(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102929894 A
(43)申请公布日 2013.02.13

(21)申请号 201110230978.3

(22)申请日 2011.08.12

(71)申请人 中国人民解放军总参谋部第五十七研究所
地址 610041 四川省成都市 393 信箱 C62

(72)发明人 金烨 徐诗恒

(51)Int.Cl.
*G06F 17/30*(2006.01)

权利要求书 3 页 说明书 8 页 附图 4 页

(54)发明名称

一种文本在线聚类可视化方法

(57)摘要

一种文本在线聚类可视化方法，属于属于计算机学科下的智能信息处理领域。本发明的目的在于，通过引入用户对类别特征词汇标注信息，实现对聚类过程的约束和优化，提升文本聚类结构的清晰度和可理解性；并且设计了文本在线式聚类技术，实现对文本数据流的增量聚类，保持聚类结构的总体稳定，并自适应更新模型。本发明设计了一种在线式高维数据降维布局方法，能够适应大规模数据或数据流环境；通过对聚类后的文本类别分布向量进行降维布局，实现对文本数据的增量式可视化，在二维或三维欧氏空间中实现对文本数据及其类别结构的可视化展示。

CN 102929894 A

1. 一种文本在线聚类可视化方法，其特征在于，包括基于词汇标注的文本在线聚类、在线式高维数据降维可视化两大步骤：

所述的基于词汇标注的文本在线聚类步骤为：

步骤 a，用户设置聚类数目，并对其中部分或者全部类别提供若干特征词汇；

步骤 b，统计初始文本集合中的单词词频信息，采用 LDA 模型对数据进行建模，并利用标注的类别特征词汇对 LDA 模型进行约束，采用 Gibbs Sampling 技术求解模型参数；

步骤 c，模型参数中的文档类别分布 θ 用于文本类别的预测，模型参数中的词汇-类别分布频次 n(w，z) 将作为约束参数，用于增量聚类过程；

步骤 d，在线聚类时，新文本数据在已有模型参数 n(w，z) 基础上进行初始化，然后按照步骤 b 和步骤 c 进行建模运算，计算完成后，新文本实现增量聚类，模型参数实现自动更新；

所述在线式高维数据降维可视化步骤为：

步骤 e，对文本聚类得到的高维类别分布向量，计算任意两向量间的相似性，同时随机产生对应低维向量初始值，计算任意两低维向量间的相似性；

步骤 f，利用 KL 距离 (Kullback-Leibler Divergence) 度量高维向量相似性集合与低维向量相似性集合间的差异；

步骤 g，通过最优化方法迭代搜索步骤 f 中相似性集合间差异的最小值，同时不断更新低维向量，达到设定误差范围时停止迭代，利用可视化工具对低维向量可视化；

步骤 h，在线式处理时，对新到来的高维向量降维利用了已产生的低维向量信息，在迭代搜索时已产生的低维向量不再更新，只对新到来的高维向量按照步骤 e、步骤 f 和步骤 g 作增量式处理；

2. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 a 中，对于用户设置的类别数目 K，用户可以选择从中任意标注若干个类别；对于所选的类别，用户只需提供少量特征词汇，也可以提供标注文本。

3. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 b 中，单词 $w_j$ 在文本 $d_i$ 出现的频次为 $n(d_i, w_j)$，单词 $w_j$ 对于类别 $z_k$ 的采样总频次为 $n(w_j, z_k)$，文本 $d_i$ 中所有单词对于类别 $z_k$ 的采样总频次为 $n(d_i, z_k)$。

4. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 b 中，利用标注词汇对于初始化模型进行修正，计算公式为：

$$n(w_j, z_k) = \begin{cases} n(w_j, z_k) & w_j \text{非} z_k \text{的标注特征词汇} \\ n(w_j, z_k) + C & w_j \text{是} z_k \text{的标注特征词汇} \end{cases}$$

其中，C 是指标注强度系数。

5. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 b 中，Gibbs Sampling 计算公式如下：

$$P(z = k \mid z_{\neg,k}, w_j, d_i) \propto \frac{n(w_j, z_k)_{\neg} + \beta}{n(z_k)_{\neg} + W \cdot \beta} (n(d_i, z_k)_{\neg} + \alpha)$$

其中，$n(w_j, z_k)_{\neg}$ 和 $n(d_i, z_k)_{\neg}$ 表示采样点 $(d_i, w_j)$ 当前的标注状态已经从统计量中移除后的剩余频次值，α 和 β 分别表示 Dirichlet 先验分布参数。

6. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 c 中，文档 $d_i$ 对于不同类别的概率分布 θ，其计算公式如下：

$$\theta_{d_i}(z_k) = P(z_k \mid d_i) = \frac{n(d_i, z_k) + \alpha}{\sum_{k'=1}^{K} n(d_i, z_{k'}) + K \cdot \alpha}$$

7. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 d 中，新数据在原有模型参数基础上初始化，其实现方式为：首先对新数据中的词汇随机标记类别，然后统计新数据的词汇标记频次 $n(w_j, z_k)$ 和 $n(d_i, z_k)$，标记完成后，利用原模型中词汇类别分布频次，对新数据的词汇分布进行修正，修改公式如下：

$$n(w_j, z_k) = \begin{cases} n(w_j, z_k) & w_j \text{是新词汇} \\ n(w_j, z_k) + n^{(0)}(w_j, z_k) & w_j \text{是原有词汇} \end{cases}$$

其中，$n^{(0)}(w_j, z_k)$ 表示原模型中的词汇类别分布频次。

8. 根据权利要求 1 所述的基于词汇标注的文本在线聚类，其特征在于，所述的步骤 d 中，对于新文本的增量聚类可按照标准 LDA 模型求解的方法，无需固定原有模型参数，当 GibbsSampling 达到停止条件后，通过计算 θ 实现对新文本的类别判断，同时模型参数 $n(w_j, z_k)$ 也已自动进行了修正。

9. 根据权利要求 1 所述的在线式高维数据降维可视化，其特征在于，在所述步骤 e 中，对文本聚类得到的高维类别分布向量 $x_1 x_2 \ldots x_n$，$x_i$、$x_j$ 间的相似性 $p_{ij}$ 定义为：

$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$，其中 $p_{j|i} = \frac{\exp(-\| x_i - x_j \|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\| x_i - x_k \|^2 / 2\sigma_i^2)}$，$\sigma_i^2$ 为将 $x_i$ 到 $x_k$($k \neq i$) 的欧氏距离视为高斯分布的方差，记与 $x_1 x_2 \ldots x_n$ 相对应的低维数据为 $y_1 y_2 \ldots y_n$，$y_i$、$y_j$ 间的相似性 $q_{ij}$ 定义为：

$$q_{ij} = \frac{(1 + \| y_i - y_j \|^2)^{-1}}{\sum_{k \neq l} (1 + \| y_k - y_l \|^2)^{-1}}$$

