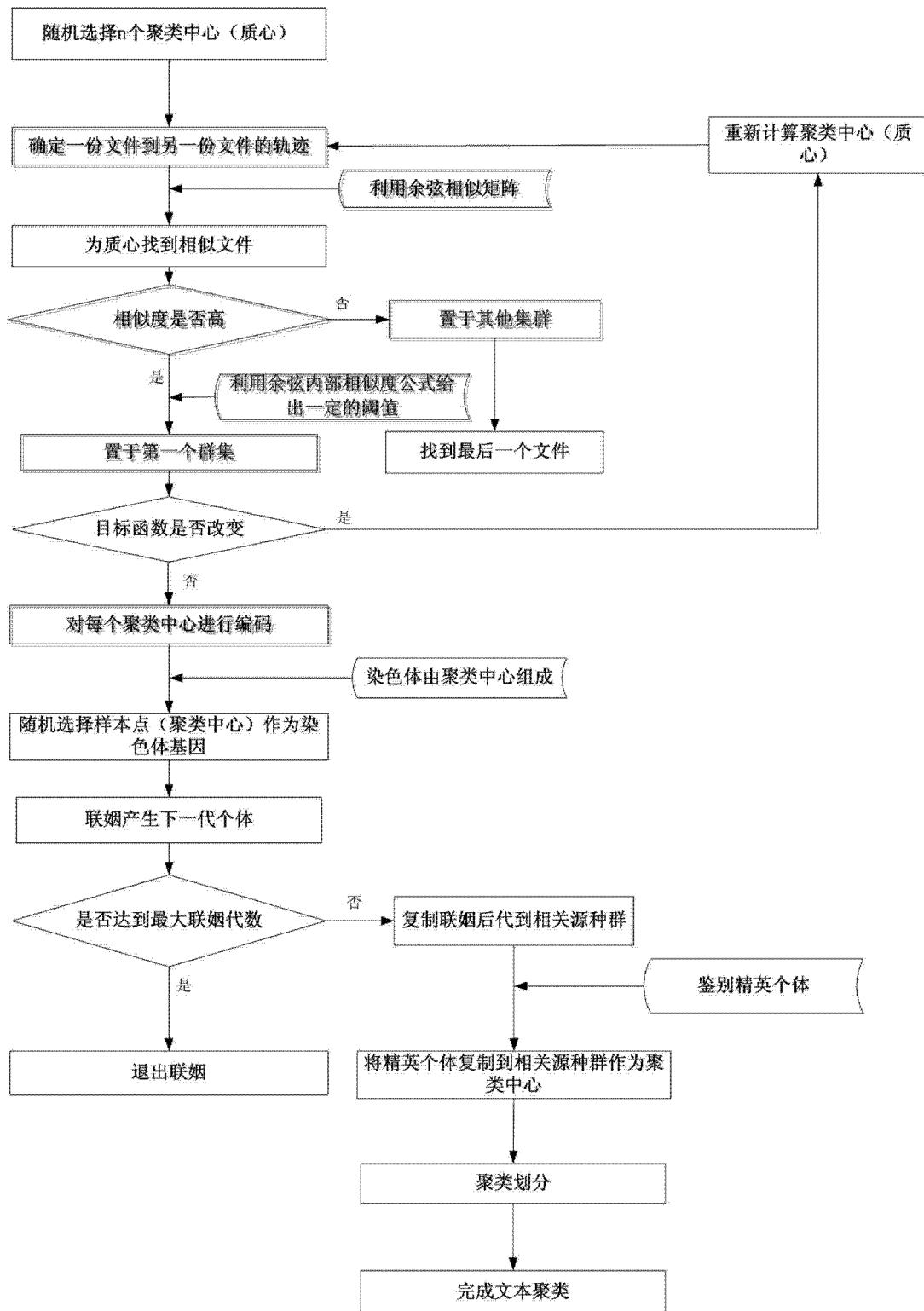


图 2