10. 根据权利要求 1 所述的在线式高维数据降维可视化，其特征在于，在所述步骤 f 中，高维数据相似性集合 $\{p_{ij}\}$ 与低维数据相似性集合 $\{q_{ij}\}$ 间的 KL 距离 $D_{KL}$ 定义为：

$$D_{KL} = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

11. 根据权利要求 1 所述的在线式高维数据降维可视化，其特征在于，在所述步骤 h 中，记 $\beta_1 \beta_2 \ldots \beta_m$ 是已经产生的低维向量，对应的高维向量为 $\alpha_1 \alpha_2 \ldots \alpha_m$，假设 $x_1 x_2 \ldots x_n$ 是需要在线式处理的高维向量，对应的低维向量为 $y_1 y_2 \ldots y_n$，对于任意 $x_i$($i = 1, 2 \ldots n$)，其与 $\alpha_j$($j = 1, 2 \ldots m$) 的相似性 $p_{j|i}$ 定义为：

$$p_{j|i} = \frac{\exp(-\| x_i - \alpha_j \|^2 / 2\sigma_i^2)}{\sum_{k=1}^{m} \exp(-\| x_i - \alpha_k \|^2 / 2\sigma_i^2)}$$

$\sigma_i^2$ 为将 $x_i$ 到 $\alpha k$($k = 1, 2 \ldots m$) 的欧氏距离视为高斯分布的方差，$x_i$ 和 $\alpha_j$ 分别对应的低维向量 $y_i$ 和 $\beta_j$ 间的相似性 $q_{j|i}$ 定义为：

$$q_{j|i} = \frac{(1+\| y_i - \beta_j \|^2)^{-1}}{\sum_{k=1}^{m}(1+\| y_i - \beta_k \|^2)^{-1}}$$

$\{p_{j|i}\}$、$\{q_{j|i}\}$ 间的 KL 距离 $D_{KL}$ 定义为：

$$D_{KL} = \sum_{j=1}^{m} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

# 一种文本在线聚类可视化方法

## 技术领域

[0001] 本发明属于计算机学科下的文本智能信息处理技术，具体涉及一种在线式的文本聚类可视化方法。

## 背景技术

[0002] 文本数据是最为重要的信息载体之一，对文本信息的浏览和处理时常见的工作场景。随着信息量的激增，用户迫切需要一种新的计算机技术，能够对源源不断到来的数据进行自动分类和管理，以方便用户按照类别浏览和查询。如果数据量进一步增大，传统文本队列就不再完全胜任对文本信息的显示要求，此时需要以二维或者三维可视图的方式，对聚类的结果进行直观显示，以方便用户更便捷的了解信息分布态势，实现对信息的准确获取。

[0003] 在文本聚类算法中，David M.Blei 提出的 Latent Dirichlet Allocation（简称 LDA）模型是一种获得广泛应用的生成模型，通过从文本的特征分析入手，探索不同数据间在特征分布上具有的共性分布，再利用贝叶斯分析技术计算这些共性分布符合的分布参数，从而实现对文本建模并依据模型参数实现对文本的聚类划分。中国专利 CN101968798.A 公开了一种对 LDA 模型进行在线算法，并用于社区推荐的方法。该方法对于新数据的处理方式是一种文本分类与在线更新的方法，即通过初始数据聚类获得模型，然后固定模型用于对新数据进行分类，再利用新数据对模型进行更新训练，因此不是一种在线聚类算法；其次，该方法没有引入用户的先验信息，所获得的聚类结构往往不符合用户对于类别的先验需求。

[0004] 在文本可视化方面，Laurens van der Maaten 和 Geoffrey Hinton 提出了 t-SNE 算法，算法假定高维文本特征向量空间符合 Gaussian 分布，降维后低维欧氏空间中对应的坐标点符合 t- 分布，算法采用 KL 散度函数来评估高维数据和低维数据间分布的差异性，并且通过最小化 KL 散度函数，来探索低维欧氏空间的一组坐标点，使得这组坐标点能够尽量保持与高维数据同样的分布结构。L 和 G 等人所提出的 t-SNE 算法只能对批量式数据进行处理，算法对数据的容量较为有限，不能支持对文本数据流进行在线式处理。

## 发明内容

[0005] 本发明的主要目的在于改进 LDA 模型，使之能够接受用户以词汇标注方式提供的先验信息，从而提高聚类结构对于用户的实用性；同时提出一种在线聚类方法，能够完成对文本数据流在线式聚类并自动更新模型；另一方面，还提出一种文本在线可视化方法，能够对聚类结构进行增量降维布局显示。

[0006] 本发明的目的是通过如下技术方案实现的：

[0007] 一种文本在线聚类可视化方法，包括基于词汇标注的文本在线聚类、在线式高维数据降维可视化。

[0008] 所述的基于词汇标注的文本在线聚类步骤为：

[0009] 步骤 a，聚类任务设置，用户根据任务需要设置聚类的数目 K，如果用户有明确定

义的类别，允许用户提供少量特征词汇（通常是 5 ～ 20 个词汇）以标示类别；

[0010] 步骤 b，文本预处理，对于集合 D 中的文本，统计文本中的词汇出现频次（如果是中文数据，则需要先进行中文分词处理），以 $d_i$ 表示集合中的第 i 个文本，以 $w_j$ 表示集合中所有词汇形成的词汇表 W 中的第 j 个词，以 $n(d_i, w_j)$ 表示第 j 个词 $w_j$ 在第 i 个文本 $d_i$ 中出现的频次，以 N 表示集合中文本总数，以 M 表示集合词汇表词汇总数，以 Z 表示类别；

[0011] 步骤 c，采用 LDA 模型对集合中文本进行建模，并利用类别特征词汇对模型进行约束和优化，再利用 Gibbs Sampling 进行模型求解运算，实现文本聚类，具体过程如下：

[0012] 步骤 c1，随机初始化，为 D 中每一份文本 d 的每一个词汇 w($w \in W$)，随机标注一个类别 z($z \in Z$)；然后统计：$n(d_i, z_k)$，表示文本 $d_i$ 中标注为第 k 个类别的词频总数；$n(d_i)$，文本 $d_i$ 词汇总数（计重复）；$n(w_j, z_k)$，表示词汇 $w_j$ 在所有文本中被标注为第 k 个类别的总频数；$n(z_k)$，所有词汇被标注为第 k 个类别的总频数；

[0013] 步骤 c2，标注信息约束初始化，利用标注词汇对于初始化模型参数进行修正，计算公式为：

[0014]
$$n(w_j, z_k) = \begin{cases} n(w_j, z_k) & w_j \text{非} z_k \text{的标注特征词汇} \\ n(w_j, z_k) + C & w_j \text{是} z_k \text{的标注特征词汇} \end{cases}$$

[0015] 其中，C 是指标注强度系数，通常取 50 ～ 5000 间的整数，并修改 $n(z_k) = \sum_{j=1}^{M} n(w_j, z_k)$；

[0016] 步骤 c3，采样，利用 Gibbs sampling 技术对文本中的词汇进行随机采样，具体公式如下：

[0017]
$$P(z = k \mid z_{\neg,k}, w_j, d_i) \propto \frac{n(w_j, z_k)_{\neg} + \beta}{n(z_k)_{\neg} + W \cdot \beta} (n(d_i, z_k)_{\neg} + \alpha)$$

[0018] 其中，$n(w_j, z_k)_{\neg}$ 和 $n(d_i, z_k)_{\neg}$ 表示采样点 ($d_i$, $w_j$) 当前的标注状态已经从统计量中移除后的剩余频次值，α 和 β 分别表示 Dirichlet 先验分布参数；

[0019] 步骤 c4，模型参数计算，Gibbs sampling 达到停止条件后，按照下面公式计算文本 d 对于类别的分布概率 $\theta_d$，同时保存所有词汇的类别采样频次 n(w, z)：

[0020]
$$\theta_{d_i}(z_k) = P(z_k \mid d_i) = \frac{n(d_i, z_k) + \alpha}{\sum_{k'=1}^{K} n(d_i, z_{k'}) + K \cdot \alpha}$$

[0021] 步骤 c5，文本类别判断，对于文本 d，取 $\theta_d$ 的最大分量所在类别为 d 的判断类别，即 $z(d) = \arg\max_k \theta_d(z_k)$；

[0022] 步骤 d，新数据在线聚类，利用已有模型词汇类别分布频次 $n^{(0)}(w, z)$，实现对新文本数据聚类的增量式聚类，以及 $n^{(0)}(w, z)$ 的自动更新，具体过程如下：

[0023] 步骤 d1，对新文本数据进行预处理，然后按照 c1 步骤进行随机初始化；

[0024] 步骤 d2，对新文本数据中的词汇 w，如果在原模型已经存在，则按照下面的公式修改 w 的类别分布频次，否则不进行改动：

[0025]

$$n(w_j,z_k)=\begin{cases} n(w_j,z_k) & w_j\text{是新词汇} \\ n(w_j,z_k)+n^{(0)}(w_j,z_k) & w_j\text{是原有词汇} \end{cases}$$

[0026] 步骤 d3，按照 c3-c5 对新文本数据进行聚类；

[0027] 步骤 d4，对新的词汇类别分布频次 n(w,z)，如果 w 是新出现的词汇，则添加到原始模型中，如果 w 不是新词汇，则用 n(w,z) 替换原有分布频次 $n^{(0)}(w,z)$。

[0028] 所述在线式高维数据降维可视化步骤为：

[0029] 步骤 e，对文本聚类得到的高维类别分布向量 $x_1x_2...x_n$（其中 $x_i$ 即文本 $d_i$ 聚类得到的类别分布向量 $\theta_{d_i}$），计算任意两向量 $x_i$、$x_j$ 间的相似性 $p_{ij}$，同时随机产生对应的低维向量初始值 $y_1y_2...y_n$，计算任意两向量 $y_i$、$y_j$ 间的相似性 $q_{ij}$。$p_{ij}$ 和 $q_{ij}$ 计算方法如下：

[0030] $$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}$$

[0031] $$q_{ij}=\frac{(1+\|y_i-y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k-y_l\|^2)^{-1}}$$

[0032] 其中 $p_{j|i}=\frac{\exp(-\|x_i-x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i-x_k\|^2/2\sigma_i^2)}$，$\sigma_i^2$ 为将 $x_i$ 到 $x_k$（$k\neq i$）的欧氏距离视为高斯分布的方差；

[0033] 步骤 f，利用 KL 距离度量 $\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的差异，$\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的 KL 距离 $D_{KL}$ 定义为：

[0034] $$D_{KL}=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

[0035] 步骤 g，通过最优化方法（梯度下降法）寻找 $D_{KL}$ 的最小值，同时不断更新低维向量 $y_1y_2...y_n$。梯度下降法按照 $D_{KL}$ 的负梯度方向逐步搜索最优解，将 $D_{KL}$ 对 $y_i$ 求偏导数，有

[0036] $$\frac{\partial D_{KL}}{\partial y_i}=4\sum_j(p_{ij}-q_{ij})(y_i-y_j)(1+\|y_i-y_j\|^2)^{-1}$$

[0037] 记 $Y=(y_1,y_2...y_n)$，$Y^{(t)}=(y_1^{(t)},y_2^{(t)}...y_n^{(t)})$ 为经过 t 次迭代后的解，梯度下降法采用下式更新 Y：

[0038] $$Y^{(t)}=Y^{(t-1)}+\eta\frac{\partial D_{KL}}{\partial Y}+\alpha(t)(Y^{(t-1)}-Y^{(t-2)})$$

[0039] 其中 η 为步长，通常取 50 ～ 500 间的整数，α(t) 为加速因子，在 0 ～ 1 间取值。迭代若干次后，即可得到满足预定误差要求的低维向量 $y_1y_2...y_n$，利用可视化工具对 $y_1y_2...y_n$ 可视化。

[0040] 步骤 h，在线式降维处理时，随机选择已经产生的低维向量 $\beta_1\beta_2...\beta_m$，对应的高维向量为 $\alpha_1\alpha_2...\alpha_m$。假设 $x_1x_2...x_n$ 是需要增量式处理的高维向量，对应的低维向量为 $y_1y_2...y_n$。对于任意 $x_i$（$i=1,2...n$），其与 $\alpha_j$（$j=1,2...m$）的相似性 $p_{j|i}$ 定义为：

[0041] $$p_{j|i}=\frac{\exp(-\|x_i-\alpha_j\|^2/2\sigma_i^2)}{\sum_{k=1}^{m}\exp(-\|x_i-\alpha_k\|^2/2\sigma_i^2)}$$

[0042] $\sigma_i^2$ 为将 $x_i$ 到 $\alpha_k$（$k=1,2...m$）的欧氏距离视为高斯分布的方差。$x_i$ 和 $\alpha_j$ 分别

对应的低维向量 $y_i$ 和 $\beta_j$ 间的相似性 $q_{j|i}$ 定义为：

[0043] $$q_{j|i}=\frac{(1+\|y_i-\beta_j\|^2)^{-1}}{\sum_{k=1}^{m}(1+\|y_i-\beta_k\|^2)^{-1}}$$

[0044] $\{p_{j|i}\}$、$\{q_{j|i}\}$ 间的 KL 距离 $D_{KL}$ 定义为：$D_{KL}=\sum_{j=1}^{m}p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}$ 同样利用梯度下降法可以求出 $y_1y_2...y_n$，将 $y_1y_2...y_n$ 添加到可视化图形中。

## 附图说明

[0045] 图 1 是文本在线聚类可视化技术流程图。

[0046] 图 2 是基于词汇标注的文本在线聚类流程图。

[0047] 图 3 是文本在线聚类随词汇标注系数的性能变化图。

[0048] 图 4 是文本在线布局可视化技术流程图。

[0049] 图 5 是文本在线聚类可视化技术二维效果图。

[0050] 图 6 是文本在线聚类可视化技术三维效果图。

## 具体实施方式

[0051] 为了使本发明的目的和优点更清楚，下面结合附图和具体实施方式对本发明作进一步描述。

[0052] 如图 1 所示，为应用本发明对文本进行在线聚类可视化的系统示意图。系统首先搜集一定数量的历史数据作为初始文本数据，这些数据无需标注文本的类别；系统对初始数据进行聚类，以获得初始文本数据的文本类别分布向量参数和词汇类别分布频次参数，前者为高维数据降维布局方法初始布局提供数据来源，并计算获得布局模型参数，后者作为模型数据作为约束参数，用于文本在线聚类；处理在线文本数据时，系统对每次获取得到的数据进行在线聚类，获得的文本类别分布向量，再由降维布局计算低维空间坐标，最后根据用户的要求，以二维或者三维的方式进行可视化展示。

[0053] 本实例涉及的实验数据，均来自复旦中文文本数据语料库，数据共分为 10 类，总共有约 2800 份文本，其中最多的类别约有 500 份文本，最小的类别约有 200 份。

[0054] 本发明需要由用户预先设置聚类类别数目，通常根据用户对数据的熟悉程度，在一个较为合适的数值范围内选择；对于有明确先验信息的类别，允许用户为这些类别分别提供若干关键词，关键词数目通常在 5 ～ 20 之间，也可以更多，这些词汇能够较好的表述类别的主题内容，允许不同类别之间出现同样的词汇。

[0055] 本发明中文本预处理过程需要利用分词软件实现对中文文本的切分，我们主要采用中科院计算所提供的 ICTCLAS 软件，并且软件标注的词性中筛选名词和动词作为文本的实意词汇。这些内容并不在本发明的范围内。

[0056] 下面对文本在线聚类过程进行详细说明。

[0057] 如图 2 所示，文本在线聚类过程分为初始聚类和在线聚类两部分，初始聚类利用词汇标注信息对 LDA 模型进行约束，实现对初始数据的聚类，并获取模型参数；在线聚类保持初始聚类类别结构，实现对文本数据流的在线聚类，并实时更新模型参数。图 3 是词汇标

注的聚类与无标注聚类的性能对比图。

[0058] 初始聚类的详细过程如下：

[0059] 步骤 101，初始数据预处理，通过将中文数据进行分词和词性挑选，并进一步删除只在一份文本中出现过的词汇，将文本转换为由词汇频次表征的词频向量；

[0060] 经过预处理后，定义聚类数据各种符合为：D 表示文本集合，共有 N 份文本，$d_i$ 表示其中第 i 份文本；W 表示由 D 中词汇生成的词汇表，共有 M 个词汇，$w_j$ 表示其中第 j 个词汇；Z 表示类别集合，共有 K 个类别，$z_k$ 表示其中第 k 个类别；$n(d_i, w_j)$ 表示 $w_j$ 在 $d_i$ 中出现的频次；

[0061] 步骤 102，初始化参数，首先对文本向量中的词汇，从 Z 中随机挑选类别进行标注，并统计：$n(d_i, z_k)$，表示文本 $d_i$ 中所有单词对于类别 $z_k$ 的采样总频次，$n(d_i)$，表示文本 $d_i$ 的计重复词汇总数，$n(w_j, z_k)$，表示单词 $w_j$ 对于类别 $z_k$ 的采样总频次，$n(z_k)$，表示类别 $z_k$ 获得的词汇采样总频次；其次，利用用户在 Z 中若干类别所标注的词汇信息，对 $n(w_j, z_k)$ 进行修正，计算公式为：

[0062]

$$n(w_j, z_k) = \begin{cases} n(w_j, z_k) & w_j \text{非} z_k \text{的标注特征词汇} \\ n(w_j, z_k) + C & w_j \text{是} z_k \text{的标注特征词汇} \end{cases}$$

[0063] 其中，C 是指标注强度系数，通常取 50 ～ 5000 间的整数，并修改 $n(z_k) = \sum_{j=1}^{M} n(w_j, z_k)$；

[0064] 步骤 103，迭代采样，我们通常采用单链条采样，链条长度一般设置为 1000 ～ 2000；其中，一次采样的实施过程具体如下：

[0065] 步骤 103a，设置文本下标 i 从 1 到 N：

[0066] 步骤 103b，设置词汇下标 j 从 1 到 M：

[0067] 步骤 103c，如果 $w_j$ 没有在 $d_i$ 中出现，则转步骤 103b；否则，提取，当前词汇的标注类别为 z，将 $n(d_i, z)$、$n(w_j, z)$ 和 $n(z)$ 的数值各自减 1；

[0068] 步骤 103d，设置类别下标 k 从 1 到 K，分别计算词汇从当前类别 z 到类别 k 的转移概率，计算公式如下：

[0069] $$P(z = k \mid z_{\neg,k}, w_j, d_i) \propto \frac{n(w_j, z_k)_{\neg} + \beta}{n(z_k)_{\neg} + W \cdot \beta} (n(d_i, z_k)_{\neg} + \alpha)；$$

[0070] 其中，α 和 β 分别表示 Dirichlet 先验分布参数，一般我们固定取 α 为 0.1，β 为 0.01；

[0071] 步骤 103e，根据前一步计算得到的 $P(z = k \mid z_{\neg,k}, w_j, d_i)$，依比例进行随机采样，将得到的类别 z′ 作为当前词汇的新采样状态，将 $n(d_i, z')$、$n(w_j, z')$ 和 $n(z')$ 的数值各自加 1；

[0072] 步骤 104，最后 100 次采样时，每一次采样完成后，根据当前的采样状态计算模型参数，计算公式如下：

[0073] $$\theta_{d_i}^{(n)}(z_k) = P(z_k \mid d_i) = \frac{n(d_i, z_k) + \alpha}{\sum_{k'=1}^{K} n(d_i, z_{k'}) + K \cdot \alpha}；$$

[0074] $Mod^{(n)}(w_j, z_k) = n(w_j, z_k)$

[0075] $\theta_{d_i}$表示文本 $d_i$ 的类别分布概率，$Mod(w_j)$ 表示词汇 $w_j$ 的类别采样频次。

[0076] 步骤 105，所有采样完成后，取最后 100 次状态计算的参数的平均值为最终的模型参数，即有$\theta_{d_i}(z_k) = E_n[\theta_{d_i}^{(n)}(z_k)]$，$Mod(w_j, z_k) = E_n[Mod^{(n)}(w_j, z_k)]$；取$\theta_{d_i}$的最大分量所在类别为 $d_i$ 的判断类别，即$z(d_i) = \arg\max_k \theta_{d_i}(z_k)$。

[0077] 在线聚类的详细过程如下：

[0078] 步骤 106，按照步骤 101 的方法进行文本预处理，按照步骤 102 中方法对新文本中词汇进行随机初始化，并统计新文本的 n(d, z)，n(d)，n(w, z)，n(z)；在新文本中出现的词汇，如果在原有模型中已经出现，则修改 n(w, z) 和 n(z)，否则保持不变，具体计算方法如下：

[0079]

$$n(w_j, z_k) = \begin{cases} n(w_j, z_k) & w_j\text{是新词汇} \\ n(w_j, z_k) + Mod(w_j, z_k) & w_j\text{是原有词汇} \end{cases};$$

[0080] 步骤 107，按照步骤 103 和步骤 104 进行新文本的采样迭代，并实现对新文本类别的判断；

[0081] 步骤 108，更新原有模型，对于新文本获得的词汇类别分布频次 Mod(w, z)，如果词汇 w 在原模型中没有出现，则将 w 的频次分布添加进原模型，如果 w 是原模型中的已有词汇，则将新获得的 Mod(w, z) 替换到原模型中的频次分布。

[0082] 下面对在线式高维数据降维可视化过程进行详细说明。

[0083] 在线式高维数据降维可视化过程如图 4 所示，下面详细介绍在线式高维数据降维可视化过程。

[0084] 步骤 101，输入文本聚类过程产生的高维类别分布向量 $x_1x_2...x_n$（其中 $x_i$ 即文本 $d_i$ 聚类得到的类别分布向量$\theta_{d_i}$），判断是否首次实施降维可视化，若是转步骤 102，否则转步骤 106；

[0085] 步骤 102，计算任意两高维向量 $x_i$、$x_j$ 间的相似性 $p_{ij}$，计算方法如下：

[0086] $$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

[0087] 其中，

[0088] $$p_{j|i} = \frac{\exp(-\| x_i - x_j \|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\| x_i - x_k \|^2 / 2\sigma_i^2)}$$

[0089] $\sigma_i^2$ 为将 $x_i$ 到 $x_k$（$k \neq i$）的欧氏距离视为高斯分布的方差；

[0090] 步骤 103，随机产生 $x_1x_2...x_n$ 对应的低维向量初始值 $y_1y_2...y_n$，计算任意两低维向量 $y_i$、$y_j$ 间的相似性 $q_{ij}$，计算方法如下：

[0091] $$q_{ij} = \frac{(1+\| y_i - y_j \|^2)^{-1}}{\sum_{k \neq l} (1+\| y_k - y_l \|^2)^{-1}}$$

[0092] 步骤 104，利用 KL 距离度量 $\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的差异，$\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的 KL 距离 $D_{KL}$ 定义为：

[0093] $$D_{KL}=\sum_{i}\sum_{j}p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

[0094] 步骤 105，利用梯度下降法迭代寻找 $D_{KL}$ 的最小值，同时更新低维向量 $y_1y_2...y_n$，具体过程如下：

[0095] 步骤 105a，按照 $D_{KL}$ 的负梯度方向逐步搜索最优解，将 $D_{KL}$ 对 $y_i$ 求偏导数，有

[0096] $$\frac{\partial D_{KL}}{\partial y_i}=4\sum_{j}(p_{ij}-q_{ij})(y_i-y_j)(1+\|y_i-y_j\|^2)^{-1}$$

[0097] 步骤 105b，记 $Y=(y_1, y_2...y_n)$，$Y^{(t)}=(y_1^{(t)}, y_2^{(t)}...y_n^{(t)})$ 为经过 t 次迭代后的解，梯度下降法采用下式更新 Y：

[0098] $$Y^{(t)}=Y^{(t-1)}+\eta\frac{\partial D_{KL}}{\partial Y}+\alpha(t)(Y^{(t-1)}-Y^{(t-2)})$$

[0099] 其中 η 为步长，通常取 50 ～ 500 间的整数，α(t) 为加速因子，在 0 ～ 1 间取值。迭代若干次后，即可得到满足预定误差要求的低维向量 $y_1y_2...y_n$，然后转步骤 112；

[0100] 步骤 106，随机选择首次降维可视化产生的一组低维向量 $\beta_1\beta_2...\beta_m$ 作为基准低维向量，对应的高维向量 $\alpha_1\alpha_2...\alpha_m$ 作为基准高维向量；

[0101] 步骤 107，判断 $x_1x_2...x_n$ 中是否有未处理的向量，若有转步骤 108，否则转步骤 112；

[0102] 步骤 108，取第一个未处理向量 $x_i$，采用下式计算 $x_i$ 与基准高维向量 $\alpha_j$(j = 1, 2...m) 的相似性 $p_{j|i}$：

[0103] $$p_{j|i}=\frac{\exp(-\|x_i-\alpha_j\|^2/2\sigma_i^2)}{\sum_{k=1}^{m}\exp(-\|x_i-\alpha_k\|^2/2\sigma_i^2)}$$

[0104] $\sigma_i^2$ 为将 $x_i$ 到 $\alpha_k$(k = 1, 2...m) 的欧氏距离视为高斯分布的方差。

[0105] 步骤 109，随机产生 $x_i$ 对应的低维向量初始值 $y_i$，采用下式计算 $y_i$ 和基准低维向量 $\beta_j$(j = 1, 2...m) 间的相似性 $q_{j|i}$：

[0106] $$q_{j|i}=\frac{(1+\|y_i-\beta_j\|^2)^{-1}}{\sum_{k=1}^{m}(1+\|y_i-\beta_k\|^2)^{-1}}$$

[0107] 步骤 110：利用 KL 距离度量 $\{p_{j|i}\}$ 与 $\{q_{j|i}\}$ 间的差异，$\{p_{j|i}\}$ 与 $\{q_{j|i}\}$ 间的 KL 距离 $D_{KL}$ 定义为：

[0108] $$D_{KL}=\sum_{j=1}^{m}p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}$$

[0109] 步骤 111，利用梯度下降法迭代寻找 $D_{KL}$ 的最小值，同时更新低维向量 $y_i$，具体过程如下：

[0110] 步骤 111a，按照 $D_{KL}$ 的负梯度方向逐步搜索最优解，将 $D_{KL}$ 对 $y_i$ 求偏导数，有

[0111] $$\frac{\partial D_{KL}}{\partial y_i}=2\sum_{j}(p_{j|i}-q_{j|i})(y_i-\beta_j)(1+\|y_i-\beta_j\|^2)^{-1}$$

[0112] 步骤 111b，记 $y_i^{(t)}$ 为经过 t 次迭代后的解，梯度下降法采用下式更新 $y_i$：

[0113] $$y_i^{(t)} = y_i^{(t-1)} + \eta \frac{\partial D_{KL}}{\partial y_i} + \alpha(t)(y_i^{(t-1)} - y_i^{(t-2)})$$

[0114] 其中 η 为步长，通常取 50 ～ 500 间的整数，α (t) 为加速因子，在 0 ～ 1 间取值。迭代若干次后，即可得到满足预定误差要求的低维向量 $y_i$ ；

[0115] 步骤 112，对降维后的低维向量 $y_i$ 进行可视化展示。实际中我们常把文本类别分布向量降维到二维欧氏空间中，即 $y_i$ 为二维坐标，对于二维可视化要求，可以直接输出所有文本降维后的坐标，如图 5 所示；对于三维可视化要求，我们常取文本文本类别分布向量中的最大分量数值作为第三维的坐标值，结合 $y_i$ 进行三维展示，如图 6 所示。

[0116] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

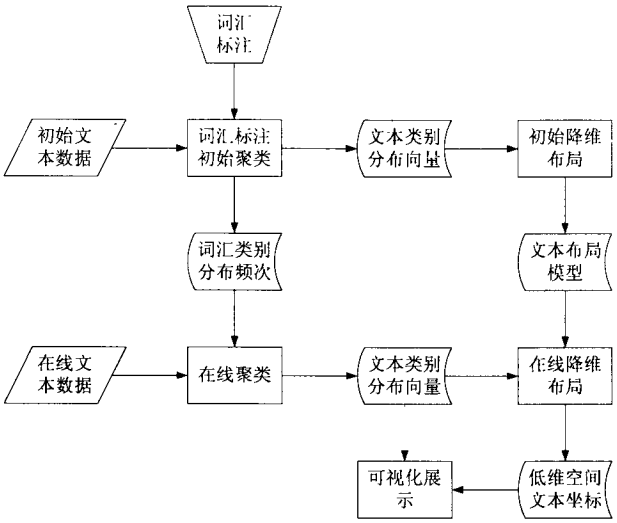
词汇
标注
初始文
本数据
词汇标注
初始聚类
文本类别
分布向量
初始降维
布局
词汇类别
分布频次
文本布局
模型
在线文
本数据
在线聚类
文本类别
分布向量
在线降维
布局
可视化展
示
低维空间
文本坐标

图 1

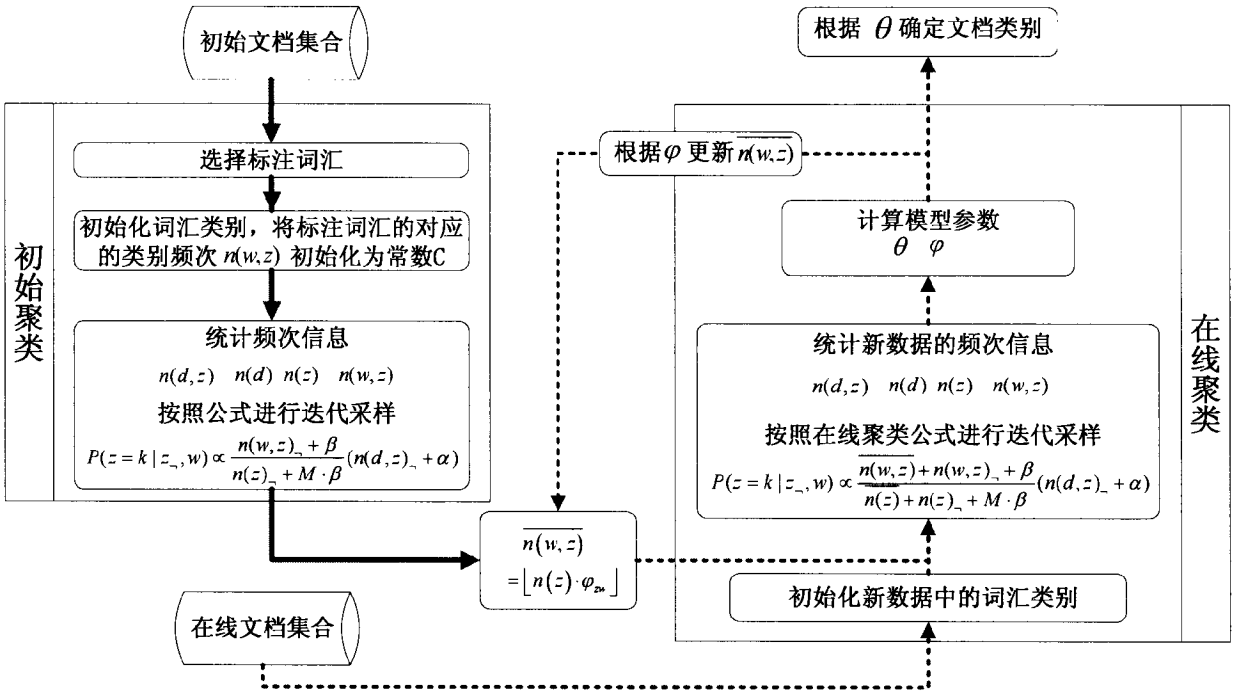
初始文档集合
初始聚类
选择标注词汇
初始化词汇类别，将标注词汇的对应的类别频次 n(w,z) 初始化为常数C
统计频次信息
n(d,z) n(d) n(z) n(w,z)
按照公式进行迭代采样
$P(z=k|z_{\neg},w)\propto\frac{n(w,z)_{\neg}+\beta}{n(z)_{\neg}+M\cdot\beta}(n(d,z)_{\neg}+\alpha)$
在线文档集合
$\overline{n(w,z)}=\lfloor n(z)\cdot\varphi_{zw}\rfloor$
根据 θ 确定文档类别
根据φ 更新 $\overline{n(w,z)}$
计算模型参数
θ φ
统计新数据的频次信息
n(d,z) n(d) n(z) n(w,z)
按照在线聚类公式进行迭代采样
$P(z=k|z_{\neg},w)\propto\frac{\overline{n(w,z)}+n(w,z)_{\neg}+\beta}{n(z)+n(z)_{\neg}+M\cdot\beta}(n(d,z)_{\neg}+\alpha)$
初始化新数据中的词汇类别
在线聚类

图 2

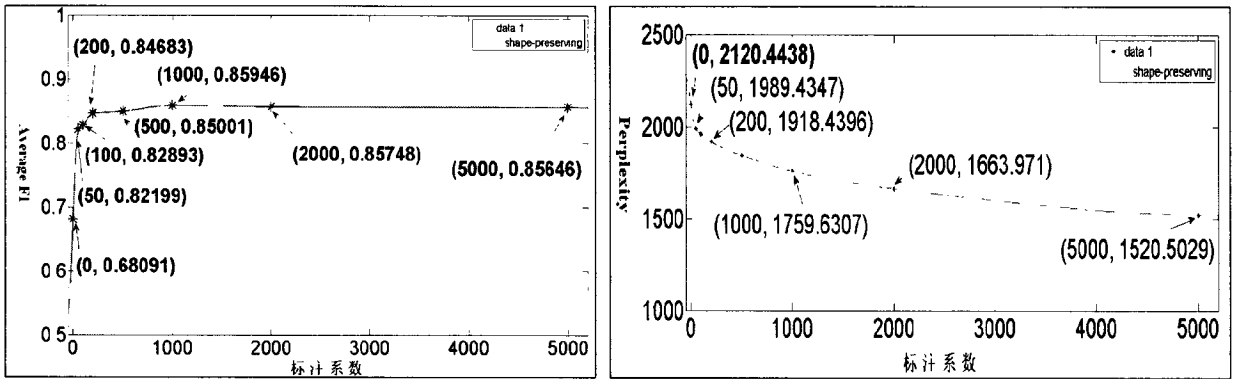
data 1
shape-preserving
(200, 0.84683)
(1000, 0.85946)
(500, 0.85001)
(100, 0.82893)
(2000, 0.85748)
(5000, 0.85646)
(50, 0.82199)
(0, 0.68091)
Average
标计系数
(0, 2120.4438)
(50, 1989.4347)
(200, 1918.4396)
(2000, 1663.971)
(1000, 1759.6307)
(5000, 1520.5029)
Perplexity

图 3

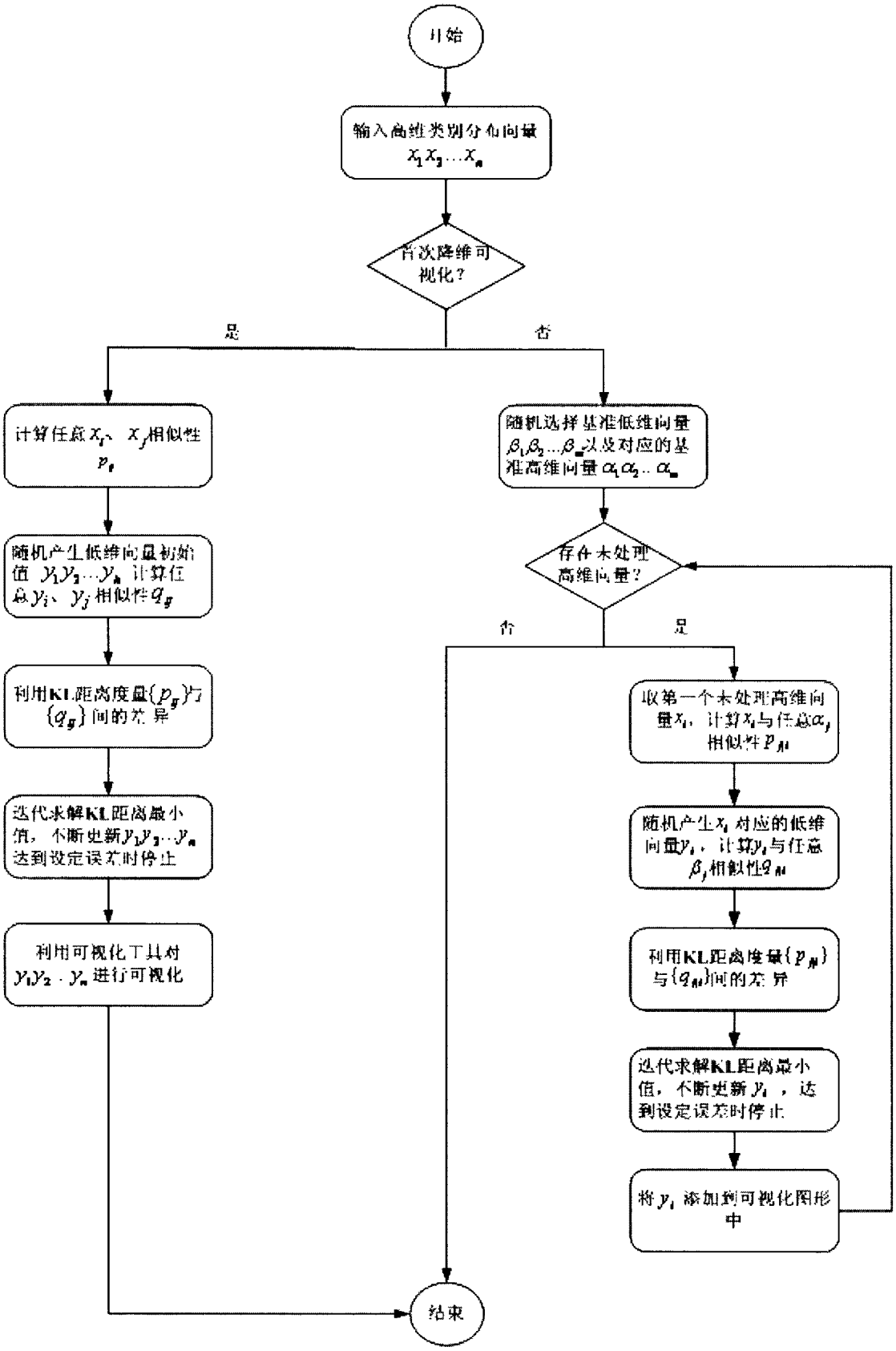
开始
输入高维类别分布向量 $x_1 x_2 \cdots x_n$
首次降维可视化？
是
否
计算任意 $x_i$、$x_j$ 相似性 $p_{ij}$
随机产生低维向量初始值 $y_1 y_2 \cdots y_n$ 计算任意 $y_i$、$y_j$ 相似性 $q_{ij}$
利用KL距离度量 $\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的差异
迭代求解KL距离最小值，不断更新 $y_1 y_2 \cdots y_n$ 达到设定误差时停止
利用可视化工具对 $y_1 y_2 \cdots y_n$ 进行可视化
随机选择基准低维向量 $\beta_1 \beta_2 \cdots \beta_m$ 以及对应的基准高维向量 $\alpha_1 \alpha_2 \cdots \alpha_m$
存在未处理高维向量？
否
是
取第一个未处理高维向量 $x_i$，计算 $x_i$ 与任意 $\alpha_j$ 相似性 $p_{ij}$
随机产生 $x_i$ 对应的低维向量 $y_i$，计算 $y_i$ 与任意 $\beta_j$ 相似性 $q_{ij}$
利用KL距离度量 $\{p_{ij}\}$ 与 $\{q_{ij}\}$ 间的差异
迭代求解KL距离最小值，不断更新 $y_i$，达到设定误差时停止
将 $y_i$ 添加到可视化图形中
结束

图 4

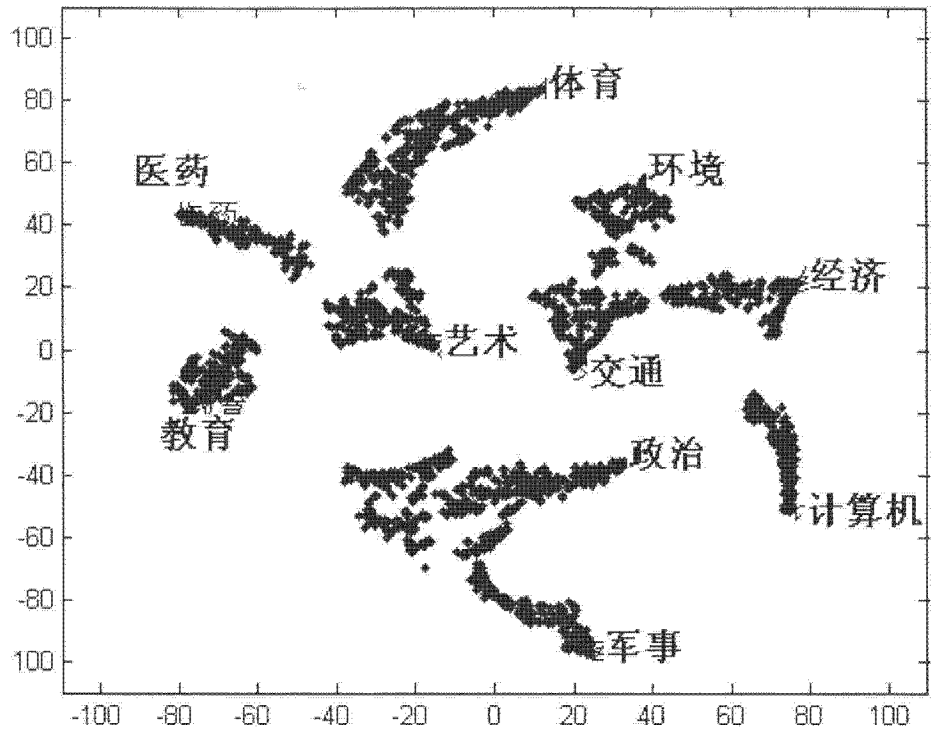
体育
医药
环境
经济
艺术
交通
教育
政治
计算机
军事
100
80
60
40
20
0
-20
-40
-60
-80
100
-100 -80 -60 -40 -20 0 20 40 60 80 100

图 5

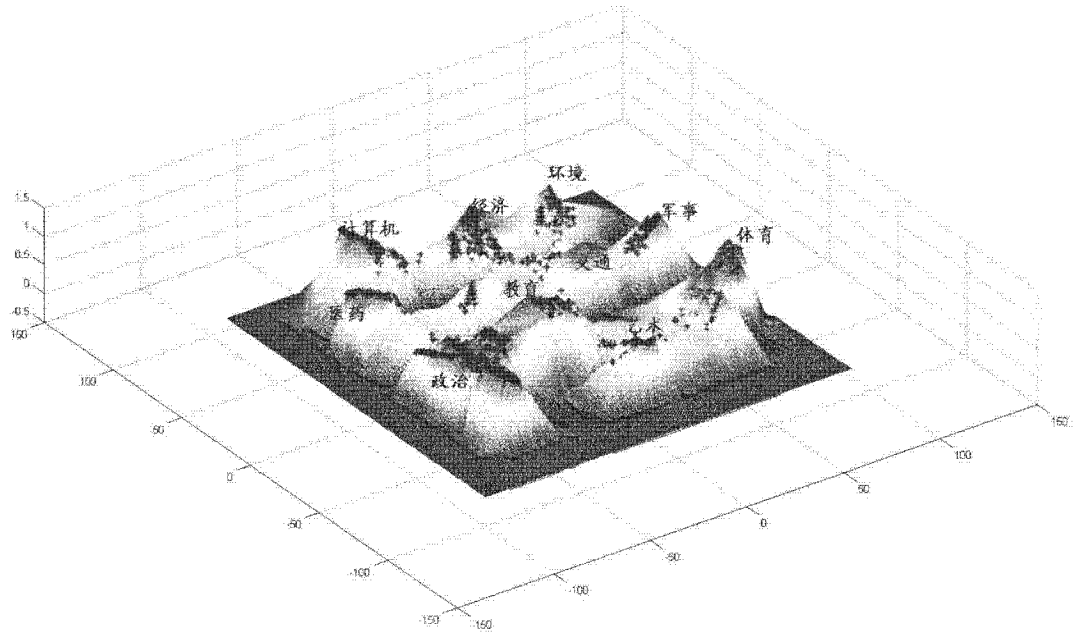
环境
经济
军事
计算机
体育
交通
教育
医药
艺术
政治

图 6